THE BELL SYSTEM TECHNICAL JOURNAL

FEBRUARY 1979
VOL. 58, NO. 2

ISSN0005-8580

THE BELL SYSTEM TECHNICAL JOURNAL

THE BELL SYSTEM TECHNICAL JOURNAL is published monthly, except for the May-June and July-August combined issues, by the American Telephone and Telegraph Company, J. D. deButts, Chairman and Chief Executive Officer; C. L. Brown, President; W. G. Burns, Vice President and Treasurer; F. A. Hutson, Jr., Secretary. Editorial enquiries should be addressed to the Editor, The Bell System Technical Journal, Bell Laboratories, 600 Mountain Ave., Murray Hill, N.J. 07974. Checks for subscriptions should be made payable to The Bell System Technical Journal and should be addressed to Bell Laboratories, Circulation Group, Whippany Road, Whippany, N.J. 07981. Subscriptions $20.00 per year; single copies $2.00 each. Foreign postage $1.00 per year; 15 cents per copy. Printed in U.S.A. Second-class postage paid at New Providence, New Jersey 07974 and additional mailing offices.

THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 58 February 1979 Number 2

A Queuing Model for a Hybrid Data Multiplexer

By R. R. ANDERSON, G. J. FOSCHINI, and B. GOPINATH

(Manuscript received March 3, 1978)

There are several instances in a data network where a communication line is shared by two or more types of data. In this paper, we analyze the performance of a buffer used to multiplex two types of data. Sporadic short messages, like inquiries from terminals, share the same channel as relatively steady synchronous data, like trunk traffic or long messages from computer data bases. To the authors' knowledge, previous studies have been limited to an ad hoc approximation to the probability distribution of interest. We solve for the equilibrium distribution of number of units of data in the buffer. The delay distribution easily follows. Numerical results are also presented which can be used as a guide to determining how much of each type of traffic can be sustained simultaneously.

I. INTRODUCTION

We analyze the performance of a buffer that is used to multiplex two types of data. Sporadic short bursts of data, like inquiries from terminals, share the same communication channel with relatively steady streams of data like digitized voice, lengthy messages from computers, data bases, or traffic from a busy trunk. A line-switched network, one that provides a dedicated channel for each connection, is preferred for lengthy steady messages. A packet-switched network, on the other hand, is efficient for messages that are short and bursty. In a packet-switched network, messages are forwarded from node to node in the form of packets of data that include addressing information. In such a network, there is no necessity for a dedicated channel for each connection.

A data network could accommodate both types of traffic by dividing the transmission facilities into two fixed parts—one part exclusively for line-switched traffic, the other for packet-switched traffic. The subframe switching concept introduced in Ref. 1 is an example of a temporal division of capacity. The model developed here can be used for the analysis of such a system. In any system where a resource such as a transmission line is shared between two or more types of users, the performance guaranteed to each individual type of customers has to be met. Packet delay and the probability of losing packets are the two measures of performance we consider.

Kummerle[2] proposed a model for multiplexing line and packet-switched data and derived, using an ad-hoc approximation, formulas relating the performance measures to the traffic intensity and transmission capacity. A similar problem was analyzed using a diffusion approximation in Ref. 3. In Ref. 4, an $M/D/N$ model is used for an approximate analysis. Reference 5 looks at a related continuous-time problem where the arrival mechanism (rather than the service) has a periodic component. An integral equation is derived that can be solved using Wiener-Hopf techniques. In this paper, we formulate a model for the multiplexer and solve it exactly. We then describe the computational method used to derive the numerical results and display them to illustrate the tradeoff involved in performance and line utilization.

II. DESCRIPTION OF THE MULTIPLEXING SYSTEM

The sources of data that are connected to a generic node in the network are divided into two groups—synchronous sources and asynchronous sources. Both these sources generate messages randomly. However, when a synchronous source generates a message, the message is generated at a constant rate and is much longer than the messages generated by asynchronous sources. To describe the system, we use an example. Let sources A, B in Fig. 1 be two synchronous sources transmitting at ½ and ⅙ the line (marked LINE in Fig. 1) rate, respectively. Packets are assumed to be of fixed size, and the unit of time is normalized to be the time required for the line to transmit one packet. The output of source A is assembled into packets by the line buffers as shown in Fig. 1. Then the output of this buffer, connected to source A, will produce one packet every two units of time. As soon as these packets are ready, they are transmitted by the line even if packets from asynchronous terminals, marked T in Fig. 1, are waiting to be transmitted in the packet buffer. Similarly, the output of the line buffer connected to source B produces a packet every six units of time. However, the output of B is so synchronized that A and B packets do not have to be transmitted at the same time on the line. The asyn-

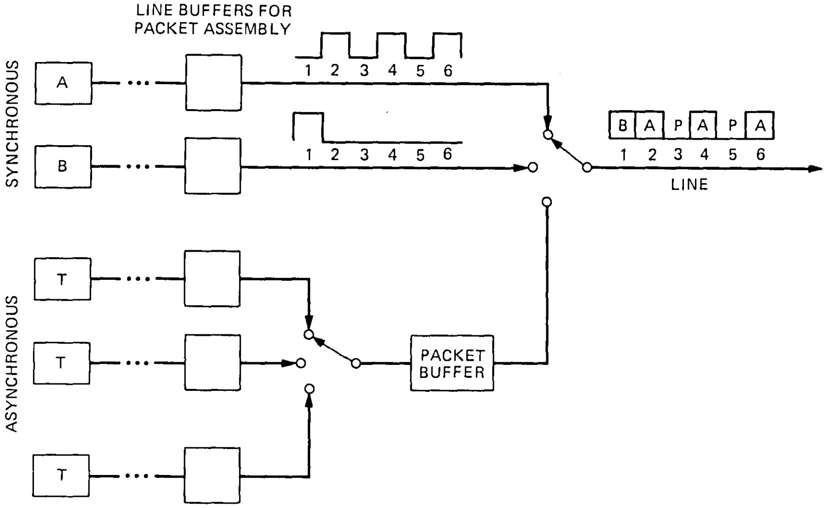


Fig. 1—Example of multiplexing system.

chronous sources have their own line buffers doing the packet assembly, but the output of these buffers are not necessarily synchronous with the line. As soon as these sources produce packets, they are inserted into the packet buffer and there await transmission on the output line. Notice that this synchronous method of transmitting sources A and B allows us to discard addressing information in all except the first packets of a message from these sources. In practice, if a packet from an asynchronous source arrives at the buffer when it is full, the packet is lost. In order to guarantee satisfactory performance for the asynchronous sources, we must keep the probability of such a loss small (requirements in the $10^{-5} - 10^{-7}$ range are typical). It is mathematically convenient to work with an infinite rather than a finite buffer model and solve for the probability that queue size exceeds a given threshold. The probability that queue size in such a buffer exceeds a level B is an upper bound for the probability that a finite buffer of size B overflows.

III. THE MATHEMATICAL MODEL

The model considered here applies to systems where there are one or more packet buffers (associated with asynchronous sources) and, of course, many synchronous sources. The packet buffers and synchronous sources may be served by the line (it is available to accept a packet for transmission) in any fixed periodic pattern. The distribution of the number of packets in any given packet buffer is only influenced by the asynchronous traffic connected to it and the pattern in which the line becomes available to it. During the time slots that the line is not available to the given packet buffer, it is immaterial whether another packet buffer or a synchronous source is being served.

Hence, the mathematical model described below analyzes a single-server queue in discrete time with the server being absent according to a fixed periodic pattern.

The number of packets in the packet buffer at the end of nth unit of time is denoted by b_{n+1}. The number of packets that the asynchronous sources collectively generate in the nth unit of time is denoted by x_n. The sequence of integers $\{x_n\}$ is assumed to be samples of independent, identically distributed, random variables. The frame length denoted by M is the period of the pattern of serving packet buffers and the synchronous sources. In the example of Fig. 1, the frame length is 6. There are 6 slots per frame and slots 1, 2, 4, 6 are dedicated to synchronous sources A and B. Slots 3 and 5 are used for transmitting packets from asynchronous sources whenever there are any to be transmitted. In general, let $\mathscr{S}$ denote the set of indices of slots in which the given packet buffer is not served. In the example of Fig. 1, $\mathscr{S} = \{1, 2, 4, 6\}$. Then b_{n+1}, the number of packets in the buffer at the end of the $(n + 1)^{\text{st}}$ unit of time, is given by

$$b_{n+1} = (b_n - u_n)^+ + x_n, \tag{1}^*$$

where

$$u_n = \begin{cases} 0 \text{ if } n \equiv \mathscr{S}(M) \\ 1 \text{ otherwise.} \end{cases}$$

Whenever $n \equiv \mathscr{S}(M)$, no packets from the packet buffer can be transmitted in the nth unit of time. Therefore, for such an n,

$$b_{n+1} = b_n + x_n.$$

On the other hand, if $n \not\equiv \mathscr{S}(M)$, a packet from the packet buffer can be transmitted in the nth unit of time (if there were any left at the end of the $(n - 1)$st time interval). Hence,

$$b_{n+1} = (b_n - 1)^+ + x_n$$

if $n \not\equiv \mathscr{S}(\mathrm{M})$. Because of the time-varying nature of u_n, it is clear that $\{b_n\}$ itself has no stationary distribution. However, the vector process $\mathbf{b}_m = (b_{mM+1}, b_{mM+2}, \cdots b_{(m+1)M})^t$ indexed by m has a stationary distribution because of the periodic nature of u_n. We will show that we can find a relationship between the marginal distributions of the components of b_m that uniquely specify the equilibrium distribution of the b_{mM+i}, $i = 1, 2, \cdots M$.

The process $\{b_n\}$ is Markov process with the state space being the

* $x^+ = x$ if $x > 0$, $x^+ = 0$ if $x \leq 0$, $n \equiv \mathscr{S}(\mathrm{M})$ if $n \equiv i \bmod M$ for some $i \in \mathscr{S}$.

nonnegative integers. The transition probability matrix P_n at time n is determined by whether $n \equiv \mathscr{S}(\mathrm{M})$ or not; $P_n = P_m$ if $n \equiv m$ mod M. Let P_i denote P_{mM+i}, $i = 1, 2, \cdots, M$. For each i, the process b_{mM+i}, indexed by m, is again a Markov process with the associated transition probability matrix

$$Q_i = P_{i-1}P_{i-2}\cdots P_1 P_M P_{M-1}\cdots P_{i+1}P_i.$$

It is easily shown that the Markov chain associated with Q_i is aperiodic and irreducible. We will now show that, if the average number of packets arriving in the packet buffer is less than the number of slots available for transmission during a frame, this Markov chain is positive recurrent and hence, for each i, b_{mM+i} has a limiting stationary distribution as $m \uparrow \infty$. Let J denote the number of slots in a frame during which no packets from the packet buffer can be transmitted.

Lemma 1: For each $i = 1, 2, \cdots, M$ the Markov chain associated with Q_i is positive recurrent if

$$MEx_n < M - J.$$

Proof: Repeated application of (1) shows that, for $m = 0, 1, 2, \cdots,$

$$b_{(m+1)M+i} = (\cdots((b_{mM+i} - \delta_1)^+ + x_{mM+i} - \delta_2)^+ \cdots + x_{(m+1)M+i-2} - \delta_M)^+ + x_{(m+1)M+i-1},$$

where $\delta_j = 1$ if $j \not\equiv \mathscr{S}(M)$ and $\delta_j = 0$ otherwise. Therefore, if $b_{mM+i} \geq M$, then

$$b_{(m+1)M+i} = b_{mM+i} + \left\{ \sum_{\ell=0}^{M+1} x_{mM+i+\ell} \right\} - (M - J);$$

hence,

$$E(b_{(m+1)M+i} \mid b_{mM+i} = j) = b_{mM+i} + MEx_n - (M - J)$$

for all $j \geq M$. Let $Q_i(\ell, j)$ be the (ℓ, j)th element of Q_i; then, if $MEx_n < M - J$

$$\sum_{\ell=0}^{\infty} Q_i(\ell, j)\ell < j \qquad \text{for} \quad j \geq M.$$

Hence, using Theorem 2 of Ref. 6, the Markov chain corresponding to Q_i is positive recurrent.

IV. CALCULATION OF STEADY-STATE DISTRIBUTION

In this section, we show how to calculate the steady-state distribution of the components of $\mathbf{b}_m$. As we mentioned earlier, the distribution of b_n has no limiting value as n tends to ∞. However, we showed that the distribution of b_{mM+i} for $i = 1, 2, \cdots, M$ approaches a limiting value when $MEx_n < M - J$, which is obviously the condition for stability of the queuing process. Let $\phi_i(s)$ denote the generating function corresponding to the limiting distribution of b_{mM+i} for $i = 1, 2, \cdots, M$. Then

$$\phi_i(s) = \lim_{m \to \infty} Es^{b_{mM+i}} \; i = 1, 2, \cdots, M.$$

Calculating the generating functions of both sides of (1) and then letting m tend to ∞, we can derive equations for $\phi_i(s)$. Let $\chi(s) = Es^{x_n}$, and note that $\chi(s)$ factors out on the right-hand side of (1) since x_n is independent of b_n. Then we have for $i = 0, \cdots, M - 1$,

$$\begin{aligned} \phi_{i+1}(s) &= \phi_i(s)\chi(s) \qquad \text{for} \quad i \equiv S(M) \\ \phi_{i+1}(s) &= [s^{-1}\phi_i(s) + (1 - s^{-1})p_{io}]\chi(s) \qquad \text{for} \quad i \not\equiv S(M), \end{aligned} \tag{2}$$

where

$$p_{i0} = \lim_{n \to \infty} \Pr\{b_{nM+i} = 0\} \quad \text{and} \quad \phi_0 = \phi_M. \tag{3}$$

We can write the above equations (2) in matrix form as follows. Let ϵ_i, δ_i be s and 0 respectively if $i \equiv S(M)$. For $i \not\equiv S(M)$, let $\epsilon_i = \delta_i = 1$. Then

$$\begin{bmatrix} \phi_1(s) \\ \phi_2(s) \\ \cdot \\ \cdot \\ \cdot \\ \phi_M(s) \end{bmatrix} = \frac{\chi(s)}{s} \begin{bmatrix} 0 & 0 & \cdots & 0 & \epsilon_M \\ \epsilon_1 & 0 & \cdots & 0 & 0 \\ 0 & \epsilon_2 & \cdots & 0 & 0 \\ & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & 0 & 0 \\ 0 & 0 & \cdots & \epsilon_{M-1} & 0 \end{bmatrix} \begin{bmatrix} \phi_1(s) \\ \phi_2(s) \\ \\ \\ \phi_M(s) \end{bmatrix} + (1 - s^{-1})\chi(s) \begin{bmatrix} p_{M,0}\delta_M \\ p_{1,0}\delta_1 \\ \cdot \\ \cdot \\ \cdot \\ p_{M-1,0}\delta_{M-1} \end{bmatrix} \tag{4}$$

Using the symbol ϕ for the vector $(\phi_1, \cdots, \phi_M)^t$, $\mathbf{p}$ for $(\delta_M p_{M,0}, \cdots, \delta_{M-1} p_{M-1,0})^t$, $A(s)$ for the matrix consisting of entries either 1, 0, or s, and I for the identity matrix, we can rewrite (4) as

$$\left[I - \frac{\chi(s)}{s} A(s) \right] \phi(s) = (1 - s^{-1})\chi(s)\mathbf{p}. \tag{5}$$

For every s such that the matrix on the left is invertible, let $B(s)$

denote its inverse. Represent $B(s)$ by its rows $B_i(s)$ for $i = 1, 2, \cdots,$ M. Then we can derive the following equation for B_1.

$$B_1 = \frac{s^{M-J}}{s^{M-J} - \chi^M(s)} \times \left[1, \prod_{j=2}^{M} \left[\frac{\epsilon_j}{s} \right] \chi^{M-1}(s), \cdots, \prod_{j=M-1}^{M} \left[\frac{\epsilon_j}{s} \right] \chi^2(s), \frac{\epsilon_M}{s} \chi(s) \right]. \quad (6)$$

The other rows of B can be recursively computed as follows: Let $\ell_i = \left(0, 0, \cdots, \overset{i}{1}, 0, \cdots, 0\right)$ then

$$B_{i+1} = \ell_{i+1} + \chi(s) \frac{\epsilon_i}{s} B_i \qquad i = 1, 2, \cdots, M-1. \quad (7)$$

Since $\chi(1) = 1$, $s^{M-J} - \chi^M(s) = 0$ for $s = 1$. It can be shown by Rouché's theorem that $s^{M-J} - \chi^M(s)$ has $M - J - 1$ distinct roots strictly inside the unit disk.[7] Each of the generating functions ϕ_i satisfies

$$\phi_i(s) = (1 - s^{-1}) \chi(s) B_i \mathbf{p} \quad (8)$$

for every s for which $B_i(s)$ is well defined ($B_i\mathbf{p}$ is the product of B_i and $\mathbf{p}$ viewed as matrices). Since $\phi_i(s)$ is analytic on the unit disk, the representation (8) extends to the roots of $s^{M-J} - \chi^M(s)$ that lie within the unit disk if $\mathbf{p}$ is such that the singularities of $[s^{M-J} - \chi^M(s)]^{-1}$ are removed. We illustrate this in detail by using an important special case of the model presented here. This is the case when, out of the M slots, the last J slots are needed for line switched or steady traffic. Only the first $M - J$ slots are used for transmitting asynchronous traffic. In this case, the matrix $A(s)$ has the form:

$$\begin{bmatrix}
0 & 0 & \cdots & & & \cdots & 0 & s \\
1 & 0 & \cdots & & & \cdots & 0 & 0 \\
0 & 1 & \cdots & & & \cdots & 0 & 0 \\
\cdot & \cdot & & & & & & \\
\cdot & & \cdot & & & & & \\
\cdot & & & \cdot & & & & \\
0 & & & 1 & & \cdots & 0 & 0 \\
& & & & & & & \\
0 & & \cdots & s & & \cdots & 0 & 0 \\
0 & & \cdots & & s & \cdots & 0 & 0 \\
\cdot & & & & \cdot & & & \\
\cdot & & & & & \cdot & & \\
\cdot & & & & & & \cdot & \\
0 & & \cdots & & & \cdots & s & 0
\end{bmatrix}$$

and the vector $\mathbf{p} = (0, p_{1,0}, p_{2,0}, \cdots, p_{M-J,0}, 0, \cdots, 0)^t$.

Let $\xi_1 = 1, \xi_2, \xi_3, \cdots, \xi_{M-J}$ denote the roots of $s^{M-J} - \chi^M(s)$ that are in the unit disk. At these roots, the first row of B, excluding the scalar multiplier in (6), can be expressed as

$$\left[1, \frac{\chi^{M-1}(\xi_i)}{\xi_i^{M-J-1}}, \cdots, \chi^J(\xi_i), \chi^{J-1}(\xi_i), \cdots, \chi^2(\xi_i), \chi(\xi_i) \right].$$

Hence a choice of $\mathbf{p}$ that will cancel the singularities of the scalar multiplier in (6) satisfies the following equation

$$\begin{bmatrix} 1 & \chi(\xi_1)/\xi_1 & \chi^2(\xi_1)/\xi_1^2 & \cdots & \chi^{M-J-1}_{(\xi_1)}/\xi_1^{M-J-1} \\ 1 & \chi(\xi_2)/\xi_2 & \chi^2(\xi_2)/\xi_2^2 & \cdots & \chi^{M-J-1}_{(\xi_2)}/\xi_2^{M-J-1} \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ 1 & \cdots & & & \end{bmatrix} \begin{bmatrix} p_{M-J,0} \\ \cdot \\ \cdot \\ \cdot \\ p_{2,0} \\ p_{1,0} \end{bmatrix} = \begin{bmatrix} \rho \\ 0 \\ 0 \\ \\ \\ 0 \end{bmatrix}, \quad (9)$$

where $= \sum_{i=1}^{M-J} p_{i,0}$. Except for the first equation, the relations in (9) express the fact that the numerator of (8) vanishes at $1, \xi_2, \cdots \xi_{M-J}$. The vector $(p_{1,0}, p_{2,0}, \cdots p_{M-J,0})^t$ appears reversed in (9), so the coefficient matrix can be written as a Vandermonde matrix, which we denote $V = V(\xi_1, \xi_2, \cdots \xi_{M-J})$. The nonsingularity of V follows from the distinctness of the $\{\xi_j\}_1^{M-J}$

Therefore, if the scalar ρ is known, the vector $\mathbf{p}$ is determined uniquely from (9). Differentiating (2) with respect to s and letting s approach 1 gives $\rho = (M - J) - MEx_n$. Hence $\mathbf{p}$ and then $\phi(s)$ can be determined uniquely from (5).

Theorem 1: *Let $MEx_n < (M - J)$, then there always exists constants $\{p_{i,0}\}_{i=1}^{M}$ such that the functions $\{\phi_i(s)\}_{i=1}^{M}$ are analytic in the unit circle. Moreover, these constants are uniquely determined by the condition $\phi_1(1) = 1$.*

In the general case, the components of $\mathbf{p}$ that are zero depend on the set $\mathscr{S}$. Let $i_1 < i_2 < \cdots < i_{M-J}$ represent the indices not included in $\mathscr{S}$. By definition, $\delta_j = 0$ whenever $j \neq$ some i_m. Hence the unknown constants in eq. (8) are $p_{i_m,0}$, $m = 1, 2, \cdots, M - J$. Once again, using the arguments above, we can arrive at the $(M - J)$ equations that the $p_{i_m,0}$ satisfy in order that $\phi_i(s)$ have no singularities inside the unit circle. These correspond to (9) and will be denoted by (9′); however, the matrix appearing on the left-hand side of (9′) is no longer Vandermonde. Let V′ be the matrix associated with (9′):

$$\sum_{m=1}^{M-J} V'_{1m}\, p_{i_m,0} = \rho$$

$$\sum_{m=1}^{M-J} V'_{\ell m}\, \rho\; i_{m},_0 = 0 \qquad \text{for} \quad M - J \geq \ell > 1 \tag{9'}$$

and the elements $V'_{\ell m}$ are

$$V'_{\ell m} = \left[\prod_{j=i_m}^{M} \frac{\epsilon_j}{s} \right] \chi(s)^{M-i_m+1} \qquad m = 1, 2, \cdots, M.$$

There must be at least one solution for these equations whenever $MEx_n < M - J$, since the invariant distribution corresponding to Q_i, which exists by Theorem 1, satisfies (9′) and $\phi_i(s)$ has no singularities inside the unit circle. Moreover, we will show that any solution of (9′) gives the unique invariant distribution corresponding to Q_i. Corresponding to any solution of (9′), we can find functions $\hat{\phi}_i(s)$ from (8) such that the associated sequences $\{\hat{p}_{ij}\}$ with $\sum_{j=0}^{\infty} \hat{p}_{ij} s^j = \hat{\phi}_i(s)$ are absolutely summable, since $\hat{\phi}_i(s)$ will have no singularities inside the unit disk. Inspection of (2) shows that the vectors $\hat{\pi}_i = \{\hat{p}_{ij}\}_{j=0}^{\infty}$ satisfy

$$\hat{\pi}_{i+1} = P_i \hat{\pi}_i \qquad i = 1, 2, \cdots, M-1$$

and

$$\hat{\pi}_1 = P_M \hat{\pi} M.$$

Hence, from the definition of Q_i,

$$\hat{\pi}_i = Q_i \hat{\pi}_i.$$

Since the Markov chain corresponding to Q_i is positive recurrent, the (ℓ, m) element of Q_i^n, the nth power of Q_i, tends to $p_{i\ell}$ the ℓth component of the invariant distribution corresponding to Q_i. From the above equation,

$$\hat{p}_{i\ell} = \sum_m Q_i^n(\ell, m)\, \hat{p}_{im}.$$

Since the sequence $\{\hat{p}_{ij}\}$ is absolutely summable taking limits of both sides and interchanging limits, we have

$$\hat{p}_{i\ell} = p_{i\ell} \sum_m \hat{p}_{im}.$$

However, from the first of equation of (9′) we can show that $\sum_{m=0}^{\infty} \hat{p}_{im} = 1$. Hence $\hat{p}_{i\ell} = p_{i\ell}$, the unique invariant density corresponding to Q_i. Hence we have shown that (9′) has a unique solution whenever $MEx_n < M\text{-}J$. Q.E.D.

Remark: Usually, the queue size β at a "random" time is of interest. The generating function of β, denoted by $\psi((s)$, is the average of $\phi_i's$, i.e.,

$$\psi(s) = M^{-1} \sum_1^M \phi_i(s).$$

V. SOME SPECIAL CASES

In the special case $J = 1$, $M = 2$, it is easy to express $\phi_i(s)$ in closed form. For Poisson arrivals, we have

$$\begin{bmatrix} \phi_1(s) \\ \phi_2(s) \end{bmatrix} = \frac{e^{\lambda(s-1)}(s-1)(1-2\lambda)}{s - e^{2\lambda(s-1)}} \begin{bmatrix} e^{\lambda(s-1)} \\ 1 \end{bmatrix}.$$

So

$$\psi(s) = \frac{(1-2\lambda)}{2} \frac{e^{\lambda(s-1)}(s-1)}{s - e^{2\lambda(s-1)}} (e^{\lambda(s-1)} + 1).$$

The probability of an empty buffer is

$$p_0 = \psi(0) = \frac{(1-2\lambda)}{2}(1 + e^{\lambda})$$

and, of course,

$$\lim_{\lambda \to (1/2)} p_0 \text{ and } \lim_{\lambda \to 0} p_0 = 1.$$

The mean buffer content

$$\bar{\beta} = \psi(s) \Big|_{s=1} = \frac{3}{2}\lambda + \frac{2\lambda^2}{1-2\lambda}.$$

As is intuitively obvious,

$$\lim_{\lambda \uparrow (1/2)} \bar{\beta} = \infty \text{ and } \lim_{\lambda \downarrow 0} \bar{\beta} = 0.$$

For the variance of the buffer content, we have

$$\begin{aligned} \text{Var}(\beta) &= \psi'(\text{s}) + \psi'(\text{s}) - (\psi'(\text{s}))^2 |_{s=1} \\ &= \frac{5}{8}\lambda^2 + \frac{13}{6}\frac{\lambda^3}{1-2\lambda} + \frac{2\lambda^2}{(1-2\lambda)^2} + \bar{\beta}^2 - \bar{\beta}, \end{aligned}$$

and again

$$\lim_{\lambda \uparrow (1/2)} \text{Var}(\beta) = \infty$$

while

$$\lim_{\lambda \downarrow 0} (\beta) = 0.$$

Another simple case is when $M \to \infty$ with J fixed, so the relative time when the server is absent tends to zero. Asymptotically, the system behaves like a discrete time $M \mid \mathrm{D} \mid 1$ queue with time quantum equal to one service time. The analysis of this queue is given in the excellent survey paper of Ref. 8. The generating function of β is

$$\left[\frac{s-1}{s-\chi(s)}\right]\chi(s)(1-\lambda).$$

So, for the Poisson case,

$$p_0 = (1-\lambda),$$

$$\bar{\beta} = \lambda\left[1 + \frac{\lambda}{2(1-\lambda)}\right],$$

and

$$\operatorname{Var}\beta = \lambda^2 + \frac{4}{3}\left[\frac{\lambda^3}{1-\lambda}\right]\frac{\lambda^4}{2(1-\lambda)^2} + \bar{\beta} - \bar{\beta}^2.$$

A related queuing problem is introduced in Ref. 7. They undertake a discrete time analysis of the waiting room occupancy in a situation where a shuttle visits every M time units, whereupon up to a maximum of K occupants are removed. If less than K occupants confront the arriving shuttle, all are removed. As in our analysis, the arrivals are arbitrary i.i.d. variables. The generating function $\xi(s)$, of the equilibrium density of waiting room occupancy as seen by the arriving shuttle, is determined (see Ref. 9 for more detail). In the special case where $K = 1$, if we set $J = M - 1$ in our analysis, then $\xi(s)$ is the same as $\phi_0(s)$.

Another variation of the process analyzed in the last section occurs when the synchronous packets are also queued and hence are subject to delay and loss. Let the resulting buffer process be denoted by b'_n. Then

$$b'_{n+1} = (b'_n - 1)^+ + x_n + (1 - u_n).$$

Starting with $\mathrm{b}_0 = b'_0 = 0$, we can show by induction that for each arrival stream realization, b_n and b'_n agree to within one packet, that is, with probability one $\mid b_n - b'_n \mid \leq 1$ uniformly in n. Thus the analysis in the previous section aids in estimating the jitter suffered by the synchronous input stream. Such jitter considerations, which are basic to the emerging topic of packetized speech, will not be explored here.

VI. EXTENDING THE RESULTS

6.1 Obtaining delay densities from the buffer densities

So far in this paper, we have focused on the problem of obtaining the buffer density; however, in many applications the density of delay is also of importance. It is reasonable to expect several system requirements to be in effect, as, for example

$$\begin{aligned}&\Pr[\text{buffered packets} > 32] < 10^{-7}\\&\Pr[\text{waiting time} > 500 \text{ ms}] < 10^{-1}\\&E[\text{waiting time}] < 250 \text{ ms.}\end{aligned}$$

In some applications, satisfying the first objective obviates the other two. However, when the delay requirements are more crucial, it is easy to obtain the delay density numerically as it is simply related to $\phi(s)$, as we now show.

Since we are using a discrete time model, we consider all arrivals to occur at integer times. The discrete time model of arrivals can be considered to arise from a continuous-time arrival process in which all arrivals on $]n, n + 1]$ are associated with a time of arrival $n + 1$. For the purpose of computing delay, it is essential to retain an order relationship for the arrivals at time n.

The packets to be served ahead of a typical arrival y in $]n, n + 1]$ can be partitioned into three groups:

(i) The packets already in the buffer at time n.

(ii) The number of asynchronous packets arriving in $]n, n + 1]$ ahead of y, denoted by x_n^*. For example, if x_i is Poisson, an arrival occurs according to a uniform distribution so the generating function of x_n^* is

$$\chi^*(z) = \int_0^1 e^{\lambda\tau(z-1)}\, d\tau = \frac{e^{\lambda(z-1)}-1}{\lambda(z-1)}.$$

(iii) The synchronous packets arriving during the transmission of asynchronous packets in the system before the transmission of y.

Let E_n denote the sum of (i) and (ii) above. Note the generating function of E_n is the product of the generating functions of b_n and x_n^*. The delay caused by interruptions, as mentioned in (iii) above, can be derived easily from the nature of $\mathscr{S}$ and depends only on E_n and the slot in the frame that corresponds to n. For a "random" arrival, the delay density is found by averaging over the densities corresponding to the M slots. Of course, an additional delay unit must be included to account for the service of the packet whose delay distribution is being calculated. In conclusion, when delay performance is a dominant consideration, the delay density functions can also be computed from $\phi(s)$.

Once the delay density and/or the density of buffered packets is computed for the parameter range of interest, the network designer can determine the tradeoffs among

Delay and/or buffer size.
Trunk capacity.
Average packet arrival rate.
Percentage of transmission facilities devoted to line switching (or batch service).

Such information, along with market and revenue forecasts and resource cost estimates, allows the designer to determine an optimum packet service-line service mix for the projected environment.

6.2 Extension to batch arrival model

The process $\{x_n\}_{-\infty}^{\infty}$ of packet arrivals on the nth time interval can be considered to be realized from an underlying message arrival process $\{m_n\}_{-\infty}^{\infty}$ where each message contains a random number of packets $\{l_n\}_{-\infty}^{\infty}$. So both $\{m_n\}_{-\infty}^{\infty}$ and $\{l_n\}_{-\infty}^{\infty}$ are processes of i.i.d variables independent of each other and with (different) underlying densities with values in the nonnegative integers. If $M(z)$ and $L(z)$ are the corresponding generating functions of m_n and l_n, then $M(L(z))$ is the generating function of x_n.

In some applications, fast switching may be employed so that a fraction of a packet is the basic unit switched (e.g., a byte in a system employing fixed-size 1024-bit packets). To analyze such situations, one can take the basic time quantum in the mathematical model to be the time required to transmit a subpacket. Then a packet corresponds to a message and a subpacket to a packet in the above discussion. For the example cited, $L(z) = z^{128}$.

So the model is accommodating whether we view arrivals as single packets or batches of packets. The most useful case is when the arrivals are Poisson and a geometric number of packets is associated with each arrival. In this case,

$$M(z) = \exp \lambda(z-1) \quad \text{and} \quad L(z) = z(1-p)/(1-pz).$$

So

$$\chi(z) = \exp \lambda[(z-1)/(1-pz)].$$

The mean number of arrivals per time slot is $\lambda(1-p)^{-1}$. For purposes of obtaining the delay density in the Poisson case, we need $\chi^*(z)$ which, by a straightforward integration, is

$$\chi^*(z) = \frac{(1-pz)}{\lambda(z-1)} e^{\lambda(z-1)/(1-pz)} - 1.$$

6.3 A more general problem

Here our objective is to point out a seemingly more complex queuing situation where our methods still apply. The type of queuing situation we discuss plays an important role in the analysis of subframe packet switching in Ref. 2.

Consider the situation in Fig. 2. The server visits the various queues according to some fixed periodic pattern. The input to various queues are independent. During every visit, the server completes servicing one job if there are any jobs waiting. Each input stream represented by an arrow can have batch arrivals according to any arbitrary distribution function (it is permissible for different distributions to correspond to different inputs). The question is: What is the buffer and delay distribution as perceived by one of the inputs (indicated by the only dashed arrow inputting the starred queue)?

A little reflection reveals that, from the perspective of one input, the problem is no different than the one we have already solved. The equivalence would be immediate if the dashed arrow denoted the sole

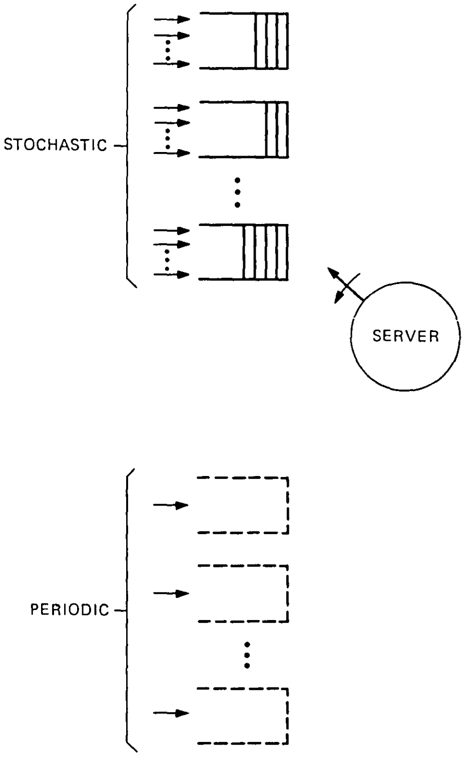


Fig. 2—Generalized system.

input to its queue. After all, from the perspective of the input in question, the structure of what the server does while he visits the unstarred queues is irrelevant to the buffer and delay performance. All that matters is the periodic pattern of the service availability and unavailability to the starred queue. When the starred queue has other inputs, there is no real complication. To obtain the buffer density, simply replace the parallel inputs by a single input stream choosing a single batch density representing the various batch densities with their corresponding frequencies. A similar method accommodates delay with the adjustment that the time the batch, whose delay is being computed, spends in service is represented as a random choice from the distribution of batch sizes from the dotted input.

Paul Lue of Bell Laboratories in Holmdel, who had an earlier version of this paper, has formulated a further generalization involving multiple periodic parallel servers for Fig. 2. He plans to report his successful analysis in the future.

VII. DISCUSSION OF NUMERICAL WORK

7.1 The computer program

For input distributions which have a generating function of the form

$$\chi(s) = e^{[\lambda(s-1)/1-ps]},$$

a program for obtaining the equilibrium density of buffer size and delay as a function of (λ, M, J, p) was developed. The core of the program is the computation of $(\phi_1(s), \phi_2(s), \cdots, \phi_M(s))$ and a generating function inversion routine.

Double precision was used throughout the program since at the present time overflow probabilities of the order of 10^{-7} are of interest. The determination of the $\phi_m(s)$ includes finding the location of the "apparent" poles in the unit disk. While a straightforward search of the disk for poles does the trick, the Newton Raphson routine suggested in Ref. 2 is preferred for its speed.

For inverting a generating function, we employ a fast Fourier transform (FFT) program. The use of the FFT in this situation stems from the observation that, if we replace s by $e^{i\omega}$, the generating function is then a Fourier series on the boundary of the unit disk. The Fourier coefficients are the probabilities of interest. In using the FFT, the generating function is represented by its values at the discrete sample points

$$\{e^{\sqrt{-1}(2\pi\ell/L)}\}_{\ell=0}^{L},$$

where L is taken to be sufficiently large to obtain the accuracy required.

As indicated in the previous sections, once the $\phi_m(s)$ are known, the analytical determination of the delay distribution is also possible. Programmed implementation of the procedure for obtaining the delay distributions from the $\{\phi_n\}_1^M$ is straightforward.

Usually the program output routine is set only to provide the densities (buffer or delay) averaged over an entire frame as the more refined intraframe densities are of secondary interest.

7.2 A peculiarity of the numerical data

With each irreducible fraction $r = p/q$ in (0, 1), associate a frame of size q in which the first p time slots for asynchronous data are followed by $q - p$ time slots for synchronous data (see Fig. 3). Let $r_n = p_n/q_n$ be such that $\lim_{n\to\infty} \mathrm{r}_n = 1/2$ and $\lim_{n\to\infty} q_n = \infty$. Mean buffer size $\bar{\beta}$ and mean delay $\bar{\alpha}$ are discontinuous functions of r. Indeed, the buffer size and delay of those packets arriving in the first half of the last $q_n - p_n$ slots are going to infinity since they receive no service in the second half of the last $q_n\, p_n$ slots.

The above argument points out that two frame organizations can be arbitrarily close in terms of the relative number of time slots devoted to packet switching yet the mean delay and buffer sizes of both systems can differ by an arbitrarily large number. The preceding discussion also shows that interpolation of statistics to an intermediate r value is a perilous calculation. However, interpolation to an r point from points with the same denominator (frame size) can be reasonable.

7.3 Numerical results

The program for determining the distribution of β was exercised for numerous cases with $M \leq 16$ and $\rho_{\mathrm{eq}} \triangleq \lambda M/(M - J)(1 - p) < 1$, the latter inequality being required for stability. For illustrative purposes, Tables I and II summarize the statistics associated with a wide variety of examples. The parameters for Tables I and II differ only in that, in I, single packet messages are assumed, while, in II, the messages are of random (geometric) size with mean five (i.e., $p = 0.2$). The 10^{-x} headers

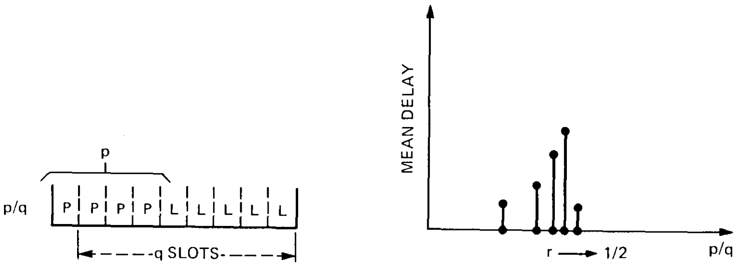


Fig. 3—Discontinuity of buffer and delay statistics.

Table I—Buffer size and delay statistics for cases where all slots but one in a frame are allotted for asynchronous data

		$\lambda = 0.2$						$\lambda = 0.4$						$\lambda = 0.6$					
		Mean	Var	10^{-2}	10^{-5}	10^{-6}	10^{-7}	Mean	Var	10^{-2}	10^{-5}	10^{-6}	10^{-7}	Mean	Var	10^{-2}	10^{-5}	10^{-6}	10^{-7}
$M = 2$	Buffer	0.433	0.497		7	9	10	2.200	5.65		27	32	37	Unstable					
$J = 1$	Delay	1.17	3.05	5				4.50	22.4	22									
$M = 3$	Buffer	0.328	0.352		6	7	8	0.974	1.38		12	15	17	4.83	23.6				
$J = 1$	Delay	0.63	0.793	5				1.42	2.97	8				6.98	50.2	34			
$M = 4$	Buffer	0.297	0.315		5	6	7	0.800	1.02		10	12	14	2.34	5.80		27	32	38
$J = 1$	Delay	0.48	0.572	4				0.99	1.67	6				2.89	10.1	15			
$M = 5$	Buffer	0.282	0.297		5	6	7	0.729	0.897		9	11	13	1.84	3.72		21	25	29
$J = 1$	Delay	0.41	0.487	4				0.82	1.27	5				2.06	5.55	11			
$M = 6$	Buffer	0.272	0.285		5	6	7	0.690	0.832		9	11	12	1.62	2.97		19	22	26
$J = 1$	Delay	0.36	0.436	4				0.721	1.08	5				0.70	4.02	10			

of the columns refer to an upper bound on the probability that the random variable (buffer size or delay) takes a value larger than the number entered in the column. For example, in the first subtable

$$\Pr[\text{delay} > 5 \text{ packets}] < 10^{-2}$$

and

$$\Pr[\text{buffer size} > 11 \text{ packets}] < 10^{-7}.$$

We use "packets" as the unit of delay since this is readily converted to time, because in our model we assumed that the duration of one time slot is the transmission time for one packet.

To emphasize that more refined statistics are easily obtained, Table III presents the intraframe details for a specific case.

Figures 4 and 5 stem from Tables I and II and are used below in a pair of hypothetical examples we include to show how a designer could make use of the available numerical capability.

Example I. Consider a situation in which a 56-kb/s trunk is available for transmission of 1024 bit packets. The packet arrival process is Poisson, and one packet is associated with each arrival. A 32-packet buffer is available. It is required that the probability of a lost packet not exceed 10^{-6}. If $\lambda = 0.4$ (23 packets per second), the question is how

Table II—Second-order statistics for buffer and delay for cases where all slots but one in a frame are allotted for asynchronous data. Multiple packet messages of mean size five (p^{-1})

		$\frac{\lambda}{1-p} = 0.2$		$\frac{\lambda}{1-p} = 0.4$		$\frac{\lambda}{1-p} = 0.6$	
		Mean	Var	Mean	Var	Mean	Var
$M = 2$	Buffer	3.072	35.4	11.9	13.75	Unstable	
$J = 1$	Delay	12.3	132.7	176.2	227.8		
$M = 3$	Buffer	2.02	21.6	6.53	86.5	12.7	223.5
$J = 1$	Delay	8.85	76.5	13.1	146.5	11.6	230.8
$M = 4$	Buffer	1.74	18.1	5.17	65.5	12.0	176
$J = 1$	Delay	7.58	61.5	11.2	114	14.8	203
$M = 5$	Buffer	1.60	16.5	4.59	56.7	10.6	151
$J = 1$	Delay	6.95	53.1	10.1	97.9	14.5	177
$M = 6$	Buffer	1.52	15.6	4.27	51.9	9.76	137
$J = 1$	Delay	6.58	48.2	9.45	88.2	13.8	161

Table III—Intraframe buffer size statistics (frame size six, server absent for four slots, $\lambda = 0.2$, $\rho = 0.6$)

	Time Slot						
	1	2	3	4	5	6	$M\|D\|1$
Mean	1.361	0.850	0.561	0.761	0.961	1.161	1.05
Var	1.712	1.331	.912	1.112	1.312	1.512	1.43
10^{-6}	15	15	14	14	15	15	

much capacity can we devote to line switching. Figure 4 shows that the answer is 50 percent. If, at a subsequent date, we have that λ has increased to 0.6, then only 25 percent of capacity can be devoted to line switching.

Figure 5 shows that, in the case λ = 0.6, the 99-percent delay is about 250 ms, while the mean is about 55 ms.

Example II: The ability to compute the density of β (J, M, λ, p)

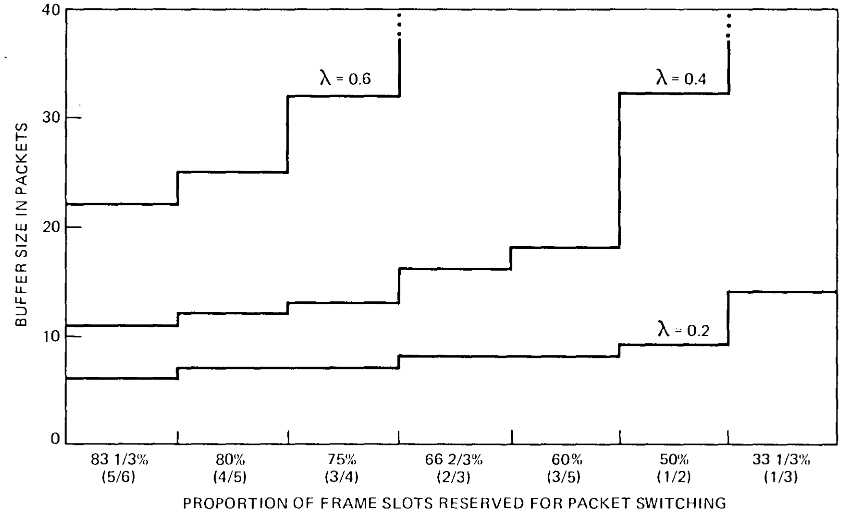


Fig. 4—10^{-6} loss threshold for various mixes of packet and line switching.

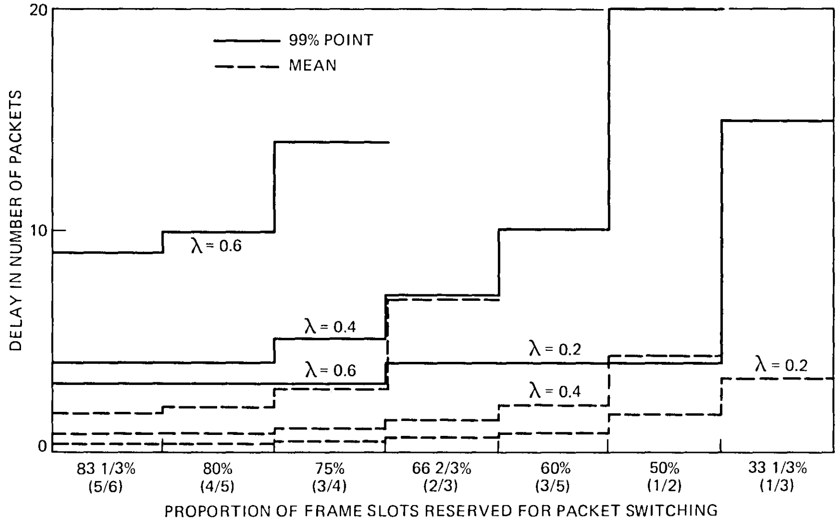


Fig. 5—Delay (mean and 99-percent point) for various mixes of packet and line switching.

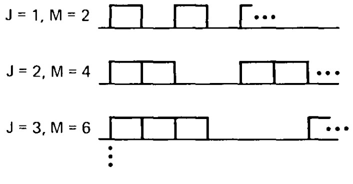
also offers useful information relating to switching considerations. To see this, consider a hypothetical situation in which a designer wishes to devote 50 percent of capacity to packet switching and the remaining for line switching. The question of organizing the frame arises. Any of the patterns shown in Fig. 6a will suffice. Intuitively, one would expect buffer occupancy statistics to degrade as one moves down the list. Yet the reduced load on the processor for attending to switching among the two types of customers would make the longer frames attractive. For λ = 0.4, the degradation is illustrated in Fig. 6b. The numerical capability reported here enables one to determine the optimal operating point on the basis of projected switching costs and sensitivity of revenue to performance.

This example and the discussion in Section 7.2 point out that the *amount* of trunk capacity devoted to asynchronous traffic of a specified intensity is not enough information for one to determine the buffer and delay distributions. The details of frame organization can be essential for obtaining accurate statistics.

7.4 The role of an M | D | 1 model in computations

Previous analysis of such systems used simulation or approximation to obtain buffer and delay statistics. Yet simulations are usually too expensive for obtaining the extremal statistics preferred by the requirements engineer. On the other hand, the accuracy of "approximations" such as using an $M \mid D \mid 1$ model could not be appraised. It is reasonable to require that the $M \mid D \mid 1$ model have the same utilization $\rho_{\text{eq}} = [M/(M-J)]\,\lambda$ and the same throughput as the hybrid model. Indeed, ρ_{eq} and throughput uniquely determine an $M \mid D \mid 1$ model.

With the hybrid multiplexor şolution in hand, one can evaluate the above nonexact methods in the parameter range of interest. While a thorough exploration of this issue is beyond the scope of this paper, we shall include some comparisons that were made for the $J = 1$ case. For $M = 2$, the errors range as high as 28 percent and then decrease as M increases, as one would expect. It appears that, insofar as the tail probabilities which only register order of magnitude are concerned,



(a)

	J/M				
	1/2	2/4	3/6	4/8	5/10
MEAN	2.200	2.265	2.351	2.450	2.557
VAR	2.126	2.173	2.243	2.341	2.453

(b)

Fig. 6—(a) A sequence of possibilities for attaining a 50-percent mix ($J = M/2$). (b) Central moments of buffer size vs frame length for 50-percent mix ($J = M/2$).

one might as well use $M|D|1$ formulas. Of course, if a much more finely resolved graph is meaningful, the $M|D|1$ approximation may no longer suffice. In initial stages of the evaluation of computer networks, it is unusually unrealistic to expect to know the arrival rate to a tighter tolerance than 10 percent, and so the $M|D|1$ analysis of overflow for the 1 out of M cases provides a useful simplified model.

Nonetheless, there are parameter ranges where the $M|D|1$ approximation is useless. To see this, fix λ and take $J = M/2$. Consider what happens as M increases. Note the $M|D|1$ approximation has $\rho_{eq} = 2\lambda$ and the throughput is independent of M. For the hybrid model, the packets arriving in the third quarter of a frame must wait out at least the last quarter before they are eligible for service. So, as $M \rightarrow \infty$, the mean buffer size and mean delay increase without bound and the error in using an $M|D|1$ approximation goes to infinity. This example is by no means pathological, as it addresses precisely those cases that arise in the switching study mentioned in Example II. The dotted line in Fig. 7 gives the $M|D|1$ result.

The $M|D|1$ model is useful in comparing the hybrid system with a system providing separate dedicated facilities for synchronous and asynchronous data. For example, Fig. 7 compares the performance between hybrid and dedicated implementations, and an $M|D|1$ model is used to provide numbers for the latter. With reference to Example II in Section 7.3, the dotted line of Fig. 7 indicates the average delay performance of a competitive system using dedicated trunks. In Fig. 7 we see a region where dedicated trunks of a given capacity do not perform as well as a hybrid system that devotes the same capacity to asynchronous traffic.

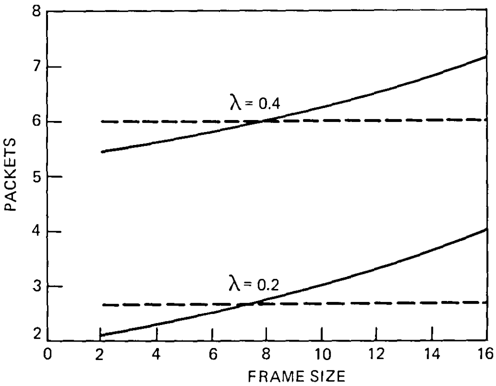


Fig. 7—Mean delay for various realizations of a 50-percent packet service rate.

REFERENCES

1. G. J. Foschini, B. Gopinath, and J. F. Hayes, unpublished work.
2. K. Kummerle, "Multiplexor Performance for Integrated Line and Packet-Switched Traffic," ICCC Stockholm Second International Conference on Computer Communication, Stockholm, 1974, pp. 507–515.
3. M. J. Fischer, "Analysis and Design of Loop Service Systems Via a Diffusion Approximation," Defense Communications Engineering Center Report.
4. M. J. Fisher and T. C. Harris, "A Model for Evaluating the Performance of an Integrated Circuit- and Packet-Switched Multiplex Structure," IEEE Trans. Commun., *COM-24* (February 1976), pp. 195–202.
5. Izzet Sahin and U. Narayan Bhat, "A Stochastic System with Scheduled Secondary Inputs," Operations Research, *19* (1971), pp. 436–446.
6. F. G. Foster, "On the Stochastic Matrices Associated with Certain Queueing Problems," Ann. Math. Statist. *24* (1953), pp. 355–360.
7. P. E. Boudreau, J. S. Griffin, and M. Kac, "An Elementary Queueing Problem," American Mathematical Monthly, October 1962, pp. 713–724.
8. H. Kobayashi and A. G. Konheim, "Queueing Models for Computer Communications System Analysis," IEEE Trans. Commun. (Special Issue on Computer Communications), *COM-25*, No. 1 (January 1977), pp. 2–28.
9. A. G. Konheim and R. Meister, "Service in a Loop System," J. Ass. Comput. Mach., *19*, No. 1 (January 1972), pp. 92–108.

Copyright © 1979 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 58, No. 2, February 1979
Printed in U.S.A.

On the Required Tap-Weight Precision for Digitally Implemented, Adaptive, Mean-Squared Equalizers

By R. D. GITLIN and S. B. WEINSTEIN

(Manuscript received May 3, 1978)

An analysis is made of the degree of precision required in a digitally implemented adaptive equalizer to achieve a satisfactory level of performance. Considering both the conventional synchronously spaced equalizer and the newer fractionally spaced equalizer, insight is provided into the relationship between the tap-weight precision and the steady-state, mean-squared error. It is demonstrated why the number of adaptive tap weights should be kept to a minimum (consistent with acceptable steady-state performance), both from convergence and precision requirements. A simple formula is given that displays the tradeoff among the equalizer mean-squared error, the number of taps, the channel characteristics, and digital resolution. For typical basic-conditioned voiceband channels operating at 9.6 kb/s, and neglecting the effects that limited resolution might have on timing and carrier phase tracking, analysis and simulation both indicate that the required tap-weight resolution is of the order of 11 or 12 bits. Moreover, the minimum precision is only weakly dependent on the quality of the channel.

I. INTRODUCTION

State-of-the art adaptive equalizers for voiceband modems are digitally implemented and strive to minimize the equalized mean-squared error.[1] An important consideration in assessing the complexity of such an adaptive digital equalizer is the number of bits required to represent the stored signal samples and the equalizer tap weights. Gitlin, Mazo, and Taylor[2] have shown that the precision required for successful adaptive operation, via the estimated-gradient algorithm,[3] can be significantly greater than that required for static or fixed equalization. The purpose of this paper is to determine the precision required in the tap-updating circuitry so that the equalizer mean-squared error can attain an acceptable level.

In the well-known estimated-gradient tap adjustment algorithm,[3] the tap weights are incremented by a term proportional to the product of the instantaneous output error and the voltage stored in the corresponding delay element. When this correction term is less than half a tap-weight quantization interval, the algorithm ceases to make any further substantive adjustment. To determine the minimum number of bits needed to achieve an acceptable performance level (mean-squared error), an appropriate proportionality constant, or step size, must be determined for use in the algorithm. From pure analog, or infinite precision considerations, a relatively large step size is desirable to accelerate initial convergence,[2-5] while a small step size is needed to reduce the residual mean-squared error (that part of the error in excess of the minimum attainable mean-squared error). If the channel is stationary, then in the converged mode the analog algorithm should use a vanishingly small step size to provide almost no fluctuation about the minimum obtainable mean-squared error. However, in a digitally implemented algorithm, a decrease in the step size can actually degrade performance unless there is a compensating increase in the precision of the tap weights. This occurs when the error is so small that an increased number of bits are needed in order that the proportionately smaller corrections be "seen" by the equalizer.

A useful compromise is to choose a step size that provides a slight increase in the steady-state mean-squared error to a level which can be attained by a digital equalizer of reasonable precision. This precludes the choice of an unrealistically small step size—with its concomitant requirement of excessive precision—and provides a mechanism for the analytical determination of the necessary level of precision. Our objective, then, is to be able to directly calculate the precision required to achieve an acceptable performance level.

In Section II, assuming parameters with infinite precision, we determine the step size which produces a specified increase in the mean-squared error. This result is combined in Section III with digital considerations to determine the precision required in the equalizer coefficients. Simulation results are presented in Section IV to illustrate our results for both the conventional synchronous and the newer fractionally spaced equalizers.[6-7]

II. ANALOG CONSIDERATIONS

In this section, we review the basic equalized data communications system in an analog, or infinite precision, environment and determine the steady-state step size associated with a fractional increase in the residual mean-squared error above the minimum attainable error.

2.1 System model

For simplicity, we consider the baseband-equivalent data transmission system of Fig. 1 with received samples

$$r(nT') \triangleq r_n = \sum_m a_m x(nT' - mT) + \nu(nT'), \tag{1}$$

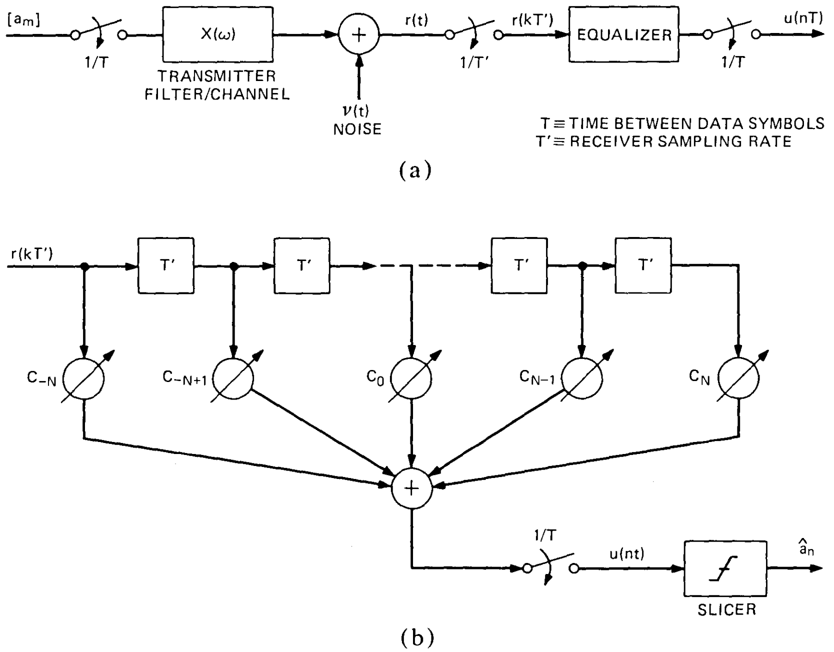
where $\{a_m\}$ are the discrete-valued data symbols, $x(\cdot)$ is the pulse shape at the receiver input, $\{\nu(nT')\}$ are independent noise samples, $1/T$ is the symbol rate, and $1/T'$ is the receiver sampling rate. For the conventional synchronous equalizer, $T' = T$, while for fractionally spaced equalizers[6,7] (FSEs), $1/T'$ will exceed twice the highest frequency component in $x(t)$. The equalizer output is computed only every T seconds and is given by

$$u(nT) = \sum_{m=-N}^{N} c_m r(nT - mT'), \tag{2}$$

where the equalizer has $(2N + 1)$ tap weights. The standard performance measure is the mean-squared error (MSE) at the equalizer output,

$$\mathscr{E} = \langle (u(nT) - a_n)^2 \rangle = \langle (\mathbf{c}'\mathbf{r}_n - a_n)^2 \rangle, \tag{3}$$

where $\langle \cdot \rangle$ denotes the ensemble average, $\mathbf{c}$ and $\mathbf{r}_n$ are, respectively,



(a)

(b)

Fig. 1—(a) Simplified baseband-equivalent PAM data transmission system. (b) Tapped delay line equalizer.

the vector of equalizer tap weights and the vector of samples stored in the equalizer at the nth output sampling instant, and the prime denotes vector transpose. Carrying out the indicated expectation in (3) for binary-valued independent data symbols gives

$$\mathscr{E} = \mathbf{c}'A\mathbf{c} - 2\mathbf{c}'\,\mathbf{x} + 1, \tag{4}$$

where $A = \langle \mathbf{r}_n \mathbf{r}_n' \rangle$ is the channel correlation matrix and $\mathbf{x} = \langle a_n r_n \rangle$ is the truncated channel-sample vector; minimizing $\mathscr{E}$ with respect to the tap weights gives the familiar optimum quantities:[1-2]

$$\mathbf{c}_{\text{opt}} = A^{-1}\mathbf{x} \tag{5}$$

$$\mathscr{E}_{\text{opt}} = 1 - \mathbf{x}'A^{-1}\mathbf{x}. \tag{6}$$

2.2 Estimated-gradient algorithm

A well-known, and frequently implemented, algorithm for the iterative adaptive determination of the optimal tap weights is

$$\mathbf{c}_{n+M} = \mathbf{c}_n - \Delta_n e_n \mathbf{r}_n, \qquad n = 0, M, 2M, \cdots, \tag{7}$$

where $e_n = u_n - a_n$ is the error signal,* Δ_n is the step size, and $\mathbf{c}_{n+M}$ is the vector of tap weights at time $(n + M)\,T$. The algorithm is obtained from the (gradient) steepest-descent algorithm by replacing the gradient of $\mathscr{E}$, with respect to $\mathbf{c}$, by the convenient unbiased estimate, $e_n\mathbf{r}_n$. The scaling of the correction term is provided by the step size Δ_n.

Under the assumption of tap-weight adjustments infrequent enough (i.e., M large enough) so that the tap-voltage vectors $\{\mathbf{r}_n\}$ are mutually independent, eqs. (4) to (7) can be used to show that

$$\mathscr{E}_n = \mathscr{E}_{\text{opt}} + \langle \boldsymbol{\epsilon}_n' A \boldsymbol{\epsilon}_n \rangle,$$

where

$$\boldsymbol{\epsilon}_n = \mathbf{c}_n - \mathbf{c}_{\text{opt}} \tag{8}$$

is the corresponding tap-weight error. In practice, it is observed that adjustments at the symbol rate ($M = 1$) result in a comparable mean-squared error, even though the independence of the $\{\mathbf{r}_n\}$, used in deriving (8), is clearly not valid.†

If we let

$$q_n \equiv \langle \boldsymbol{\epsilon}_n'\, A \boldsymbol{\epsilon}_n \rangle \tag{9}$$

denote the excess mean-squared error, then it is known[2] that q_∞ decreases as Δ_∞ decreases, while the rate of convergence (ROC) increases as Δ_n increases up to half the stability limit. Thus, from analog considerations alone, the choice of step size is important in achieving a balance between rate of convergence and the steady-state error.

* It can be assumed that an initial training data sequence is known to the receiver.

† Current work by J. E. Mazo appears to explain this anomaly.

Previous studies[2,5] have concentrated on finding the best sequence $\{\Delta_n\}$ to maximize the rate of convergence. However, the choice of the final Δ_∞ appropriate for the steady state has not received much attention, perhaps because from analog considerations alone it is clear that Δ_∞ should be as small as possible consistent with a moderate tracking capability. Difficulties can ensue, however, in a digitally implemented equalizer, where too small a value of Δ_∞ can result in an increased steady-state mean-squared error.[2] With this in mind, a reasonable compromise is to accept an $\mathscr{E}_\infty$ which is somewhat larger than $\mathscr{E}_{\text{opt}}$; it will be shown that this implies a finite value of Δ_∞ from which the required tap coefficient precision can be determined.

2.3 An iterative relation for the residual MSE

As a compromise, the step size can be selected such that the residual excess mean-squared error (9) is an acceptable fraction of the minimum attainable steady-state error, (6); i.e., let

$$q_\infty = \gamma \mathscr{E}_{\text{opt}}, \tag{10}$$

where $0 \leq \gamma \leq 1$. With this range for γ there is at most a 3-dB increase in the steady-state MSE, due to a finite value of Δ. To proceed further, we diagonalize the channel-correlation matrix and write

$$A = P \Lambda P', \tag{11}$$

where Λ is a $(2N + 1)$ by $(2N + 1)$ diagonal matrix whose entries are the eigenvalues, λ_i, of A, and P is an orthogonal matrix composed of the eigenvectors, $\mathbf{p}_i$, of A. If we denote the rotated tap-error vector by

$$\mathbf{y}_n = P \boldsymbol{\epsilon}_n, \tag{12}$$

then

$$q_n = \langle \boldsymbol{\epsilon}_n' A \boldsymbol{\epsilon}_n \rangle = \langle \mathbf{y}_n' \Lambda \mathbf{y}_n \rangle = \sum_{i=-N}^{N} \lambda_i \langle y_{ni}^2 \rangle, \tag{13}$$

where y_{ni} is the ith component of $\mathbf{y}_n$.

Using the above definitions, we can investigate the dynamic behavior of the rotated tap-error vector, $\mathbf{y}_n$, by subtracting $\mathbf{c}_{\text{opt}}$ from both sides of (7) to obtain

$$\begin{aligned} \boldsymbol{\epsilon}_{n+M} &= \boldsymbol{\epsilon}_n - \Delta \mathbf{r}_n \, (\boldsymbol{\epsilon}_n' \mathbf{r}_n + \mathbf{c}_{\text{opt}}' \mathbf{r}_n - a_n) \\ &= \boldsymbol{\epsilon}_n - \Delta \mathbf{r}_n \, [\boldsymbol{\epsilon}_n' \mathbf{r}_n + e_n(\text{opt})] \\ &= [I - \Delta \mathbf{r}_n \mathbf{r}_n'] \, \boldsymbol{\epsilon}_n - \Delta \mathbf{r}_n \, e_n(\text{opt}), \end{aligned} \tag{14}$$

where $e_n(\text{opt}) = \mathbf{c}_{\text{opt}}' \mathbf{r}_n - a_n$ is the instantaneous error when the taps are at their optimum settings. From (14) we obtain

$$\mathbf{y}_{n+M} = [I - \Delta P \mathbf{r}_n \mathbf{r}_n' P'] \, \mathbf{y}_n - \Delta e_n(\text{opt}) \, P \mathbf{r}_n, \tag{15}$$

as the iterative equation satisfied by the rotated tap-error vector.

For small M, determination of the behavior of the residual MSE, using (15), remains one of the most difficult and frustrating problems in data transmission. We shall avoid this problem by making the following assumptions:[3]

(i) The interval M (in symbol intervals) between equalizer adjustments is large enough so that the received vectors $\{\mathbf{r}_{n+\ell M}\}$ are mutually independent.

(ii) The minimum error e_n(opt) is effectively statistically independent of all received vectors $\mathbf{r}_m$.

The support for these assumptions is the following: If the channel memory is less than M symbol intervals (MT seconds), then successive received vectors stored in the delay line will be independent, since they depend upon totally disjoint data symbols. Since we are concerned with steady-state equalizer properties, rather than convergence rate, infrequent adjustment does not adversely affect our results. The independence of e_n(opt) and $\mathbf{r}_{n+\ell M}$ is supported by the following observations. First,

$$\begin{aligned}\langle e_n(\text{opt})\mathbf{r}_n\rangle &= \langle[\mathbf{r}'_n\mathbf{c}_{\text{opt}} - a_n]\mathbf{r}_n\rangle \\ &= \langle \mathbf{r}_n\mathbf{r}'_n A^{-1}\mathbf{x}\rangle - \mathbf{x} = \langle \mathbf{r}_n\mathbf{r}'_n\rangle\, A^{-1}\,\mathbf{x} - \mathbf{x} = 0. \qquad (16)\end{aligned}$$

This equation expresses the well-known fact that, at the optimum tap setting, the error signal is uncorrelated with the current received sample vector. Using the assumption of independent sample vectors, and since $\langle \mathbf{r}_n\rangle = 0$, it follows that $\langle e_n(\text{opt})\mathbf{r}_{n+\ell M}\rangle = 0$ for any ℓ. The statistical *independence* of e_n(opt) and $\mathbf{r}_n$ depends on their higher order moments as well but simulation results indicate that the steady-state squared error is a rather insensitive function of the received samples.

To determine the steady-state step size, we use the above assumptions to derive an iterative relation for $q_n \equiv \langle \mathbf{y}'_n\ \Lambda\,\mathbf{y}_n\rangle$. We first present results for a synchronous equalizer and then modify those results for a fractionally spaced equalizer. From (13) and (15) with the rotated received vector, $\mathbf{s}_n$, defined by

$$\mathbf{s}_n = P\mathbf{r}_n, \qquad (17)$$

we have

$$\begin{aligned}q_{n+M} &= \mathbf{y}'_{n+M}\,\Lambda\,\mathbf{y}_{n+M} \\ &= \langle[\mathbf{y}'_n\,(I - \Delta\mathbf{s}_n\mathbf{s}'_n) - \Delta\, e_n(\text{opt})\mathbf{s}'_n] \\ &\quad \Lambda\,[(I - \Delta\mathbf{s}_n\mathbf{s}'_n)\,\mathbf{y}_n - \Delta\, e_n(\text{opt})\mathbf{s}_n]\rangle. \qquad (18)\end{aligned}$$

To simplify (18) and to obtain a first-order linear recursion, we note the following:

(iii) By virtue of the definition of e_n(opt), $\langle e_n(\text{opt})\mathbf{r}_n\rangle = 0$, so that

$$\langle e_n(\text{opt})\mathbf{s}_n\,\Lambda\,(I - \Delta\mathbf{s}_n\mathbf{s}_n)\mathbf{y}_n\rangle = 0. \qquad (19a)$$

(*iv*) By using the familiar eigenvalue bounds* and (*i*), we have

$$\begin{aligned}\langle \mathbf{y}_n'\mathbf{s}_n\mathbf{s}_n' \, \Lambda \, \mathbf{s}_n\mathbf{s}_n'\mathbf{y}_n\rangle &\leq \lambda_M \, \langle \mathbf{y}_n'\mathbf{s}_n\mathbf{s}_n'\mathbf{s}_n\mathbf{s}_n'\mathbf{y}_n\rangle \\ &= \lambda_M(2N+1) \, \langle r_n^2\rangle \, \langle \mathbf{y}_n' \, \Lambda \, \mathbf{y}_n\rangle, \qquad (19\text{b})\end{aligned}$$

where $\langle r_n^2\rangle$ is the variance of any element in the vector $\mathbf{r}_n$, and the automatic gain control is accurate enough, so that $\mathbf{s}_n' \, \mathbf{s}_n = (2N+1) \, \langle r_n^2\rangle$ is a good approximation for any data sequence.

(*v*) By virtue of (*ii*), we have

$$\langle e_n^2(\text{opt}) \, \mathbf{s}_n' \, \Lambda \, \mathbf{s}_n\rangle = \mathscr{E}_{\text{opt}} \, \langle \mathbf{r}_n'A\mathbf{r}_n\rangle \leq \lambda_M \, \mathscr{E}_{\text{opt}} \, (2N+1) \, \langle r_n^2\rangle. \qquad (19\text{c})$$

The bounds (19b) and (19c) are relatively tight, since the bulk of the eigenvalues will, in practice, be comparable to λ_M.

(*vi*) The term in (18) which contributes a negative sign,

$$\langle \mathbf{y}_n'\mathbf{s}_n\mathbf{s}_n' \, \Lambda \, \mathbf{y}_n\rangle = \langle \mathbf{y}_n' \, \Lambda^2 \, \mathbf{y}_n\rangle = \sum_{i=-N}^{N} \lambda_i^2 \, \langle y_{ni}^2\rangle, \qquad (19\text{d})$$

has a very significant influence on both convergence and steady-state behavior and must be treated more delicately. What is needed is a good lower bound on (19d); however, the most direct lower bound,

$$\langle \mathbf{y}_n' \, \Lambda^2 \, \mathbf{y}_n\rangle \geq \lambda_m \, \langle \mathbf{y}_n' \, \Lambda \, \mathbf{y}_n\rangle, \qquad (20)$$

which involves the minimum eigenvalue, λ_m, is (in general) too loose, since just one small eigenvalue will drastically reduce the magnitude of this term.

There is, unfortunately, no tighter lower bound, since if the only significant component $\langle y_{ni}^2\rangle$ is associated with the smallest eigenvalue, as is possible when there are no restrictions on these components, then the bound (20) can be achieved. In practice, this is an extremely unlikely event (the mean-squared tap errors, $\langle y_{ni}^2\rangle$, are pretty much equal in value), and (20) is unduly pessimistic in suggesting the choice of a steady-state step size.

We therefore choose to approximate rather than (lower) bound (19d). If, as suggested above, the $\langle y_{ni}^2\rangle$ are relatively uniform for all i, then a reasonable approximation is

$$\langle \mathbf{y}_n' \, \Lambda^2 \, \mathbf{y}_n\rangle \approx \bar{\lambda}\langle \mathbf{y}_n' \, \Lambda \, \mathbf{y}_n\rangle = \bar{\lambda} q_n \qquad (21)$$

where $\bar{\lambda}$ is defined as either the average eigenvalue

$$\bar{\lambda} = \frac{1}{2N+1} \sum_{i=-N}^{N} \lambda_i, \qquad (22\text{a})$$

[*] If A is a symmetric matrix, then $\lambda_m \mathbf{z}'\mathbf{z} \leq \mathbf{z}'A\mathbf{z} \leq \lambda_M \mathbf{z}'\mathbf{z}$, where λ_m and λ_M are the minimum and maximum eigenvalues of A respectively, and $\mathbf{z}$ is any vector.

or the RMS eigenvalue

$$\bar{\lambda} = \left(\frac{1}{2N+1} \sum_{-N}^{N} \lambda_i^2 \right)^{1/2} \tag{22b}$$

Using (19) to (22) in (18), we have the key iterative relation,

$$q_{n+M} \lesssim [1 - 2\Delta\bar{\lambda} + \lambda_M \Delta^2 (2N+1) \langle r_n^2 \rangle] q_n + \lambda_M (2N+1) \langle r_n^2 \rangle \Delta^2 \mathscr{E}_{\text{opt}}, \tag{23}$$

for the excess mean-squared error.

To apply the above equation to systems which use a fractionally spaced equalizer (FSE), some of the terms appearing in (23) must be appropriately interpreted. In systems which use a FSE, the received signal is sampled at the rate $1/T'$, where $1/T'$ is greater than twice the highest frequency component of the baseband signal. Note that if the time span of an FSE is kept constant, the number of tap weights is in inverse proportion to T'. The channel correlation matrix, A, which is Toeplitz for a synchronous equalizer, is no longer Toeplitz for a FSE. It is shown in the appendix that, for $T' = T/2$ and an infinitely long FSE, half the eigenvalues are zero and the other half tend to follow a uniform sampling of the aliased magnitude-squared channel characteristic. The ith eigenvector corresponding to the nonzero eigenvalues is given approximately as a sinusoid of frequency $\omega_i = [2i/(2N+1)] (\pi/T)$, $i = 0, 1, \cdots, N$. The eigenvectors corresponding to the zero eigenvalues have most of their spectral energy concentrated near $1/T$ Hz.

In the light of this information, we wish to determine if the bounds (19b) and (19c) are still reasonably tight for a suitably long FSE. Recall that (19b) was obtained by using the bound

$$\sum_{-N}^{N} \lambda_i s_i^2 = \mathbf{s}' \Lambda \mathbf{s} \leq \lambda_M \sum_{-N}^{N} s_i^2 .$$

Since half the eigenvalues will be quite small, we have as a tight bound that

$$\sum_{-N}^{N} \lambda_i s_i^2 \approx \sum_{N/2}^{N/2} \lambda_i s_i^2 \leq \lambda_M \sum_{N/2}^{N/2} s_i^2 ,$$

where the indices greater than $N/2$ will be associated with the zero eigenvalues. We can, however, recover the full summation by noting that s_i, a component of $\mathbf{s} = P\mathbf{r}$, is given by the convolution of the input samples and the ith eigenvector. For $|i| > N/2$, this convolution is equivalent to passing the received bandlimited signal through a narrow-band filter centered at $1/T$ Hz, and is thus close to zero. We can conclude that

$$\sum_{-N}^{N} \lambda_i s_i^2 \leq \lambda_M \sum_{N/2}^{N/2} s_i^2 \cong \lambda_M \mathbf{s}'\mathbf{s} = \lambda_M \mathbf{r}'\mathbf{r},$$

and hence (19b) and (19c) remain valid. We now reconsider the discussion which precedes (20) and consider the term

$$\begin{aligned} \langle \mathbf{y}' \Lambda^2 \mathbf{y} \rangle = \sum_{-N}^{N} \lambda_i^2 \langle y_i^2 \rangle &\approx \sum_{N/2}^{N/2} \lambda_i^2 \langle y_i^2 \rangle \\ &\approx \bar{\lambda} \sum_{N/2}^{N/2} \lambda_i \langle y_i^2 \rangle \\ &\approx \bar{\lambda} \sum_{-N}^{N} \lambda_i \langle y_i^2 \rangle = \bar{\lambda} \langle \mathbf{y}' \Lambda \mathbf{y} \rangle = \bar{\lambda} q, \end{aligned} \tag{24}$$

where $\bar{\lambda}$ is an average eigenvalue over the set of *significant* eigenvalues of the channel covariance matrix. In obtaining (24), we have again assumed that the $\langle y_i^2 \rangle$ are fairly uniform (in contrast to the s_i^2, which depend critically on the index i), and we interpret $\bar{\lambda}$ as the average of the "nonzero" eigenvalues of the channel correlation matrix.

In practice, it is not difficult to estimate $\bar{\lambda}$ for a FSE, as the eigenvalues, λ_i, tend to approach zero quite rapidly. A reasonable criterion is the average eigenvalue over the partial set of eigenvalues containing all but a small fraction (perhaps 5 percent) of the eigenvalue mass. With this discussion in mind, we can apply (23) to both synchronous and fractionally spaced equalizers.

2.4 Choice of initial and steady-state step sizes

We first investigate the conditions under which the excess MSE will decrease with time. Now in order for the mean-squared error to decay it is clear from (23) that

$$| 1 - 2\Delta\bar{\lambda} + \lambda_M \Delta^2 (2N + 1) \langle r_n^2 \rangle | < 1$$

or

$$\Delta \leq \Delta_{\text{MAX}} = \frac{2\bar{\lambda}}{\lambda_M} \frac{1}{(2N + 1)} \frac{1}{\langle r_n^2 \rangle}. \tag{25}$$

Even with all the bounds and approximations which have been made in reaching (23), a significant difference is readily apparent in the maximum allowable step size for the known-gradient[2,3] algorithm and the estimated-gradient algorithm. From Refs. 2 and 3, we know that, for the known-gradient algorithm to converge, it is required that $| 1 - \Delta\lambda_i | < 1$, or equivalently $0 \leq \Delta \leq 2/\lambda_M$. The fact that the maximum step size for the estimated-gradient algorithm is considerably smaller than that for the known-gradient algorithm is deduced from (22a) and (25):†

$$\Delta_{\text{MAX}} = \frac{2}{\lambda_M} \frac{\bar{\lambda}}{(2N+1)} \frac{1}{\langle r_n^2 \rangle}$$

$$= \frac{2}{\lambda_M} \frac{\bar{\lambda}}{\sum\limits_{i=N}^{N} \lambda_i} < \frac{2}{\lambda_M (2N+1)}. \tag{26}$$

Thus the maximum permissible step size for the estimated-gradient algorithm is reduced, by a factor on the order of the number of tap weights, from the maximum step size permitted in the steepest descent (known-gradient) algorithm.

By differentiating the right-hand side of (23) with respect to Δ, we obtain the step size, Δ_n^*, which provides the maximum rate of convergence (relative to the bound (23)):

$$\Delta_n^* = \frac{\bar{\lambda} q_n}{\lambda_M (2N+1) \langle r_n^2 \rangle} \cdot \frac{1}{[q_n + \mathscr{E}_{\text{opt}}]}. \tag{27a}$$

Note that Δ_n^* is a function of time, n, and the generally unknown (to the receiver) quantities q_n and $\mathscr{E}_{\text{opt}}$. During the early stages of convergence, $q_n \gg \mathscr{E}_{\text{opt}}$, so that (27) becomes the constant value

$$\Delta_0^* = \frac{\bar{\lambda}}{\lambda_M (2N+1) \langle r_n^2 \rangle} = \frac{1}{2} \Delta_{\text{MAX}}, \tag{27b}$$

and q_n converges exponentially towards a steady-state value. Thus a useful rule is: *The initial step size should be half the maximum permissible step size.* Equations (25) to (27) are similar to those proposed by Ungerboeck,[5] except for the channel-dependent factor, $\bar{\lambda}/\lambda_M$, appearing in our equations. This factor suggests a reduction in the step size for most rapid convergence with highly distorted channels.

As convergence nears completion, the steady-state step size, Δ, resulting in a specified mean-squared error, $\mathscr{E}_{\text{opt}} + q_\infty$, is found by equating the two sides of (23). Substituting the constraint, (10), into this relation gives

$$\Delta = \frac{2\bar{\lambda}}{\lambda_M} \cdot \frac{1}{(2N+1)\langle r_n^2 \rangle} \cdot \frac{q_\infty}{q_\infty + \mathscr{E}_{\text{opt}}}$$

$$= 2 \cdot \frac{\bar{\lambda}}{\lambda_M} \cdot \frac{\gamma}{1+\gamma} \cdot \frac{1}{(2N+1)\langle r_n^2 \rangle} = \frac{\gamma}{1+\gamma} \Delta_{\text{MAX}}, \tag{28}$$

† We use the fact that the trace of $A = (2N+1)\langle r_n^2 \rangle = \sum_{-N}^{N} \lambda_i$.

as the formula for the steady-state step size. Thus the steady-state step size ranges from 0 to $\frac{1}{2}\Delta_{\text{MAX}}$ as γ varies from zero to unity. Ideally, the step size should vary between the initial value, $\Delta_0^* = \frac{1}{2}\Delta_{\text{MAX}}$, and the steady-state value, $\Delta = [\gamma/(1+\gamma)]\,\Delta_{\text{MAX}} < \Delta_0^*$, in accord with (27a). In practice, the step size is generally changed in discrete steps between Δ_0^* and Δ.

In summary, (28) provides an approximation to the required steady-state step size in terms of the number of taps, the *effective eigenvalue ratio* (which depends implicitly on the number of taps), the power of the received samples, and the acceptable residual mean-squared error. Note that, as $\gamma \to 0$, we require a vanishingly small step size, and that increasing the number of taps (to achieve the desired level of $\mathscr{E}_{\text{opt}}$) also requires a diminished Δ.

III. DIGITAL CONSIDERATION

In this section, we first review[2] the effects of digital implementation on the estimated-gradient tap-adjustment algorithm, and we then combine our analog and digital results to compute the minimum precision necessary to achieve acceptable performance.

3.1 Digital cutoff of the algorithm

In Fig. 2 we sketch the evolution of the mean-squared error in high- and low-precision equalizers. In the low-precision equalizer, the steady-

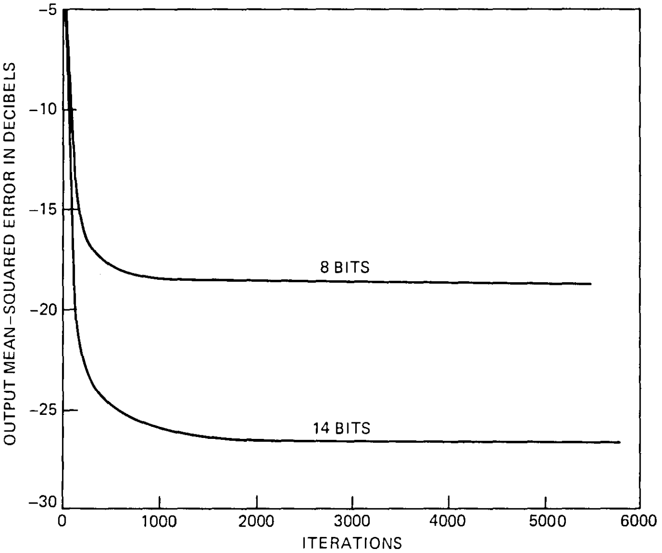


Fig. 2—Evolution of mean-squared error for 32-tap equalizers with tap weights of 8 bits and 14 bits resolution. Steady-state step size = 0.00021.

state, mean-squared error is constrained by the impossibility of changing a tap weight when the correction term in (7) decreases below half a quantization interval. Some corrections will be possible until the peaks of the correction terms fall below the critical level, i.e., until

$$|\Delta e_n r_n| \leq \tfrac{1}{2}\delta, \tag{29}$$

where δ is the interval between quantization levels, or conversely, the algorithm continues to adapt if

$$|\Delta e_n r_n| \geq \tfrac{1}{2}\delta = 2^{-B}\alpha, \tag{30}$$

where B is the number of bits (including sign) used to represent the equalizer tap weights and $(-\alpha, \alpha)$ is the range covered by the (uniform) quantizer.

The above stopping condition can be approximated by replacing the magnitude in (30) by its peak value which is assumed to be $\sqrt{2}$ times its RMS value, i.e., adaptation continues if

$$\sqrt{2}\,\Delta \cdot \sqrt{\langle e_n^2 \rangle} \cdot \sqrt{\langle r_n^2 \rangle} \geq 2^{-B}\alpha; \tag{31}$$

where the MSE which satisfies (31), with equality, will be called the digitally limited MSE. In a passband equalizer for which (30) applies separately to the in-phase and quadrature parts of the tap increment, the condition equivalent to (31) is $\Delta\sqrt{\langle |e_n|^2 \rangle \langle |r_n|^2 \rangle /2} \geq 2^{-B}\alpha$.

Two important consequences of (31) are:

(*i*) Any attempt to make Δ arbitrarily small, for the purpose of reducing q_∞, will ultimately *increase* the steady-state MSE so that (31) is satisfied.

(*ii*) The ratio of the mean-squared error of an *adaptive* digital equalizer to the MSE due to quantizing the $\{c_n\}$ to within a LSB of their optimum values in a non-adaptive equalizer grows linearly with the number of taps and the "effective" eigenvalue ratio (see Appendices II and III in Ref. 2). In other words, considerably more precision is required for adaptation of the tap weights than for filtering the received signal (performing the equalizer convolution).

Some further observations follow immediately from the above discussions and from numerical substitutions in (31):

(*i*) The number of adaptive parameters should be kept to a minimum consistent with achieving the desired steady-state MSE, since an increase in the number of taps calls for a decreased Δ, which in turn increases the precision required [from (31)].

(*ii*) The excess MSE (associated with a finite step size) evaluated in the last section can be traded against the required precision.

(*iii*) Highly dispersive channels have a larger eigenvalue spread than do good channels and thus require more precision to achieve the same MSE. However, the increased precision will be shown typically to be only 1 bit for channels of moderate distortion.

(*iv*) Digital word size will be restricted to a reasonable value if the digital equalizer is designed to allow an appropriate excess MSE on the order of $\mathscr{E}_{\text{opt}}$.

3.2 Required precision

We now presume that the steady-state step size Δ, determined from (28), is used in a *digitally implemented* equalizer. The analog parameters are assumed such that the steady-state MSE is equal to the digitally limited MSE. Thus performance will be determined by the available precision. Hence the digital word length, B bits, needed in the tap weights to achieve $\mathscr{E}_{\infty}$ can be determined by substituting (28) in the digital stopping condition (31):

$$2\frac{\bar{\lambda}}{\lambda_M}\frac{\gamma}{1+\gamma}\frac{1}{(2N+1)}\cdot\frac{1}{\langle r_n^2\rangle}\cdot\sqrt{\langle e_n^2\rangle\langle r_n^2\rangle} \geq 2^{-B}\alpha, \tag{32}$$

which reduces to the important relation

$$2^{-B}\alpha \leq 2\frac{\bar{\lambda}}{\lambda_M}\frac{\gamma}{1+\gamma}\cdot\frac{1}{(2N+1)}\cdot\frac{1}{\sqrt{\text{SNR}\cdot\rho}}, \tag{33}$$

where

$$\text{SNR} = \frac{S_{\text{out}}}{\langle e_n^2\rangle} \tag{34}$$

is the (equalized) output signal-to-noise ratio, and where

$$\rho = \frac{\langle r_n^2\rangle}{S_{\text{out}}} \tag{35}$$

is the ratio of input signal power to output (baseband) signal power.† Note that the equalized mean-squared error appearing in (32) can be written as $\langle e_n^2\rangle = \mathscr{E}_{\text{opt}} + q_{\infty} = (1+\gamma)\mathscr{E}_{\text{opt}} = (1+\gamma)\,[1 - \mathbf{x}'A^{-1}\mathbf{x}]$, where the last equality follows from (6). Thus for a given (or known) channel, all the terms (33) can be readily computed. We now use (33) to estimate the required digital word length under various conditions typical of 9.6-kb/s data transmission.‡

(*i*) *Operation on a Good Channel* (Fig. 3a). A good channel is one for which $\bar{\lambda}/\lambda_M \cong 1$, and the number of equalizer taps can be quite small. In practice, however, a synchronously spaced equalizer will have a fixed number of taps, typically 32. With a passband equalizer[1] having 32 complex tap pairs, the effective eigenvalue ratio is 0.93 for the

† For the complex passband equalizer, the factor on the right-hand side of (33) is $\sqrt{2}$ instead of 2, and $\langle |r_n|^2\rangle$ replaces $\langle r_n^2\rangle$ in (34).

‡ It should be noted that the word lengths derived in the following examples are considerably longer than the precision indicated from using a variance equal to $\sigma^2 = \delta^2/12$ for the quantization error in each tap weight.

"good" channel using a near-optimum sampling epoch. We assume that the maximum quantization level is $\alpha = 1$, that $\gamma = ½$ (corresponding to a 1.1-dB degradation in output s/n ratio), and an output s/n ratio of 25.7 dB is observed (down 1.1 dB from that observed with effectively infinite resolution and vanishing step size). An AGC setting is assumed such that $\rho = 2$. We find from (33) that $B \approx 11$ bits.

(*ii*) *Operation on a Severely Distorted Channel* (Fig. 3b). This severely distorted channel, which just meets the requirements of basic voice-grade line conditioning, has been found to have an effective eigenvalue ratio $\bar{\lambda}/\lambda_M$ (for the best sampling phase) of 0.5. Using this latter value and with the parameters of the previous sample, except for an output s/n ratio of 23.4 dB associated with the distorted channel, we find that $B \approx 11.5$ bits.

(*iii*) *Operation with a Fractionally Spaced Equalizer on the Distorted Channel.* With channel samples taken at $T/2$ intervals, where T is a symbol interval, a 64-tap equalizer is appropriate. The effective eigenvalue ratio† is 0.58, and we find that $B \approx 12$ bits for the same 1.1-dB degradation used in the above examples.

In the next section, the precision predictions of these three examples are compared with simulation results.

IV. SIMULATIONS

The same channels for which eigenvalue ratios were derived for the examples of the last section were used in a simulation program for a QAM data communication system[1] operating at 9600 bps with a baud of 2400/s. Only the tap weights were quantized (rounded to the nearest quantization level); all other variables had the IBM 370 single-precision resolution of roughly 24 bits. T|he magnitude of the largest real tap weight was about 0.5 in a full quantization range of (−1, 1). Timing and phase references were ideal and not subject to statistical fluctuation. The steady-state step size used in each case was computed from (28). The equalizer was either a 32 (complex) tap synchronous (tap spacing = T = symbol interval) structure, or a 64 (complex) tap $T/2$ structure.

Some simulation runs were made with a gear shifting sequence of adaptation step sizes[4] to reach convergence within a reasonable number of iterations. However, great care was taken to reach the smallest (steady-state) step size well before complete convergence. This is because, with a larger step size, digital equalizer performance can possibly be better for a transient period than that corresponding to the chosen steady-state value. Deterioration of this "good" performance, once it is achieved, depends upon large signal and/or noise

† The "average" tap weight was the average over the 26 tap weights which collectively contained 95 percent of the tap-weight mass.

values, and may not be observed over the short duration of a simulation run.

Curves A in Fig. 4, for operation with the synchronous equalizer on the "good" channel, indicate a degradation of about 1.5 dB for the step size of 0.001 calculated from (28). This is not far from the 1.1 dB ($\gamma = 0.5$) used in that formula; however, the s/n ratio degrades another 0.6 dB when the predicted digital word size of 11 bits is used instead of infinite resolution. The source of this additional degradation has not been investigated, but ordinary quantization noise will make a contribution, and probably also a slowed rate of adaptation, just before the

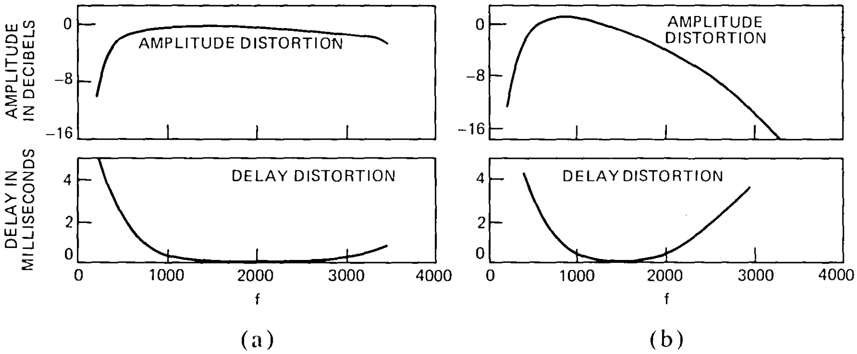


Fig. 3—(a) "Good" and (b) "distorted" channels used in analytical and simulation examples.

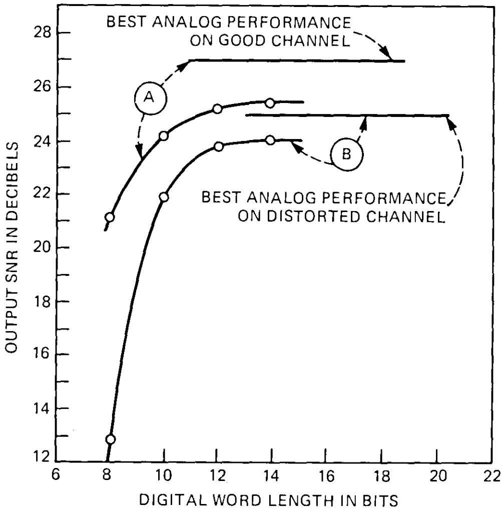


Fig. 4—Measured output s/n ratio vs digital word length (of equalizer tap weights) for (A) a "good" channel (eigenvalue ratio = 0.93), $\Delta = 0.001$; and (B) a distorted channel (eigenvalue ratio = 0.5), $\Delta = 0.0005$. A 32-tap synchronous equalizer was used in both cases.

digital stopping condition prevails. Curve B, for operation on the distorted channel, is down 1.1 dB for a digital word length very close to the predicted 11.5 bits.

Figure 5 illustrates a similar curve for operation on the distorted channel with the 64-tap $T/2$ equalizer, as described in example (*iii*). The digital resolution of about 12 bits predicted from (27) for 1.1-dB degradation is consistent with the simulation results. Figure 6 presents another view of the performance of the $T/2$ equalizer on the distorted channel. This is a curve of output s/n ratio vs adaptation step size for the 12-bit resolution determined from (33). The step size of 0.0003 determined from (28) and used in the experiments represented as points on the curve of Fig. 4, corresponds to a near-peak value of output s/n ratio. This again supports the analysis of Section III as providing a useful formula for deciding on the steady-state step size to be used in a digitally implemented equalizer.

V. DISCUSSION AND CONCLUSIONS

We have proposed a criterion for determining the number of bits needed to represent the tap weights in a digitally implemented equalizer. For a given steady-state adaptation step size, with its attendant increase in steady-state mean-squared error, this criterion is that the word size used be just large enough to "match" this increase without further degrading performance. The word length is a function of the output s/n ratio (the ratio of output signal power to steady-state mean-squared error), the fractional increase in the mean-squared error over

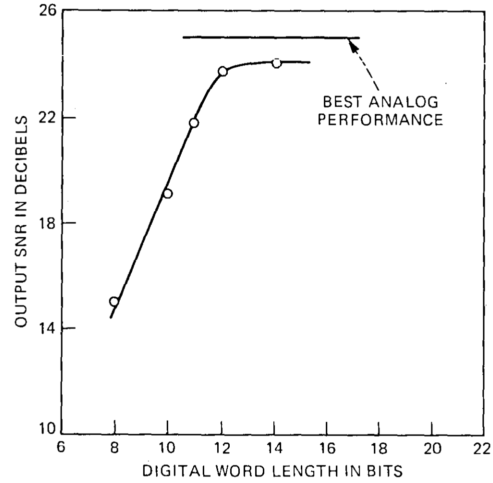


Fig. 5—Output s/n ratio vs digital word length (of equalizer tap weights) for a 64 $T/2$ tap equalizer operating on the severely distorted channel of Fig. 1, with step size = 0.0003 determined from (28).

that of an ideal infinite-resolution receiver, the number of taps, and the effective eigenvalue ratio of the channel-covariance matrix.

For data transmission at 9.6 kb/s, it can be concluded, on the basis of the representative voice-grade channels simulated in this study and neglecting the effects of limited tap-weight resolution on timing recovery and carrier phase tracking as well as the (relatively minor) degradations resulting from limiting the resolution of variables other than the tap weights, that a digital word length of 12 bits is adequate to represent the tap weights for updating purposes in a 32-tap synchronous equalizer or a 64-tap $T/2$ equalizer.

APPENDIX

Asymptotic Eigenvalue Distribution for the Correlation Matrix of Synchronous and Fractionally Spaced Equalizers

In this appendix, we describe the eigenvalues of the correlation matrix for infinitely long synchronous and fractionally spaced equalizers.

A.1 Synchronous equalizer

From the definition $A = \langle \mathbf{r}_n \mathbf{r}'_n \rangle$, we note that A is a Toeplitz matrix, and that the eigenvalue equation is given by

$$\sum_{\ell=-N}^{N} A_{k-\ell} p_\ell = \lambda p_k \qquad -N \le k \le N, \tag{36}$$

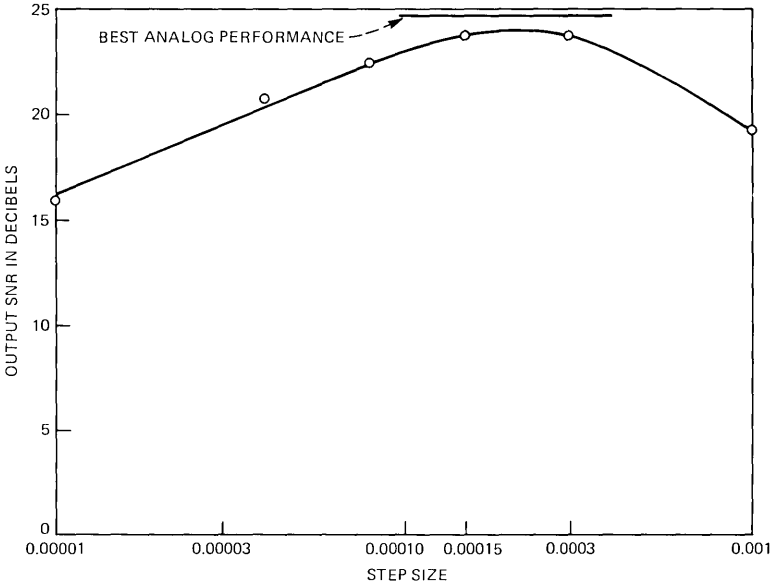


Fig. 6—Output s/n ratio vs adaptation step size for 64 $T/2$ tap equalizer, tap weights quantized to 12 bits, and severely distorted basic-conditioned channel.

where $A_{k-\ell}$ is the $k\ell$th element of A, λ is an eigenvalue, and $\mathbf{p}' = (p_{-N}, \cdots, p_0, \cdots, p_N)$ is the associated eigenvector. As $N \rightarrow \infty$, taking the Fourier transform of both sides of (36) yields

$$A(\omega)P(\omega) = \lambda P(\omega), \qquad |\omega| \leq \frac{\pi}{T}, \tag{37}$$

where

$$A(\omega) = \left|\sum_k X\left(\omega + k\frac{2\pi}{T}\right)\right|^2 + \sigma^2 = |X_{\text{eq}}(\omega)|^2 + \sigma^2,$$

$$|\omega| \leq \frac{\pi}{T}. \quad (38)\dagger$$

For very large N, the discrete asymptotic approximation to (37) is

$$A(\omega_i)P(\omega_i) = \lambda P(\omega_i), \qquad \omega_i = \frac{2\pi i}{(2N+1)T}, \qquad -N \leq i \leq N. \tag{39a}$$

Unless $A(\omega_i)$ has the same value for two or more values of the index i, the only way (39a) can be satisfied is for $P(\omega)$ to be concentrated at a single frequency, i.e., be a sinusoid. Repeated values of $A(\omega_i)$ correspond to repeated eigenvalues and an eigenvector subspace which can be spanned either by distinct sinusoids or by combinations of sinusoids. Then the solution to (37) is

$$\begin{aligned} \lambda_i &= A(\omega_i) \\ & \qquad\qquad\qquad -N \leq i \leq N. \\ P_i(\omega) &= \delta(\omega - \omega_i) \pm \delta(\omega + \omega_i) \end{aligned} \tag{39b}$$

Thus for a synchronous equalizer, the asymptotic ($N \rightarrow \infty$) eigenvalues uniformly sample the folded-channel-plus-noise spectrum, and the eigenvectors are the corresponding sinusoids.

A.2 Fractionally spaced equalizers

Here the channel-correlation matrix while symmetric is *not* Toeplitz; thus Fourier transform techniques do not yield the eigenvalues and eigenvectors in the above short order. For convenience, we consider the noiseless situation, and the eigenvalue equation is

$$\sum_{\ell=-N}^{N} A(kT', \ell T')p(\ell T') = \lambda p(kT') \qquad -N \leq k \leq N, \tag{40}$$

† The Nyquist-equivalent spectrum, $X_{\text{eq}}(\omega)$, is defined as $X_{\text{eq}}(\omega) = \sum_k X[\omega + k(2\pi/T]$, where the receiver sampling phase is incorporated into $X(\omega)$.

where the $k\ell$th element of A is given by

$$A(kT', \ell T') = \sum_m x(mT - kT')x(mT - \ell T'). \qquad (41)$$

With $T' = T/2$, we write (40) for even and odd values of k,

$$\sum_{\substack{\ell \text{even} \\ |\ell| \le N}} A\left(k\frac{T}{2}, \ell\frac{T}{2}\right)p\left(\ell\frac{T}{2}\right) + \sum_{\substack{\ell \text{odd} \\ |\ell| \le N}} A\left(k\frac{T}{2}, \ell\frac{T}{2}\right)p\left(\ell\frac{T}{2}\right) = \lambda p\left(k\frac{T}{2}\right), \quad k \text{ even} \qquad (42)$$

$$\sum_{\substack{\ell \, even \\ |\ell| \le N}} A\left(k\frac{T}{2}, \ell\frac{T}{2}\right)p\left(\ell\frac{T}{2}\right) + \sum_{\substack{\ell \text{odd} \\ |\ell| \le N}} A\left(k\frac{T}{2}, \ell\frac{T}{2}\right)p\left(\ell\frac{T}{2}\right) = \lambda p\left(k\frac{T}{2}\right), \quad k \text{ odd}. \qquad (43)$$

Now (42) and (43) can be written as

$$\sum_{\ell=-N/2}^{N/2} A(kT, \ell T)p(\ell T) + \sum_{\ell=-N/2}^{N/2} A\left(kT, \ell T + \frac{T}{2}\right)p\left(\ell T + \frac{T}{2}\right) = \lambda p(kT), \qquad -N/2 < k < N/2 \qquad (44)$$

$$\sum_{\ell=-N/2}^{N/2} A\left(kT + \frac{T}{2}, \ell T\right)p(\ell T) + \sum_{\ell=-N/2}^{N/2} \cdot A\left(kT + \frac{T}{2}, \ell T + \frac{T}{2}\right)p\left(\ell T + \frac{T}{2}\right) = \lambda p\left(kT + \frac{T}{2}\right), \qquad (45)$$

respectively, where both equations hold for N integer values of k, and more importantly the various component matrices are now all Toeplitz.† If we consider the situation where $X(\omega)$ has less than 100 percent excess bandwidth,[8] then it is useful to introduce the four spectra

$$\begin{aligned} X_{\text{eq}}(\omega) &\triangleq X(\omega) + X\left(\omega - \frac{2\pi}{T}\right) + X\left(\omega + \frac{2\pi}{T}\right) \\ \tilde{X}_{\text{eq}}(\omega) &\triangleq X(\omega) - X\left(\omega - \frac{2\pi}{T}\right) - X\left(\omega + \frac{2\pi}{T}\right), \qquad |\omega| \le \frac{\pi}{T} \qquad (46) \\ P_{\text{eq}}(\omega) &\triangleq P(\omega) + P\left(\omega - \frac{2\pi}{T}\right) + P\left(\omega + \frac{2\pi}{T}\right) \end{aligned}$$

† For example, $A(kT, \ell T + (T/2)) = \sum_m x(mT - kT)x(mT - \ell T + (T/2)) = \sum_n x(nT)x(nT + (k - \ell)T + (T/2))$.

$$\tilde{P}_{\text{eq}}(\omega) \triangleq P(\omega) - P\left(\omega - \frac{2\pi}{T}\right) - P\left(\omega + \frac{2\pi}{T}\right), \qquad |\omega| \leq \frac{\pi}{T}, \tag{47}$$

where we note that the discrete Fourier transform of the eigenvector $p(\ell(T/2))$, $P(\omega)$, is given by

$$\begin{aligned} P(\omega) &\triangleq \sum_{\ell=-N}^{N} p\left(\ell \frac{T}{2}\right) e^{-j\omega_j \ell(T/2)} \\ &= P_{\text{eq}}(\omega_j) + e^{-j\omega_j(T/2)} \tilde{P}_{\text{eq}}(\omega_j), \qquad \omega_j = \left(\frac{j}{N}\right)\left(\frac{\pi}{T}\right), \ -N \leq j \leq N. \end{aligned} \tag{48}$$

Taking the *synchronous* Fourier transform (i.e., with respect to the T-sec sampling interval) of (44) and (45) gives as an approximation (exact as $N \rightarrow \infty$)

$$\begin{aligned} &|X_{\text{eq}}(\omega_j)|^2 P_{\text{eq}}(\omega_j) + X_{\text{eq}}(\omega_j) \tilde{X}^*_{\text{eq}}(\omega_j) \tilde{P}_{\text{eq}}(\omega_j) = \lambda P_{\text{eq}}(\omega_j) \\ &\qquad\qquad\qquad\qquad \omega_j = \left(\frac{2j}{N}\right)\left(\frac{\pi}{T}\right), \ -N/2 \leq j \leq N/2. \\ &\tilde{X}_{\text{eq}}(\omega_j) X^*_{\text{eq}}(\omega_j) P_{\text{eq}}(\omega_j) + |\tilde{X}_{\text{eq}}(\omega_j)|^2 \tilde{P}_{\text{eq}}(\omega_j) = \lambda \tilde{P}_{\text{eq}}(\omega_j). \end{aligned} \tag{49}$$

Note that $p(kT + (T/2))$ has the discrete Fourier transform exp $(-j\omega_j(T/2)\ \tilde{P}_{\text{eq}}(\omega_j))$, where the synchronous transform of $p(kT)$ is $P_{\text{eq}}(\omega_j)$. Arguing as we did for the synchronous equalizer, we see that the ith eigenvectors $P_i(\omega_j)$ and $\tilde{P}_i(\omega_j)$ must again be delta functions at $\omega_i = (2i/N)(\pi/T)$. Consequently, the eigenvalues, λ_i, must satisfy

$$\lambda_i^2 - \lambda_i[\,|X_{\text{eq}}(\omega_i)|^2 + |\tilde{X}_{\text{eq}}(\omega_i)|^2] = 0, \tag{50}$$

and thus the eigenvalues are

$$\begin{aligned} \lambda_i^{(1)} &= 0 \\ \lambda_i^{(2)} &= |X_{\text{eq}}(\omega_i)|^2 + |\tilde{X}_{\text{eq}}(\omega_i)|^2 = \sum_k \left| X\left(\omega_i + \frac{k2\pi}{T}\right)\right|^2, \\ &\qquad\qquad\qquad\qquad -\frac{N}{2} \leq i \leq \frac{N}{2}. \end{aligned} \tag{51}$$

In contrast to (38), which applies to the synchronous equalizer, half the eigenvalues are exactly zero, while the other half are samples of the aliased magnitude-squared channel transfer function. Not surprisingly, the eigenvalues are independent of both the channel phase characteristics and the receiver sampling phase, and if $|X(\omega)|^2$ is Nyquist, then all the eigenvalues are unity. Since the eigenvalues are determined, we can now solve for the eigenvectors. Since $p(\ell T)$ has the transform $P_{\text{eq}}(\omega)$ and $p(\ell T + (T/2))$ has the transform $\tilde{P}_{\text{eq}}(\omega)\ e^{-j\omega(T/2)}$, the eigenvectors associated with the zero eigenvalue are constructed as

$$p_i\left(n\,\frac{T}{2}\right) = \begin{cases} \tilde{X}_{\text{eq}}(\omega_i)e^{j\omega_i n\,(T/2)}, & n \text{ even} \\ -X_{\text{eq}}(\omega_i)e^{j\omega_i n\,(T/2)}, & n \text{ odd,} \end{cases} \tag{52}$$

while the eigenvector associated with the nonzero eigenvalue is

$$p_i\left(n\,\frac{T}{2}\right) = \begin{cases} X_{\text{eq}}(\omega_i)e^{j\omega_i n\,(T/2)}, & n \text{ even} \\ \tilde{X}_{\text{eq}}(\omega_i)e^{j\omega_i n\,(T/2)}, & n \text{ odd.} \end{cases} \tag{53}$$

At this point, we remark that when ω_i is not in the rolloff region, then $X_{\text{eq}}(\omega_j) = \tilde{X}_{\text{eq}}(\omega_i)$, and (53) describes a sinusoid of frequency ω_i, since the even and odd portions of $p_i(n(T/2))$ mesh together in a continuous manner (i.e., $p_i(n(T/2)) = X_{\text{eq}}(\omega_i)e^{j\omega_i n\,(T/2)}$). However, (52) describes a function which changes sign and oscillates almost a full cycle in T seconds. Consequently, $p_i(n(T/2))$, as given by (52), will have most of its spectral energy concentrated near $1/T$ Hz. When ω_i is in the rolloff region, the frequency content of (52) and (53) will differ somewhat from the above extreme cases, but the general results will still be as above. Numerical evaluations have confirmed the above.

This completes our discussion concerning the nature of the eigenvalues and eigenvectors for a fractionally spaced equalizer.

REFERENCES

1. D. D. Falconer, "Jointly Adaptive Equalization and Carrier Recovery in Two-Dimensional Digital Communication Systems," B.S.T.J., *55,* No. 3 (March 1976), pp. 317–334.
2. R. D. Gitlin, J. E. Mazo, and M. G. Taylor, "On the Design of Gradient Algorithms for Digitally Implemented Adaptive Filters," IEEE Trans. Circuit Theory, *CT-20,* No. 2 (March 1973).
3. B. Widrow et al, "Noise Cancellation Using the LMS Algorithm," Proc. IEEE, December 1975.
4. R. R. Anderson, R. D. Gitlin, and S. B. Weinstein, unpublished work.
5. G. Ungerboeck, "Theory on the Speed of Convergence in Adaptive Equalizer for Digital Communication," IBM J. Res. Develop., November 1972, pp. 546–555.
6. G. Ungerboeck, "Fractional Tap-Spacing Equalizer and Consequences for Clock Recovery in Data Modems," IEEE Trans. Commun., *COM-24* (August 1976), pp. 856–864.
7. S. U. H. Qureshi and G. D. Forney, Jr., "Performance and Properties of a $T/2$ Equalizer," Conference Record NTC 1977, December 1977.
8. R. W. Lucky, J. Salz, and E. J. Weldon Jr., *Principles of Data Communication,* New York: McGraw-Hill, 1968.

Copyright © 1979 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 58, No. 2, February 1979
Printed in U.S.A.

Wear of Gold Electrodeposits: Effect of Substrate and of Nickel Underplate

By M. ANTLER and M. H. DROZDOWICZ

(Manuscript received May 8, 1978)

The adhesive and abrasive wear of electrodeposited gold was studied and the effect of a ductile hard underplate (nickel) and a hard substrate (beryllium copper) determined. Wearing conditions experienced by connector contacts were modeled with rider-on-flat apparatus. It was found that: (i) both the adhesive and abrasive wear of gold plate can be markedly reduced by hard substrates and underplates, (ii) thin film lubrication is effective when adhesive wear predominates, but is of little value in abrasive wear, (iii) there is a load-dependent minimum thickness of gold plate that will not wear through because of a change in mechanism which occurs during unlubricated sliding in the adhesive regime, (iv) sliding at mild conditions, as in the presence of a good boundary lubricant, occurs with lateral flow of gold which can close as-plated pores and wear-induced breaks produced early in a run, (v) pure (soft) gold platings may be able to resist abrasive wear at loads where hard golds fail by brittle fracture, and (vi) the prow formation mechanism adequately describes the adhesive wear of gold plate. Nickel underplate may be desirable on connector contacts to minimize the wear of gold plate, especially at high loads, with thin golds, where it is impractical to use lubricants and where abrasion is the dominant mechanism.

I. INTRODUCTION

This investigation concerns the wear of gold electrodeposits such as are used on the contacts of electronic connectors. The wear of contact finishes is important to connector reliability because, if base metal is exposed, oxide or corrosion films form which may lead to unacceptably high contact resistance.

Connector contact surfaces generally consist of gold platings hardened with from 0.1 to 0.5 weight percent of codeposited cobalt or nickel and having a thickness in the range of 0.5 to 5 μm. Underplatings of copper or ductile nickel are often used, chosen for their expected value in retarding substrate diffusion and for their effect in mitigating the formation of corrosion and tarnish films which can develop in aggressive environments at pores in the gold deposit. Wrought pure gold and silver-gold alloy weldments or claddings are also employed, generally on only one member of a contact pair. Among the common substrates are copper and copper alloys, as in the edge contacts of printed circuit boards and the spring elements of connectors. Lubricants may also be used to lower insertion forces and to reduce wear.

Prior research in the wear of gold has been summarized.[1] A role of underplate was found by Holden[2] and Solomon and Antler,[3] who recognized that contact finishes having low deformability are desirable and that the effective hardness of a multi-layer finish can be increased by the use of some underplatings, such as nickel. Studies with connectors have borne out their predictions,[4,5] but it is also common experience that nickel underplating may not change the wear of gold in some cases and, when associated with poor finishing practices which lead to nonadherent coatings or nodular surfaces, may even degrade wear behavior.

It was the objective of this work to determine the mechanisms by which nickel underplate and substrate metals control the wear of gold plate. If the role of underplate and substrate could be understood, the selection of contact materials might be better made than at present.

The wear of gold in connectors occurs by adhesion, abrasion, and brittle fracture.

Adhesive damage occurs when there is metal transfer. Adhesive bonds form that are stronger than the cohesive strength of the metal. These bonds lead to transfer and wear as sliding continues.

Wear generated by plowing of a surface by an opposing member which is rough and substantially harder is termed two-body abrasion. Loose hard particles between sliding surfaces causes three-body abrasion of one or both members. Abrasive wear of connector contacts originates in misalignment, parts having burrs or similar defects, work-hardened transfer and loose particles from adhesive wear, and foreign materials such as dusts and debris from connector structures or printed circuit boards. The abrasive wear of gold has been little studied.

Fracture wear occurs with brittle platings, especially when the substrate is deformable. The surface develops cracks during sliding and may result in catastrophic loss of the coating. Since the common cobalt- and nickel-hardened gold electrodeposits have elongations of less than one percent,[6] brittle fracture can be expected to contribute to the wear of gold contacts.

The experimental approach of this investigation was to model wearing conditions experienced by connector contacts using a rider-on-flat apparatus. It will be shown that: (*i*) both the adhesive and abrasive wear of gold plate can be markedly reduced by nickel underplate and by hard substrates, although small thicknesses of nickel may be ineffective, (*ii*) thin film lubrication is effective for wear reduction when adhesive wear predominates, but is of little value in abrasive wear, (*iii*) a load-dependent minimum thickness of gold plate will not wear through because of a change in mechanism which occurs during unlubricated sliding in the adhesive regime, (*iv*) sliding at mild conditions, as in the presence of a good boundary lubricant, occurs with significant lateral flow of gold which can close as-plated pores and wear-induced breaks produced early in a run, (*v*) pure (soft) gold platings may be able to resist abrasive wear at loads where hard golds fail by brittle fracture, and (*vi*) the prow formation mechanism adequately describes the adhesive wear of gold plate.

II. EXPERIMENTAL

2.1 *Materials*

2.1.1 *Riders*

In the study of adhesive wear, hemispherically ended smooth solid gold riders having a diameter of 3.2 mm were used. The purity of the gold was 99.99 percent, with a typical hardness of 65 KHN$_{25}$. A newly turned rider was employed for each run.

Riders used in three-body abrasion studies were 3.2-mm diameter solid rhodium rods, 99.9 percent pure, having a hardness of 370 KHN$_{250}$. The riders, which initially had hemispherical ends, were installed in the wear apparatus and then abraded with silicon carbide metallographic paper placed on the specimen table in place of the flat until a circular area of 0.5-mm diameter on the rider had been produced. Rhodium was used instead of gold to minimize embedding of powdered abrasive in the rider surface, and having flat mating surfaces assured that abrasive particles would be trapped between the specimens.

Runs in two-body abrasion were with a 90-degree conical diamond rider having a rounded tip with a radius of 0.1 mm.

2.1.2 *Flats*

The flats were a soft and a hard material, oxygen-free copper and beryllium copper alloy, respectively, and are described in Table I. They were randomly abraded on metallographic paper prior to plating.

Table I—Substrates

	Copper, Oxygen Free	Beryllium Copper
Size, cm	1.3 × 3.8	1.3 × 3.8
Thickness, mm	3.2	4.9
Composition, wt. %	99.95 Cu	97.9 Cu, 1.9 Be, 0.2 Ni or Co
Hardness, kg/mm^2 (cross-sections)	40–60 KHN$_{25}$	266 KHN$_{25}$
Roughness after plating, μm CLA	0.4 ± 0.08	0.2 ± 0.005

Platings were cobalt-hardened gold from an acid cyanide bath, pure gold from a citrate-buffered acid solution of $KAu(CN)_2$, and nickel from a sulfamate bath (Table II). The cobalt golds in three thicknesses and the pure gold were plated both with and without nickel underplate. The golds had low intrinsic porosity and thus were suitable to a chemical method (electrography) for assessing wear-through of the deposit. The cobalt gold contained codeposited polymer,[7] and was typical of platings used on connector contacts.

Immediately prior to a run, specimens were cleaned by immersion in multiple baths of reagent grade 1,1,1-trichloroethane and methanol.

2.1.3 Lubrication

Lubricant was a liquid polyphenyl ether, Monsanto OS-124.[8] It was applied to the cleaned flats by immersion and withdrawal at room temperature from a 0.5-percent solution in 1,1,1-trichloroethane. The solvent quickly evaporated, leaving a thin residual film of lubricant.

2.1.4 Abrasive

In three-body sliding, graded boron carbide powder consisting of equiaxed particles, approximately 50 μm across, was applied by sifting particles onto lubricated flats. About 50 percent of the specimen surface was covered with abrasive powder.

2.2 Apparatus

Two rider-flat machines were used. They were similar, involving dead-weight loading of a stationary rider against the moving flat. The riders were free to move vertically so as to accommodate wear debris, roughness, and other irregularities.

The machine for studying adhesive wear and three-body abrasive wear has been described earlier,[9] and was used in reciprocating mode. Contact members were observed during sliding at 30 diameters with a stereomicroscope. Test conditions were: tracks, slightly curved, with a

Table II—Electrodeposits

	Co Gold	Pure Gold	Nickel (Underplate)
Avg. thickness μm*	0.75, 2.0, 3.3	3.3	1.5, 2.5, 4.0
Composition, wt. % metals†	0.26 Co 0.20 K ‡	‡	
Carbon§	0.26 C		
Hardness, kg/mm^2¶	180 KHN$_{25}$	90 KHN$_{25}$	550 KHN$_{10}$

* Thickness determined from metallographic sections. Variability among replicate samples was ±15 percent.

† Analysis by atomic absorption method.

‡ Other metals not detected, or present in only trace amount.

§ Analysis by microcombustion.

¶ Hardness determined on metallographic sections.

length of 1 cm; average velocity, 0.3 cm/s; loads, 20–500g; numbers of passes (a to-and-fro traversal is two passes), usually from 10 to 700. Some runs were for 1 or 4 passes, and others continued for up to 4000 passes.

Two-body abrasion was studied by sliding for 1 pass. Test conditions were: tracks, straight, with a length of 1 cm; velocity, 0.25 cm/s; loads, 10–500g.

One to four runs were made at each test condition, and electrographic Wear Indexes, defined below, were calculated by averaging results from the individual experiments.

2.3 Determination of wear

Four different methods were employed to determine wear: surface examination with the light and scanning electron microscopes (SEM), electrography, profilometry, and metallographic sectioning across the wear track and observation of the section with the light and the scanning electron microscopes. The electrographic and profilometric methods are described in detail.

2.3.1 Electrography

Electrography[10,11] is useful for detecting loss of gold from a surface which results in exposure of base underplate or substrate. In this method, base metal at breaks in the gold is caused to transfer to chemically treated paper that is pressed against the specimen, its location signified by colored spots which appear in the paper. The procedure is given in the appendix to this paper.

Wear tracks were observed in the electrographic print at 10 diameters. Decorations consisted of discrete spots along the wear track and grew in number and size with increasing wear, e.g., in a series of runs with increasing numbers of passes. Often there were multiple spots across the width of a track. Figure 1 shows a typical electrographic print and the test specimen from which it was made.

It was convenient in quantifying wear to determine the ratio of length of track in the electrograph having decorated features to total track length. This ratio, multiplied by 100, is defined as the "Wear Index." Determinations of Wear Index could be made quickly with a calibrated eyepiece, and reproducibility in repeat observations of the same print was about 3 percent. This is less than the variability of wear from replicate runs.

2.3.2 Profilometry

Wear occurs with loss or displacement of metal, or, if transfer predominates, an increase in surface roughness. Since exposure of base metal is of most relevance in contact studies, a profilometric analysis

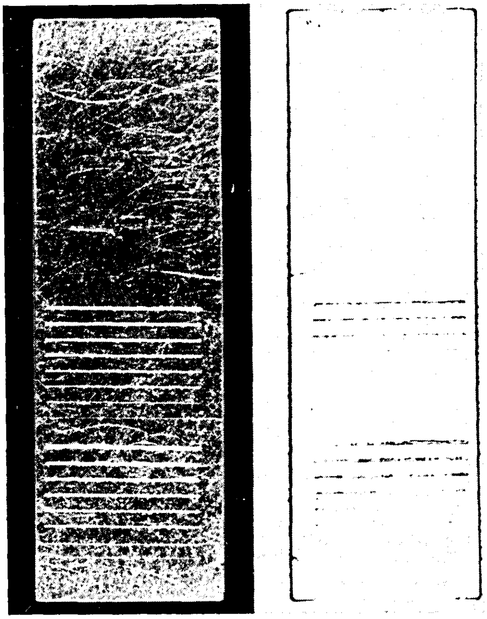


Fig. 1—Typical specimen and electrographic print.

was made which emphasized the maximum depth in the wear tracks compared to the level of unworn surface.

Three traversals across each wear track were made with a stylus instrument and charts of stylus excursions obtained. The average of the maximum depths in the three tracks with respect to the levels of the original specimen surface was then calculated. This is termed "depth of wear."

It was found that wear analysis by profilometry gave nearly the same results as analysis based on the electrographic Wear Index, as shown later for a typical case in unlubricated adhesive wear. Accordingly, profilometry was not used routinely because of its relative complexity.

III. OBSERVATIONS

3.1 Adhesive wear

3.1.1 Unlubricated

Solid gold riders and flats plated with several thicknesses of cobalt gold electrodeposits on copper, with and without nickel underplate, were used. The effect of nickel underplate was also determined with pure gold plate on copper and with cobalt gold plate on beryllium copper. The analysis of wear by electrography is given in Figs. 2, 3, and 4.

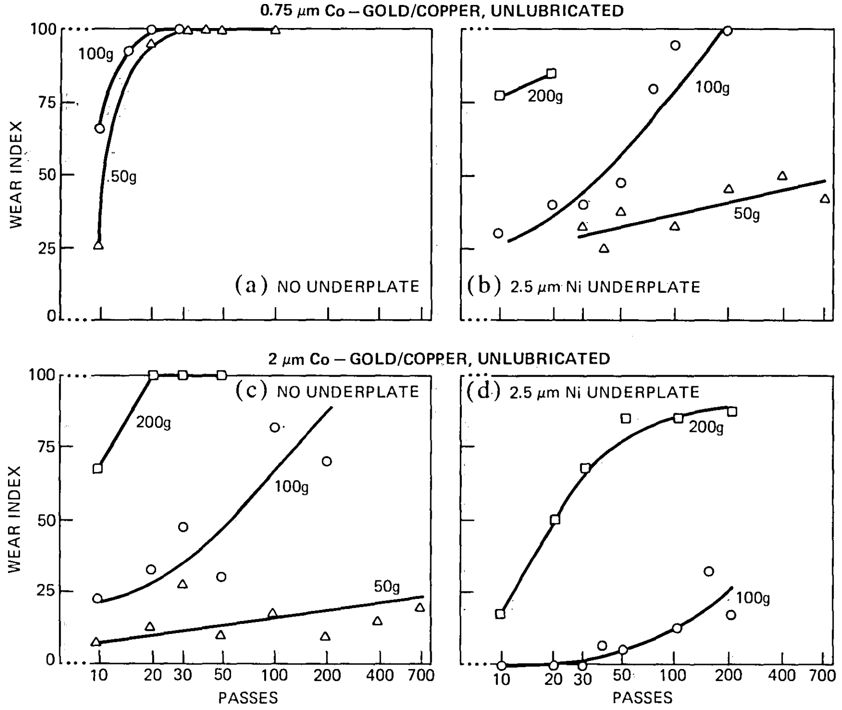


Fig. 2—Electrographic Wear Indexes from unlubricated adhesive wear runs with 0.75 and 2 μm cobalt gold electrodeposits on copper with and without 2.5 μm nickel underplate.

Figures 2a and 2b illustrate the large reduction in wear of 0.75-μm cobalt gold deposits obtained with nickel underplate. For example, a Wear Index of 100 occurs at 100g within 20 passes without the nickel, while 200 passes are required to produce equivalent wear with 2.5 μm of nickel. Figure 2c shows that roughly comparable wear results with 2 μm of cobalt gold plated directly on copper and with 0.75 μm of cobalt gold having nickel underplate, and Fig. 2d shows nickel is able to effect an equally dramatic reduction in the wear of 2-μm cobalt gold deposits.

Figure 3 continues this analysis with 3.3 μm of cobalt gold with nickel thickness being varied, including 1.5, 2.5, and 4-μm underplatings. In Fig. 3a, thick cobalt gold plate on copper is found to be more durable than thinner golds (Figs. 2a and 2c), but again nickel underplate is further able to reduce wear, although the improvement is less dramatic than when nickel is used with thinner gold finishes. Thus, at 200g a Wear Index of 50 is achieved in about 60 passes without nickel (Fig. 3a), and in about 120 passes with 2.5 μm of nickel (Fig. 3c). The sample having 1.5 μm of nickel underplate (Fig. 3b) is not substantially better at 200 and 300g than when nickel-free, although there is im-

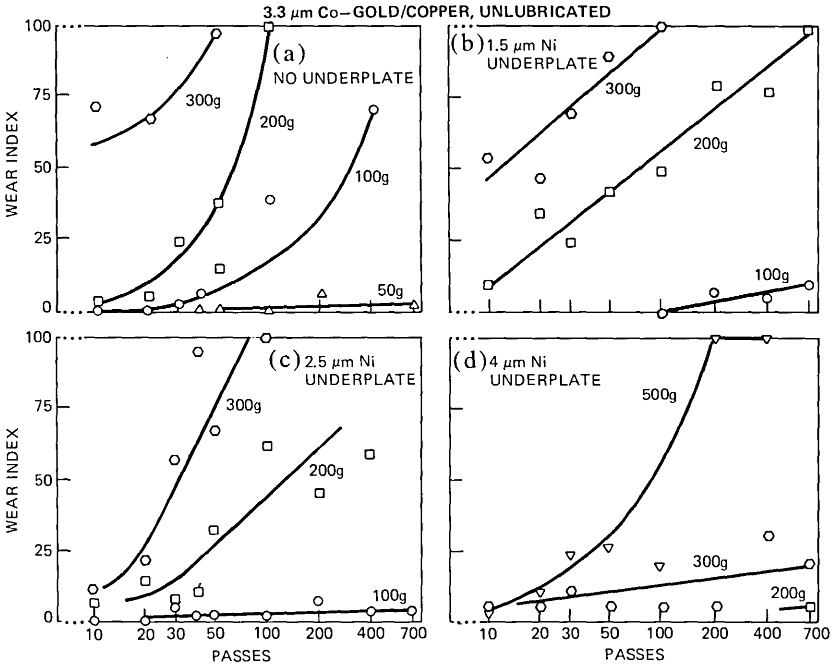


Fig. 3—Electrographic Wear Indexes from unlubricated adhesive wear runs with 3.3 μm cobalt gold electrodeposits on copper with various thicknesses of nickel underplate.

provement at 100g. On the other hand, the sample with the thickest nickel underplate in this study, 4 μm (Fig. 3d), is the most durable of the series with copper substrate; it was possible to slide at 500g, a load which caused catastrophic seizure when attempted with many of the other platings described in Figs. 2 and 3.

In Fig. 4a, the hard substrate, beryllium copper, is used with 2 μm of cobalt gold plate, and the results can be compared with Fig. 2c, cobalt gold on copper. Beryllium copper improves the wear of gold even more dramatically than does the intermediate (2.5 μm) thickness of nickel underplate (Fig. 2d) when on a copper substrate. Thus, Fig. 2c shows that at 200g a Wear Index of 100 was obtained within 20 passes with the copper substrate, while 700 passes were required for this to occur with beryllium copper. Most striking of all, 2.5 μm of nickel underplate on beryllium copper was able to reduce the Wear Index nearly to zero, even at 300g in runs to 700 passes (Fig. 4b).

Thick soft gold electrodeposits[12] have wear behavior indistinguishable from pure wrought gold, and the practice of making both contacts of pure gold has long been recognized to be impractical for unlubricated connectors, giving unacceptably high wear and friction. Figure 4c, with

3.3 μm of pure gold plate on copper, confirms that experience, a Wear Index of 100 occurring within 20 passes at 100 and 200g. On the other hand, 2.5 μm of nickel underplate was able to reduce wear considerably (Fig. 4d), a Wear Index of 100 at 200g not developing for up to 200 passes.

Historically, the original incentive to find golds for contact applications that were more durable than pure electrodeposits resulted in the development of the cobalt golds and similar hard platings. However, it is noteworthy that cobalt gold on nickel underplate is only slightly better than pure gold with the same underplate (Figs. 3c and 4d).

Figure 5 presents profilometric wear data for comparison with that from electrographic analysis, Fig. 2c, with 2-μm cobalt gold on copper. Results by both methods are strikingly similar. An additional finding, although expected, is that Wear Index does not rise sharply with numbers of passes until the depth of wear exceeds the thickness of the gold deposit. Subsequent analysis of sectioned samples revealed that a wear mechanism (prow formation with back transfer) was operating

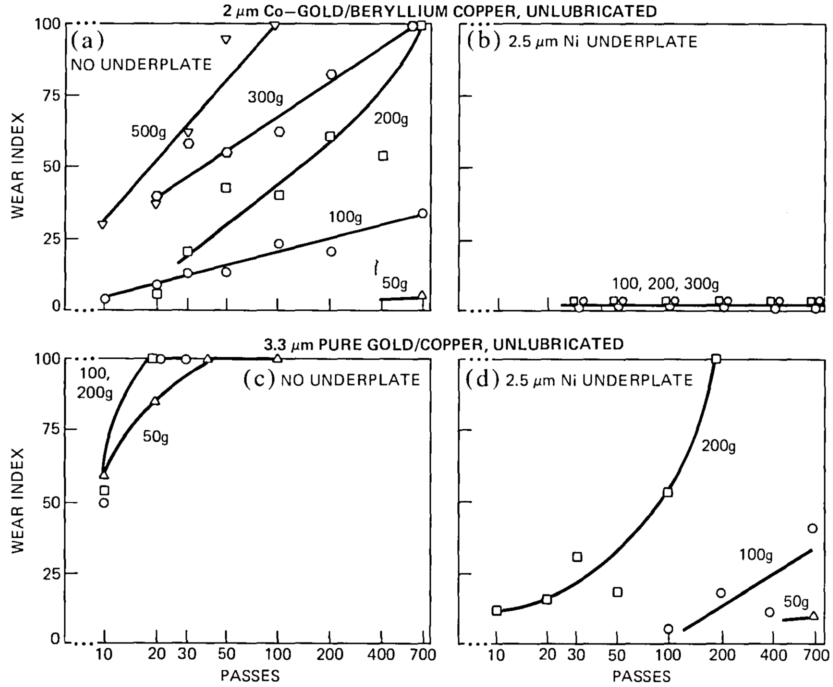


Fig. 4—Electrographic Wear Indexes from unlubricated adhesive wear runs. (a), (b) 2 μm cobalt gold electrodeposits on beryllium copper, with and without 2.5 μm nickel underplate. (c), (d) 3.3 μm pure gold electrodeposits on copper, with and without 2.5 μm nickel underplate.

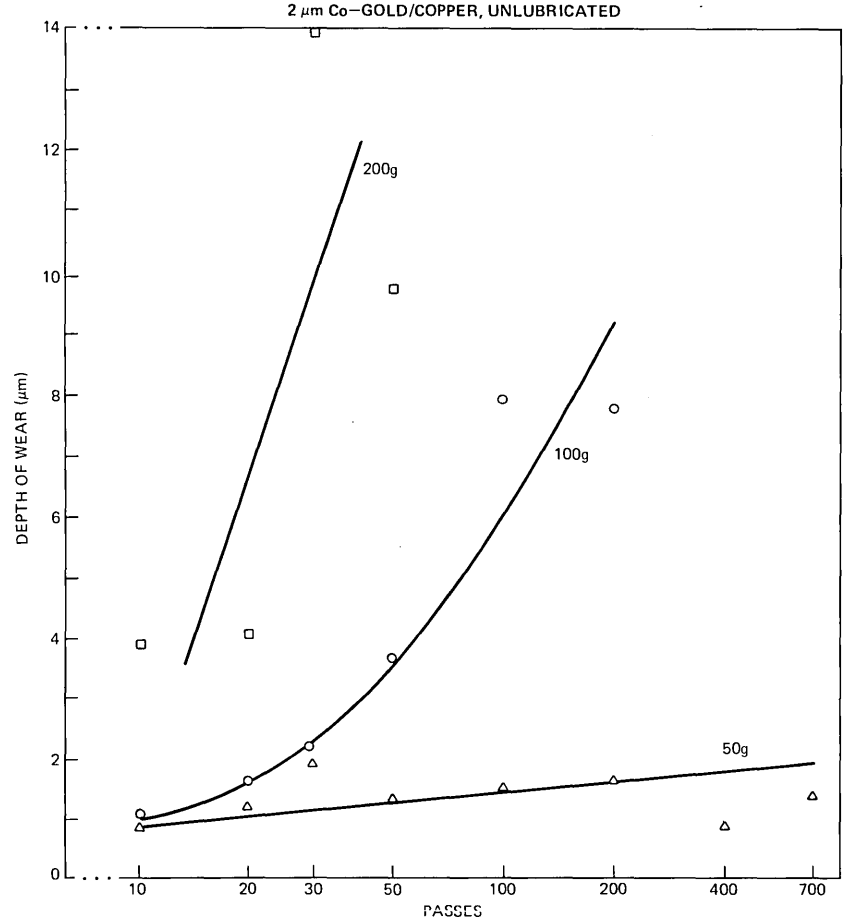


Fig. 5—Wear determined by profilometry from unlubricated adhesive wear runs with 2 μm cobalt gold electrodeposits on copper.

which permitted gold to persist in grooves in the wear track which are three to four times as deep as the thickness of the gold. This explains in part how Wear Indexes below 100 could be obtained in cases of severe roughening.

3.1.2 Lubricated

Adhesive wear was examined with a polyphenyl ether connector contact lubricant used in the telephone industry, as well as for connectors in computer and general-purpose applications. Boundary or elastohydrodynamic lubrication[13] prevailed, and there was good metallic contact and low contact resistance during sliding.[8] A high load, 500g, was used so as to promote adhesive wear because of the outstanding effectiveness of this lubricant.

Figure 6 presents the results with 3.3 μm of pure gold and 2 μm of cobalt gold, both on copper with and without 2.5 μm of nickel underplate. The data points are from independent runs. The overall conclusion is that the wear of all specimens is low in sliding for 700 passes. Additional runs with 2 μm of cobalt gold on copper at 500g for 1200 passes, and with 0.75 μm of cobalt gold on 2.5 μm nickel underplate at 200g for 4000 passes, discussed later, also gave little wear from microscopic examinations after test.

An unexpected result was the fall of the Wear Index with increasing numbers of passes following a rise early in sliding. This is most pronounced with the cobalt gold deposit (Fig. 6b) where the Wear Index was greater after 20 passes than after 700 passes. As shown later, this is attributable to self-healing which originates in the ability of the plating to flow laterally and to become burnished during sliding.

3.2 Abrasive wear

3.2.1 Two-body

Two-body abrasive wear of plated samples was studied with the pointed diamond rider in one pass at various loads, and the results

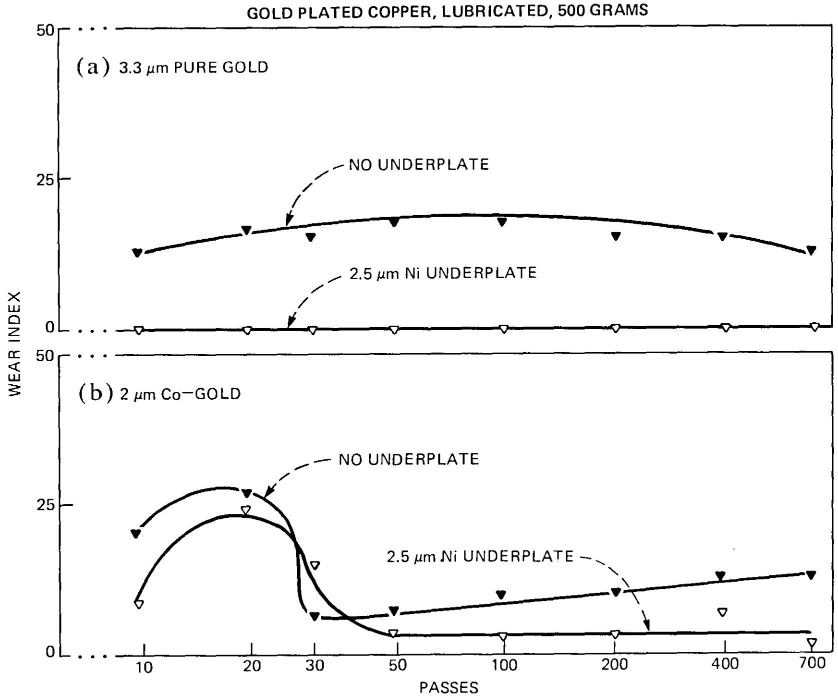


Fig. 6—Electrographic Wear Indexes from sliding at 500g with thin film of polyphenyl ether contact lubricant. (a) 3.3 μm pure gold electrodeposits on copper. (b) 2 μm cobalt gold electrodeposits on copper.

from electrographic analyses are given in Fig. 7. All curves have the same shape, a region in which the Wear Index increases little with increasing load until there is a precipitous break-through of the deposit. The point of inflection in the curves occurs at a Wear Index of 50 and a load that we term the "critical load for abrasive wear."

Increasing the thickness of the cobalt gold increases the critical load, for example, from 45 to 70g comparing 2 with 3.3-μm deposits on copper (Figs. 7a and 7b). Nickel underplate, 2.5-μm thick, increases the critical load from 45 to 90g with 2 μm of cobalt gold (Fig. 7a) and from 70 to 150g with 3.3 μm of cobalt gold plating (Fig. 7b). Likewise, a hard substrate, beryllium copper, increases the critical load from 45 to 150g with 2-μm cobalt gold deposits (Figs. 7a and 7c), and a further increase, from 150 to 200g, is obtained with nickel underplate on beryllium copper (Fig. 7c).

Critical load for 3.3 μm of pure gold on copper (Fig. 7d) was low, but this result is not representative; subsequent analysis, discussed later, revealed that the deposit was poorly adherent to the substrate. Most striking of all are the results with 3.3 μm of pure gold on 2.5-μm nickel underplate on copper (Fig. 7d), which gave the highest critical load, 250g, fully 100g greater than the value for the same thicknesses of cobalt gold and nickel underplate on copper.

Figure 7b shows the results with increasing thickness of nickel underplate on copper. There is no significant change of critical load with thin nickel, 1.5 μm, and the critical loads for 2.5 and 4 μm of

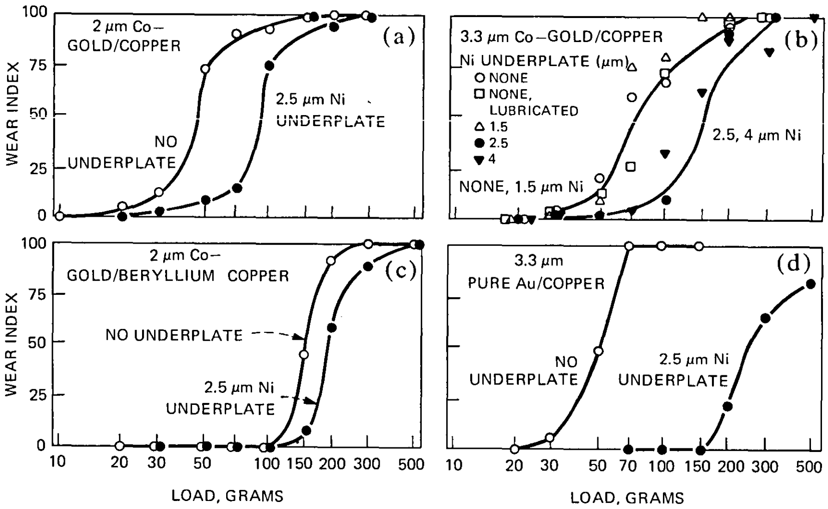


Fig. 7—Electrographic Wear Indexes from two-body abrasive wear runs with various gold electrodeposits, substrates, and underplates. Figure 7b shows no effect on wear by thin film lubricant, and 7d shows that results on specimen without underplate are poor because gold was non-adherent (see Fig. 15).

underplate are approximately the same. Finally, Fig. 7b shows that two-body abrasive wear is little affected by lubrication.

3.2.2 Three-body

Three-body abrasion was studied with 3.3-μm cobalt gold deposits, with and without 2.5 μm of nickel underplate. Runs at 50g were made to 200 passes, at which both samples had a Wear Index of 100 (Fig. 8). Control runs without abrasive gave a Wear Index of zero at 200 passes.

Nickel underplate markedly improved the durability of gold, five to ten times the sliding being required to attain comparable Wear Indexes.

IV. ANALYSIS AND DISCUSSION

4.1 Adhesive wear

The adhesive wear of gold has been extensively studied,[1] including the solid wrought metal, thick and thin electrodeposits in dry sliding, when adventitiously contaminated[7] and with liquid and solid lubri-

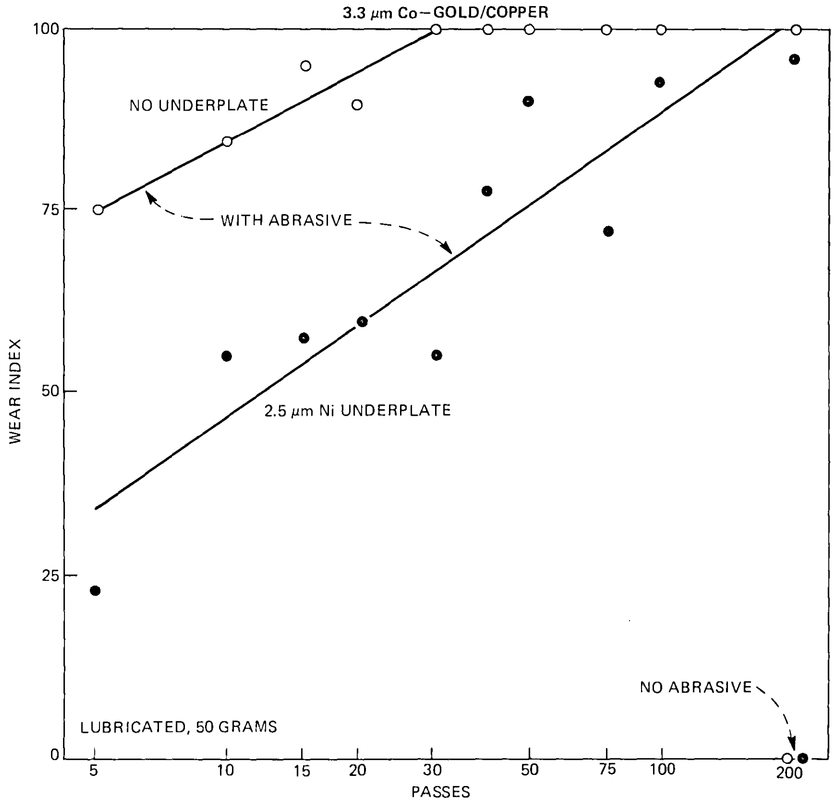


Fig. 8—Top: Electrographic Wear Indexes from runs with 3.3 μm cobalt gold electrodeposits on copper; three-body abrasion with coarse boron carbide, thin film lubricated.

cants. The present study of underplate and substrate variables shows that the mechanical properties of the basis material can have a profound effect on wear rate, although the mechanisms are the same. Details of sliding wear will be discussed, and as background, it is useful to summarize the present understanding of the adhesive wear of gold.

In connectors and other hardware, contact members are usually not identical in size and shape. Further, rubbing is generally localized to a small area on one surface and spread out on the other contact. In such situations, the unlubricated sliding of gold, and of many other metals, is characterized by unsymmetrical transfer, virtually all of the metal moving initially from one member to the other and not in both directions.

In a second step, the transferred metal is removed by back-transfer to the original part or is lost as loose debris.

These processes are readily observed with rider-flat apparatus in which the sliding members have the idealized geometry of a hemispherically ended rider contact that is pressed against a flat contact attached to a turntable. At the onset of rubbing, a lump of metal, called a prow, that comes from a flat, forms between and separates the specimens, as shown in Fig. 9. Sliding now continues at the junction between them. The prow projects against the direction of movement and gouges the opposing member. Routed solid becomes attached to the prow so that it grows in length.

A plausible explanation for the origin of the prow and why it is always located on the rider is that it is related to the difference in size of rubbing areas of the members. The population density of transfer particles (formed by breaking at other than the original interface of asperities of the specimens that have cold-welded together) is greater on the smaller part. Since compressive forces are large, the particles readily weld to each other and to the rider to form a prow. The prow is harder than either original surface because of the extreme degree to which it is worked in its transfer and growth.

Prows form on the smaller part in any arrangement of sliding members. For example, with edge connector contacts and printed circuit board fingers, prows form on the contact that corresponds to the rider in rider-on-flat systems.

When the rider traverses the same track repetitively, prow formation eventually ceases and is replaced by "rider wear," a process in which the member with the smaller surface involved in sliding loses metal. This is due to the accumulation of sufficient back-transfer prows on the flat to increase its hardness in all places to the level attainable by extreme work hardening. When the surface of the flat reaches the

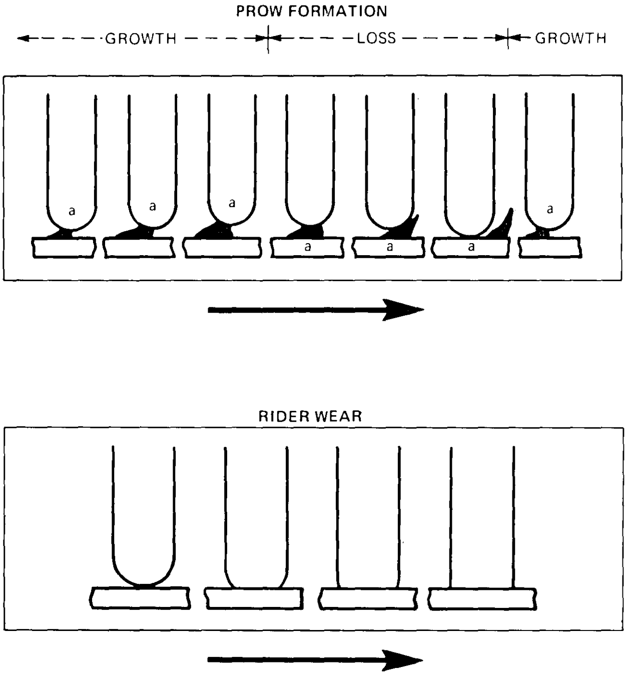


Fig. 9—Schematic representation of prow formation mechanism of adhesive wear. One of the processes by which rider loses prow and cold-welds to flat is shown. Letter "a" designates surface to which prow adheres. Arrow indicates direction of movement of the flat. Bottom: Schematic representation of rider wear mechanism.

hardness of the prow on the rider, routing of the flat ceases and a wear land appears on the rider. Figure 9 schematically shows progressive rider wear, with the rider shrinking in length as its hemispherical end wears away.

The number of passes to the transition from prow formation to rider wear is related to both track length (in unidirectional or reciprocating sliding) and to load. The shorter the track or higher the load, the more quickly will the transition come.

Past the transition, the flat gains mass and the rider loses metal by transfer to the flat or as loose debris.

Prow formation occurs with dissimilar metals, provided that the unworn flat is not excessively harder than the unworn rider. The flat provides prow metal which wears the flat in the usual way. Systems have also been observed with soft riders on hard flats in which transfer early in the run is by smearing of rider metal on the flat. Eventually, however, on repeat passes, the transfer metal is picked up by the rider and becomes so worked with resulting increase in its hardness that it begins to rout the flat. Prow formation ensues.

Figure 10 illustrates the prow formation mechanism and the effect of nickel underplate. The letter designations in the figure refer to vertical rows of photomicrographs: (a) to (d), 2 μm cobalt gold plate on copper and (e), 2 μm cobalt gold plate and 2.5 μm nickel underplate on copper.

Figure 10a, from initial sliding at a light load (20g, 1 pass), shows cold welding of asperities of the contact members and transfer of gold from rider (soft) to flat (hard). After a few additional passes, the direction of metal transfer reverses with the initiation of prow formation, as in Fig. 10b (20g, 15 passes) where minute fissures now appear in the gold plate with attendant loosening of a particle of the deposit. It should be noted that, at higher loads, prow formation occurs at the initiation of sliding, without preliminary wiping of rider metal on the flat. In Fig. 10c, prow formation continues with increasing loss of metal from the flat and the development of porosity (center of track) in the deposit. A secondary wear process involves plowing or abrasion of the flat by severely work-hardened prows which persist on the rider. This gives elongated wear features, often for the full length of the track (100g, 100 passes). Figure 10d shows further development of the process in (c) resulting in deep grooving of the track. The rider has acquired a large prow, and the wear track is extensively covered with coarse back-transferred metal (former prows). Both rider and flat have loose matter of varied size and shape, including equiaxed, plate-like, and roller-shaped wear particles. Not shown in the figure is the advanced stage of wear when prow formation ceases and rider wear occurs. The effect of nickel underplate on sliding in Fig. 10e is to reduce the scale of transfer and wear, with smaller prows, less debris, and lower Wear Indexes. The photographs in Fig. 10e were taken at sliding conditions and numbers of passes identical to those in (d). Hard underplates and substrates can also increase the effectiveness of marginal lubricants and of adventitious contamination.

A graphic illustration of the value of nickel underplate in increasing the durability of gold plate is given in Figure 11, based on data presented in Figs. 2 and 3. Figure 11a shows that the resistance to wear-through of gold plate increases dramatically when its thickness exceeds a particular value, about 3.5 μm for cobalt gold plated copper at 100g (unlubricated) calculated for a Wear Index of 50. This thickness increases with increasing load and for smaller Wear Indexes. With 2.5 μm of nickel underplate, the thickness of cobalt gold at which durability shows a sharp increase is only slightly in excess of 2 μm. In Fig. 11b, passes to the Wear Index of 50 is plotted against thickness of nickel underplate for 3.3 μm of cobalt gold. Again, a sharp rise in durability occurs at a characteristic nickel thickness, which becomes greater the

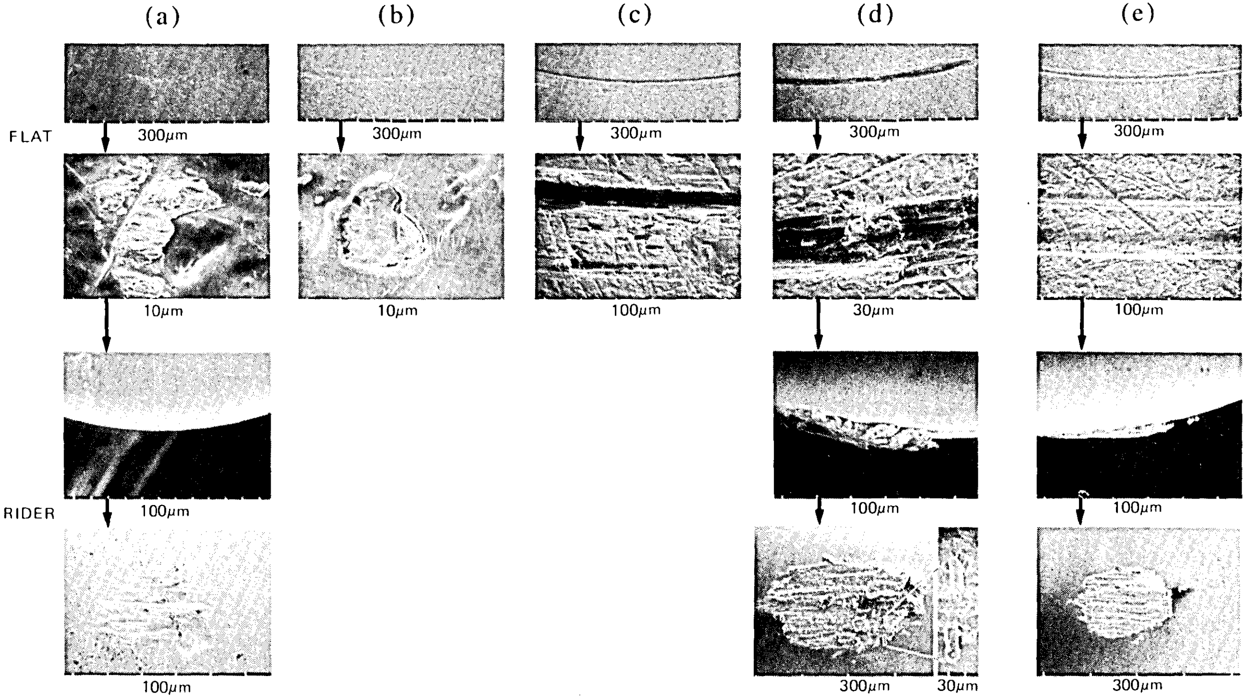


Fig. 10—Worn specimens from sliding in the absence of lubricants. Flats: 2 μm cobalt gold electrodeposits on copper. (a)-(d) Plated directly on substrate. (e) Plated with 2.5 μm nickel underplate. Riders: solid gold (views are at right angles and normal to the worn surface). (a) to (d) Runs of increasing severity. (d) and (e) At identical conditions of sliding, illustrating similarity of wear processes with significant attenuation in severity when nickel underplate is used. Scanning electron microscope numbers are distances between adjacent white markers.

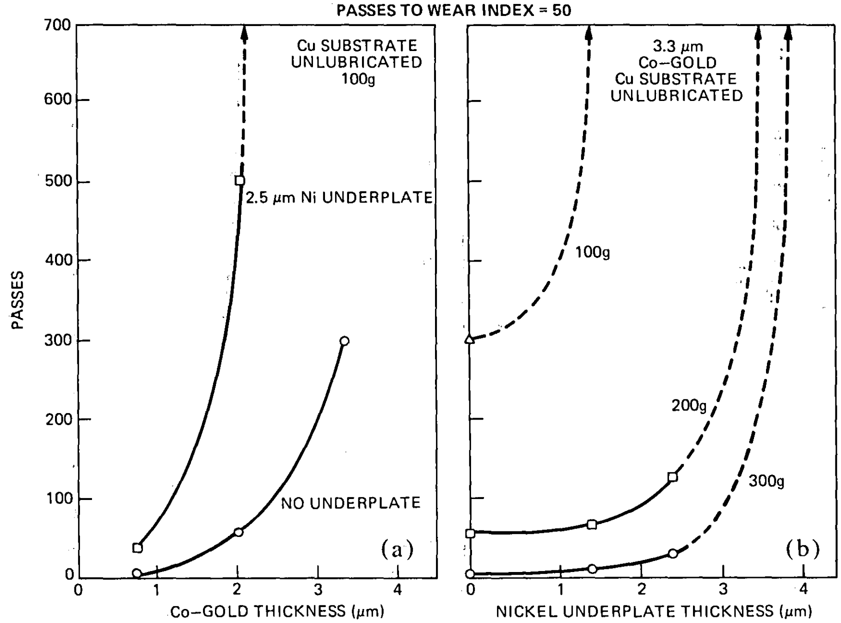


Fig. 11—Unlubricated sliding at adhesive wear conditions (passes to Wear Index = 50). (a) 0.75, 2, and 3.3 μm cobalt gold electrodeposits on copper. (b) 3.3 μm cobalt gold electrodeposits on copper with 0, 1.5, 2.5, and 4 μm nickel underplate. Arrows at ends of dashed curves indicate runs at 700 passes without achieving a Wear Index of 50.

larger the load. From Figs. 11a and 11b, the less the gold thickness, the thicker the nickel underplate has to be to obtain a given level of protection, e.g., against reaching a prescribed Wear Index at a designated load.

It has been shown that sliding occurs initially with considerable transfer, roughening, and wear of the flat. These processes diminish in severity to an equilibrium level as sliding continues.[14] The early stage of sliding is where most connectors operate during their lifetimes, common service requirements being 200 insertions and withdrawals (400 passes). The use of nickel underplate is a way to permit the high wearing conditions of early sliding to occur without excessive loss of gold, or with a lesser thickness of gold than otherwise would be required.

4.2 Lubricated sliding and burnishing

Lubricants are desirable for sliding contacts because they reduce the interaction of surfaces and, thereby, adhesive transfer and wear. If only a few scattered asperities continue to touch, contact resistance during sliding would be little different from that of stationary contacts. Connectors ordinarily do not carry current during engagement, and thus do not have the critical requirements for low wear and electrical noise of instrument slip rings and many other sliding metallic contacts.

It has been shown[8] that lubricants can be effective through a wide range of sliding conditions, and that adventitious contamination, although variable, may be adequate for some applications.[7]

Figure 12 illustrates surface changes from sliding with a good lubricant, a liquid polyphenyl ether, characterized by little transfer and wear and a marked tendency for the surfaces to become burnished. The photomicrographs are from copper flats plated with 2 μm of cobalt gold. In Fig. 12a, there are signs of burnishing of the flat with the simultaneous appearance of a few fine scratches and a small amount

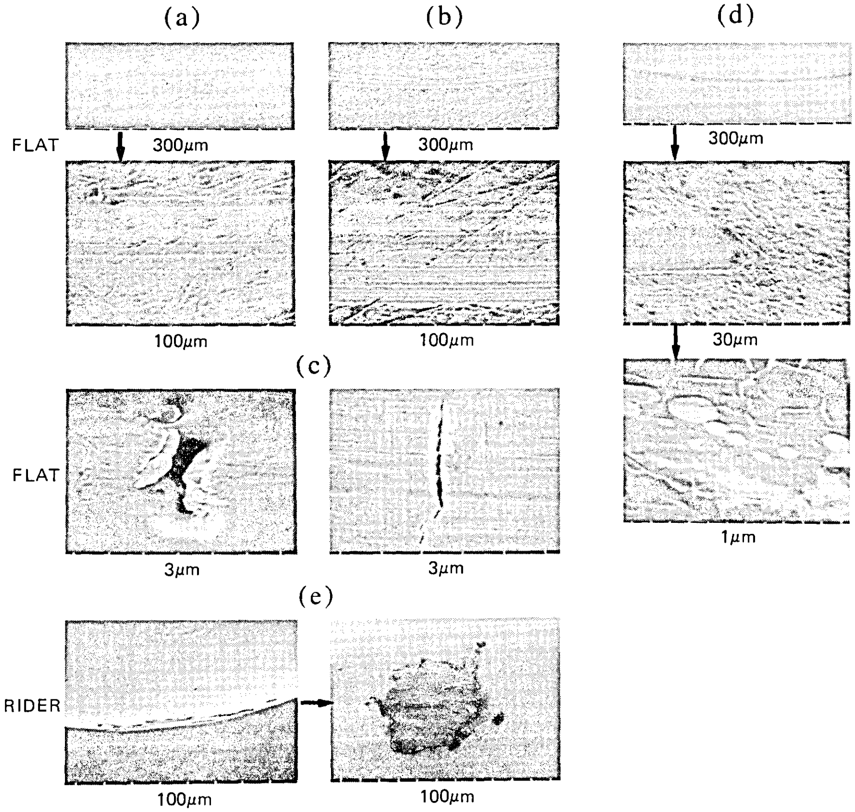


Fig. 12—Worn specimens from sliding with a thin film of liquid polyphenyl ether lubricant. (a), (b), (c) 2 μm cobalt gold electrodeposits on copper flats at 500g (a = 4 passes, b = 1200 passes, c = 10 passes). (d) 0.75 μm cobalt gold on 2.5 μm nickel underplate on copper flats (200g, 4000 passes). (e) gold rider from sliding against 2 μm cobalt gold plated copper flat (500g, 100 passes). Figures 12a, 12b, and 12d illustrate burnished surfaces due to lateral movement of metal on which are superimposed very fine scratches. Figure 12d has a nearly continuous gold layer (from electrographic testing), but so thin that contrast differences appear due to the underplate from the electron beam which has penetrated the gold. Figure 12c (left and right) are concurrent processes: left is burnishing in which a pit or pore initially present in the deposit is being closed, and right is a tiny fissure (rare) that has developed in the deposit. Figure 12e shows a small prow on the rider (views at right angles and normal to the tip).

of wear-induced porosity (500g, 4 passes). On further sliding (Fig. 12b), burnishing continues, with marked thinning of the gold plate evidenced by the grey striations in the track attributable to the copper substrate from SEM electron beam penetration of the overlying gold (500g, 1200 passes).

Figure 12c illustrates concurrent wear processes in gold plate: left, lateral flow of metal (burnishing) which has partially closed over a micropit or pore in the as-plated deposit; and right, a fine crack or tear (500g, 10 passes).

These phenomena are qualitatively the same with nickel underplate. Figure 12d, with 0.75 μm of cobalt gold on 2.5 μm nickel after prolonged sliding (200g, 4000 passes), shows the gold to be still nearly continuous, proven by little response in an electrographic test. The center photomicrograph shows that a small amount of fine loose debris has formed and accumulated mainly at the ends of the track. The wear track is concave along its width, most of the wear and metal flow occurring in the center because of the hemispherical shape of the rider. Figure 12d (bottom), the middle of the wear track (along its width) at high magnification, has a thin but continuous gold layer. The light patches are thicker gold in micro depressions of the unworn surface.

Figure 12e shows the gold rider after 100 passes at 500g against a 2 μm cobalt gold-plated copper flat, thinly coated with polyphenyl ether. Lubricant transferred to the rider was not removed prior to SEM study and appears as a dark stain. A small prow on the rider accounts in part for the small loss of metal by the flat and is the cause, along with loose debris, of the fine scratches in the cobalt gold plate.

It was found that burnishing and widening of the track proceeds more rapidly in the absence of intentional lubrication, if the samples are intrinsically low wearing. An example is 3.3 μm of cobalt gold on 2.5 μm of nickel underplate (Fig. 3c). A section was made across the wear track from a run for 40 passes at 200g, and 171 measurements of gold thickness were obtained at fixed increments, 45 outside the wear track and 126 within its boundaries. A probability plot (Fig. 13) of the two sets of measurements clearly reveals the extent of burnishing that has occurred; gold is displaced from the high spots to the low spots in the surface (inset), with some measurements less than 1 μm compared to the rather uniform 3.3 μm original deposit, and with measurements at the other end of the distribution greater than 5 μm. It can be seen, at the 50-percent point, the median gold thickness decreased in sliding by 0.1 μm, or 3 percent of the unworn value. The worn gold appeared as loose debris and transfer matter. The photomicrographs also show no change in thickness of nickel underplate during sliding. Fresh surface was not created by lateral flow of metal, since the gold deposit is dense and featureless.

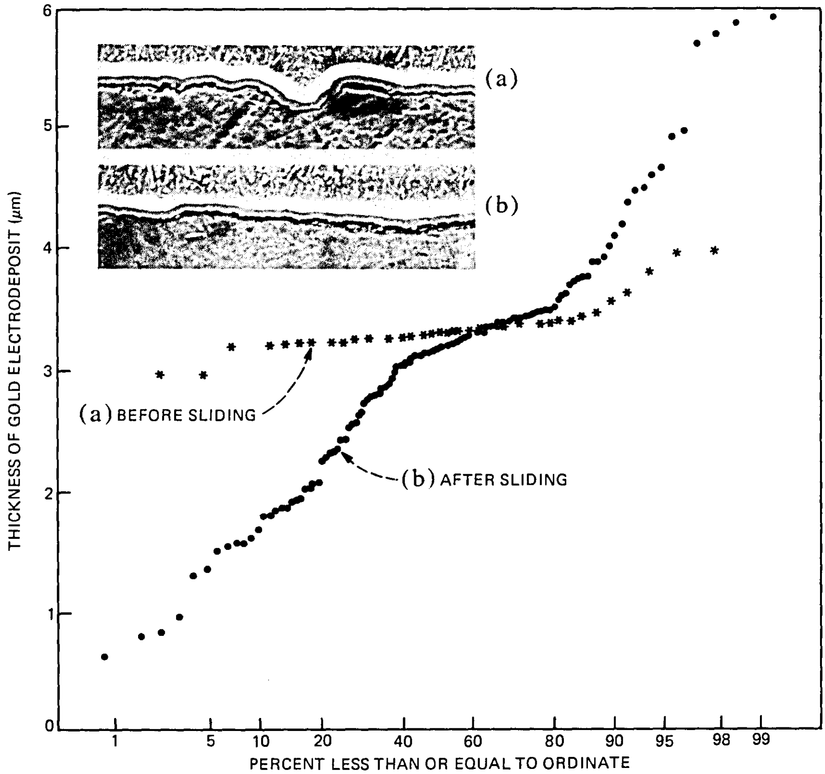


Fig. 13—Thickness distribution of gold plate. Cobalt gold electrodeposit on nickel underplate (3.3 and 2.5 μm average thicknesses) on copper. Section a, unworn flat, and section b across wear track from sliding at 100g, 40 passes with solid gold rider. The inset is an SEM photograph of the section showing a portion of the wear track. The gold plate is the light band, and the dark band below it is nickel underplate.

The evidence of burnishing in Figs. 12 and 13 show how the Wear Index can fall with continued sliding (Fig. 6). Remaining to be explained, however, is the increase in Wear Index from the initiation of the run, before significant burnishing has developed. The rise of the Wear Index is due to a brief period of severe wear, during which small-scale prow formation occurs. As the surfaces run in, contact pressure diminishes and conditions become more favorable for a hydrodynamic contribution to sliding by the contact lubricant.[13] Wear rate then becomes less and burnishing begins to predominate. The balance between adhesive wear, with attendant roughening and high friction, and burnishing having the opposite effect is delicate and can shift in the course of a single run.

4.3 Abrasive wear

4.3.1 Two-body

Figure 14 illustrates surface damage which occurs in two-body abrasion. The samples are arranged in order of decreasing surface damage (left to right) and increasing load (top to bottom). The samples are plated with 3.3 μm of gold: (a) cobalt gold on copper; (b) cobalt gold on 2.5 μm of nickel underplate, and (c) pure gold on 2.5 μm of nickel underplate. At light loads, relatively smooth depressions having horizontal striations appear which can completely alter the topography of the surface. Tensile cracks in the deposit develop at higher loads, characteristic of the material, and grow in size and number as the load is further increased. Porosity in the gold originates in these cracks, as well as from removal of gold by a cutting action which exposes the underlying metal.

Although the width of the wear tracks at a given load is greater with pure gold compared to those from an equal thickness of cobalt gold on

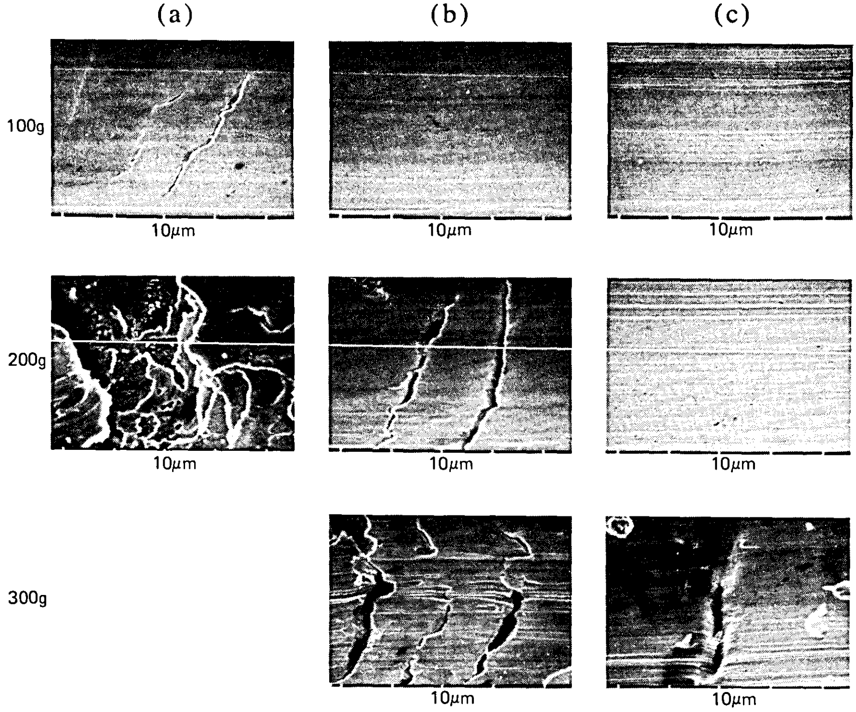


Fig. 14—Worn gold plated flats from single pass sliding with conical diamond rider (two-body abrasion) at various loads. (a) 3.3 μm cobalt gold electrodeposit on copper. (b) same as (a) with 2.5 μm ductile nickel underplate. (c) 3.3 μm pure gold electrodeposit on 2.5 μm nickel underplate on copper. Sliding in horizontal direction. Wear appears as smooth depressions in surface with tears in gold coating. Nickel underplate (b) reduces extent of tearing. Pure gold (c) resists tearing better than cobalt gold due to its greater ductility.

the same substrate because it is softer, the pure gold sample has significantly less porosity (see also Figs. 7b and 7d for 3.3 μm of gold plate on 2.5 μm nickel). This is attributable to its greater ductility with an elongation estimated to be 2 to 3 percent, compared to less than 1 percent for the cobalt gold.[6] It is apparent also that a hard ductile underplate (nickel) on a softer substrate (copper and beryllium copper) provides significant protection for the gold deposit (Figs. 7a, 7b, and 7c). Thus, both hardness of the composite of gold and underplate and the ductility of the gold control the resistance of the finish to two-body abrasive wear.

With pure gold plated directly on copper (Fig. 7d), the load range in which there is a steep rise of Wear Index is abnormally low compared to the load range for cobalt gold on copper (Fig. 7b). This was probably due to poor adhesion of the deposit on this particular sample which led to premature stripping, as shown in Fig. 15.

4.3.2 Three-body

To obtain three-body abrasion, particles were used that were substantially harder (2750 kg/mm^2) than the plating and too coarse to be buried in the valleys between the high spots of the specimen surface. As in two-body abrasion, the thin lubricant film probably did not

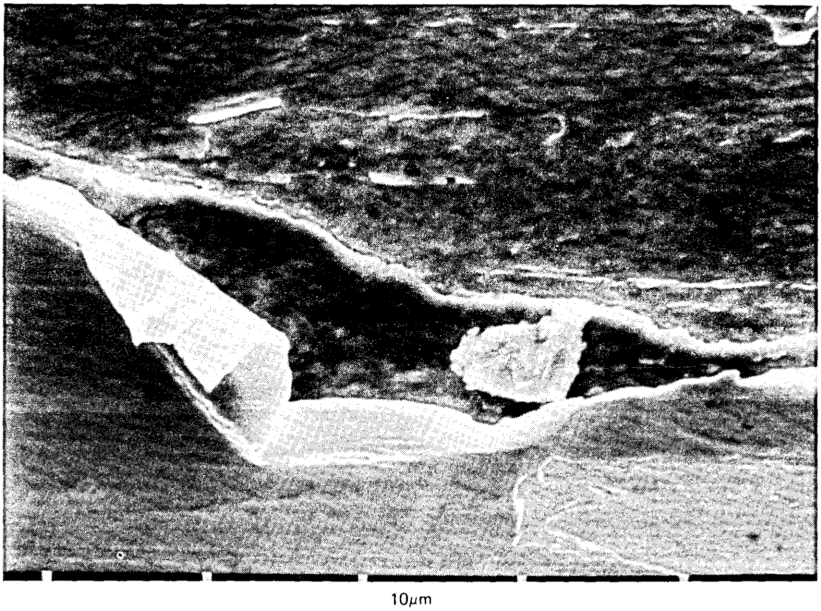


Fig. 15—Effect of sliding on poorly adherent deposit. Plating is disrupted and porous. Two-body abrasion: 3.3 μm pure gold electrodeposit on copper. Diamond rider, 1 pass at 100g (see Fig. 7d).

mitigate wear. The wear processes of two-body and three-body abrasion are identical, due to penetration of the gold surface under load with the development of porosity by plowing and tearing of the gold. A hard ductile underplate, nickel, is effective in reducing wear.

The rate of increase of Wear Index with numbers of passes (Fig. 8) is small, Wear Index doubling from about 45 to 90 between 10 and 100 passes with gold on nickel underplate. This is probably due to extensive rolling of the loose abrasive during which it would be ineffective in causing wear.

4.4 Effect of hardness and ductility of contact material on its wear

The hardness of a multilayer contact material reflects the hardness of the individual layers, depending on their thickness, load, and the shape and size of the indenter. In the present study, deposit thicknesses, loads, dimensions of the riders, and conditions of sliding were chosen so as to realistically model the wear processes of separable electronic connectors.

It was found that increasing the hardness of any of the layers of the contact material—gold plate, underplate, or substrate—could be beneficial in reducing wear. To a first approximation, the hardness of the samples determined normal to the surface with a Knoop indenter at a low load permitted ranking the composite contact materials in inverse order of their wear resistance in unlubricated sliding according to Figs. 2 to 4. Some measured hardnesses are given in Table III.

The dependence of adhesive wear on hardness is explained by all the common theories of wear. For example, the "adhesion" theory, attributed to Holm based on an atomic model, and the Archard version, based on asperity interactions,[15] give the relationship:

$$\text{Wear Volume } \alpha \frac{\text{(distance slid) (load)}}{\text{hardness}}.$$

Physically, an increase in hardness reduces the real area of contact and thus the numbers of junctions which can weld. Prow formation is a special case of adhesive wear.

In the lubricated adhesive wear experiments (Fig. 6), there was a small amount of metal transfer with prow formation early in the runs at high load, shown in Fig. 12e with cobalt gold-plated copper. Prow formation is attenuated by increasing the hardness of the composite contact material.

In the case of abrasive wear, increasing the hardness of the material will, in general, also improve its wear resistance because the depth of penetration of the surface is reduced, and with it the volume of metal carried away by plowing. However, ductility is also a factor in wear because even when the gold layer is not removed, it can crack if it is

Table III—Hardness of plated contact materials*

Co-Gold (μm)	Ni Underplate (μm)	Substrate	KHN_{25}	Depth of Penetration of Indenter (μm)†
2	None	Cu	57	2.6
2	2.5	Cu	64	2.5
3.3	2.5	Cu	95	2.0
3.3	4	Cu	134	1.7
2	2.5	Be-Cu	294	1.1

* Hardness determined normal to surface.

† Depth of penetration $= \frac{\text{length of diagonal, Knoop}}{30.53}$.

unable to yield plastically under the loaded slider. Thus, pure soft gold may be better able to resist abrasive wear than cobalt gold, as shown in Fig. 7 with nickel underplate. In the adhesive wear studies, prows that persisted on the rider were able to cause further wear in repeat pass sliding by secondary abrasion.

Although primary attention has been directed to the flat in this investigation, prior[1] observations permit statements to be made concerning wear behavior were the rider made of a material other than solid gold, such as cobalt gold-plated copper alloy. When wear of the flat occurs with the formation of prows, the material of the opposing contact is unimportant, the rider merely acting as a holder for the prow. The prow appears, grows, and breaks off as sliding continues but a new one immediately forms without damage to the rider. Eventually, however, when the transition to rider wear occurs, the rider begins to lose metal. The rider should be made of a hard material to reduce its wear rate, and nickel underplate is desirable.

A few connectors are designed so that their mating contacts have identical geometry. In this case, either member can accept prows with the other one wearing, and the role (rider or flat) played by a contact may change several times as sliding continues.

4.5 Application to hardware

An objective of this investigation was the determination of the role of nickel underplate and of substrates in the wear of gold plate so that guidelines could be developed in contact materials selection for electronic connectors. This study has shown that wear, determined by the development of porosity, can be reduced when a hard ductile underplate such as pure nickel is used. Increasing substrate hardness has the same beneficial effect, although in practice the substrate material is selected for reasons other than hardness, such as spring properties, formability, and conductivity. Thus, the underplate remains the only way that the hardness of a gold contact finish can be increased.

Advantages to the use of nickel underplate were demonstrated for all wear mechanisms: adhesion, two-body abrasion, three-body abrasion, and brittle fracture. It was also found that at the conditions of this study there was little advantage in using thin, i.e., 1.5 μm, nickel plating with thick (3.3 μm) gold, but that nickel deposits 2.5 μm or greater in thickness could be highly desirable. The benefits in using nickel underplate were found for a wide range of gold thickness, but were most striking with the thinnest golds.

It does not follow, however, that a satisfactory gold contact cannot be designed without a hard underplate. Depending on the numbers of insertions which are required, the level of porosity in the gold layer that can be tolerated and on other factors, performance may be adequate in its absence. For example, there is little reason to use nickel underplate with very thick (say, greater than 5 μm) gold deposits at loads typical of electronic connectors. Another example would be when abrasive wear is unlikely to occur, and good contact lubricants are employed to minimize adhesive wear. At the other extreme, when lubrication is not used or when particulate contamination of the contact occurs, as with glass fibers from the leading edge of a glass-epoxy printed circuit board, nickel underplate is especially desirable because of its ability to control both the adhesive and abrasive wear of gold plate.

When the application of the connector or printed circuit board cannot be controlled or is unknown, the best course is to specify nickel underplate. This recommendation is especially applicable when thin (below 1 μm) gold coatings are used.

V. ACKNOWLEDGMENT

Dale E. Heath contributed significantly to this investigation by his excellent metallographic work.

The assistance of the Sel-Rex Division, Oxy Metal Industries Corporation is acknowledged for plating the samples used in this study.

APPENDIX

Electrography

The procedure in this work was to use Kodak Dye Transfer Paper F that had been soaked for 15 to 45 minutes in an electrolyte-indicator solution of 20g of the disodium salt of dimethylglyoxime and 20g of sodium chloride in 1 liter of water. Excess liquid was removed by squeezing the paper between rubber rollers, and the paper pressed against the specimen at 70 kg/cm^2 between titanium electrodes. Current was applied from a constant voltage source at 2.0V dc for 1 minute. The papers were peeled from the samples and dried in an oven at 50°C. Colored spots (red for nickel and green for copper) signified exposed base metal. Spots were, however, somewhat larger than worn areas in the specimen to which they could be related due to spreading of the chemicals in the paper. The electrographic prints were made within 1 day of the wear runs.

Electrography can be made more or less severe by varying applied voltage. However, at about 4-V dc, the gold is stripped from the specimen.

It was found that there is no significant change in the surface of a worn deposit as a result of electrographic testing. Distinguishing surface features in a track were photographed with the SEM at various magnifications from 30 to 5000 diameters both before and after obtaining an electrograph. This finding is in agreement with an earlier[11] optical study at 13 to 75 diameters.

REFERENCES

1. M. Antler, "Tribological Properties of Gold for Electric Contacts," IEEE Trans. on Parts, Hybrids, and Packaging, *PHP-9* (1973), pp. 4–14.
2. C. A. Holden, "Wear Study of Electroplated Coatings for Contacts," Proc. Holm Seminar on Electrical Contacts, Illinois Inst. Tech., Chicago, Ill. (1967), pp. 1-19.
3. A. J. Solomon and M. Antler, "Mechanisms of Wear of Gold Plate," Plating, *57* (1970), pp. 812–816.
4. G. Horn and W. Merl, "Friction and Wear of Electroplated Hard Gold Deposits for Connectors," Proc. Sixth International Conference on Electrical Contact Phenomena, Illinois Inst. Tech., Chicago, Ill. (1972), pp. 65–72.
5. W. Burt and R. Zimmerman, "Wear Capabilities of Gold Over Copper and Nickel Underplates," Proc. Third Annual Connector Symposium, Electronic Connector Study Group, Inc., Cherry Hill, N.J. (1970), pp. 435–444.
6. J. M. Dueber and G. R. Lurie, "The Strength and Ductility of Some Gold Electrodeposits," Plating, *60* (1973), pp. 715–719.
7. M. Antler, "Wear of Gold Plate: Effect of Surface Films and Polymer Codeposits," IEEE Trans. on Parts, Hybrids, and Packaging, *PHP-10* (1974), pp. 11–17.
8. M. Antler, "The Lubrication of Gold," Wear, *6* (1962), pp. 44–65.
9. M. Antler, "Wear and Contact Resistance," Ch. 21 in *Properties of Electrodeposits—Their Measurement and Significance,* R. Sard, H. Leidheiser, Jr., and F. Ogburn, eds., Princeton: The Electrochemical Soc., 1975, pp. 353–373.
10. H. W. Hermance and H. V. Wadlow, "Electrography and Electrospot Testing," Ch. 25 in *Standard Methods of Chemical Analysis,* 6th ed., W. W. Scott, N. H. Furman, and F. J. Welcher, eds., New York: Van Nostrand, 1962, Vol. 3, Part A, pp. 500-520.
11. H. J. Noonan, "Electrographic Determination of Porosity in Gold Electrodeposits," Plating, *53* (1966), pp. 461–470.
12. M. Antler, "Wear of Electrodeposited Gold," ASLE Trans., *11* (1968), pp. 348-360.
13. W. E. Campbell, "The Lubrication of Electrical Contacts," IEEE Trans. on Components, Hybrids, and Manufacturing Technology, *CHMT-1* (1978), pp. 4–16.
14. M. Antler, "Stages in the Wear of a Prow-Forming Metal," ASLE Trans., *13* (1970), pp. 79–86.
15. E. R. Braithwaite, ed., *Lubrication and Lubricants,* New York: Elsevier, 1967, p. 58.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 58, No. 2, February 1979
Printed in U.S.A.

A Model Relating Measurement and Forecast Errors to the Provisioning of Direct Final Trunk Groups*

By R. L. FRANKS, H. HEFFES, J. M. HOLTZMAN, S. HORING, and E. J. MESSERLI

(Manuscript received June 12, 1978)

This paper describes a mathematical model of the provisioning of direct final trunk groups with forecasting and measurement errors. This model can be used to study the effects of applying standard trunking formulas to possibly inaccurate load forecasts. An important consideration in the process is the degree to which the trunk forecast is actually followed. This so-called provisioning policy is modeled parametrically to allow consideration of a range of strategies, from following the forecast precisely to complete reluctance to remove trunks when indicated by the trunk forecast. When load forecast errors are combined with a reluctance to remove trunks, there will be a net reserve capacity on the average, i.e., more trunks than would be needed if the loads were known exactly. Using the mathematical model, a set of curves known as Trunk Provisioning Operating Characteristics is calculated. These relate percentage of reserve capacity to service (as measured by the fraction of trunk groups with blocking exceeding 0.03). The accuracy of the estimate of the traffic load defines the curve on which one is constrained to operate. The degree of reluctance to remove trunks together with the traffic growth rate determines the operating point. Improved estimation accuracy corresponds to a more desirable operating characteristic. The accuracy of the forecast load estimate is influenced by many factors, such as data base errors (e.g., measuring the wrong quantity due to wiring or other problems), recording errors (e.g., key punch errors), and projection ratio errors. This type of modeling may be useful both in evaluating the potential effects of proposed improvements in measurement or forecasting accuracy, and in studying the effects of changes in provisioning policy. A discussion is given of trunk provisioning process issues based on the viewpoint presented in this paper.

* A version of part of this paper was presented at the 8th International Teletraffic Congress, Melbourne, November 1976.

I. INTRODUCTION

Traffic measurements in the Bell System are used as the basis of those efforts aimed at planning an efficient network by providing appropriate quantities of trunking and switching equipment. They also form the basis of many efforts aimed at efficiently administering the network as well as serving as primary inputs for purposes of evaluating network performance.

New traffic measurement systems typically bring with them a variety of benefits such as more accurate and detailed data and more automated and convenient collection and processing of the raw data, together with possible new uses for the data which these improvements allow. Of course, to prove in economically, these advantages must offset the costs associated with installing and operating the system. Clerical savings associated with data collection and processing represent a good example of a relatively easily quantifiable advantage. Other advantages are not so simply equated to dollar savings.

To quantify the traffic-related benefits of improved measurements, models of the trunk provisioning process are required. We describe here a mathematical model of the provisioning of direct final trunk groups* with measurement and forecasting errors. An important consideration which ultimately determines the number of installed trunks is the degree to which the trunk forecast is actually followed. This so-called provisioning policy is modeled parametrically to allow consideration of a range of possibilities, from following the forecast precisely to complete reluctance to remove trunks when indicated by the trunk forecast. Errors in the load forecast will result in some trunk groups having more trunks than required while others will not meet the service criterion. When these errors are combined with a reluctance to remove trunks, there will be a net reserve capacity on the average, i.e., more trunks than would be needed if the loads were known exactly.

Using the mathematical model as a building block, a set of curves known as the Trunk Provisioning Operating Characteristics (TPOCS) is developed. These relate percent reserve capacity (i.e., the amount of trunks in service in excess of what would be required if the load were known perfectly) to service (as measured by the fraction of trunk groups with blocking >0.03). Figure 1 illustrates a typical set of TPOC curves. The accuracy of the estimate of the traffic load defines the curve on which one is constrained to operate. The reluctance to remove trunks together with the traffic growth rate determines the operating point.† Improved estimation accuracy corresponds to a more desirable operating characteristic.

* Also called full direct or nonalternate route groups.

† Traffic growth and reluctance may be related (e.g., high growth may cause high reluctance).

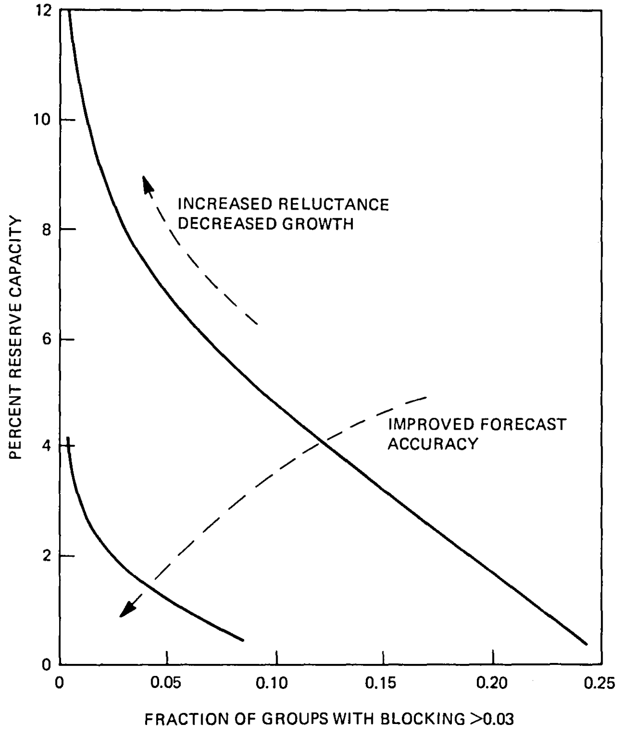


Fig. 1—Trunk provisioning operating characteristics.

This type of modeling can be useful in evaluating the potential effects of proposed changes in measurement or forecasting accuracy as well as studying the effects of changes in provisioning policy. However, as the results which are presented in this paper are based on a simplified model of the provisioning process for the case of direct final trunk groups only, they are intended to serve primarily as a way of viewing the problem with focus on some of the important trade-offs which are present. Much work remains to be done to develop and more fully exploit these ideas.

The paper is broken down as follows: Section II describes the mathematical models which we use to characterize the trunk provisioning process and generate the TPOC curves. Basically, they consist of a procedure for estimating the traffic load for the next busy season coupled with an engineering rule describing the provisioning policy for converting this estimate into the number of trunks provided. The mathematical details associated with these models are presented in the appendix.

Section III considers the sensitivity of the results to a number of the important submodels and assumptions which are used in the development. The TPOC curves are shown to be quite insensitive to the particular model of provisioning policy used as well as to the traffic growth factor and the distributions which characterize the various measurement and forecasting errors (though both the reluctance and growth affect the operating point). The curves do depend somewhat on trunk group size (corresponding to true offered load), however. While specific results are not presented in this paper, the curves also depend upon the mean values of measurement or forecasting errors.

Section IV is a discussion of perspectives on the trunk provisioning process based on the viewpoint presented in this paper. Section V summarizes and discusses further work which is needed.

II. A PROVISIONING MODEL FOR DIRECT FINAL GROUPS

2.1 Basic model concepts

The trunk provisioning process actually used in practice is very complex based only on what is explicitly recommended and standardized.* Overlaid on this are decisions made on a judgment or discretion basis which take into account a multitude of practical considerations. The following steps summarize the highly simplified version of a portion of the trunk provisioning process for direct finals which we will analyze. In Section IV, we discuss the trunk provisioning process in more detail.

(*i*) Traffic measurements—typically consisting of usage, offered attempts (peg count), and overflow—are used to estimate the base offered load during the last busy season for each trunk group. Usage may include a component due to maintenance, as trunks removed from service for investigation and repair are often measured as being busy.

(*ii*) The estimates of offered load for the past busy season are projected ahead—using "projection ratios" that reflect anticipated growth—to yield estimates of base offered loads for subsequent busy seasons.

(*iii*) The estimated number of trunks required to meet a grade of service criterion for the upcoming busy season is determined from an appropriate engineering rule. Typically, trunk groups are sized to produce an average blocking of 0.01 during 20 consecutive business days of the busy season.

(*iv*) The trunk servicer uses the trunks forecast and the trunks in service as the primary information in deciding on the number of trunks to be available for the next busy season. If the forecast number of trunks is less then the number currently in service, there is a reluctance

* There is a large amount of literature dealing with aspects of the trunk provisioning process, generated both within and without the Bell System. Some recent introductory material is given in Refs. 1 to 3. References 4 and 5 also contain related material.

to take trunks out of service unless this frees up equipment needed for other purposes. A number of factors can justify reluctance as a prudent policy. In a growth environment, trunks not needed next year might be needed in the following year. The desire to avoid rearrangement costs also encourages leaving trunks as they are.

A mathematical model capturing the essence of these steps is discussed in the remainder of this section. Though simplified, this model is rich enough to provide valid insights into the trunk provisioning process.

2.2 A model of provisioning policy

It is useful to begin the model development by introducing a model for the provisioning policy. This policy relates the current number of trunks, N_i, and the estimated requirements for the next period, $\hat{N}_{i+1}$, to the number of trunks to be provided for the next period, N_{i+1}. This is modeled parametrically by,

$$N_{i+1} = \max\left[\hat{N}_{i+1}, N_i - \beta\,(N_i - \hat{N}_{i+1})\right], \qquad 0 \leq \beta \leq 1, \qquad (1)^*$$

where β is a parameter which we introduce to provide a measure of the reluctance to provision down. If $\beta = 0$, trunks will never be removed while, at the other extreme, $\beta = 1$ corresponds to the situation where exactly the estimated number of trunks is always provided, even if that requires removing a large number of the trunks currently installed. Intermediate values of β correspond to intermediate provisioning policies.

It is convenient to rewrite (1) in normalized form by dividing through by M_{i+1}, the number of trunks required to handle the true traffic load in year $i + 1$. This yields

$$\frac{N_{i+1}}{M_{i+1}} = \max\left[\frac{\hat{N}_{i+1}}{M_{i+1}}, \beta\frac{\hat{N}_{i+1}}{M_{i+1}} + (1-\beta)\frac{1}{\Delta_i}\frac{N_i}{M_i}\right], \qquad (2)$$

where $\Delta_i = M_{i+1}/M_i$ represents the growth in trunks required to handle the true traffic load in going from year i to year $i + 1$.† The equilibrium solution to (2) has been determined both from (approximate) analytic techniques and from simulation. Specifically, important characteristics of the distribution of the quantity N/M are found. The mean value of this variable is a direct measure of the reserve capacity of the trunk groups under consideration, while the grade of service provided by a collection of trunk groups is related to the distribution of $(N - m)/M$ where m is the number of maintenance trunks (e.g., the fraction of

* Another model of reluctance was also used and was found to give similar results. This is discussed further in Section III.

† The results which follow are in terms of the traffic growth.

groups with blocking ≥ 0.03 is given by prob$[(N-m)/M < \alpha]$, where α is a suitably chosen constant which depends upon group size). Figure 2 qualitatively shows this relationship. In fact, the TPOC is a plot of $E[N/M] - [1 + (Em/M)]$, the reserve capacity, vs. prob$[(N-m)/M < \alpha]$.

Inspection of eq. (1) or (2) indicates that the various forecasting uncertainties which enter into the provisioning process are all summarized in the distribution of $\hat{N}_{i+1}$, which represents our estimate of the number of trunks which will be required in year $i+1$. This is the reason that the provisioning policy model was introduced first.

2.3 A model of trunk forecasting

Having discussed the model which converts the estimated trunk requirement, $\hat{N}_{i+1}$, into the number of trunks provided, N_{i+1}, and a simplified discussion of the way it leads to the tradeoff curves, we now turn to the process of estimating the trunks required from the traffic measurements.

It is necessary to consider a specific estimation model. This model is defined by the following equations.

$$\hat{a}_i = \frac{\hat{U}_i}{1-\hat{B}_i}, \tag{3}$$

$$\tilde{a}_{i+1} = \hat{g}_i \hat{a}_i, \tag{4}$$

and

$$B(\hat{N}_{i+1}, \tilde{a}_{i+1}) = 0.01. \tag{5}$$

In these equations, $\hat{U}_i$ represents the mqasured carried load (which typically consists of the sum of the measured traffic load plus the measured maintenance usage), $\hat{B}_i$ is the blocking estimate (determined

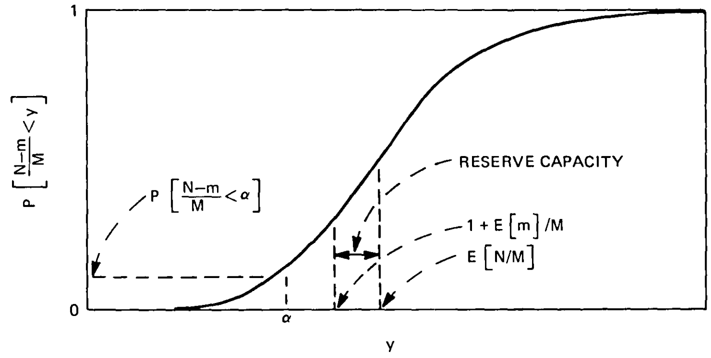


Fig. 2—A distribution of N/M indicating reserve capacity and $P[(N-m)/M < \alpha]$, the fraction of groups with poor service.

from measured peg count and overflow), $\hat{a}_i$ is the estimate of offered load, $\hat{g}_i$ is the projection ratio (which is simply an estimate of the traffic growth on the trunk group), and $B(N, a)$ is the Erlang blocking formula.

As can be seen from (5), the distribution of the estimated number of required trunks is determined by the distribution of the estimated offered load, $\tilde{a}_{i+1}$. The (normalized) variance of the estimation error associated with $\tilde{a}_{i+1}$ turns out to play a major role in results which we will soon describe. If we denote this normalized variance by σ_x^2, then the first order analysis of (3) and (4) given in the appendix (which assumes statistically independent, zero-mean error components) yields

$$\sigma_x^2 = \sigma_{U_i}^2 + \sigma_{B_i}^2 + \sigma_{g_i}^2, \tag{6}$$

where*

$\sigma_x^2 = \text{var}\left(\dfrac{\tilde{a}_{i+1} - a_{i+1}}{a_{i+1}}\right)$ = normalized variance of error in forecast load,

$\sigma_{U_i}^2 = \text{var}\left(\dfrac{\hat{U}_i - U_i}{u_i}\right)$ = normalized† variance of total carried usage error,

$\sigma_{B_i}^2 = \text{var}\left(\dfrac{\hat{B}_i - B_i}{1 - B_i}\right)$ = normalized variance of error in blocking estimate, and

$\sigma_{g_i}^2 = \text{var}\left(\dfrac{\hat{g}_i - g_i}{g_i}\right)$ = normalized variance of error in traffic growth factor.

To further facilitate the identification of the various error sources which contribute to our total estimation error, we may (approximately) decompose $\sigma_{U_i}^2$ into its component parts as follows:

$$\sigma_{U_i}^2 = \sigma_{d_i}^2 + \sigma_{s_i}^2 + \sigma_{m_i}^2. \tag{7}$$

In the above expression, $\sigma_{d_i}^2$ represents the variance of the data base and data handling induced errors in usage, e.g., due to measurement system deficiencies, $\sigma_{s_i}^2$ represents the statistical‡ errors in usage and $\sigma_{m_i}^2$ represents the variance of the fraction of trunks plugged busy for maintenance purposes (the trunking engineer, in general, does not

* These terms are defined more fully in the appendix.

† u_i represents the mean traffic usage and thus does not have a maintenance component.

‡ The statistical error term represents the effects of basing an estimate on a fixed number of samples of the busy-idle state of each trunk over a finite time interval.

know how much of the total usage is attributable to trunks plugged busy for maintenance purposes* and how much is traffic usage). Each of these quantities is normalized to the group traffic usage.

To compute the reserve capacity and the fraction of groups with blocking greater than 0.03, it is necessary to know the expected traffic growth, the number of trunks actually required, and the statistical behavior of the fraction of trunks plugged busy for maintenance. The form of the model output is shown in Fig. 3. The model parameters used to generate the curve apply to trunk groups of nominal size 25, a traffic growth rate of 5 percent annually, $\sigma_m = 0.02$ and $\sigma_x = 0.2$.† Each point on the curve corresponds to choosing one value of the provisioning parameter β and then computing the corresponding percent reserve capacity for the groups and the fraction of groups which have blocking greater than 0.03. Several values of β are indicated on the curve. Thus this curve, which we shall refer to as the Trunk Provisioning Operating Characteristic (TPOC), indicates the tradeoff which exists between the percent reserve trunk capacity and the fraction of groups with blocking exceeding 0.03.

Figure 4 shows the same tradeoff curve as Fig. 3, but also includes the tradeoff curve for a system which has $\sigma_x = 0.1$. The operating point on the tradeoff curve depends on the amount of reluctance to service down, but the tradeoff curve depends on the σ_x of the provisioning system. Improving the accuracy of the forecast (which depends in part on measurement accuracy) for the trunk provisioning process causes a decrease in σ_x. That is, it places the operating point on a more desirable tradeoff curve. The service improvement or trunk savings realized by such an improvement depends on the location of the operating point on the curve.

One additional point: a perfect measuring system corresponds to a nonzero σ_x due to nonmeasurement errors, such as statistical errors. This is indicated in Fig. 4 by the shaded area which contains TPOCs which cannot be reached by only decreasing measurement errors.‡

A number of studies were conducted to assess the sensitivities of the tradeoff curves to the model assumptions. These are discussed in the next section.

III. SENSITIVITY RESULTS FOR TRADEOFF CURVES

The provisioning policy model used is an approximation to the actual provisioning decisions which occur in the provisioning process.

* Maintenance usage may or may not be directly measured; even if measured, it varies from year to year.

† Normality is assumed; sensitivities to distribution (and to other factors) are considered in the next section.

‡ In the forecasting process, other factors, such as day-to-day variations, also contribute to a minimum σ_x. These factors are discussed, for example, in Refs. 6 to 8.

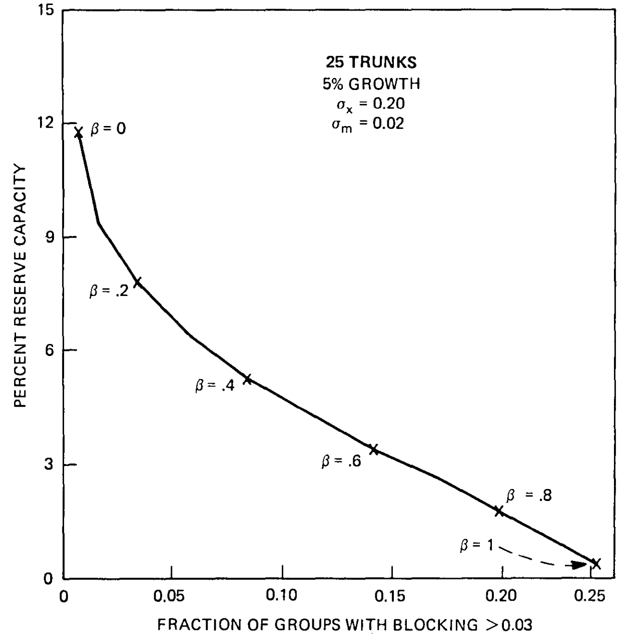


Fig. 3—Typical model output, TPOC tradeoff curve.

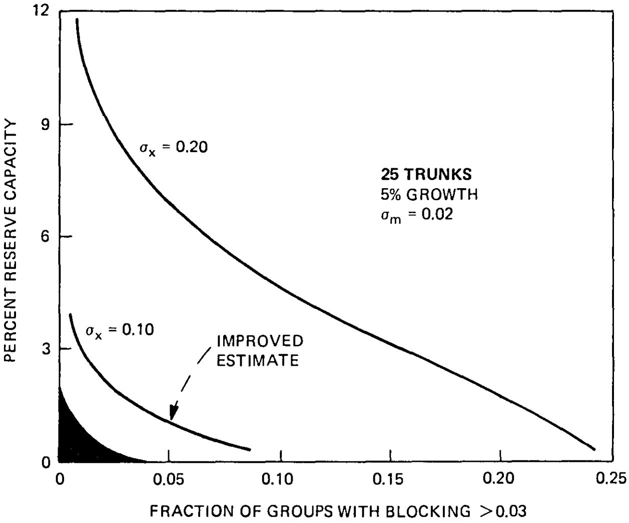


Fig. 4—Typical model output, TPOC tradeoff curves.

In fact, there probably is no universally applicable model since many decisions are largely based on judgment. The important question is whether a different model for provisioning policy would significantly change the tradeoff curves. To investigate the sensitivity of the curves to the assumed model of provisioning policy, we considered a different model for the reluctance to remove trunks.

This model (the ω-model) is given by

$$N_{i+1} = \max\{\hat{N}_{i+1}, \omega N_i\}, \qquad (8)$$

where ω is between zero and one. When ω is zero, the ω-model always provides the estimate of the trunks required, just as the β-model, eq. (1), does when $\beta = 1$. Also, when $\omega = 1$ the ω-model never removes trunks, just as with the β-model when $\beta = 0$. Thus, the two models are identical when there is no reluctance to remove trunks, $\omega = 0$ and $\beta = 1$, and when there is complete reluctance, $\omega = 1$ and $\beta = 0$. However, the models are quite different in the case of an intermediate amount of reluctance. The ω-model will freely remove trunks down to ωN_i, but

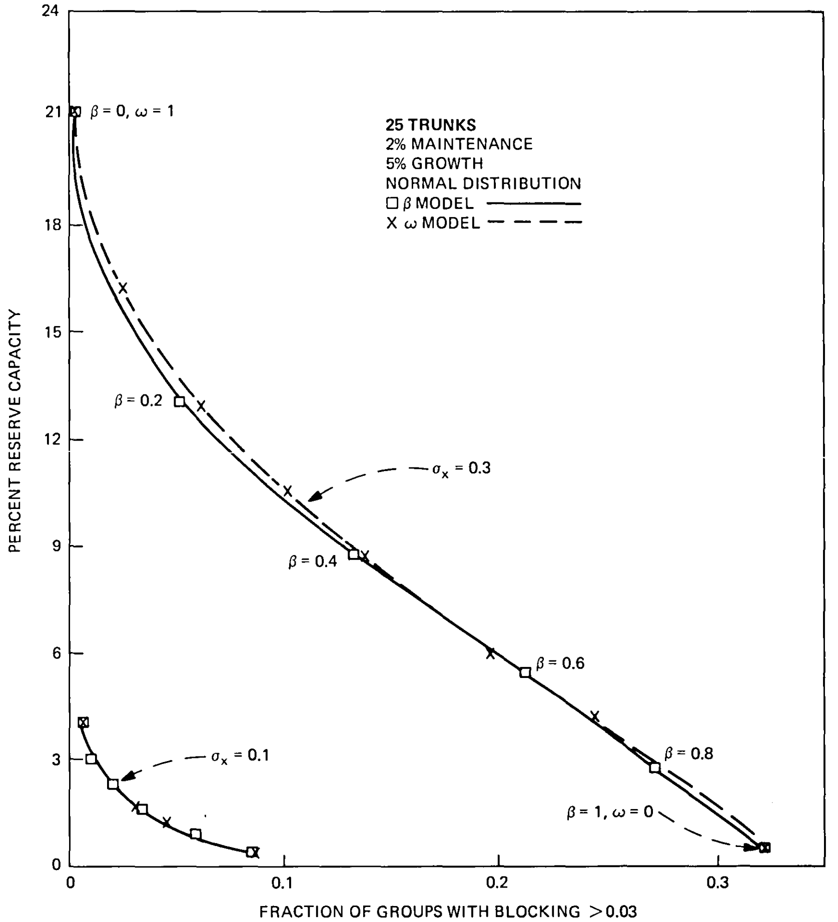


Fig. 5—Sensitivity to reluctance model.

won't remove any beyond that, while the β-model removes a fixed fraction of any excess of $\hat{N}_{i+1}$ over N_i.

TPOCs for each of these models are shown in Fig. 5 for nominal parameters of trunk group size 25, a 5-percent growth rate, fraction of maintenance busy trunks (0.02) and values of σ_x of 0.1 and 0.3 (standard deviation of load forecast error). As expected, the end points of these plots are identical, since for $\omega = 0$, $\beta = 1$ and $\omega = 1$, $\beta = 0$ the models are identical. Furthermore, the curves are seen to be close throughout the entire range. We thus observe that the two reluctance functions provide different mechanisms for tracing out approximately the same tradeoffs.

Figure 6 is a plot of the tradeoff curves for several different growth

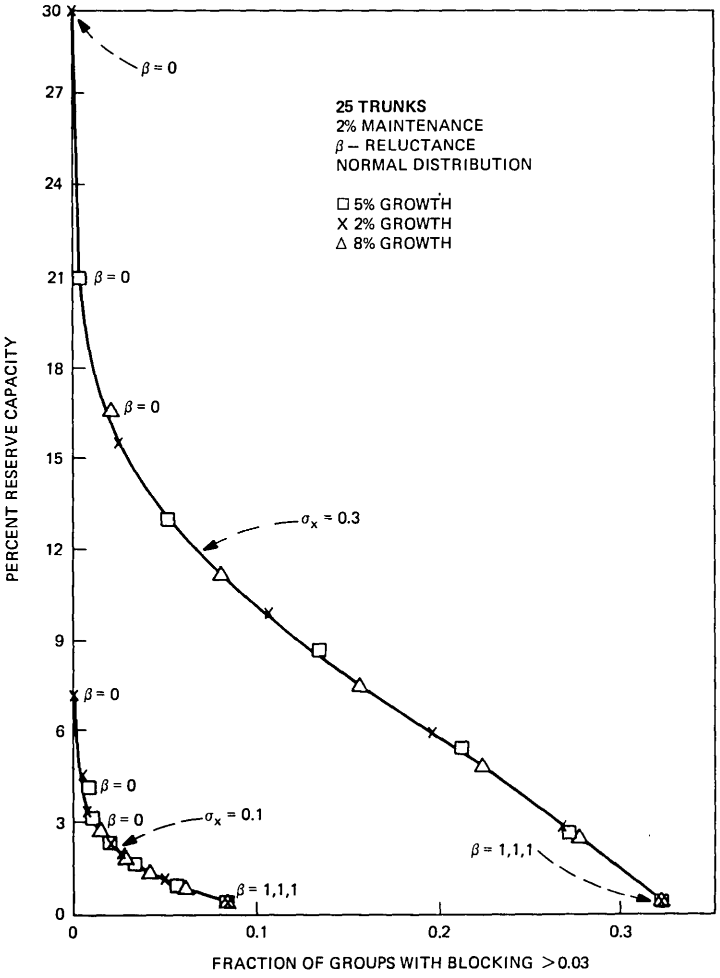


Fig. 6—Sensitivity to growth.

rates (2, 5, and 8 percent). We note that the curves are quite insensitive to the particular traffic growth rate, although the specific operating point, for a given reluctance, can be quite sensitive. Also, the extent of the curves are sensitive to growth. We noted earlier that, for a given growth, increasing reluctance caused one to ride up a tradeoff curve. Here we observe that, for a given reluctance, increasing the growth rate causes one to ride down a tradeoff curve. This is illustrated by the $\beta = 0$ (complete reluctance) points indicated on the figure. We are thus observing, as expected, higher growth rates tending to reduce reserve capacity. The results were also shown to be relatively insensitive to the distribution function of the forecast of trunk requirements and the level of maintenance busies; however, trunk group size* and mean forecasting error did influence the tradeoff curves.

IV. APPLICATION OF THE MODEL TO THE PROVISIONING PROCESS

4.1 Provisioning process issues

Trunk provisioning process issues can be broken down into the general areas:

(*i*) Input data quality.
(*ii*) Improving the use of data.
(*iii*) Monitoring the performance of the provisioning process.

Figure 7 shows a simplified functional view of the trunk provisioning process which we shall use to discuss these issues.

Actual traffic in the network is sensed by a measurement, data collection, and processing system. The data from this system enter a trunk servicing system which estimates current requirements and service and a trunk forecasting system which, with other information, forecasts trunk requirements for 1 to 5 years out. The processed data from the trunk servicing and forecasting systems, together with the source data for these systems, are the focus for assessing input data quality. Typical questions which often arise in this context include:

(*i*) How big are the errors in these data, and what are the main contributors?
(*ii*) To what extent can these errors be controlled by better technology or maintenance?
(*iii*) What are the benefits of such improvements?

The outputs of the trunk forecasting system are used in facility and equipment planning and for trunk servicing, i.e., for deciding on the actual trunks to be provided for the next busy season. Trunk servicers base these decisions primarily on the forecast, but they factor in

* This sensitivity to trunk group size occurs if σ_x is the relative error in the load, as it has been defined in this paper. However, it may be shown that, if the relative error in the number of trunks is used, this sensitivity becomes small.

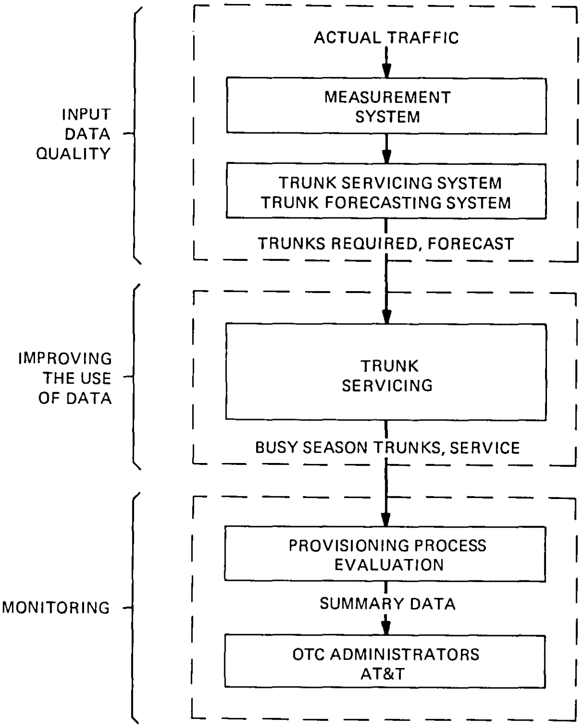


Fig. 7—Simplified view of the trunk provisioning process.

current trunks installed, recent trunk servicing data, facility and equipment availability, etc. This decision-making process is the focus for improving the use of data. Typical questions include:

(*i*) How can all available data be used so that the final judgment on required trunks in the decision-making process (i.e., the "effective" forecast) is an improvement on the quality of the nominal forecast (i.e., the "published" forecast)?

(*ii*) What provisioning policies (e.g., disconnect policies) should be followed, given that all uncertainties in forecasting cannot be removed?

(*iii*) What are the benefits from such improvements in the use of data?

The quality of the input data together with the actions based on the data determine the overall performance of the provisioning process. Data on service, trunks provided, network activity, etc., is gathered by the operating telephone companies, with summary results provided to AT&T for use in assessing and establishing system policies. This evaluation process is the focus of the monitoring issue. Typical questions here include:

(*i*) What data should be used to monitor the provisioning process at the local level; that is, to aid servicers and forecasters directly involved in provisioning, and at the global (or network) level, that is, to aid operating telephone company administrators and AT&T in overseeing the process?

(*ii*) How should the data be displayed and interpreted so that the causes underlying observed results become evident?

(*iii*) How can the monitoring results be used to improve the process?

The major issues which we have outlined can be directly related to the provisioning model developed in Section II. Input data quality is concerned primarily with the "nominal" forecast accuracy, i.e., the accuracy of the published forecast. The key parameters to quantify forecast accuracy are the forecast bias and the forecast standard deviation. Improving the use of data is concerned with decisions which determine operating points along a TPOC, and, perhaps more important, with methods to move to preferred TPOCs by improving on the nominal forecast accuracy (by using all available information). For example, generating a short-term forecast based on the most recent servicing information (traffic data) and combining it with the nominal forecast can result in an improved "effective" forecast. Monitoring is concerned primarily with assessing trunks and service results. The TPOC model clearly shows that one-dimensional measures of performance, such as service, do not provide sufficient information. From the TPOC point of view, methods which exploit knowledge of the reserve capacity and service tradeoff are desirable.

Relating provisioning process issues to the TPOC framework is important for several reasons. First, this framework helps to identify potential improvements to the provisioning process. Perhaps more important, this framework provides a basis for quantifying the benefits associated with such improvement. The ultimate dollar value of any of the improvements considered requires detailed specifications and application studies and consideration of the coupling between trunk servicing and facility and equipment provisioning, which are beyond the scope of this paper.

4.2 Input data quality

The key parameters for assessing input data quality are the nominal forecast bias, a measure of average error, and the nominal forecast standard deviation, a measure of the spread of errors. When analyzing a collection of trunk groups at an office, an error in the office growth factor would tend to show up as a bias in the forecast error on those groups. In practice, this can occur in any given year due, for example, to errors in commercial forecasting of main stations and loads per main station, which are inputs to the trunk forecasting process. For example, in the case of an unexpected economic downturn, which normally results in reduced calling, trunking requirements would tend to be overforecast *on average.* To the extent that this bias is influenced by unexpected exogenous economic factors, it may not be controllable by better measurement technology.

Forecast standard deviation is primarily influenced by inherent variability of traffic (for example, day-to-day variability), measurement errors, and forecasting methodology (for example, the procedure by which aggregate traffic growth is allocated to individual trunk groups). Of these factors, measurement error can be an important, though technologically controllable, factor. For example, it has been generally accepted that wiring errors, which can cause individual trunk usage to be attributed to the wrong group, can occur in conventional Traffic Usage Recorders (TURS). Manual wiring changes in a TUR must be made when trunks are added to, or deleted from, a group, or if trunking rearrangements are made. The buildup of wiring errors is generally controlled by periodic audits. In contrast, an Electronic Switching System records group usage via the same trunk to group association used in the switching process. Thus, it is to be expected that measurement error and forecast standard deviation should be lower for ESS offices as compared to TUR offices. Limited data from a small number of offices supported this expectation, and suggested that the better accuracy of ESS technology may, in some cases, lead to a reduction of several percent in trunks in service, without a consequent degradation in customer service. This benefit can also be achieved by upgrading of conventional measurement technology in electromechanical systems. For example, the Engineering and Administrative Data System (EADAS), when equipped with Individual Circuit Usage Recording (ICUR), maintains the trunk-to-group mapping in software, with a variety of checks to guard against mapping errors.

4.3 *Improving the use of data*

The goal of improving the use of data, as suggested by the TPOC framework, is to control the effective (as opposed to the nominal) forecast accuracy driving the process, and to improve provisioning policies. As in any large scale process involving the acquisition, transfer, and processing of data to produce outputs—with perhaps both manual as well as automated segments—various irregularities can occur in trunk forecasting and servicing processes. Some simple data validation and screening techniques which can help trunk servicers control the effective forecast accuracy exploit: (*i*) incomplete data, e.g., a pattern of missing usage data which may suggest a faulty TUR, (*ii*) inconsistent data, e.g., apparent high blocking on a group with low usage, indicative of a possible data error, (*iii*) data outliers, e.g., unusually large changes in trunk forecasts, and (*iv*) redundant data, e.g., usage data from both ends of a group can be compared for consistency.

As a specific illustration of a method for improving the use of data suggested by the TPOC model, consider short-term forecasting. At the latest time a servicer must make a trunk group provisioning decision, i.e., just prior to an upcoming busy season, available data normally include the forecast trunks based on the last busy-season base load, and a recent traffic profile for the group. This latter data may be appropriately projected ahead to produce a "short-term forecast," essentially independent of the regular forecast. By linearly combining those two forecasts with appropriate weights (see Ref. 9), thus reducing variance, a significantly improved forecast can be constructed. As illustrated by Fig. 8, which assumes a regular and short-term forecast of comparable accuracy, a substantial savings in trunks turned up for service can potentially be achieved. The improvement obviously depends on the relative accuracies of these forecasts which require further studies to quantify.

The actual provisioning decisions by servicers, or more generally, provisioning policies, should reflect the fact that data are imperfect. The TPOC model shows that disconnect policy is a particularly important factor in the performance of the provisioning process and permits

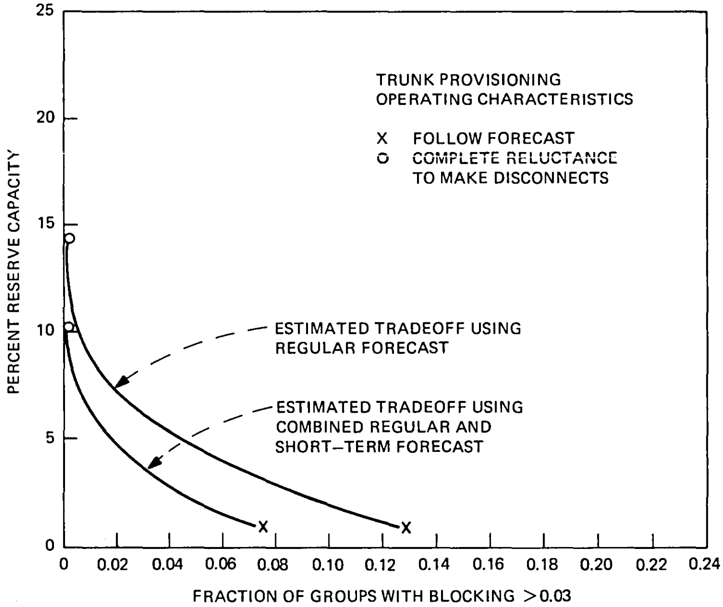


Fig. 8—Potential benefit of short-term forecast [regular and (assumed) short-term forecast standard deviation = 15 percent in trunks].

the study of various policies. In addition to reserve capacity and service, a prudent disconnect policy should factor in the cost of rearrangements and the possibility of unnecessarily disconnecting trunks which are needed in the next year or so. A modification of the TPOC model has been used by N. E. Kalb to examine various disconnect policies and define an appropriate one for Bell System use.

4.4 Monitoring

The goal of monitoring is to track the performance of the provisioning process, and to allow the underlying factors to be identified which contribute to that performance. For displaying performance results for particular administrative units, e.g., an operating division, the TPOC perspective indicates the importance of considering both reserve capacity and service, as illustrated with assumed data in Fig. 9. We note that comparing these assumed data points by service could be very misleading, i.e., the "best" service is obtained at the price of very high reserve capacity, while the "worst" service corresponds to a moderate level of reserve capacity. If one now further exploits the TPOC model and views each point as representing a particular operating point along a TPOC tradeoff curve, the data point that looks most like a weakspot (i.e., appears to lie on the worst tradeoff curve) does not have the worst service or reserve capacity. Similarly, the entity that might be considered to have the best performance does not correspond to either the

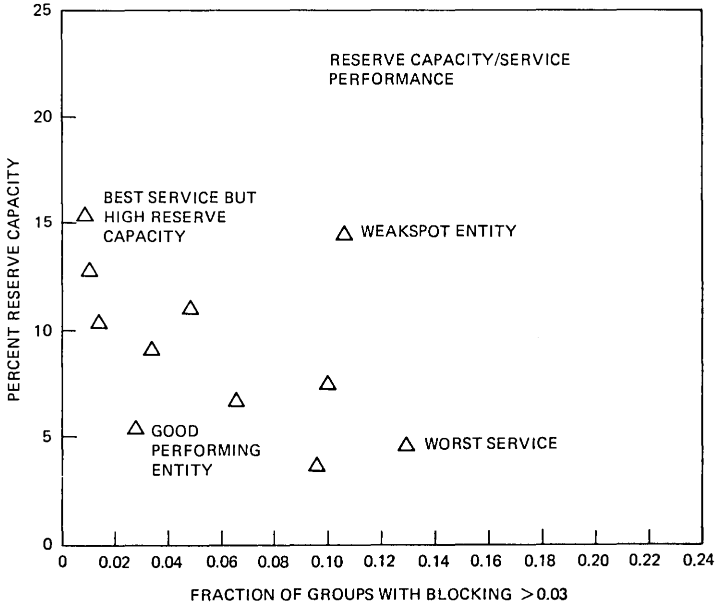


Fig. 9—Reserve capacity vs service display.

lowest reserve capacity or the best service. In both these cases, the actually operating points may have resulted from prudent disconnect policies that achieved a reasonable balance between reserve capacity and service, avoiding the extremes of no disconnects or of blindly following the forecast.

The preceding perceptions about performance, and the need to encourage an appropriate balance between reserve capacity and service, might be emphasized by subdividing the reserve capacity and service plane into appropriate regions, based on preferred operating points. This could result in a performance monitoring approach such as shown by Fig. 10, which clearly facilitates the comparison and ranking of overall performance of the various assumed data points. The shape of the performance regions shown in Fig. 10 reflect the undesirability of a disconnect policy which works at the tradeoff extremes. The performance partitions—which ideally would be designed to reflect equally desirable operating points, with cost factors considered—provide motivation both to improve forecast accuracy and to control disconnect policy. Further study would be required to determine an actual partitioning into performance regions.

In addition to allowing cross-comparisons of performance, monitoring should allow the underlying factors which contributed to the performance of a particular office to be identified. Again, the TPOC model suggests an approach. By supplementing the data needed to

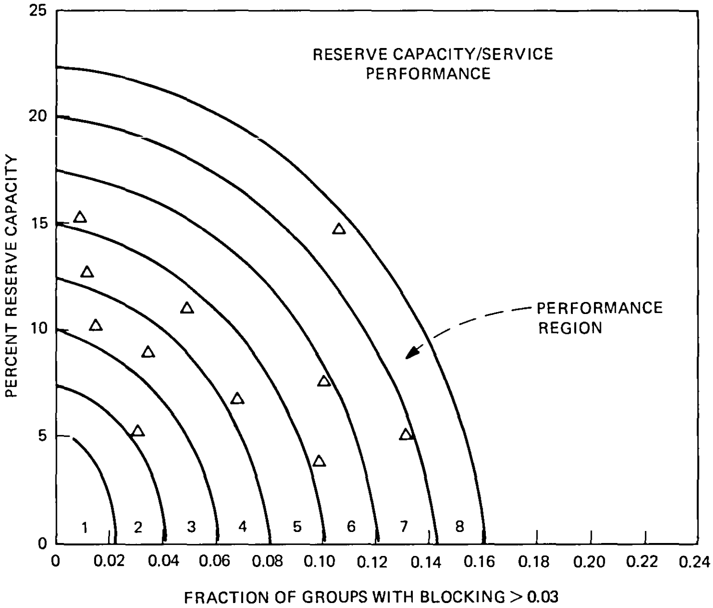


Fig. 10—Reserve capacity vs service display with (assumed) performance regions.

estimate service and reserve capacity with information on the previous busy season trunks, and the trunks forecast for the busy season, the operating points that would have corresponded to complete reluctance to disconnect and follow forecast can be computed.* By displaying these points with the actual operating point, as illustrated in Fig. 11, significant information is obtained. For example, it is immediately clear that reluctance to disconnect trunks could potentially have contributed about 12 percent in trunks to reserve capacity (labeled potential reluctance). However, the actual operating point indicates a prudent disconnect policy, with net reluctance adding only about 4 percent to reserve capacity. In contrast, the bias (perhaps due to a growth slowdown) has added about 5 percent to reserve capacity and is a dominant factor. The end result of this type of performance analysis is a decomposition of reserve capacity into basic components,† due to forecast bias plus actual reluctance, together with a relative measure of disconnect policy, i.e., actual disconnects/potential disconnects. One may also view the final service as being decomposed according to: follow forecast service minus service benefit of actual reluctance. Such graphical analysis, together with associated quantifications from the data (e.g., forecast standard deviation), could provide significant insight into provisioning performance and suggest where improvement is needed.

In addition to comparative evaluations and to evaluation of key factors on the global, or network, levels, it is important to provide information to servicers and forecasters responsible for decisions on individual trunk groups. Simple graphical displays using the same raw data needed for global evaluations can be used to provide servicers, forecasters, and trunk administrators with clear pictures of performance, and of the role of forecast accuracy, disconnect policy, and other aspects of the process. For example, Fig. 12 illustrates a graphical aid which can be used to display and quantify forecasting accuracy; a similar display using trunks in service and required can be used to display reserve capacity and service. Figure 13 illustrates a graphical aid which can be used to examine disconnect policy. Such displays also quickly point out unusual actions, such as the addition of trunks when the forecast calls for removals. This type of action may have been initiated if the ongoing servicing data from the group indicated higher traffic levels than those forecast.

* It turns out that over a fairly wide range of group sizes, for example ≥10 trunks, the theoretical equivalent to P.03 service for direct final groups corresponds to an almost constant percent trunk shortage (about 13 percent).

† One may further extend this analysis (using the same data) to account for cases where all forecast adds may not have been made. This can occur when a growth slowdown has been recognized.

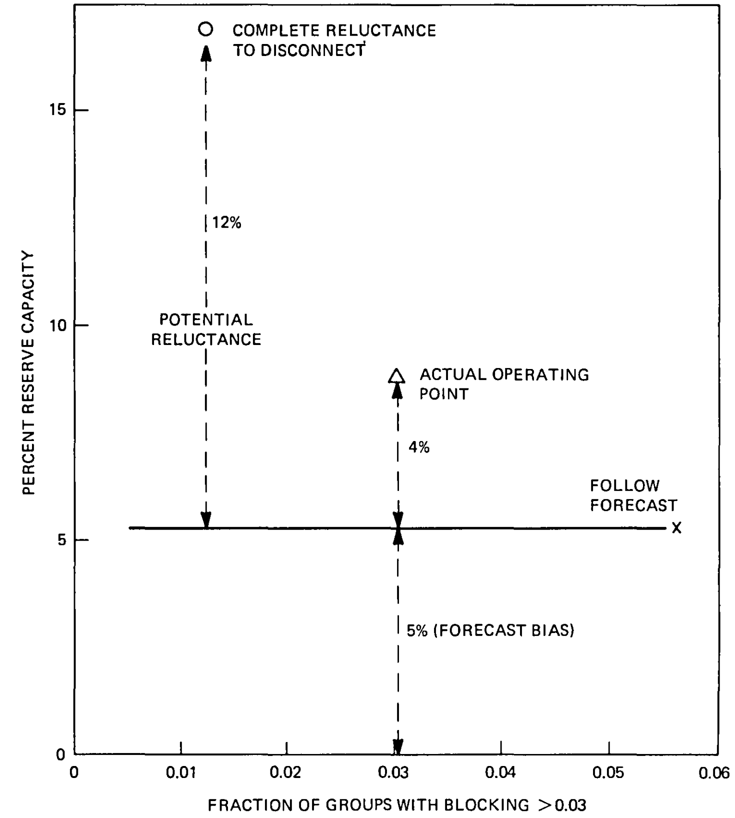


Fig. 11—Performance analysis for a particular trunking entity—with forecast and previous trunks connected information.

In summary, the TPOC model—as discussed for direct final groups—suggests a global or network monitoring technique that allows both comparative evaluations and analysis of key factors. This can be useful, for example, in determining if disconnect policy needs changing. At the same time, those responsible for decisions at the lower echelons can be provided with a set of more detailed displays that highlight the same factors, as well as any individual anomalies in the data that should be investigated. While these displays are not now available, the capability to produce such graphical aids could be an important enhancement to current trunk servicing and forecasting systems.

V. CONCLUDING REMARKS AND FURTHER WORK

In this paper, we have developed a mathematical model of the trunk provisioning process which allows one to study the relationship of traffic measurement and forecasting errors and disconnect policy to

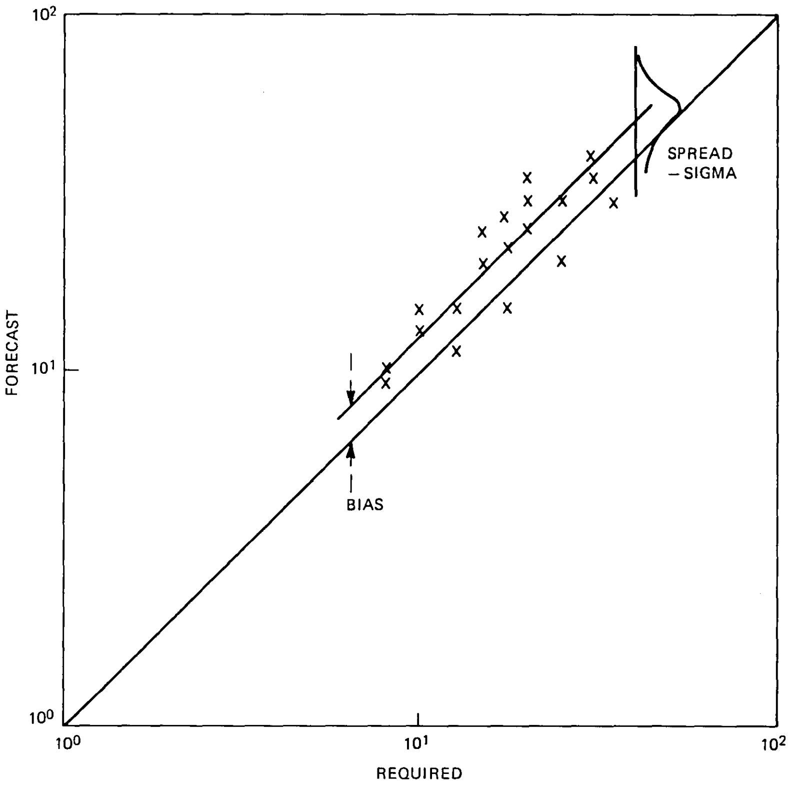


Fig. 12—Trunks forecast vs trunks required.

reserve capacity and service. We reemphasize that this is a first-cut model based on many simplifying assumptions.

Direct final groups were considered since they represented the simplest case.* The results were displayed as trade-off curves of percent reserve capacity vs percent of groups with blocking >3 percent. The curves were parameterized by σ_x, the normalized error in forecast load.

These TPOC curves were found to be relatively insensitive to reluctance policy, probability distribution of errors, and growth.† This robustness of the curves is useful not only from the point of view of convenience of presentation but also because of our uncertainty concerning these quantities. The basic TPOC concepts were validated using field data collected on approximately 250 direct final trunk groups. In

* It might be noted that they also represent a large fraction of the total Bell System trunks and groups.

† Where one operates on the curve depends on the reluctance policy and growth.

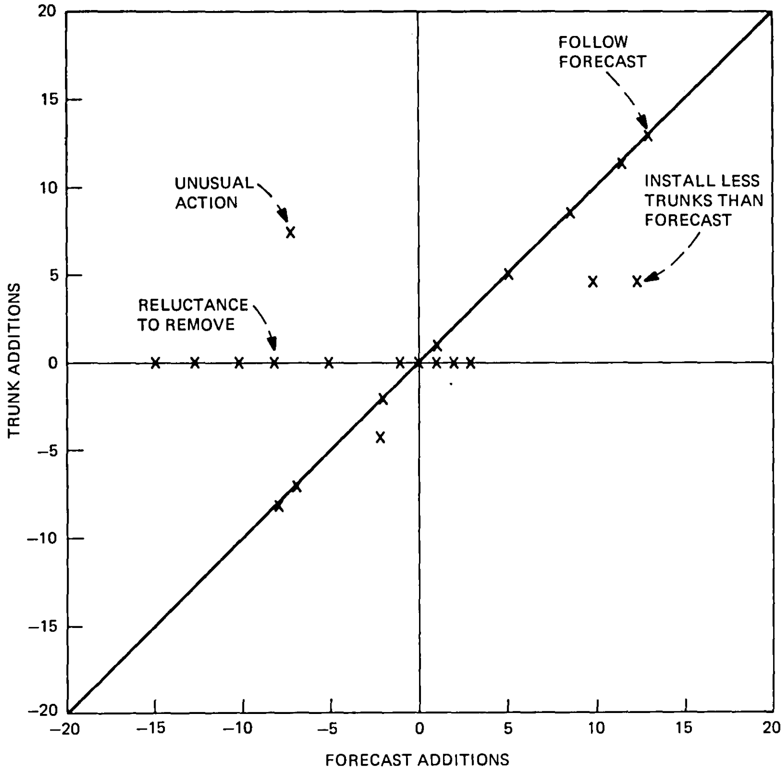


Fig. 13—Trunk additions vs forecast additions.

the course of this effort, the TPOC model was extended to include nonzero mean forecast error, and the effect of day-to-day variations and to reflect the fact that estimates of reserve capacity and service were themselves corrupted by various errors. It may be noted the nonzero mean forecast error enters the TPOC model as a persistent error for a single trunk group over a sequence of years. While the data were obtained for a set of trunk groups over one year, the model still yielded good accuracy.

Much remains to be done to further develop and extend the concepts we have described. We indicate here only a few of the many areas worthy of further study. The results we have described assume an equilibrium condition corresponding to reluctance and true growth, which are constant from year to year. These assumptions could be relaxed and transient solutions developed in more detail than we have done. Extension of the modeling approach to other probability-engineered groups would seem like an obvious next step, while the appropriate framework for viewing network clusters involving high-usage groups would require more work. In addition, further work needs to be done to characterize the dynamics of growth factor errors which can result in biases in a particular year.

To capitalize on the ideas we have discussed, some of the obvious problem areas which present themselves involve the development of improved forecasting capabilities, disconnect policies (i.e., "reluctance"), and data screening techniques. As mentioned at the end of Section 4.3, these ideas have already been used to study disconnect policies.

Finally, we note that, to draw conclusions about the value of specific proposed improvements (e.g., an improved measurement system), the overall forecast error must be decomposed into its component parts and each of these components must be quantified for each alternative under consideration. In addition, it should be kept in mind that casual attempts to plot reported trunking data on existing TPOC curves can be misleading because of factors that might be influencing the data but left out of the curves (trunks installed as a result of new machine cutovers, etc.).

Additional work (e.g., further modeling and statistical analysis of usage errors) is needed to more accurately quantify the effect of improvements in measurements (or forecasting) when one is dealing with a nonuniform* collection of trunk groups.

Even with the additional work that is warranted, it is seen from Section IV that the TPOC model is a powerful aid to identify issues and to suggest improvements and ways of evaluating them.

VI. ACKNOWLEDGMENTS

We thank Anne Carlson, Joanne Darcy, Nancy Kalb, Dave Jagerman, and Sharon Miller for their valuable contributions. We also received much valuable information from a number of other people at Bell Labs and AT&T and Pacific Telephone (from whom we obtained data for validation).

APPENDIX

Model Development

A.1 Model and definitions

Our model for the provisioning of direct final groups with no day-to-day load variation assumes that peg count, overflow, and usage are

* Nonuniformity can result from dealing with groups with different measurement devices (for example, ESS and TURS). Furthermore, even with the same measurement devices (e.g., all TUR measurements), one can expect differences in accuracies (such as well-maintained TUR vs poorly maintained TUR).

measured during a base period, such as a busy season. The offered load is estimated according to eq. (3), and the next period's (e.g., next busy season's) offered load is found by multiplying by a projection ratio. The estimated number of trunks required for the next busy season is such that the projected offered load would cause a blocking probability of 0.01. Once the number of trunks required for next year, $\hat{N}_{i+1}$, has been estimated, the number actually provided, N_{i+1}, is chosen using eq. (1), or eq. (8), depending on the provisioning policy modeling. This model assumes trunks are provided only once each period with measurements from the last period.

In order to write the model equations we need some definitions.

u_i = Mean of the true traffic usage during study period i.
$\hat{u}_i$ = Measured traffic usage during study period i.
m_i = Actual maintenance usage during study period i.
$\hat{m}_i$ = Measured maintenance usage during study period i.
U_i = Mean of the true total (maintenance plus traffic) usage during study period i.
$\hat{U}_i$ = Measured total (maintenance plus traffic) usage during study period i.
eu_i = $\hat{u}_i - u_i$, usage error due to TUR wiring, etc. with separate traffic usage measurement.
eU_i = $\hat{U}_i - u_i - m_i$, usage error due to TUR wiring, etc. with joint usage measurement.
a_i = Mean of the true offered load during study period i.
$\hat{a}_i$ = Estimate of a_i based on measurements during study period i.
$\tilde{a}_{i+1}$ = Estimate of a_{i+1} based on measurements during study period i.
M_i = Number of traffic trunks required in study period i. (e.g. $B(M_i, a_i) = 0.01$).
$\hat{M}_{i+1}$ = Estimate of M_{i+1} based on measurements during study period i.
N_i = Number of trunks in place in study period i (includes maintenance).
$\hat{N}_{i+1}$ = Number of trunks estimated as required for period $i + 1$.
g_i = a_{i+1}/a_i, the traffic growth.
$\hat{g}_i$ = The estimate of g_i.
$\hat{B}_i$ = Measured average fraction overflowing during study period i.
B_i = Mean of the fraction overflowing during study period i.
$eB_i = \hat{B}_i - B_i$.
$eg_i = \hat{g}_i - g_i$.

A.2 Estimating $\hat{N}$ from lumped traffic and maintenance usage

Estimating $\hat{N}$, in terms of its mean and variance, requires estimating a_i. The estimate of a_i depends on whether or not maintenance usage is measured separately.* If it is not, then $\hat{U}_i$ is measured and we estimate a_i by

$$\hat{a}_i = \frac{\hat{U}_i}{1 - \hat{B}_i} = \frac{m_i + u_i + eU_i}{1 - [B(N_i - m_i, a_i) + eB_i]} \tag{9}$$

$$\approx a_i\left[1 + \frac{m_i}{u_i} + \frac{eU_i}{u_i} + \frac{eB_i}{1 - B_i}\right]. \tag{10}$$

The estimate of the next period's offered load is

$$\tilde{a}_{i+1} = \hat{g}_i\hat{a}_i \approx a_{i+1}(1 + x_i), \tag{11}$$

where

$$x_i = \frac{m_i}{u_i} + \frac{eU_i}{u_i} + \frac{eB_i}{1 - B_i} + \frac{eg_i}{g_i}. \tag{12}$$

Since $B(\hat{N}_{i+1}, \tilde{a}_{i+1}) = 0.01$ is the design criterion, expanding and regrouping terms yields

$$\frac{\hat{N}_{i+1}}{M_{i+1}} = 1 + \left(-\frac{B_2}{B_1}\ \frac{a_{i+1}}{M_{i+1}}\right)x_i = 1 + cx_i, \tag{13}$$

where

$$B_1 = \frac{\partial B}{\partial N}\bigg|_{M_{i+1}, a_{i+1}},\ B_2 = \frac{\partial B}{\partial a}\bigg|_{M_{i+1}, a_{i+1}}, \qquad c = -\frac{B_2}{B_1}\ \frac{a_{i+1}}{M_{i+1}}. \tag{14}$$

By (14), c depends only on M_{i+1}. It is 0.647 for $M_{i+1} = 10$ and increases toward 1 as M_{i+1} increases.

To find the mean of $\hat{N}/M$, we assume that all measurements and estimates are unbiased. That, together with (12), (13), and the definitions, gives

$$E\left(\frac{\hat{N}_{i+1}}{M_{i+1}}\right) = 1 + c\frac{Em_i}{u_i} = 1 + d\frac{Em_i}{M_i}, \tag{15}$$

where

$$d \triangleq c\frac{M_i}{u_i} = -\frac{B_2}{B_1}\ \frac{a_{i+1}}{u_i}\ \frac{M_i}{M_{i+1}}. \tag{16}$$

* Frequently, maintenance usage is not measured separately.

By their definitions, u_i depends only on M_i so d, which is approximately 1, depends only on M_i and M_{i+1}.

To find the variance of $\hat{N}/M$ we assume that the terms in (12) are statistically independent. That gives

$$\mathrm{Var}\left(\frac{\hat{N}_{i+1}}{M_{i+1}}\right) = c^2\sigma_x^2, \tag{17}$$

where

$$\begin{aligned}
\sigma_x^2 &= \sigma_{U_i}^2 + \sigma_{B_i}^2 + \sigma_{g_i}^2 \\
\sigma_{U_i}^2 &= \sigma_{m_i}^2 + \sigma_{u_i}^2 \\
\sigma_{m_i}^2 &= \mathrm{Var}\left(\frac{m_i}{u_i}\right), \qquad \sigma_{u_i}^2 = \mathrm{Var}\left(\frac{eU_i}{u_i}\right), \qquad \sigma_{B_i}^2 = \mathrm{Var}\left(\frac{eB_i}{1-B_i}\right), \\
\sigma_{g_i}^2 &= \mathrm{Var}\left(\frac{eg_i}{g_i}\right).
\end{aligned} \tag{18}$$

In the rest of the paper we assume that

$$\mathrm{Var}\,\frac{eU_i}{u_i} = \mathrm{Var}\,\frac{eu_i}{u_i},$$

since neither contains variation due to maintenance usage, and all the other errors have the same sources.

A similar analysis is possible for the case when the traffic usage is measured separately from the maintenance usage.

A.3. Solution of model equations

The solution of eq. (2) for the distribution of N_i/M_i has been obtained in several ways. Analytic approximations and characterizations of properties of the solution have been developed by D. L. Jagerman, while the actual TPOC curves presented in this paper were obtained by employing simulation techniques to solve (2).

REFERENCES

1. *Engineering and Operations in the Bell System,* Bell Laboratories, 1977, Chapter 14.
2. W. C. Johnson and W. P. Maguire, "The Trunk Forecasting Systems Tells What, Where, When," Bell Laboratories Record, June 1974, pp. 193–195.
3. R. J. Armstrong, R. Gottdenker, and R. L. Kornegay, "Servicing Trunks by Computer," Bell Laboratories Record, February 1976, pp. 39–44.
4. A. Parviala, "Accuracy Requirements Concerning Routine Traffic Measurements With Regard to Service Level Objectives in Telephone Network and to Certain Error and Cost Factors," Proc. Eighth Int. Teletraffic Congress, Melbourne, Australia, 1976, pp. 245/1-7.
5. J. J. O'Shaughnessy, "Traffic Data—The Need, Nature and Use," Proc. Eighth Int. Teletraffic Congress, Melbourne, Australia, 1976, pp. 241/1-8.

6. A. Elldin, "Dimensioning for the Dynamic Properties of Telephone Traffic," Ericsson Technics No. 3, 1967, pp. 315–344.
7. D. W. Hill and S. R. Neal, "Traffic Capacity of a Probability Engineered Group," B.S.T.J., *55,* No. 7 (September 1976), pp. 831–842.
8. E. J. Messerli, "An Approximation for the Variance of the UPCO Offered Load Estimate," Intl. Symp. on Measurements in Telecommunications, Lannion France, Oct. 3–4, 1977.
9. J. M. Bates and C. W. Granger, "The Combination of Forecasts," Oper. Res. Quar., *20,* No. 4, pp. 451–468.

Copyright © 1979 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 58, No. 2, February 1979
Printed in U.S.A.

A New and Interesting Class of Limit Cycles in Recursive Digital Filters

By V. B. LAWRENCE and K. V. MINA

(Manuscript received February 1, 1978)

Limit cycles oscillations often occur in recursive digital filters due to the quantization of products in the feedback section. A new and interesting class of limit cycles has been discovered and categorized for second-order sections that round either sign-magnitude or twos-complement products. These limit cycles are named rolling-pin limit cycles. They are completely defined by three integers and a simple construction rule and exist for $B_1 - B_2$ pairs lying within small rectangular regions in the $B_1 - B_2$ (coefficient) plane. Each set of integers completely defines the peak amplitude, the length, and the region of existence. The amplitude of these limit cycles can be made close to but does not exceed three times the Jackson peak estimate. Rolling-pin limit cycles often occur in filters with high Q poles located near dc or half the sampling frequency. When these large amplitude limit cycles occur, the idle channel performance of a filter is often unacceptable. Specialized techniques, requiring extra circuitry, can be used to suppress them. Alternatively, it may be less costly and more efficient to avoid the small rectangular regions within which the rolling-pin limit cycles exist in the $B_1 - B_2$ coefficient plane.

I. INTRODUCTION

Oscillations often occur in recursive digital filters as a result of the nonlinear action of quantizing the products in the feedback sections. These oscillations occur in the least significant bits of the data and are called limit cycles.* These limit cycles influence the required internal data word length and hence the cost of the filter.[1]

This paper describes a new and important class of limit cycles that exist in second-order recursive digital filters. These limit cycles often occur in filters with high Q poles located near dc or half the sampling

* These limit cycles are to be distinguished from the large limit cycles caused by overflow.

frequency. They are called rolling-pin limit cycles and derive this name from their characteristic shape when plotted in the successive value plane (Fig. 2a). The second-order section under consideration employs rounding of both feedback products in either sign-magnitude or twos-complement number format and is shown in Fig. 1. The rolling-pin limit cycles are defined by three integers, K, L, and M, and a simple construction rule. As seen in Fig. 2b, K is the constant step size in the handle of the rolling pin, L is the constant step size in the body of the rolling pin, and M is the number of steps of step size L. For each value of K, L, M, a unique set of limit cycles is completely defined. In this paper, emphasis is on the simplest case of $K = 1$.

This class of rolling-pin limit cycles is important because of its unusually large amplitude. Unusually large amplitude limit cycles can lead to severe distressing tones in idle channel conditions. It will be shown that the peak amplitude approaches three times Jackson's peak estimate.* The concept of regions of the $B_1 - B_2$ plane within which complicated limit cycles exist is important. The existence of these isolated areas within which the various K, L, M limit cycles exist presents an interesting "patchwork quilt" look in the $B_1 - B_2$ plane. A point of practical importance to note is that a small change in binary coefficient values (producing a pair of coefficients just outside the region of existence of rolling-pin limit cycles) can result in a 3:1 reduction in ac limit cycle amplitude. By ac limit cycles, we mean limit cycles with period >2.

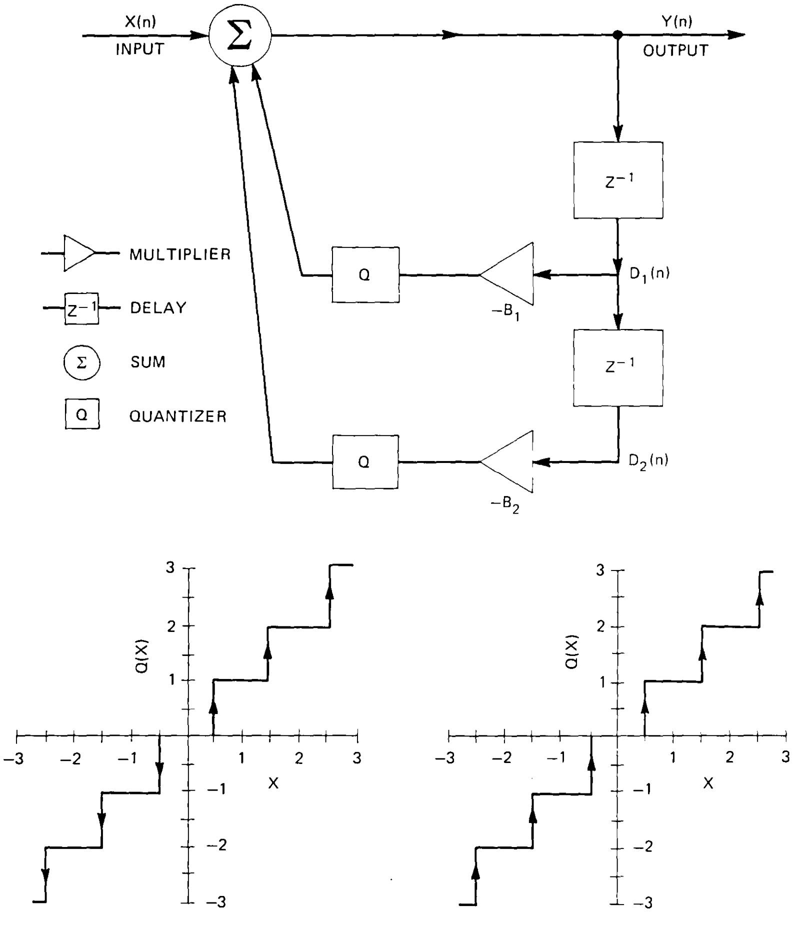
Specialized techniques requiring extra circuitry[2-6] can be used to suppress rolling-pin limit cycles. These specialized techniques may increase the roundoff noise in the presence of a signal. Alternately, it may be less costly and more efficient to avoid rolling-pin limit cycles altogether, by avoiding the small rectangular regions within which they exist in the $B_1 - B_2$ (coefficient) plane.

In the sections that follow, we develop the explicit formulas for the peak amplitude, length, and regions of existence for the 1, L, M rolling-pin limit cycles. The various amplitude bounds and estimates that have been derived[7-12] in the past are examined relative to our exact peak values. A spectral analysis of rolling-pin limit cycles is also presented. A comparison of the roundoff noise and limit cycle power of a second-order section is made.

II. PROPERTIES OF *1, L, M* ROLLING-PIN LIMIT CYCLES

It is important to get an overall view of the nature of this class of limit cycles before examining their detailed properties. For each value of L, M (the 1 will be omitted hereafter), a known number of rolling-pin limit cycles exist with identical bodies but handles that differ in

* Jackson's estimate (Ref. 7) is the integer part of $[0.5/(1 - B_2)]$.



SIGN–MAGNITUDE ROUNDING TWOS–COMPLEMENT ROUNDING

Fig. 1—(a) Second order pole section. (b) Quantization characteristics.

amplitude by one and overall length by four. For example, with L, M = 3, 4, there are three rolling-pin limit cycles with identical bodies but different peak amplitudes of 13, 14, and 15 and overall lengths of 34, 38, and 42, respectively. The successive value plane plots (hereafter called the $D_1 - D_2$ plots) for these three limit cycles are shown in Fig. 3. Each has its own cell of existence in the $B_1 - B_2$ plane. The three cells are horizontally contiguous as shown in Fig. 4. The boundaries on B_2 are

$$1 - (1/2)(1/9) > B_2 \geq 1 - (1/2)(1/8), \qquad (1)$$

and on B_1 are

$$2 - (3/2)(1/15) > |B_1| \geq 2 - (3/2)(1/14) \text{ (amplitude 15)}$$

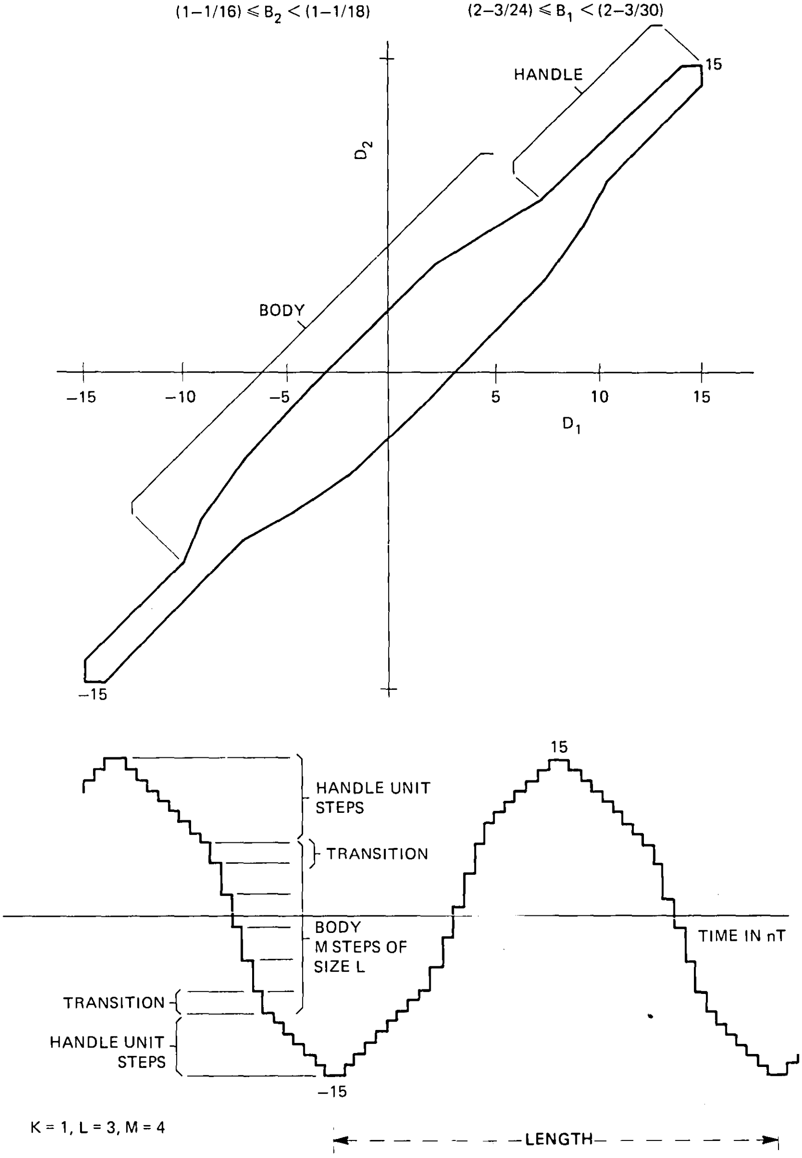


Fig. 2—(a) Successive value ($D_1 - D_2$) plot. (b) Time sequence for K, L, M = 1, 3, 4.

$$\begin{aligned} 2 - (3/2)(1/14) &> |B_1| \geq 2 - (3/2)(1/13) \text{ (amplitude 14)} \\ 2 - (3/2)(1/13) &> |B_1| \geq 2 - (3/2)(1/12) \text{ (amplitude 13).} \end{aligned} \tag{2}$$

In Fig. 4, only negative values of B_1 (low-frequency poles) are shown; however, positive values of B_1 also produce related[12] limit cycles. For convenience, we shall emphasize negative values of B_1. The boundaries

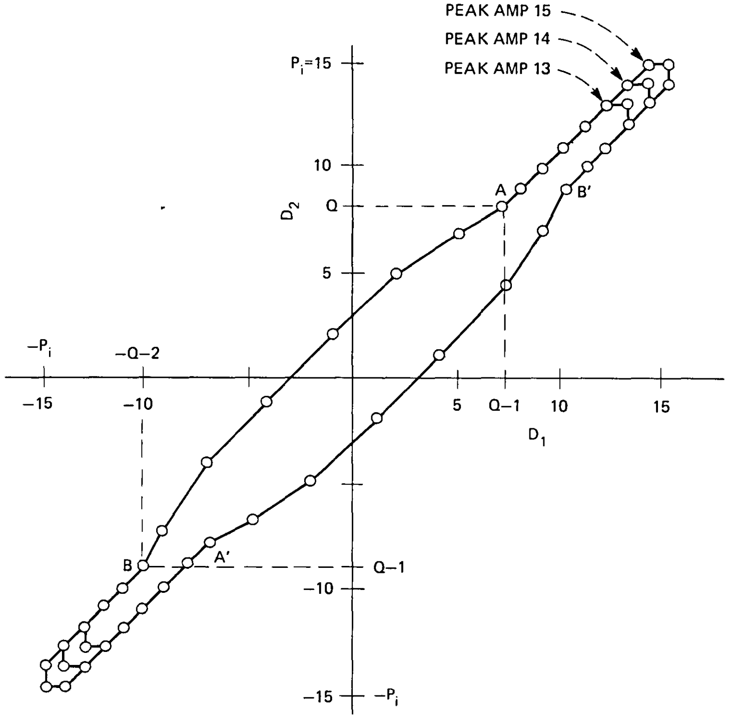


Fig. 3—Successive value ($D_1 - D_2$) plot for $L, M = 3, 4$.

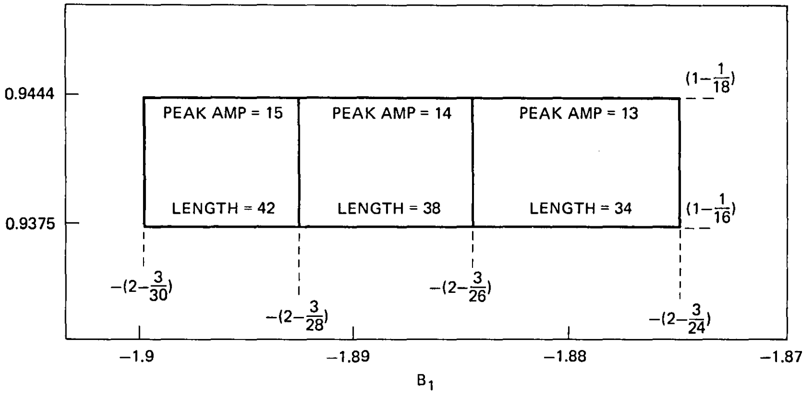


Fig. 4—Region of existence ($B_1 - B_2$ plane) for $L, M = 3, 4$.

defined in eqs. (1) and (2) are rational decimal (and rarely binary) numbers; however, in any implementation, the coefficients are binary. Accordingly, the equality signs in these constraints are usually not applicable.

For other values of L, M, similar sets of nearly identical limit cycles exist with either $3(L-2)$ or $3(L-3)$ contiguous cells of existence depending on whether the product L, M is even or odd. For each L there is a minimum value of M (but no maximum value) below which a type of degeneracy occurs in which the handle disappears into the body. Figure 5 shows the regions in the $B_1 - B_2$ plane for which each set exists for $L = 3, 4, 5, 6$ and four values of M for each L. In each case, the boundaries on B_2 are of the form:

$$1 - 1/[2(Q+1)] > B_2 \geq 1 - 1/(2Q), \tag{3}$$

and the boundaries on B_1 are

$$2 - 3/(2P_i) > |B_1| \geq 2 - 3/(2P_{i-1}) \tag{4}$$

for $i = 1, 2, \cdots, (L-2)$ or $(L-3)$. Both Q and P_i are defined in terms of L and M in Section III. P_i is the peak amplitude of the limit cycle. The allowed B_1 and B_2 values always correspond to complex conjugate poles. This can be proved using eqs. (3), (4), (6), and (12).

Whenever either B_2 or B_1 is varied so that the outer boundary of a set of limit cycles is crossed, the rolling-pin limit cycle disappears and is replaced by a "normal" ac limit cycle with a smaller amplitude. The magnitude of the limit cycle is predicted by Jackson's estimate.[7] This effect should be borne in mind during the design of recursive filters, as an infinitestimal (binary) change in coefficient (and hence transfer function) can result in dramatic changes in the ac limit cycle properties. However, dc limit cycles larger than Jackson's estimate can still exist.

This class of limit cycles has the unique feature that, given L, M, all states and properties are completely defined. In the following sections, the construction rules for rolling-pin limit cycles are given and formulas are derived for their peak amplitudes, length, and $B_1 - B_2$ boundaries in terms of L and M.

III. CONSTRUCTION RULES FOR *L, M* ROLLING-PIN LIMIT CYCLES

The construction rules for rolling-pin limit cycles are best understood by reference to Fig. 6, where a time sequence of an arbitrary rolling-pin limit cycle is shown. The 1, L, M rolling-pin limit cycles have amplitudes with half-period odd symmetry, i.e.,

$$Y\left(n + \frac{N}{2}\right) = -Y(n), \tag{5}$$

where N is the period (or length) of the limit cycle and is an even number. The rules differ slightly for even and odd values of the product L, M; thus they are discussed separately.

3.1 Even L, M

As evident in Fig. 6, a rolling-pin limit cycle consists of sections of constant slope (i.e., constant first difference) separated by smooth

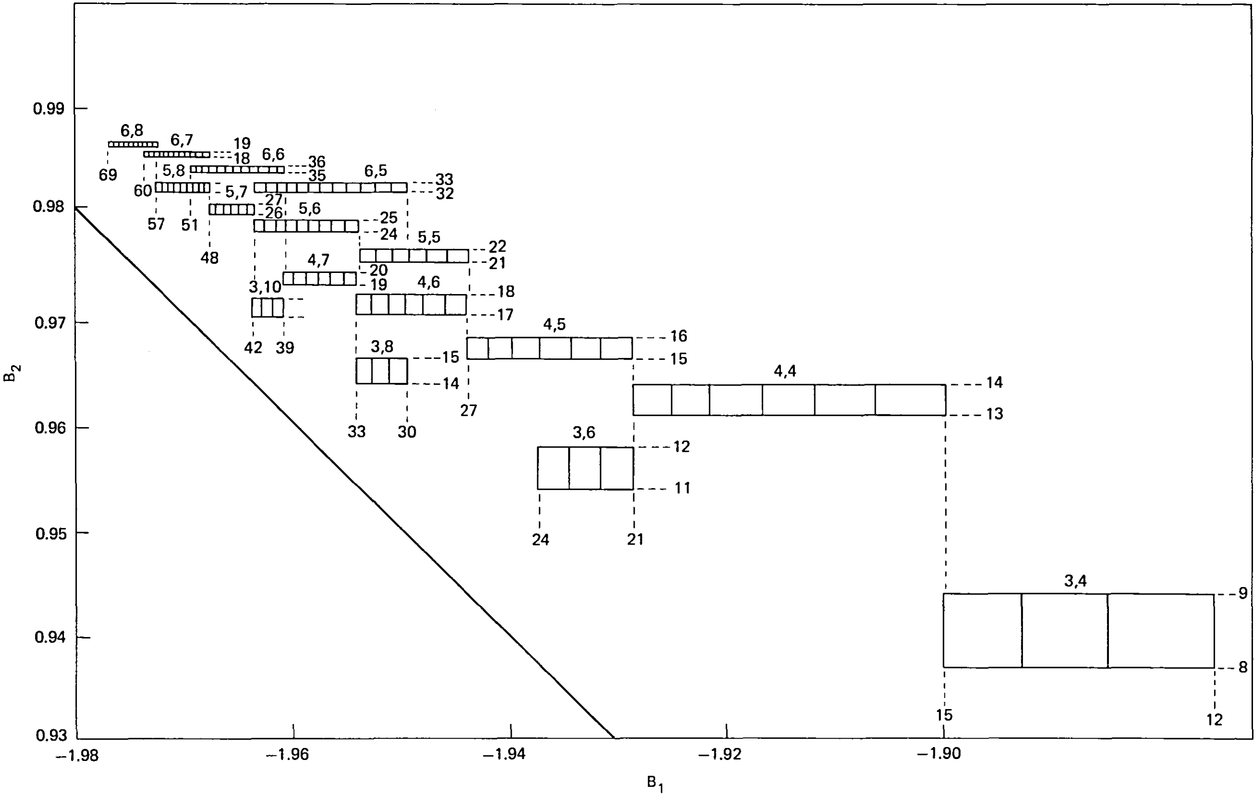


Fig. 5—Region of existence ($B_1 - B_2$ plane) for 4 sets of rolling-pin limit cycles. The bounds on B_1 are 2-3/(2i) and on B_2 1-1/(2i) where the four appropriate values of i are shown by each rectangle.

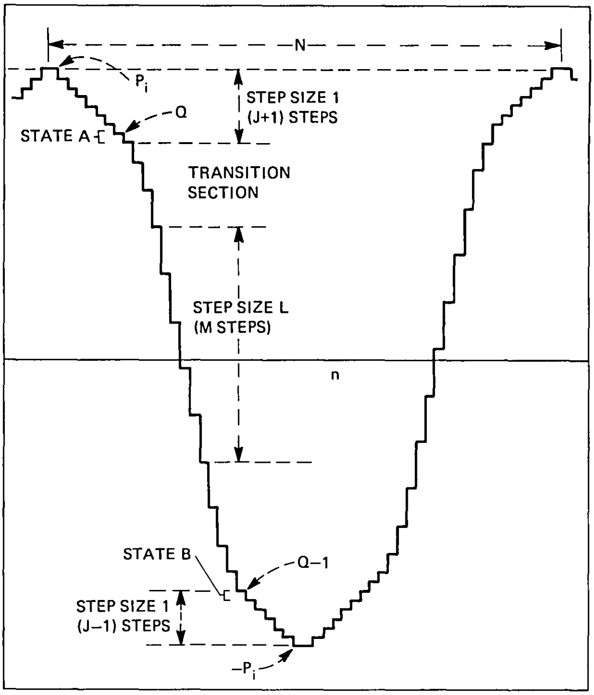


Fig. 6—Time sequence of 1, L, M limit cycle.

transition regions (of constant second difference). Due to the odd symmetry, we need consider only half a cycle of the time sequence. Starting from the positive peak, we find that a zero crossing occurs during the decreasing section of constant step size L. This section has exactly M steps and is preceded and succeeded by smooth transitions to sections of constant unit step size. The transition section is of length $L - 2$. The handle and the body of the rolling pin are pictorially obvious in the $D_1 - D_2$ plot in Fig. 2a; however, certain ambiguities arise in the discussion of the time sequence. First, every point in the $D_1 - D_2$ plot corresponds to a pair of values in the time sequence. Second, the transition region may be associated with either the handle or the body. The following convention will be used. The transition region is in the body of the rolling pin. The intersection of the handle and the transition region is also on the body. A value in the time sequence will be referred to as being on a certain part only when it is unambiguous. This would occur when both $D_1 - D_2$ points associated with the sample values were in the same part.

For a given L, M, the bodies of all the constituent limit cycles are identical. It is the handle that varies in amplitude. Any single point on the body can be used to start the construction. We choose a value, Q, which is associated with the intersection of the handle and the body (state A in Fig. 7). This value Q, labeled in the time sequence in Fig. 6, is the next to last point on the unit step section immediately after the positive peak. For *all* even L, M,

$$Q = LM/2 + L(L-1)/2 - 1. \tag{6}$$

The point Q defines the B_2 bounds within which the limit cycle exists by

$$1 - \frac{1}{2(Q+1)} > B_2 \geq 1 - \frac{1}{2Q}. \tag{7}$$

Thus, for any even L, M we have defined the body of the rolling pin as well as the values of B_2 for which it exists. Continuing on the decreasing portion of the limit cycle, the intersection of the body and the handle reoccurs at point B in Fig. 7. The D_1, D_2 values of B are $-Q-2$, $-Q-1$. To verify this, we start at A, with $D_1^A = Q - 1$ and counting to B using the construction rule just outlined, we have

$$D_1^B = D_1^A - 2 - 3 \cdots - (L-1) - LM \\ - (L-1) - (L-2) \cdots - 2 - 1$$

$$D_1^B = Q - 1 \left(-2 \sum_{i=1}^{L-1} i \right) + 1 - LM$$

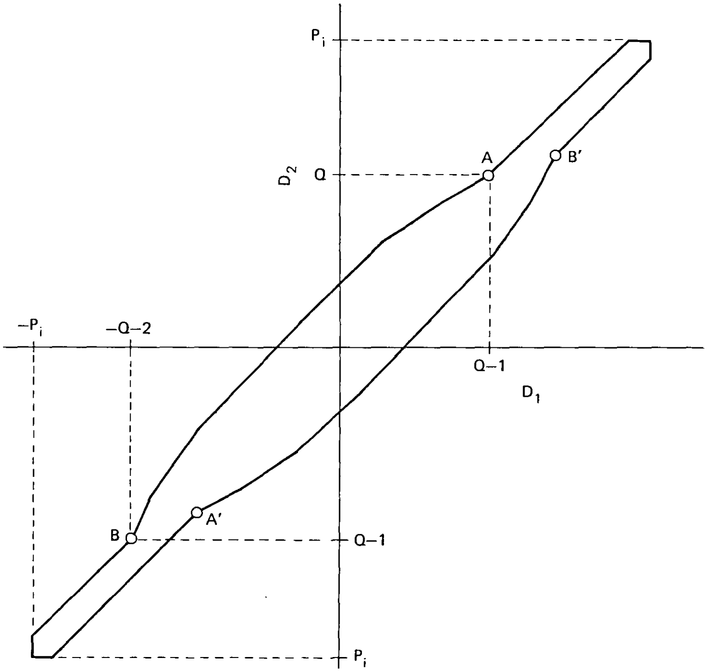


Fig. 7—$D_1 - D_2$ plot.

$$D_1^B = \frac{LM}{2} + \frac{L(L-1)}{2} - 2 - LM - L(L-1)$$

$$D_1^B = -LM - \frac{L(L-1)}{2} + 1 - 2$$

$$D_1^B = -Q - 2.$$

By symmetry, the states A' and B' also defined in Fig. 7 are $-Q + 1$, $-Q$ and $Q + 2$, $Q + 1$ respectively. For the case of L, $M = 3, 4$ (the smallest set of rolling-pin limit cycles), $Q = 8$ and the bounds on B_2 are

$$1 - 1/18 > B_2 \geq 1 - 1/16.$$

The common partial sequence in descending order is:

$$12, 11, 10, 9, \underbrace{\overset{Q}{8}}_{\text{state } A}, 7, 5, 2, -1, -4, -7, \underbrace{-9, -10}_{\text{state } B}, -11, -12.$$

Having defined the body, we now define the allowed lengths of the handle (which fixes the peak amplitude of the limit cycle) and the bounds on B_1. We do this by introducing the index J_i. J_i varies by unit steps from $J_{\max}$ to $J_{\min}$ inclusively, where

$$J_{\min} = LM - \frac{L(L-1)}{2} - 3L + 5 \tag{8}$$

$$J_{\max} = LM - \frac{L(L-1)}{2} - 2. \tag{9}$$

The peak amplitude of the limit cycle is given by

$$P_i = Q + J_i. \tag{10}$$

The number of nearly identical rolling-pin limit cycles is:

$$J_{\max} - J_{\min} + 1 = 3(L - 2). \tag{11}$$

Equation (11) has to be positive for a rolling-pin limit cycle to exist; hence, the minimum value of L is 3. The range on the peak amplitude P_i is between $P_{\max}$ and $P_{\min}$, where

$$P_{\max} = Q + J_{\max} = 3\left(\frac{LM}{2} - 1\right) \tag{12}$$

$$\begin{aligned} P_{\min} &= Q + J_{\min} \\ &= 3\,\frac{LM}{2} - 3L + 4. \end{aligned} \tag{13}$$

For L, $M = 3, 4$

$$5 \le J_i \le 7$$
$$13 \le P_i \le 15.$$

The bounds on B_1 for each of the L, M rolling-pin limit cycles are

$$2 - \frac{3}{2P_i} > -B_1 \ge 2 - \frac{3}{2P_{i-1}} \quad \text{for} \quad i = 1, 2, \cdots, 3(L-2), \tag{14}$$

where

$$P_i = P_o + i \tag{15}$$

and

$$P_o = 3(LM/2 - L + 1). \tag{16}$$

Crossing either the bounds on B_2 or the outer bounds on B_1 results in elimination of the rolling-pin limit cycle. The amplitude of the resulting largest ac limit cycle will then be predicted by the Jackson estimate.

The length of the limit cycle is N, where

$$N = (4J_i + 4L - 6 + 2M). \tag{17}$$

Since N contains the term $4J_i$ and J_i varies by 1 for each sucessive limit cycle in the set, N varies by 4.

The example used throughout this section has been for L, M = 3, 4 (the smallest of the set of the rolling-pin limit cycles). An example of a larger rolling-pin is L, M = 7, 8. This set has $3(L-2) = 15$ possible rolling-pin limit cycles. The peak amplitudes vary from 67 to 81. The values of Q and P_o are 48 and 66 respectively. The index J_i varies from 19 to 33. Figure 8 shows a magnified portion of the $B_1 - B_2$ plane where this set exits, and in Fig. 9 the time sequence of the largest and smallest constituent limit cycles are tabulated.

3.2 *Summary of construction rules and properties of the rolling-pin limit cycles*

3.2.1 *Even L, M*

The results for even L, M rolling-pin limit cycles can be summarized as follows:

(*i*) There are $3(L-2)$ distinct limit cycles for any even L, M. Each has a time sequence with a half-period odd amplitude symmetry

$$Y\left(n + \frac{N}{2}\right) = -Y(n).$$

(*ii*) On the descending* half of the time sequence, there are

(*a*) $J_i + 1$ steps of unit step size.

(*b*) A smooth transition region (i.e., constant second difference).

* I.e., we start from the positive peak and scan the limit cycle.

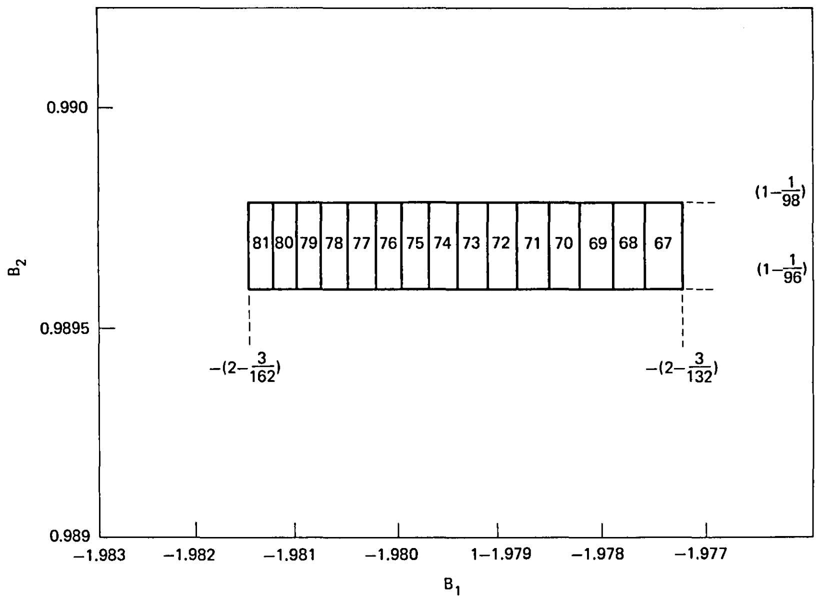


Fig. 8—Region of existence ($B_1 - B_2$ plane) for $L, M = 7, 8$.

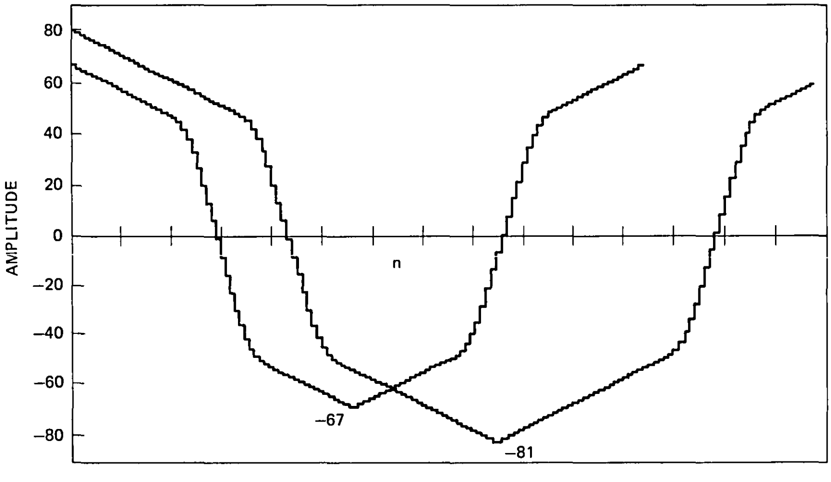


Fig. 9—Time sequences of $L, M = 7, 8$.

(*c*) M steps of step size L.

(*d*) A smooth transition region (also constant second difference).

(*e*) $J_i - 1$ steps of unit step size.

(*iii*) The peak amplitudes of the limit cycles are given by

$$P_i = J_i + Q,$$

where

$$Q = LM/2 + L(L-1)/2 - 1,$$

$$J_{\min} = LM - L(L-1)/2 - 3(L-2) - 1,$$

$$J_{\max} = LM - L(L-1)/2 - 2.$$

(*iv*) The region of existence of even L, M rolling-pin limit cycles in the $B_1 - B_2$ plane is defined by:

$$1 - \frac{1}{2(Q+1)} > B_2 \geq 1 - \frac{1}{2Q},$$

$$2 - \frac{3}{2P_i} > -B_1 \geq 2 - \frac{3}{2P_{i-1}} \quad \text{for} \quad i = 1, 2, \cdots, 3(L-2),$$

$$P_i = P_o + i,$$

where

$$P_o = 3(LM/2 - (L-2) - 1).$$

3.2.2 Odd L, M

Similar construction rules and bounds on B_1 and B_2 exist for odd values of L, M. These are simply summarized below.

(*i*) There are $3(L-3)$ distinct limit cycles for any odd L, M. Each has a half-amplitude period odd symmetry,

$$Y\left(n + \frac{N}{2}\right) = -Y(n).$$

(*ii*) On the descending half of the time sequence, there are

(*a*) $J_i + 1$ steps of unit step size.

(*b*) A smooth transition region (i.e., constant second difference).

(*c*) M steps of step size L.

(*d*) A smooth transition region (i.e., constant second difference).

(*e*) $J_i - 2$ steps of unit step size.

(*iii*) The peak amplitude of the limit cycle is given by

$$P_i = J_i + Q,$$

where

$$Q = (LM-1)/2 + L(L-1)/2 - 1 \tag{18}$$

$$J_{\min} = (LM-1) - \frac{L(L-1)}{2} - 3(L-3) - 1 \tag{19}$$

$$J_{\max} = (LM-1) - \frac{L(L-1)}{2} - 2. \tag{20}$$

(*iv*) The region of existence of odd L, M rolling-pin limit cycles in the $B_1 - B_2$ plane is defined by

$$1 - \frac{1}{2(Q+1)} > B_2 \geq 1 - \frac{1}{2Q},$$

$$2 - \frac{3}{2P_i} > -B_1 \geq 2 - \frac{3}{2P_{i-1}},$$

for $i = 1, 2, \cdots, 3(L-3)$, where

$$P_i = P_o + i,$$

$$P_o = 3\left(\frac{(L\,M-1)}{2} - (L-3) - 1\right). \tag{21}$$

A point to note is that, for an odd L, M rolling-pin limit cycle to exist, we must have $L \geq 5$.

The length of the limit cycle is N where

$$N = (4J_i + 4L + 2M - 8). \tag{22}$$

Comparison of the equations for L, M odd with the equations for L, M even shows that the basic forms of the equations are identical if we interchange the following two quantities:

$$\frac{LM}{2} \leftrightarrow \frac{LM-1}{2}$$

$$3(L-2) \leftrightarrow 3(L-3).$$

An example of odd L, M for L, $M = 5, 7$ is given. A magnified portion of the region of existence of this set in the $B_1 - B_2$ plane is shown in Fig. 10. The time sequences of the largest and smallest members of this set are tabulated in Fig. 11.

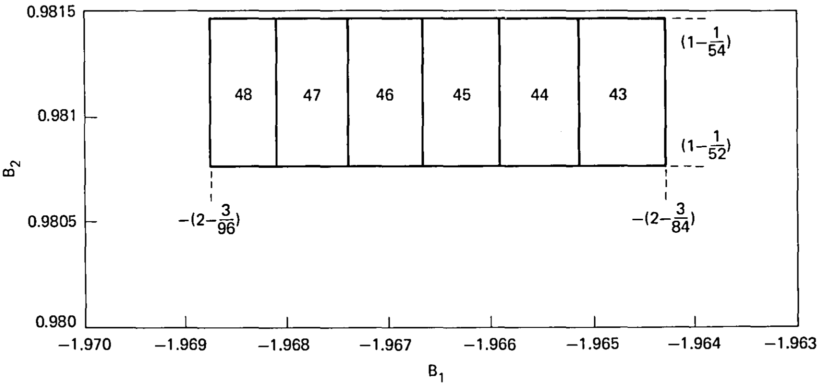


Fig. 10—Region of existence of L, $M = 5, 7$.

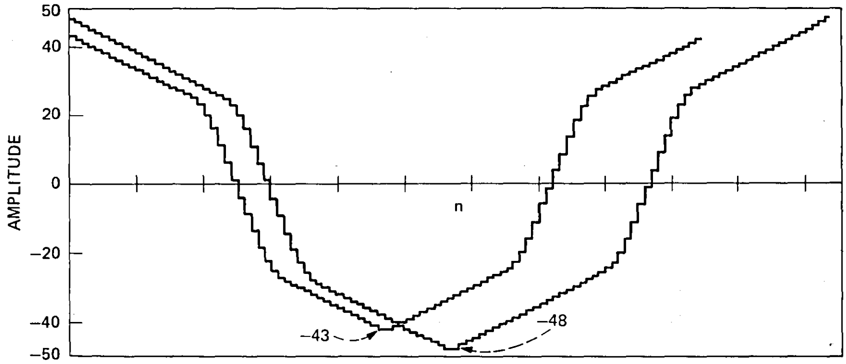


Fig. 11—Time sequences of $L, M = 5, 7$.

IV. ADDITIONAL PROPERTIES OF *1,L,M* LIMIT CYCLES

4.1 Peak amplitude comparisons

In this section, the ratio, ρ, of the peak amplitude of a rolling-pin limit cycle and Jackson's peak amplitude estimate[7] are determined. In addition, the actual peak amplitudes of the $L, M = 3, 4$ and $L, M = 3, 12$ rolling-pin limit cycles are compared to other calculated amplitude bounds or estimates.[7-11]

The peak amplitude of all rolling-pin limit cycles exceeds Jackson's estimate. It is now shown that the peak amplitude of a rolling-pin limit cycle approaches three times Jackson's estimate.* For convenience, we consider only even L, M.

Jackson's estimate is

$$JE = \frac{1}{2} \cdot \frac{1}{(1 - B_2)}. \tag{23}$$

Since this depends only on B_2, we can maximize ρ by minimizing JE. To do this, we pick the smallest value of B_2 that will cause the rolling-pin limit cycle to occur. This value is given by

$$B_2 = 1 - \frac{1}{2Q}, \tag{24}$$

so that

$$JE = Q = \frac{LM}{2} + \frac{L(L-1)}{2} - 1. \tag{25}$$

For a given L, M and a fixed $B_2 = 1 - 1/2Q$, a set of rolling-pin limit cycles exist. To maximize ρ we pick the largest amplitude case [eq. (12)],

* Parker and Hess conjectured in Ref. 8 that the limit cycle amplitude bound is three times the Jackson estimate.

$$P_{\max} = 3\left(\frac{LM}{2} - 1\right). \tag{26}$$

Then

$$\rho = \frac{JE}{P_{\max}} = 3 \cdot \frac{1}{1 + \dfrac{L(L-1)}{LM-2}}. \tag{27}$$

For a fixed L, ρ may be made arbitrarily close to 3 by increasing M. For the two examples, $L, M = 3, 4$ and $L, M = 3, 12$, ρ is 1.875 and 2.55 respectively.

Table I shows a comparison of the peak amplitudes for rolling-pin limit cycles of $L, M = 3, 4$ and $L, M = 3, 12$ with the bounds and estimates given by various authors.[7-11] As seen in this table, the actual peak amplitude is far less than the bounds predicted in Refs. 8 to 11. The Lyapunov bound[8] is the most pessimistic. The Sandberg-Kaiser peak amplitude estimate is $\sqrt{2}/2$ times their rms bound.[9] It is the best estimate for these examples. But even this bound is a factor of 6 higher for $L, M = 3, 4$ and a factor of 14 higher for $L, M = 3, 12$.

4.2 Mean-square-value comparison

Since all the states of the 1, L, M limit cycles are known, it is possible to develop an exact expression for the mean square value of the limit cycle sequence. The resulting expression is unfortunately complicated. The exact mean square value will however be compared to the Sandberg-Kaiser mean square bound for a number of examples. We define the ratio γ,

γ = Sandberg-Kaiser rms bound/actual rms value for various values of L and M. The Sandberg-Kaiser rms bound is

$$SK = \left| \frac{1}{(1 - B_2)\sqrt{1 - B_1^2/4B_2}} \right| \quad \text{if} \quad \frac{4B_2}{(1 - B_2)} > B_1. \tag{28}$$

For $P_i = \mathrm{P}_{\max}$, the above inequality on B_1 and B_2 reduces to

$$12Q > 4P + 3 \quad \text{or} \quad 2L(L-1) > 1, \tag{29}$$

which always holds for rolling-pin limit cycles. Using the values of B_1, B_2 in the lower left-hand corner of the rectangle where the limit cycle exists, we have

$$(SK)^2 = \frac{2(2P_i)^2(2Q)^2(2Q-1)}{6(2P_i)(2Q) - 9Q - 2(2P_i)^2}. \tag{30}$$

The ratio γ was calculated for the largest limit cycle in the set 1, L, M for values of L = 4, 6, 8, 10, and 12 and values of $M = M_{\min}$ to ($M_{\min}$

Table I—Comparison of various bounds, Refs. 7–11

L, M	Arbitrary Binary Coefficients Chosen Within the Cells	Actual Peak	Jackson's Peak Estimate[7]	Parker and Hess Peak Estimate[8]	Sandberg and Kaiser[9]		Long and Trick Bound[10]	Lyapunov Peak Bound[8]	Peak Bounds Requiring Knowledge of Length[10]	
					RMS Bound	Peak Estimate			Actual Length N	Length Using Angle of Pole
$L = 3$ $M = 4$	$B_2 = 0.9375$ $B_1 = 1.875$	13	8	24	64	90	125	488	$(N = 34)$ 42	$(N = 25)$ 80
$L = 3$ $M = 4$	$B_2 = 0.9375$ $B_1 = 1.890625$	14	8	24	74	104	145	648	$(N = 38)$ 55	$(N = 29)$ 94
$L = 3$ $M = 4$	$B_2 = 0.9375$ $B_1 = 1.8984375$	15	8	24	81	114	159	776	$(N = 42)$ 64	$(N = 32)$ 101
$L = 3$ $M = 12$	$B_2 = 0.9755859375$ $B_1 = 1.96875$	49	20	60	498	704	990	11.7×10^3	$(N = 146)$ 626	$(N = 76)$ 578
$L = 3$ $M = 12$	$B_2 = 0.9755859375$ $B_1 = 1.969703125$	50	20	60	538	760	1069	13.6×10^3	$(N = 150)$ 629	$(N = 82)$ 685
$L = 3$ $M = 12$	$B_2 = 0.9755859375$ $B_1 = 1.9705875$	51	20	60	585	827	1162	16×10^3	$(N = 154)$ 619	$(N = 90)$ 737

Note: For bounds which require knowledge of the length, the two Parker and Hess[8] matrix methods and the Long and Trick method[10] will produce similar results. In this table the Long and Trick method was used in calculating the peak bound. In calculating the length from information about the angle of the pole, we have made the assumption that the limit cycle has a fundamental frequency equal to the frequency that corresponds to the angle of the pole.

+ 10).* The results are plotted in Fig. 12. It can be seen that the *SK* rms bound is too high by a factor between 7 and 30. As either L or M is increased (the poles approach the unit circle), γ increases approximately linearly (pessimistically).

4.3 Comparison of roundoff noise and limit cycle power

In this section, the formula for the roundoff noise power from an isolated second-order section is derived in terms of P_i and Q. A comparison between this noise power and the limit cycle power is given for $L = 3$, 4, 5, and 6 and four values of M for each L. It is concluded from the results of the comparison that the limit cycle power is likely to dominate.

The average roundoff noise power at the output of the second-order filter section shown in Fig. 1 is given by

$$\sigma_R^2 = 2\frac{E_o^2}{12}\int H(Z)\cdot\overline{H(Z)}\,\frac{dz}{Z}. \tag{31}$$

$H(Z)$ is the transfer function of the second-order section. $\overline{H(Z)}$ is the complex conjugate of the transfer function. The quantization of products in the presence of a signal is modeled by two white noise sources and is accounted for by the $2E_o^2/12$ term where E_o is the quantization step size of the filter. Evaluating the contour integral of eq. (31), we obtain

$$\sigma_R^2 = 2\frac{E_o^2}{12}\frac{(1+B_2)}{(1-B_2)}\frac{1}{(1+B_2)^2 - B_1^2}. \tag{32}$$

Expressing eq. (32) in terms of P_i and Q, we have

$$\sigma_R^2 = 2\frac{E_o^2}{12}(4Q-1)\frac{4P_i^2\,Q^2}{P_i^2(1-8Q) + 3Q^2(8P_i - 3)}. \tag{33}$$

The limit cycle power is defined as

$$\sigma_L^2 = E_o^2\sum_{i=1}^{N}\frac{D_i^2}{N}, \tag{34}$$

where D_i are the values of the limit cycle.

A useful comparison is to calculate

$$\beta = 10\log(\sigma_L^2/\sigma_R^2). \tag{35}$$

This gives an estimate of the relative importance of errors in the presence of a signal and errors in idle channel conditions. This is only an estimate, since the limit cycles have specific frequencies while the roundoff noise has a flat spectrum. In a cascade of sections, for

* $M_{\min}$ is the minimum value of M for which the rolling-pin limit cycles exist in their nondegenerate form. See Section V.

example, succeeding poles and zeros may completely eliminate or amplify one or more of the limit cycle frequencies. Spectral analysis of the rolling-pin limit cycles has been carried out. Figure 13 shows a typical result. These results indicate that several frequencies are present. Accordingly, the estimate β is likely to be useful. In Table II, the value of β in decibels is calculated for $L = 3$, 4, 5, and 6 and four values of M. In each case, the limit cycle noise power is larger. On the

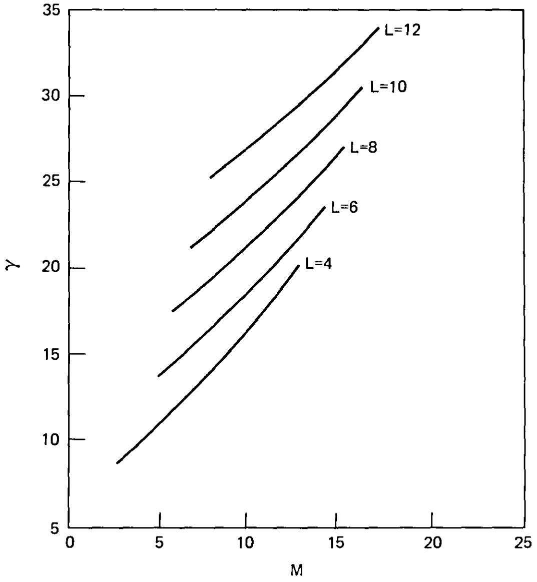


Fig. 12—Ratio of rms bounds.

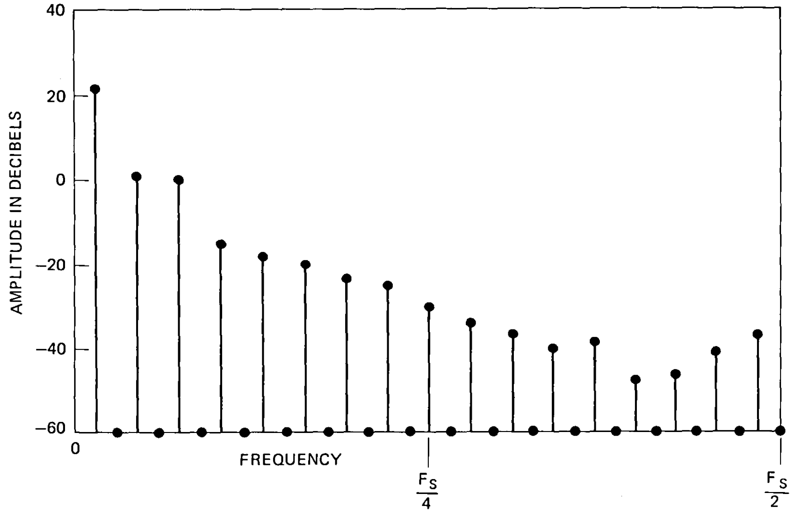


Fig. 13—Spectra of limit cycle.

Table II—Comparison of roundoff noise power and limit cycle power for various values of L, M

L, M	Roundoff Noise Power σ_R^2	Limit Cycle Power σ_L^2	$\beta = 10 \log \frac{\sigma_L^2}{\sigma_R^2}$ dB
3, 4	35.9	108.6	4.8
3, 6	108	264.8	3.9
3, 8	240.2	488.6	3.1
3, 10	450.3	780.1	2.4
4, 4	66.3	228.24	5.4
4, 5	113.1	365.5	5.1
4, 6	177.3	533.5	4.8
4, 7	261.5	732.4	4.5
5, 5	162.6	582.4	5.5
5, 6	269.8	915.8	5.3
5, 7	361.6	1173.3	5.1
5, 8	533.9	1624.9	4.8
6, 5	262.9	993.4	5.8
6, 6	388.2	1424.4	5.6
6, 7	545.9	1924.8	5.5
6, 8	739.6	2496.1	5.3

basis of these results, it is expected that limit cycle behavior will dominate in establishing the internal data word length of digital filters that generate rolling-pin limit cycles.

4.4 DC and the small ac limit cycles

A number of dc and small ac limit cycles are produced by the same second-order sections that produce the rolling-pin limit cycles. These small ac limit cycles are not accessible in the sense of Claasen et al.[12] and their peak amplitudes do not exceed the Jackson estimate. The states and form of these small ac limit cycles can also be characterized by integers. Each of these small ac limit cycles has no handles. They only have a constant step-size body* with smooth transitional paths to their peaks. These smaller limit cycles can be constructed using the method outlined in Section V for rolling-pin limit cycles.

The total number of these small ac limit cycles is $(Q - P_o/3 - 1)$, where Q and P_o can be obtained from eqs. (6) and (16), respectively. The peak amplitude of the smallest ac limit cycle is $(P_o/3 + 1)$. The largest amplitude of the small ac limit cycles is $(Q - 1)$. For $L, M = 3, 4$ we have $Q = 8$ and $P_o/3 = 4$. Therefore, there are three small ac limit cycles. The smallest has a peak amplitude of 5 and the largest a peak amplitude of 7.

The $D_1 - D_2$ plot of all the small ac and dc limit cycles for $B_1 = -1.875$ and $B_2 = 0.9375$ (binary coefficients within the region for $L, M = 3, 4$) is shown in Fig. 14. The three large rolling-pin limit cycles are

* For these various small limit cycles, the constant step size ranges from $1, 2 \cdots (L - 1)$, where L is the constant step size of the body of the associated rolling-pin limit cycle.

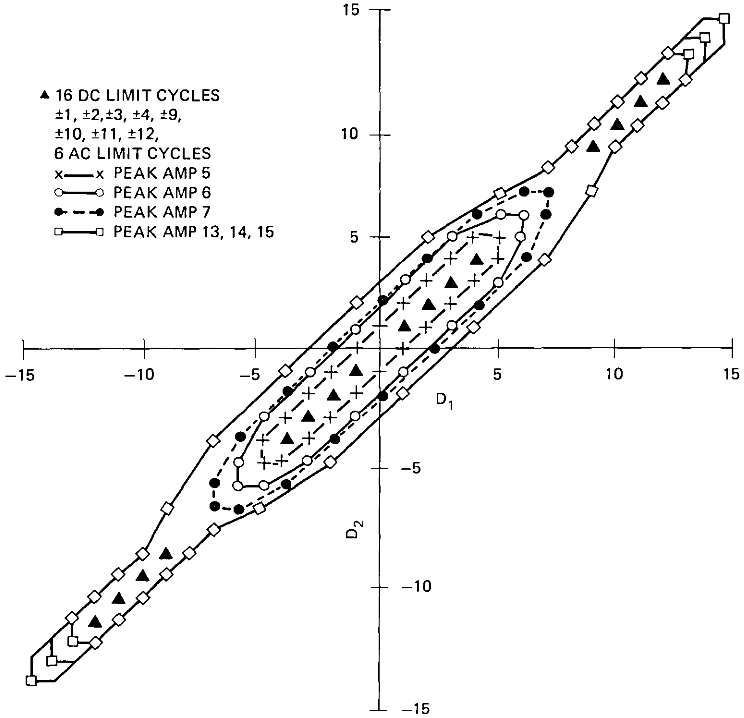


Fig. 14—Successive value (D_1 − D_2) plot small and large limit cycles.

also shown in Fig. 14. The regions of existence of these small ac limit cycles together with the three rolling-pin limit cycles are shown in Fig. 15. The rolling-pin limit cycles exist in a relatively smaller region. The small ac limit cycles exist in regions all the way up to the stability boundary $B_2 = 1$.* We see from Fig. 15 that, as the rolling-pin boundaries are crossed, the larger rolling-pin limit cycles disappear. However, the small ac and dc limit cycles still remain.

The number of small dc limit cycles can also be evaluated. This number is equal to $2P_o/3$. Any state which satisfies the equation

$$-\frac{P_o}{3} \leq D_1 = D_2 \leq \frac{P_o}{3} \tag{36}$$

is a dc state. These states represent the small dc limit cycles. The larger dc limit cycles can be evaluated from the following equation:

$$Q + 1 \leq D_1 = D_2 \leq P_o. \tag{37}$$

For $B_1 = -1.875$ and $B_2 = 0.9375$, we have $Q = 8$ and $P_o = 12$. Therefore, there are 8 small dc limit cycles, ±1, ±2, ±3, and ±4, all

* Other limit cycles may also exist in the region outside the region of the rolling-pin limit cycles. No inference should be drawn on the existence of other limit cycles.

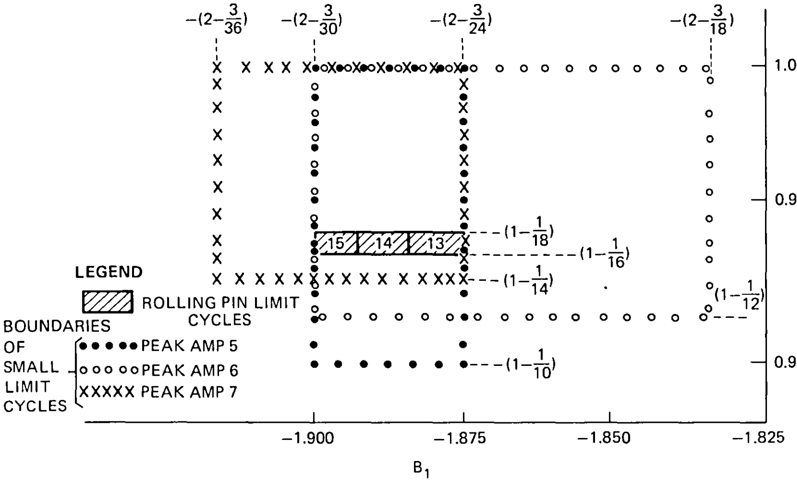


Fig. 15—Region of existence of small and large limit cycles.

satisfying eq. (36). In addition, there are 8 bigger dc limit cycles, ±9, ±10, ±11, and ±12. These values can be obtained from eq. (37). These 16 dc states are shown in Fig. 15. The state $\pm[Q, Q] = \pm[8, 8]$, which should normally correspond to the Jackson estimate is not on any limit cycle. The next state after the state ±[8, 8] is on a bigger limit cycle with peak amplitude 13. This state $\pm[Q, Q]$ is unique because it is a state from which one can spiral out onto a bigger limit cycle.

V. REGIONS OF EXISTENCE IN THE $B_1 - B_2$ PLANE

Each member of a set of rolling-pin limit cycles exists in a rectangular region. The boundaries of the regions are rational numbers. As seen in Fig. 5, the members within a set exist in horizontally contiguous regions. In any real implementation, the coefficients are binary. Strictly, all statements refering to regions of existence should refer to the binary valued points within (and occasionally on the boundary of) the regions. For simplicity, we refer only to the continuous regions to demonstrate why the regions are rectangular; we develop the explicit boundaries for rolling-pin limit cycles. We now construct the values of Q, J_i, and P_i, which were stated and used earlier. These variables, interestingly, depend on and follow from the assumed form of the body of the limit cycle which was constructed in Section III. The critical observation to make is that any state D_1, D_2 which generates the output Y [i.e., $(D_1,D_2) \rightarrow Y$] defines a region in the $B_1 - B_2$ plane. This region is an unbounded staircase of rectangles if $D_1 D_2 \neq 0$, or is a semi-infinite slab if D_1 or $D_2 \neq 0$. The rectangles for the case of $(3, 2) \rightarrow -1$ or $(-3, -2) \rightarrow 1$ are shown in Fig. 16. The height and width of each rectangle is $|1/D_2|$ and $|1/D_1|$, respectively. The center of each rectangle lies on the line defined by

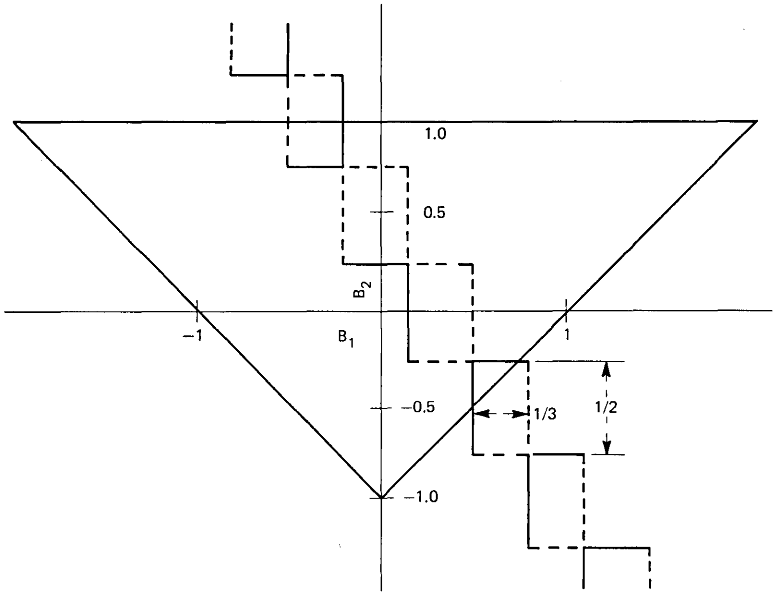


Fig. 16—Region of existence of a transition. Regions of existence include the solid, but not the dotted, boundary.

$$B_2D_2 + B_1D_1 = -Y, \tag{38}$$

with slope $-D_1/D_2$ and B_1 intercept $-Y/D_1$. The B_1 and B_2 intercepts of this line are both centers of rectangles. If $D_1 = 0$, there is a horizontal slab of height $|1/D_2|$ centered on $B_2 = -Y/D_2$. For $D_2 = 0$, there is a vertical slab of width $|1/D_1|$ centered on $B_1 = -Y/D_1$. This construction was done for sign-magnitude products for the circuit shown in Fig. 1. The rectangles include the solid boundaries but not the dotted boundaries since $\pm 1/2$ rounds to ± 1. For twos-complement rounding, the same figure applies with minor changes in the dotted and solid lines since, in twos-complement, $-1/2$ rounds to 0. Since any limit cycle is a sequence of transitions, the resulting region of existence is the region of the $B_1 - B_2$ plane common to all the transitions. This region must be a rectangle. (When independent stability arguments are applied, the stability triangle is superimposed, which may further reduce the common region. This is the only condition that can cause other than a rectangular region.)

Armed with this information, rolling-pin limit cycles were studied to determine the boundaries of their existence and to locate the critical transitions that set these boundaries. What was discovered was that the same transitions (relative to the overall shape of the successive value rolling-pin plot) *always* defined the $B_1 - B_2$ regions. Small differences exist depending on the evenness and oddness of L and M.

For convenience, we consider M even. Since for sign-magnitude rounding the states $(D_1, D_2) \rightarrow Y$ and $(-D_1, -D_2) \rightarrow -Y$ result in the same $B_1 - B_2$ region, we need only consider the half of the limit cycle from the positive peak to the negative peak. For M even, the state -1 always occurs on the limit cycle. The body of the limit cycle is symmetrical with respect to this -1 point. The critical transitions that set the B_2 boundaries are shown in Fig. 17a. Two transitions (T_u and T_u') set the maximum value of B_2 and one transition (T_L) sets the minimum value of B_2. The D_1 and D_2 states in these transitions are defined in terms of $Q = LM/2 + L(L - 1)/2 - 1$. The two transitions in the four consecutive states $Q + 1$, Q, $Q - 1$, $Q - 3$ completely determine the B_2 boundaries.

The boundaries on B_1 are more complex since a horizontally contiguous set of rectangles exist within each of which a separate rolling pin limit cycle occurs. The outer boundaries of this set of rectangles are determined by the two transitions T_u and T_L shown in Fig. 17b. The value of D_1 in the transition that sets the upper limit on B_1 is $-(1 + ML/2 - L)$. The value of D_1 in the transition that sets the lower limit on B_1 is $ML/2 - 1$. The inner B_1 boundaries that separate the nearly identical limit cycles in a set are fixed by the transitions at the peak of the limit cycle. Figure 17c illustrates this for a typical limit cycle in a set. As can be seen, the pair of consecutive transitions at either peak serves to set the inner boundaries on B_1. The peak transition of the largest limit cycle in the set sets the minimum bound on B_1. This

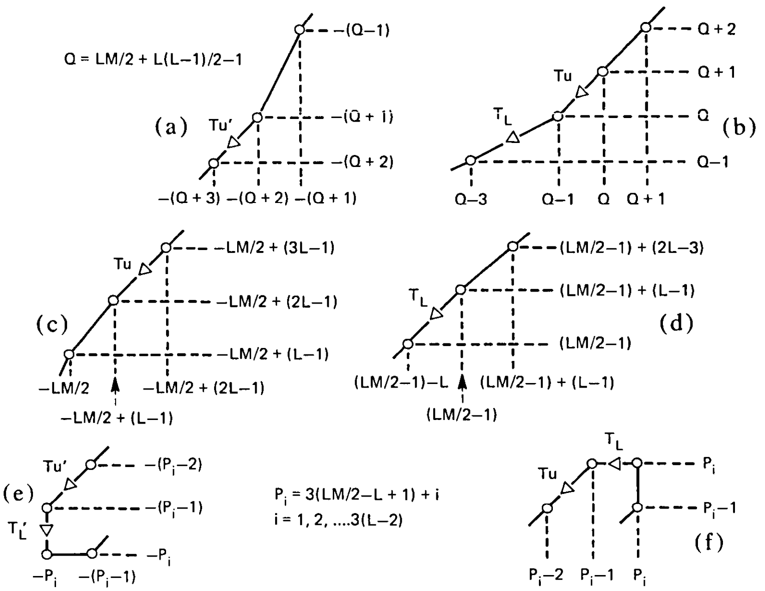


Fig. 17—Critical transitions.

bound is identical to the minimum bound on B_1 determined by the body. In fact, this is how the peak limit cycle of the set is defined. The peak transition of the smallest limit cycle in the set sets the maximum bound on B_1. This bound is identical to the maximum bound on B_1 determined by the body. All peaks are possible between and including these two. The number of limit cycles in the set is determined in this way.

When it is known which transitions set the boundaries and the values of the states in those transitions, the resulting boundaries are:

$$1 - \frac{1}{2(Q+1)} > B_2 \geq 1 - \frac{1}{2Q}$$

$$2 - \frac{3}{2P_{i+1}} > - B_1 \geq 2 - \frac{3}{2P_{i-1}}$$

for all $i = 1, 2, \cdots, 3(L - 2)$, where i indexes the consecutive states and

$$P_o = 3\left(\frac{LM}{2} - L + 1\right).$$
$$P_i = P_o + i.$$

Since Q and P_i are integers, the boundaries are rational. The existence of the 2 in the denominator is expected since the boundaries must be associated with rounding products which are an integer ±1/2. The derivations of these boundaries is in principle possible by considering each transition in the body and locating a common region in the $B_1 - B_2$ plane. In practice, this is not accomplished; rather, the boundaries have been located using programming techniques. Insight into the process can be gained using the following approach. Assume for the critical B_1 transitions in the body that B_2 is effectively unity. (This must be rechecked later, but is in fact true.) The state equation for the circuit in Fig. 1 is then

$$-D_2 + \left| D_1\left(2 - \frac{1}{2X}\right) \right|_R = Y. \tag{39}$$

The subscript R signifies sign-magnitude rounding. For the lower B_1 bound, the values of D_1, D_2, Y are

$$\frac{LM}{2} + L - 2, \quad \frac{LM}{2} - 1, \quad \frac{LM}{2} - (L + 1).$$

Using these values, eq. (39) becomes

$$\left| LM - 2 - \frac{1}{2X}\left(\frac{LM}{2} - 1\right) \right|_R = LM - 3.$$

This nonlinear equation due to the rounding is valid for the interval

$$\frac{3}{2} \geq \frac{1}{2X}\left(\frac{LM}{2} - 1\right) > \frac{1}{2}.$$

The lower B₁ bound corresponds to the minimum algebraic value of $[2 - (1/2X)]$ which occurs for the maximum value of $X = (LM/2) - 1$. This lower bound on B_1 is approachable but not attainable:

$$B_{1\,\text{lower}} = -2 + \frac{3}{2}\frac{1}{3\left(\dfrac{LM}{2} - 1\right)}.$$

In a similar fashion, the values of D_1, D_2, Y are

$$\left(2L - 1 - \frac{LM}{2}\right),\ \left(L - 1 - \frac{LM}{2}\right),\ \left(-1 - \frac{LM}{2}\right)$$

This transition leads to the attainable bound:

$$B_{1\,\text{upper}} = -2 + \frac{3}{2}\frac{1}{3\left(\dfrac{LM}{2} - L + 1\right)}.$$

Using a B_1 between these limits and considering the two transitions in Fig. 17a, the bounds on B_2 are found to be:

$$1 - \frac{1}{2(Q+1)} > B_2 \geq 1 - \frac{1}{2Q}.$$

Checking back on the assumption that B_2 is effectively unity, for $D_2 = -1 - \dfrac{LM}{2}$,

$$-\left|\left(1 - \frac{1}{2Q}\right)\left(-1 - \frac{LM}{2}\right)\right|_R = \left(-1 - \frac{LM}{2}\right) - \left|\frac{\left(1 + \dfrac{LM}{2}\right)}{2\left(\dfrac{LM}{2} + \dfrac{L(L-1)}{2} - 1\right)}\right|_{R'}.$$

Since the minimum L is 3, the R' operation yields a zero result and therefore B_2 is effectively 1. The R' operation is rounding except ±1/2 rounds to 0.

The quantity P_i is the peak value of the ith limit cycle. The remaining unknowns are $J_{\min}$ and $J_{\max}$. They follow from Q and P_i.

$$J_{\min} = P_1 - Q = LM - \frac{L(L-1)}{2} - 3\,(L-2) - 1$$

$$J_{\max} = P_{3(L-2)} - Q = LM - \frac{L(L-1)}{2} - 2.$$

There are $J_{\max} - J_{\min} + 1 = 3(L-2)$ limit cycles in each set.

This discussion can be extended to the cases "M-odd, L-even" and "M-odd, L-odd." The minor differences that occur arise from the following. For "M-odd, L-even," the value $-L/2 - 1$ is on the body of the limit cycle. This, however, leads to no changes in the equations just developed. For "M-odd, L-odd," the value $-(L+1)/2 - 1$ is on the limit cycle. This causes the minor variations in the equations summarized in Section 3.3.

The final issue discussed in this section is a degeneracy that can occur for the smallest M for a given L. Essentially what happens is that the handle disappears into the body for some of the B_1 rectangles nearest the origin. To determine if degeneracy occurs, the equation for $J_{\min}$ must be examined. For a given L, since there are $J_i - 1$ steps in the negative handle, if $J_{\min} > 1$, no degeneracy occurs. If $J_{\max} > 1$, but $J_{\min}$ is not, degeneracy occurs for that value of M. Figure 18 shows a plot of $J_{\min}$ and $J_{\max}$ as a function of M for $L = 4$ and $L = 5$. As can be seen, degeneracy occurs for $L, M = 4, 3$ and $L, M = 5, 4$.

VI. GENERALIZED *K,L,M* ROLLING-PIN LIMIT CYCLES

Only the simplest class of rolling-pin limit cycles $(1, L, M)$ has been analyzed in this paper. Other classes exist with values of K larger than 1. The difference between the classes is K, which is the step size in the handle. The construction rules for the transition regions and the body are similar for all the classes. The other properties such as (*i*) regions of existence, (*ii*) peak amplitudes, (*iii*) mean square value, (*iv*) number

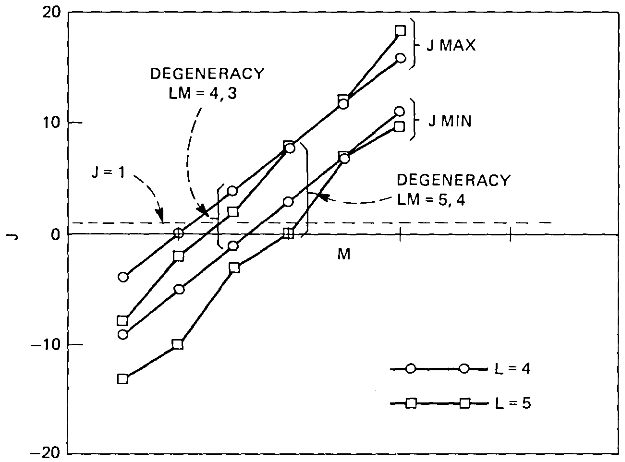


Fig. 18—Degeneracy criteria.

of cells, and (*v*) length can all be derived in terms of K, L, and M. The $D_1 - D_2$ plots, regions of existence, peak amplitudes, and number of cells for K, L, M =2, 6, 6 and K, L, M = 3, 8, 5 are shown in Figs. 19 and 20, respectively.

VII. CONCLUSIONS AND EXTENSIONS

A unique set of unusually large limit cycles has been discovered and catalogued and are called rolling-pin limit cycles. The set exists for second-order feedback sections with two quantizers that round sign-magnitude or twos-complement products. The limit cycles are completely defined by three integers K, L, and M, and a simple construction rule. The peak amplitude approaches three times Jackson's peak estimate. The limit cycles exist for $B_1 - B_2$ pairs lying within rectangular regions in the $B_1 - B_2$ plane. They occur often in filters with high Q poles near dc or half the sampling frequency.

Specialized techniques requiring extra circuitry can be used to suppress rolling-pin limit cycles. These special techniques may also increase the roundoff noise in the presence of a signal. Alternately, these limit cycles may be avoided by making small changes in the binary coefficient values to produce a pair of coefficients just outside the region of existence of rolling-pin limit cycles and yet meet specified filter characteristics.

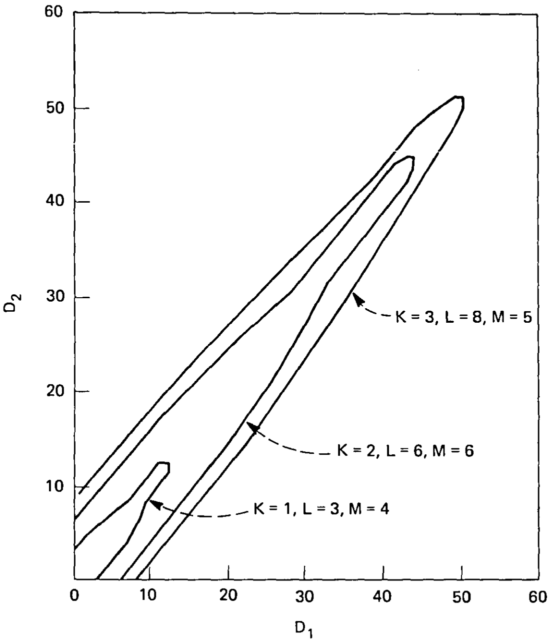


Fig. 19—Successive value plot for generalized K, L, M limit cycles.

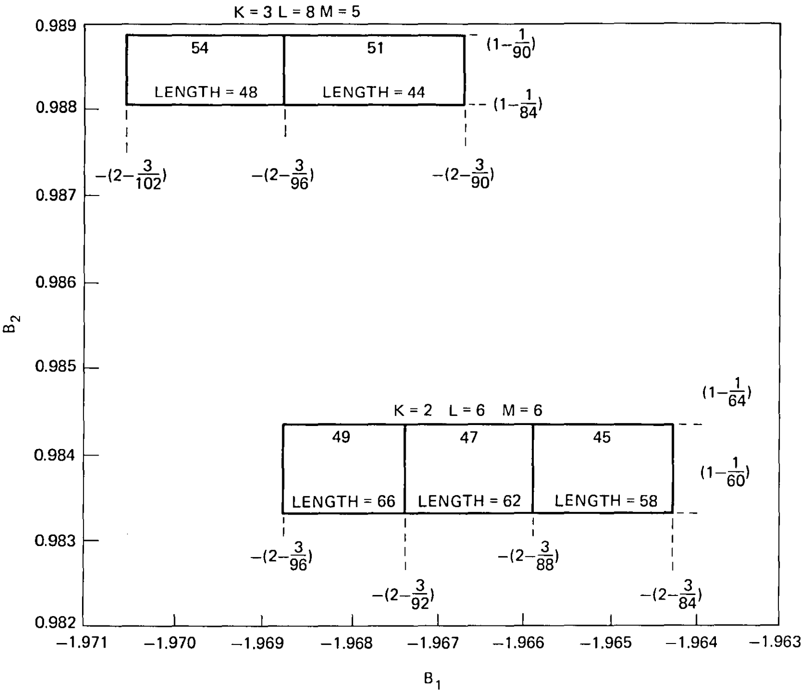


Fig. 20—Region of existence of generalized *K*, *L*, *M* limit cycles.

At present, these limit cycles are the largest (relative to Jackson's estimate) known in the region of the $B_1 - B_2$ plane where $|B_1| > 1.875$. For other regions, relatively large limit cycles have been found, but not systematized. It is expected that an approach similar to that presented will be useful. Other potentially useful extensions are for the cases of one quantizer and of truncation of products.

At present, these results can be useful to digital filter designers, whenever poles with $|B_1| > 1.875$ occur. Each design problem must be individually examined, however. A desirable goal is to incorporate these results into an automated design technique.

REFERENCES

1. S. L. Freeny, "Special-Purpose Hardware For Digital Filtering," Proc, IEEE, *63*, No. 4 (April 1975), pp. 633–648.
2. M. Buttner, "A Novel Approach to Eliminate Limit Cycles in Digital Filters with a Minimum Increase in the Quantization Noise," IEEE Trans. on Circuits and Systems, *CAS-24*, No. 6 (June 1977), pp. 300–304.
3. R. B. Kieburtz, V. B. Lawrence, and K. V. Mina, "Control of Limit Cycles in Recursive Digital Filters by Randomized Quantization," IEEE Trans. on Circuits and Systems, *CAS-24*, No. 6 (June 1977), pp. 291–299.
4. V. B. Lawrence and K. V. Mina, "Control of Limit Cycle Oscillations in Second Order Recursive Digital Filters using Constrained Random Truncation," IEEE Trans. on Acoustics, Speech and Signal Processing, *ASSP-26*, No. 2 (April 1978), pp. 127–134.

5. H. Butterweck, "Suppression of Parasitic Oscillations in Second Order Digital Filters by Means of a Controlled Rounding Arithmetic," AEU, *29* (December 1975), pp. 371–374.
6. Debasis Mitra and V. B. Lawrence, "Summary of Results on Controlled Rounding Arithmetics, For Direct Form Digital Filters, that Eliminate All Self-Sustained Oscillations," 1978 IEEE International Symposium on Circuits and Systems Proc., pp. 1023–1028.
7. L. B. Jackson, "An Analysis of Roundoff Noise in Digital Filters," Sc.D. Thesis, Stevens Institute of Technology, Hoboken, New Jersey, 1969.
8. S. R. Parker and S. F. Hess, "Limit Cycle Oscillations in Digital Filters," IEEE Trans. Circuit Theory, *CT-18*, No. 6 (November 1971), pp. 687–697.
9. I. W. Sandberg and J. F. Kaiser, "A Bound on Limit Cycles in Fixed-Point Implementations of Digital Filters," IEEE Trans. Audio and Electroacoustics, *AU-20*, No. 2 (June 1972), pp. 110–112.
10. J. L. Long and T. N. Trick, "An Absolute Bound on Limit Cycles due to Roundoff Errors in Digital Filters," IEEE Trans. Audio and Electroacoustics, *AU-21*, No. 1 (February 1973), pp. 27–30.
11. S. R. Parker and S. Yakowitz, "Computation of Bounds for Digital Filter Quantization Errors," IEEE Trans. Circuit Theory, *CT-20*, No. 4 (July 1973), pp. 391–396.
12. T. A. Claasen, W. Mecklenbrauker, and J. B. Peek, "Some Remarks on the Classification of Limit Cycles in Digital Filters," Phillips Research Reports, *28*, No. 4 (August 1973), pp. 297–303.

Copyright © 1979 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 58, No. 2, February 1979
Printed in U.S.A.

Integral Equations for Electromagnetic Scattering by Perfect Conductors With Two-Dimensional Geometry

By J. A. MORRISON

(Manuscript received June 6, 1978)

In this paper, we derive various integral equations related to the scattering of time-harmonic electromagnetic fields by perfect conductors with 2-dimensional geometry. The fields may be expressed in terms of two solutions of a scalar wave equation and decomposed into E waves and H waves. We consider the case in which part, or all, of each of the perfectly conducting cylindrical scatterers may be infinitesimally thin, and show that a standard integral equation, used in the case of H waves, does not determine the current density on the infinitesimally thin parts of the scatterers. We derive an alternate integral equation which does not suffer from this defect. This equation has been used by J. L. Blue in the numerical solution of the problem of scattering by an infinitesimally thin strip.

I. INTRODUCTION

In this paper, we consider the scattering of time-harmonic electromagnetic fields by perfect conductors with 2-dimensional geometry, in which the boundaries are independent of the z-coordinate. The z-dependence of the fields is assumed to be of the form $\exp(ik \sin \alpha\, z)$, where k is the free space wave number and $|\alpha| < \pi/2$, so that the scattering of obliquely incident plane waves may be investigated. It is known[1] that the electromagnetic fields may be expressed in terms of the longitudinal components, E_z and H_z, and that each of these two quantities satisfies the scalar wave equation, with wave number $k \cos \alpha$. Moreover, since the boundary conditions on a perfectly conducting surface imply that both E_z and the normal derivative of H_z are zero, there is no coupling between E_z and H_z, and we refer to E waves and H waves, respectively.

Integral equations for scattering problems have been considered by numerous authors. A relatively recent treatment of this topic is that of

Poggio and Miller,[2] but they give only a brief discussion of the 2-dimensional case. A useful discussion of integral equations for the scalar problem is given by Noble.[3] Poggio and Miller state that the integral equation which they derive for the surface current in the case of H waves is useless when the scatterer is infinitely thin. Noble points out that the corresponding integral equation, when applied to the problem of scattering by an elliptic cylinder, degenerates as the eccentricity tends to unity, so that the scatterer becomes an infinitesimally thin strip.

In this paper, we consider the case in which part, or all, of each of the perfectly conducting cylindrical scatterers may be infinitely thin. We derive an alternate integral equation for the current density on the scatterers, in the case of H waves, which does not degenerate on the infinitesimally thin segments. We discuss the relationship between this integral equation and one derived by Mitzner.[4] Our integral equation has been used by Blue[5] in the numerical solution of the problem of scattering by an infinitesimally thin strip. We also point out how the integral equation which does degenerate may be used to calculate H_z on both sides of the infinitesimally thin segments, once the entire current density on the scatterers is known. An integral equation which degenerates in the case of E waves is also derived, and this may be used analogously to calculate the values of the normal derivatives of E_z on both sides of the infinitesimally thin segments, once the entire current density on the scatterers is known.

In Section II we briefly derive expressions for the transverse components of the field in terms of the longitudinal components E_z and H_z, and show that the latter quantities both satisfy a scalar wave equation. We also derive the boundary conditions on a perfectly conducting surface, and give an expression for the current density on the surface. The total fields are expressed as the sum of the incident and scattered fields. In Section III we derive an integral representation for the scattered field in terms of the total field, and its normal derivative, on the scatterers. A nondegenerate integral equation for the current density on the scatterers is obtained in the case of E waves, by using this representation as a point on the boundary is approached. In the case of H waves, it is shown that the corresponding integral equation degenerates on the infinitesimally thin segments, since it contains an unknown quantity besides the current density.

In Section IV we derive representations for the transverse component of the gradient of the scattered field. By using this representation to calculate the normal derivative of the scattered field as a point on the boundary is approached, we obtain a nondegenerate integral equation for the current density on the scatterers in the case of H waves. In the case of E waves, it is shown that the corresponding integral equation degenerates on the infinitesimally thin segments. The implications of these results are discussed.

II. THE ELECTROMAGNETIC FIELDS

We first write down equations which describe the electromagnetic fields due to scattering by perfect conductors with 2-dimensional geometry. If we suppress the factor $\exp(-i\omega t)$, where ω is the angular frequency, the divergenceless electric and magnetic fields **E** and **H**, in free space, satisfy Maxwell's equations[6]

$$\nabla \times \mathbf{E} = i\omega\mu_0\mathbf{H}, \quad \nabla \times \mathbf{H} = -i\omega\epsilon_0\mathbf{E}, \tag{1}$$

where μ_0 is the permeability and ϵ_0 is the dielectric constant. The free space wave number is $k = \omega(\mu_0\epsilon_0)^{1/2}$. We consider the case of a 2-dimensional geometry in which the boundaries are independent of the coordinate z, and assume that the z-dependence of the fields is of the form $\exp(ik \sin \alpha\, z)$, where $|\alpha| < \pi/2$. This will allow us to consider scattering of obliquely incident plane waves. Accordingly, we now suppress the factor $\exp(ik \sin \alpha\, z)$, and write

$$\nabla = \nabla_t + ik \sin \alpha\, \mathbf{i}_z, \quad \mathbf{E} = \mathbf{E}_t + E_z\mathbf{i}_z, \quad \mathbf{H} = \mathbf{H}_t + H_z\mathbf{i}_z, \tag{2}$$

where $\mathbf{i}_z$ is a unit vector in the z-direction, and the subscript t refers to the transverse components.

If we split eqs. (1) into their transverse and longitudinal components, we obtain

$$\nabla_t E_z \times \mathbf{i}_z = ik \sin \alpha\, \mathbf{E}_t \times \mathbf{i}_z + i\omega\mu_0\mathbf{H}_t, \tag{3}$$

$$\nabla_t H_z \times \mathbf{i}_z = ik \sin \alpha\, \mathbf{H}_t \times \mathbf{i}_z - i\omega\epsilon_0\mathbf{E}_t, \tag{4}$$

and

$$\nabla_t \times \mathbf{E}_t = i\omega\mu_0 H_z\mathbf{i}_z, \quad \nabla_t \times \mathbf{H}_t = -i\omega\epsilon_0 E_z\mathbf{i}_z. \tag{5}$$

It is convenient to define the transverse wave number $k_t = k \cos \alpha$. Then, from (3) and (4), it follows that

$$k_t^2\mathbf{E}_t = i(k \sin \alpha\, \nabla_t E_z + \omega\mu_0\nabla_t H_z \times \mathbf{i}_z) \tag{6}$$

and

$$k_t^2\mathbf{H}_t = i(k \sin \alpha\, \nabla_t H_z - \omega\epsilon_0\nabla_t E_z \times \mathbf{i}_z). \tag{7}$$

Hence the transverse fields are expressed in terms of the longitudinal components. If we substitute these expressions for $\mathbf{E}_t$ and $\mathbf{H}_t$ into (5), we obtain

$$(\nabla_t^2 + k_t^2)E_z = 0, \quad (\nabla_t^2 + k_t^2)H_z = 0, \tag{8}$$

where we have used the relationships[7]

$$\nabla_t \times (\nabla_t V) = 0, \quad \nabla_t \times (\nabla_t V \times \mathbf{i}_z) = -\nabla_t^2 V\mathbf{i}_z. \tag{9}$$

Hence, as is known,[1] the longitudinal components of the field satisfy the scalar reduced wave equation.

The boundary conditions[8] on a perfectly conducting surface are that the tangential components of the electric field and the normal component of the magnetic field vanish, i.e.,

$$\mathbf{E} \times \mathbf{n} = 0, \quad \mathbf{H} \cdot \mathbf{n} = 0, \tag{10}$$

where $\mathbf{n}$ is a unit vector normal to the surface, directed into the scattering region. Because of the 2-dimensional geometry, $\mathbf{n} \cdot \mathbf{i}_z = 0$. From (2), (6) and (7), it follows that the boundary conditions (10) are equivalent to

$$E_z = 0, \quad \frac{\partial H_z}{\partial n} = 0 \quad \text{on a perfectly conducting surface.} \tag{11}$$

The current density on the surface[8] is $\mathbf{K} = \mathbf{n} \times \mathbf{H}$. We assume that $\mathbf{t}$, $\mathbf{n}$ and $\mathbf{i}_z$ form a right-handed set of unit vectors. Then, from (2) and (7) we find that

$$\mathbf{K} = H_z \mathbf{t} + \frac{i}{k_t^2}\left(\omega \epsilon_0 \frac{\partial E_z}{\partial n} - k \sin \alpha \frac{\partial H_z}{\partial s}\right)\mathbf{i}_z, \tag{12}$$

where s denotes arc length along the cross-sectional boundary curve, and $\mathbf{t}$ is a unit vector tangent to the curve.

We write the total fields as the sum of incident and scattered fields,

$$\mathbf{E} = \mathbf{E}^i + \mathbf{E}^s, \quad \mathbf{H} = \mathbf{H}^i + \mathbf{H}^s. \tag{13}$$

In the case of incident plane waves we have, in Cartesian coordinates (x, y, z),

$$E_z^i = E_0 \exp[ik_t(x \cos \beta + y \sin \beta)], \quad H_z^i = 0, \tag{14}$$

for an E wave, and

$$H_z^i = H_0 \exp[ik_t(x \cos \beta + y \sin \beta)], \quad E_z^i = 0, \tag{15}$$

for an H wave. The factor $\exp[i(k \sin \alpha\, z - \omega t)]$ has been suppressed.

III. INTEGRAL EQUATIONS DERIVED FROM REPRESENTATION FOR THE SCATTERED FIELD

We first derive some integral representations for the scattered fields. We suppose that

$$(\nabla_t^2 + k_t^2)\psi^i = 0, \quad (\nabla_t^2 + k_t^2)\psi^s = 0, \tag{16}$$

and set $\psi^i = E_z^i$ and $\psi^s = E_z^s$, or $\psi^i = H_z^i$ and $\psi^s = H_z^s$, corresponding to E waves, or H waves, respectively. We let

$$\psi = \psi^i + \psi^s, \tag{17}$$

and then the corresponding boundary conditions are, from (11),

$$\psi = 0, \quad \text{or} \quad \frac{\partial \psi}{\partial n} = 0, \quad \text{on a perfectly conducting boundary.} \tag{18}$$

We introduce the 2-dimensional Green's function[3]

$$G(\mathbf{r}, \rho) = \frac{i}{4} H_0^{(1)}(k_t R), \tag{19}$$

where $H_0^{(1)}$ denotes a Hankel function[9] of zero order, and

$$\mathbf{r} = x\mathbf{i}_x + y\mathbf{i}_y, \quad \rho = \xi\mathbf{i}_x + \eta\mathbf{i}_y, \quad \mathbf{R} = \mathbf{r} - \rho, \quad R = |\mathbf{R}|, \tag{20}$$

where $\mathbf{i}_x$ and $\mathbf{i}_y$ are unit vectors in the x- and y-directions. Then[3]

$$(\nabla_t^2 + k_t^2)G = 0, \quad \mathbf{R} \neq 0, \tag{21}$$

We consider the case in which part, or all, of each perfectly conducting cylinder may be infinitesimally thin, and let the cross-sectional boundary curve of the jth cylinder be denoted by $C_j = \Gamma_j \cup L_j^+ \cup L_j^-$, where L_j^+ and L_j^- denote opposite sides of the infinitesimally thin segment(s) L_j. The segments L_j may be disjoint, as may be Γ_j also, as depicted in Fig. 1. The curves C_j are assumed to be piecewise differentiable. We consider a point $\mathbf{r}$ exterior to all the curves C_j, and apply Green's theorem[10] in the region A exterior to the curves C_j, exterior to $|\rho - \mathbf{r}| = \epsilon$, and interior to $|\rho| = \tau$, as depicted in Fig. 1. Then, with $C = \cup_j C_j$,

$$-\int_{C\cup|\rho-r|=\epsilon\cup|\rho|=\tau} \left[\psi^s(\rho)\frac{\partial G}{\partial n} - G\frac{\partial \psi^s}{\partial n}(\rho)\right] ds = \int_A (\psi^s \nabla_t^2 G - G\nabla_t^2 \psi^s)\, dA = 0, \tag{22}$$

from (16) and (21). Because of our choice of $\mathbf{n}$, the normal derivatives are directed into the region A.

Now, from (19), since[9]

$$H_0^{(1)}(k_t R) = \frac{2i}{\pi}\log(k_t R) + 0(1), \quad \text{for} \quad k_t R \ll 1, \tag{23}$$

it follows that

$$\lim_{\epsilon \to 0} \int_{|\rho-\mathbf{r}|=\epsilon} \left(\psi^s \frac{\partial G}{\partial n} - G\frac{\partial \psi^s}{\partial n}\right) ds = -\psi^s(\mathbf{r}). \tag{24}$$

Also, since[11]

$$H_0^{(1)}(k_t R) \sim \left(\frac{2}{\pi k_t R}\right)^{1/2} \exp\left(i\left[k_t R - \frac{\pi}{4}\right]\right), \quad \text{for} \quad k_t R \gg 1, \tag{25}$$

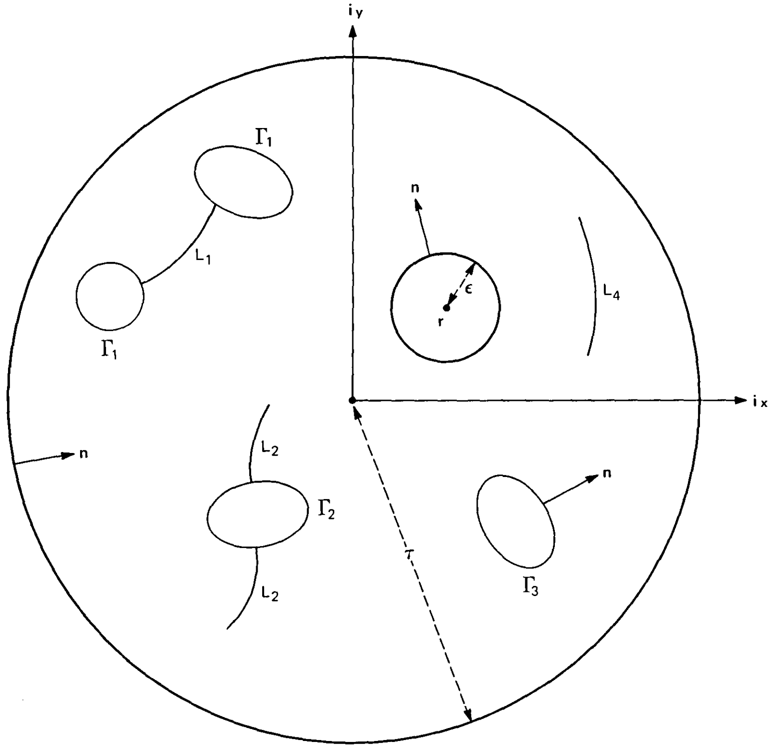


Fig. 1—Cross section of cylindrical scatterers.

$$\lim_{\tau\to\infty}\int_{|\rho|=\tau}\left(\psi^s\frac{\partial G}{\partial n}-G\frac{\partial\psi^s}{\partial n}\right)ds=0, \tag{26}$$

if the scattered fields satisfy the radiation condition[12]

$$\lim_{\rho\to\infty}\rho^{1/2}\left(\frac{\partial\psi^s}{\partial\rho}-ik_t\psi^s\right)=0,\quad(\rho=|\boldsymbol{\rho}|), \tag{27}$$

which we assume to be the case. If we let $\epsilon\to 0$ and $\tau\to\infty$ in (22), it follows from (24) and (26) that

$$\psi^s(\mathbf{r})=\int_C\left[\psi^s(\boldsymbol{\rho})\frac{\partial G}{\partial n}-G\frac{\partial\psi^s}{\partial n}(\boldsymbol{\rho})\right]ds. \tag{28}$$

If we consider the incident field ψ^i, and apply Green's theorem in the region(s) enclosed by Γ_j, and use (16) and (21), we obtain

$$\int_{\Gamma_j}\left[\psi^i(\boldsymbol{\rho})\frac{\partial G}{\partial n}-G\frac{\partial\psi^i}{\partial n}(\boldsymbol{\rho})\right]ds=0. \tag{29}$$

Also, because of the continuity of ψ^i and G, and their normal derivatives, on L_j, it follows that

$$\int_{L_j^+ \cup L_j^-} \left[\psi^i(\rho) \frac{\partial G}{\partial n} - G \frac{\partial \psi^i}{\partial n}(\rho) \right] ds = 0, \tag{30}$$

since the normals are reversed on opposite sides of L_j. Hence, since $C = \cup_j(\Gamma_j \cup L_j^+ \cup \mathrm{L}_j^-)$,

$$\int_C \left[\psi^i(\rho) \frac{\partial G}{\partial n} - G \frac{\partial \psi^i}{\partial n}(\rho) \right] ds = 0. \tag{31}$$

Consequently, with the help of (17), we may rewrite (28) in the form

$$\psi^s(\mathbf{r}) = \int_C \left[\psi(\rho) \frac{\partial G}{\partial n} - G \frac{\partial \psi}{\partial n}(\rho) \right] ds. \tag{32}$$

The advantage of doing this is that, because of the boundary conditions in (18), one of the two terms in the integrand in (32) vanishes.

We first consider the case of E waves. Then, from (11), (19), and (32), we have

$$E_z^s(\mathbf{r}) = -\frac{i}{4} \int_C H_0^{(1)}(k_t R) \frac{\partial E_z}{\partial n}(\rho)\, ds. \tag{33}$$

If we now let $\mathbf{r}$ tend to a point on C, we obtain

$$E_z^i(\mathbf{r}) = \frac{i}{4} \int_C H_0^{(1)}(k_t R) \frac{\partial E_z}{\partial n}(\rho)\, ds, \quad \mathbf{r}\ \varepsilon\ C, \tag{34}$$

since $E_z^i + E_z^s = 0$ on C. We may rewrite (34) in the form

$$\begin{aligned} E_z^i(\mathbf{r}) &= \frac{i}{4} \int_\Gamma H_0^{(1)}(k_t R) \frac{\partial E_z}{\partial n}(\rho)\, ds \\ &\quad + \frac{i}{4} \int_L H_0^{(1)}(k_t R) \left\{ \left[\frac{\partial E_z}{\partial n}(\rho) \right]_+ + \left[\frac{\partial E_z}{\partial n}(\rho) \right]_- \right\} ds, \quad \mathbf{r}\ \varepsilon\ C, \end{aligned} \tag{35}$$

where $\Gamma = \cup_j \Gamma_j$ and $L = \cup_j L_j$. This integral equation may be used to determine the current density on Γ and the total current density on L, which suffices to determine the scattered field from (33). However, (35) does not yield the separate values of the normal derivative of E_z on either side of L. We will return to this point in the next section.

We now consider the case of H waves. Since[9]

$$\frac{d}{dR} H_0^{(1)}(k_t R) = -k_t H_1^{(1)}(k_t R), \tag{36}$$

it follows from (11), (19), (20), and (32) that

$$H_z^s(\mathbf{r}) = \frac{i}{4} k_t \int_C H_1^{(1)}(k_t R) \frac{\mathbf{R} \cdot \mathbf{n}}{R} H_z(\boldsymbol{\rho})\, ds. \tag{37}$$

Let $\mathbf{r}_0$ be a point on Γ, not at a corner, and let σ be a small segment of Γ containing $\mathbf{r}_0$. Then, as seen from Fig. 2, by considering the angle $\delta\phi$ subtended by the element δs of σ, and letting $\delta s \rightarrow 0$, we obtain

$$\frac{d\phi}{ds} = \frac{\mathbf{R} \cdot \mathbf{n}}{R^2}. \tag{38}$$

Since, from (23) and (36), $k_t R H_1^{(1)}(k_t R) \rightarrow -2i/\pi$ as $k_t R \rightarrow 0$, it follows that

$$\lim_{\mathbf{r} \rightarrow \mathbf{r}_0} \frac{i}{4} k_t \int_\sigma H_1^{(1)}(k_t R) \frac{\mathbf{R} \cdot \mathbf{n}}{R} H_z(\boldsymbol{\rho})\, ds \sim \frac{H_z(\mathbf{r}_0)}{2\pi} \int_\sigma d\phi. \tag{39}$$

But since we have assumed that $\mathbf{r}_0$ is not at a corner, the angle subtended at $\mathbf{r}_0$ by σ tends to π as the length of σ tends to zero. Hence, from (37), we have

$$H_z^s(\mathbf{r}) = \frac{1}{2} H_z(\mathbf{r}) + \frac{i}{4} k_t P \int_C H_1^{(1)}(k_t R) \frac{\mathbf{R} \cdot \mathbf{n}}{R} H_z(\boldsymbol{\rho})\, ds, \quad \mathbf{r} \in \Gamma', \tag{40}$$

where P denotes the principal value of the integral, corresponding to the limit of the integral over $C - \sigma$ as the length of σ tends to zero, and Γ' denotes Γ less its corners. Since $H_z = H_z^i + H_z^s$, we may rewrite (40) in the form

$$\begin{aligned} \frac{1}{2} H_z(\mathbf{r}) = H_z^i(\mathbf{r}) &+ \frac{i}{4} k_t P \int_\Gamma H_1^{(1)}(k_t R) \frac{\mathbf{R} \cdot \mathbf{n}}{n} H_z(\boldsymbol{\rho})\, ds \\ &+ \frac{i}{4} k_t \int_L H_1^{(1)}(k_t R) \frac{\mathbf{R}}{R} \cdot \{\mathbf{n}_+ [H_z(\boldsymbol{\rho})]_+ \\ &+ \mathbf{n}_- [H_z(\boldsymbol{\rho})]_-\}\, ds, \quad \mathbf{r} \in \Gamma'. \end{aligned} \tag{41}$$

Now let $\mathbf{r}_0$ be a point on L_j, not at a corner (or endpoint), and let σ be a small segment of L_j containing $\mathbf{r}_0$. Then, from (38), if we let $\mathbf{r}$ tend to $\mathbf{r}_0$ from the L_j^+ side, we obtain

$$\begin{aligned} \lim_{\mathbf{r} \rightarrow [\mathbf{r}_0]_+} \frac{i}{4} k_t \int_\sigma H_1^{(1)}(k_t R) \frac{\mathbf{R}}{R} \cdot \{\mathbf{n}_+ [H_z(\boldsymbol{\rho})]_+ + \mathbf{n}_- [H_z(\boldsymbol{\rho})]_-\}\, ds \\ \sim \frac{1}{2\pi} \{[H_z(\mathbf{r}_0)]_+ - [H_z(\mathbf{r}_0)]_-\} \int_\sigma d\phi, \end{aligned} \tag{42}$$

since $\mathbf{n}_- = -\mathbf{n}_+$. Hence from (37), it follows that

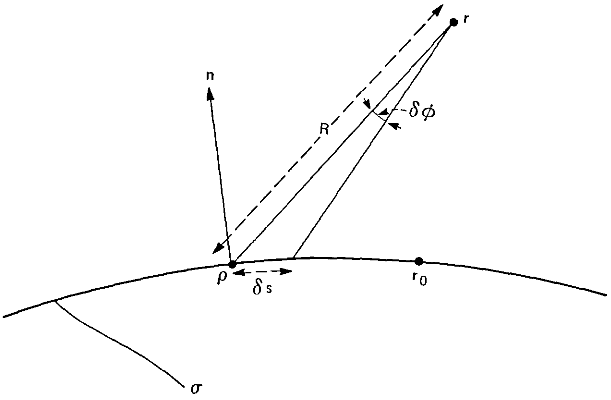


Fig. 2—Angle subtended by an element of a cross-sectional boundary curve.

$$\frac{i}{4} k_t \int_\Gamma H_1^{(1)}(k_t R)\,\frac{\mathbf{R}\cdot\mathbf{n}}{R}\,H_z(\rho)\,ds$$

$$+\frac{i}{4} k_t P \int_L H_1^{(1)}(k_t R)\,\frac{\mathbf{R}}{R}\cdot\{\mathbf{n}_+[H_z(\rho)]_+ + \mathbf{n}_-[H_z(\rho)]_-\}\,ds$$

$$= [H_z^s(\mathbf{r})]_+ - \frac{1}{2}\{[H_z(\mathbf{r})]_+ - [H_z(\mathbf{r})]_-\}$$

$$= \frac{1}{2}\{[H_z(\mathbf{r})]_+ + [H_z(\mathbf{r})]_-\} - H_z^i(\mathbf{r}), \quad \mathbf{r}\in L', \tag{43}$$

where L' denotes L less its corners (and endpoints). The same result is obtained by letting $\mathbf{r}$ tend to $\mathbf{r}_0$ from the L_j^- side, as is evident from the symmetry in (43). We have made use of the continuity of $H_z^i(\mathbf{r})$.

Now $\mathbf{n}_+ = -\mathbf{n}_-$ on L, but we note that the integral equations (41) and (43) do not determine $H_z(\mathbf{r})$ for $\mathbf{r}\in\Gamma'$, and $\{[H_z(\mathbf{r})]_+ - [H_z(\mathbf{r})]_-\}$ for $\mathbf{r}\in L'$, because of the unknown quantity $[H_z(\mathbf{r})]_+ + [H_z(\mathbf{r})]_-$ on the right-hand side of (43). Moreover, if L consists of segments of a straight line, then $\mathbf{R}\cdot\mathbf{n} = 0$ for $\mathbf{r}\in L'$, $\rho\in L'$ and $\rho\neq\mathbf{r}$, and the second integral in (43) vanishes. If, in addition, Γ is empty, then (43) reduces to

$$[H_z(\mathbf{r})]_+ + [H_z(\mathbf{r})]_- = 2H_z^i(\mathbf{r}), \quad \mathbf{r}\in L'. \tag{44}$$

This reduction was pointed out by Noble[3] in the case of an infinitesimally thin strip and by Millar[13] in the case of coplanar strips. In the next section, we derive an integral equation which does not degenerate, in the case of H waves, for $\mathbf{r}\in L'$. We remark that the integral equation (35) does not degenerate for $\mathbf{r}\in L$, and this is presumably because it was obtained by setting $E_z^s = -E_z^i$ on C. This suggests that we should derive an expression for $\partial H_z^s/\partial n$ on C', and set it equal to $-\partial H_z^i/\partial n$.

IV. INTEGRAL EQUATIONS DERIVED FROM THE GRADIENT OF THE SCATTERED FIELD

We now return to the integral representation (32) for the scattered field, and calculate the transverse component of its gradient. If we substitute the explicit form (19) of the Green's function into (32), and use (20) and (36), we obtain

$$\psi^s(\mathbf{r}) = \frac{i}{4}\int_C \left[k_t H_1^{(1)}(k_k R)\frac{\mathbf{R}\cdot\mathbf{n}}{R}\psi(\rho) - H_0^{(1)}(k_t R)\frac{\partial\psi}{\partial n}(\rho)\right] ds. \quad (45)$$

Hence, since[9]

$$\frac{d}{dR}[RH_1^{(1)}(k_t R)] = k_t R H_0^{(1)}(k_t R), \quad (46)$$

it follows that

$$\begin{aligned}\nabla_t\psi^s(\mathbf{r}) &= \frac{i}{4}k_t^2\int_C H_0^{(1)}(k_t R)\frac{(\mathbf{R}\cdot\mathbf{n})\mathbf{R}}{\mathbf{R}^2}\psi(\rho)\,ds \\ &\quad + \frac{i}{4}k_t\int_C H_1^{(1)}(k_t R)\left[\frac{\mathbf{n}}{R} - \frac{2(\mathbf{R}\cdot\mathbf{n})\mathbf{R}}{R^3}\right]\psi(\rho)\,ds \\ &\quad + \frac{i}{4}k_t\int_C H_1^{(1)}(k_t R)\frac{\mathbf{R}}{R}\frac{\partial\psi}{\partial n}(\rho)\,ds. \quad (47)\end{aligned}$$

Now, since $\partial\rho/\partial s = \mathbf{t}$, and $\mathbf{t}\times\mathbf{i}_z = -\mathbf{n}$, we have

$$\frac{\partial}{\partial s}\left[\frac{\mathbf{R}\times\mathbf{i}_z}{R^2}\right] = \frac{\mathbf{n}}{R^2} + \frac{2}{R^4}(\mathbf{R}\times\mathbf{i}_z)(\mathbf{R}\cdot\mathbf{t}). \quad (48)$$

Also,

$$\begin{aligned}(\mathbf{R}\times\mathbf{i}_z)(\mathbf{R}\cdot\mathbf{t}) &= (\mathbf{R}\cdot\mathbf{t})[(\mathbf{R}\cdot\mathbf{n})\mathbf{t} - (\mathbf{R}\cdot\mathbf{t})\mathbf{n}] \\ &= (\mathbf{R}\cdot\mathbf{n})\mathbf{R} - \mathbf{R}^2\mathbf{n}. \quad (49)\end{aligned}$$

Hence,

$$\frac{\mathbf{n}}{R^2} - \frac{2(\mathbf{R}\cdot\mathbf{n})\mathbf{R}}{R^4} = -\frac{\partial}{\partial s}\left(\frac{\mathbf{R}\times\mathbf{i}_z}{R^2}\right). \quad (50)$$

If we substitute (50) into the second integral in (47), and integrate by parts, and combine terms with the help of (49), we obtain

$$\nabla_t\psi^s(\mathbf{r}) = \frac{i}{4}k_t^2\int_C H_0^{(1)}(k_t R)\mathbf{n}\psi(\rho)\,ds$$

$$+\frac{i}{4}k_t\int_C \frac{H_1^{(1)}(k_tR)}{R}\left[(\mathbf{R}\times\mathbf{i}_z)\frac{\partial\psi}{\partial s}(\rho)+\mathbf{R}\frac{\partial\psi}{\partial n}(\rho)\right]ds. \tag{51}$$

This expression for the gradient of the scattered field is the 2-dimensional analog of that derived by Mitzner[4] in 3 dimensions. We give an alternate derivation of (51) in the appendix.

We are interested in calculating the normal derivative of the scattered field in the limit as $\mathbf{r}$ tends to a point $\mathbf{r}_0 \in \Gamma' \cup L'$, i.e. $\nabla_t\psi^s(\mathbf{r})\cdot\mathbf{n}_0$, where $\mathbf{n}_0$ is a unit normal to $\Gamma' \cup L'$ at $\mathbf{r}_0$. It will be seen that the limiting value of this quantity may be calculated with the help of (51), whereas the second integral in (47) has a singular behavior. However, it is not necessary to integrate this second integral by parts completely around C, as was done to obtain (51). If we let $\sum_0$ be a segment (or segments) of $\Gamma \cup L$ which has $\mathbf{r}_0$ as an interior point, then it suffices to integrate by parts over $\sum_0$. Since the second integral in (47) vanishes in the case of E waves, because $E_z = 0$ on the boundary, we now consider the case of H waves.

We define

$$\mathbf{n}_0 = \begin{matrix}\mathbf{n}(\mathbf{r}_0), \quad \mathbf{r}_0 \in \Gamma', \\ \mathbf{n}_+(\mathbf{r}_0) = -\mathbf{n}_-(\mathbf{r}_0), \quad \mathbf{r}_0 \in L',\end{matrix} \tag{52}$$

and choose

$$\mathbf{n} = \mathbf{n}_+ \Rightarrow d\rho/ds = \mathbf{t} = \mathbf{n}_+ \times \mathbf{i}_z, \quad \text{on } L. \tag{53}$$

We also define the tangential component of current density

$$J(\rho) = \begin{matrix} H_z(\rho), \quad \rho \in \Gamma, \\ [H_z(\rho)]_+ - [H_z(\rho)]_-, \quad \rho \in L. \end{matrix} \tag{54}$$

Then, from (11) and (47), after an integration by parts, and use of (46), (49), and (50), we obtain

$$\nabla_t H_z^s(\mathbf{r})\cdot\mathbf{n}_0 = \frac{i}{4}k_t^2\int_{\Gamma\cup L-\sum_0} H_0^{(1)}(k_tR)\frac{(\mathbf{R}\cdot\mathbf{n})(\mathbf{R}\cdot\mathbf{n}_0)}{R^2}J(\rho)\,ds$$

$$+\frac{i}{4}k_t\int_{\Gamma\cup L-\sum_0} H_1^{(1)}(k_tR)\left[\frac{\mathbf{n}}{R}-\frac{2(\mathbf{R}\cdot\mathbf{n})\mathbf{R}}{R^3}\right]\cdot\mathbf{n}_0 J(\rho)\,ds$$

$$+\frac{i}{4}k_t^2\int_{\sum_0} H_0^{(1)}(k_tR)\mathbf{n}\cdot\mathbf{n}_0 J(\rho)\,ds$$

$$+\frac{i}{4}k_t\int_{\sum_0} H_1^{(1)}(k_tR)\frac{(\mathbf{R}\times\mathbf{i}_z)}{R}\cdot\mathbf{n}_0\frac{\partial J}{\partial s}(\rho)\,ds$$

$$-\frac{i}{4}k_t\left[H_1^{(1)}(k_tR)\frac{(\mathbf{R}\times\mathbf{i}_z)}{R}\cdot\mathbf{n}_0J(\rho)\right]_{\Sigma_0} \qquad (55)$$

The contributions from all the endpoints of $\sum_0$ must be included in the last term in (55). If $\sum_0 = \Gamma \cup L$, then this last term is zero, and the first two integrals in (55) are absent.

We now consider $\mathbf{r} \rightarrow \mathbf{r}_0$, in a direction which is not tangential to $\Gamma' \cup L'$. Since $\mathbf{r}_0$ is an interior point of $\sum_0$, the first two integrals in (55) are well-behaved as $\mathbf{r} \rightarrow \mathbf{r}_0$, as are the contributions from the endpoints of $\sum_0$, represented by the last term in (55). Also, it follows from (23) that, for $\mathbf{r} = \mathbf{r}_0$, there is an integrable singularity in the third integral in (55). It remains to consider the fourth integral in (55). As depicted in Fig. 3, we take

$$\mathbf{r} = \mathbf{r}_0 + \epsilon(\cos\chi\mathbf{t}_0 + \sin\chi\mathbf{n}_0), \qquad (56)$$

where $\sin\chi \neq 0$. Also, for convenience, we take $s = 0$ at $\mathbf{r} = \mathbf{r}_0$, so that[14]

$$\rho = \mathbf{r}_0 + s\mathbf{t}_0 - \frac{1}{2}\kappa_0 s^2\mathbf{n}_0 + 0(s^3) \qquad (57)$$

for small $|s|$, where κ_0 is the curvature at $\mathbf{r}_0$. Hence,

$$\mathbf{R} = \mathbf{r} - \rho = (\epsilon\cos\chi - s)\mathbf{t}_0 + (\epsilon\sin\chi + \frac{1}{2}\kappa_0 s^2)\mathbf{n}_0 + 0(s^3), \qquad (58)$$

and

$$R^2 = |\mathbf{R}|^2 = (s - \epsilon\cos\chi)^2 + \epsilon^2\sin^2\chi + 0(\epsilon s^2) + 0(s^4). \qquad (59)$$

From (52), (53), and (58), it follows that

$$(\mathbf{R}\times\mathbf{i}_z)\cdot\mathbf{n}_0 = -\mathbf{R}\cdot\mathbf{t}_0 = (s - \epsilon\cos\chi) + 0(s^3). \qquad (60)$$

But[9]

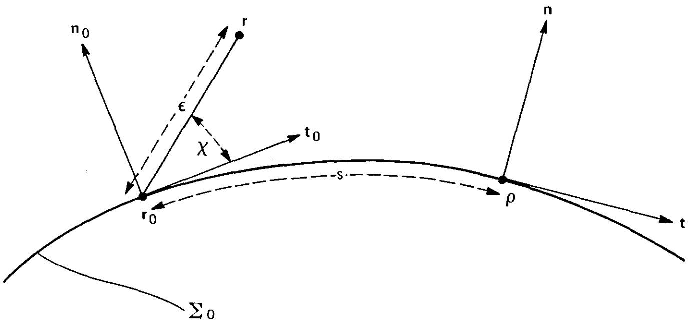


Fig. 3—Coordinates of a point in the neighborhood of a cross-sectional boundary curve.

$$H_1^{(1)}(k_tR) = \frac{-2i}{\pi k_t R} + 0[k_tR\log(k_tR)], \quad \text{for} \quad k_tR \ll 1, \tag{61}$$

and

$$\int_{-\delta}^{\delta} \frac{(s - \epsilon\cos\chi)\,ds}{[(s-\epsilon\cos\chi)^2 + \epsilon^2\sin^2\chi]}$$

$$= \frac{1}{2}\log\left[\frac{(\delta - \epsilon\cos\chi)^2 + \epsilon^2\sin^2\chi}{(\delta + \epsilon\cos\chi)^2 + \epsilon^2\sin^2\chi}\right]. \tag{62}$$

It follows from (59) to (62) that

$$\lim_{\delta\to 0}\left\{\lim_{\epsilon\to 0}\int_{-\delta}^{\delta} H_1^{(1)}(k_tR)\frac{(\mathbf{R}\times\mathbf{i}_z)}{R}\cdot\mathbf{n}_0\frac{\partial J}{\partial s}(\rho)\,ds\right\} = 0. \tag{63}$$

Hence the principal value of the fourth integral in (55) must be taken in the limit $\mathbf{r} \to \mathbf{r}_0$.

Having shown that the right-hand side of (55) is meaningful in the limit $\mathbf{r} \to \mathbf{r}_0$, we now note that the left-hand side tends to $\partial H_z^s/\partial n_0 = -\partial H_z^i/\partial n_0$, since $\partial H_z/\partial n = 0$ on the boundary, and hence its value is known. Hence the limit of (55) as $\mathbf{r} \to \mathbf{r}_0 \in \Gamma' \cup L'$ leads to the desired integral equation for $J(\rho)$, as defined in (54). We remark that $\partial J/\partial s$, as well as J, occurs in the integrand. We also remark that, when this integral equation has been solved for $J(\mathbf{r})$ for $\mathbf{r} \in \Gamma' \cup L'$, then (43) may be used to calculate $[H_z(\mathbf{r})]_+ + [H_z(\mathbf{r})]_-$ for $\mathbf{r} \in L'$, and hence the separate values of $[H_z(\mathbf{r})]_+$ and $[H_z(\mathbf{r})]_-$. We comment that we could presumably use the integral equation derived from (55) for $\mathbf{r}_0 \in L'$ only, and combine it with (41) for $\mathbf{r} \in \Gamma'$, to solve for $J(\mathbf{r})$ for $\mathbf{r} \in \Gamma' \cup L'$.

We now consider the case of E waves. Then, from (11) and (47), or equivalently (51),

$$\nabla_t E_z^s(\mathbf{r})\cdot\mathbf{n}_0 = \frac{i}{4}k_t\int_\Gamma H_1^{(1)}(k_tR)\frac{\mathbf{R}\cdot\mathbf{n}_0}{R}\frac{\partial E_z}{\partial n}(\rho)\,ds$$

$$+\frac{i}{4}k_t\int_L H_1^{(1)}(k_tR)\frac{\mathbf{R}\cdot\mathbf{n}_0}{R}\left\{\left[\frac{\partial E_z}{\partial n}(\rho)\right]_+ + \left[\frac{\partial E_z}{\partial n}(\rho)\right]_-\right\}ds. \tag{64}$$

But, from (58),

$$\mathbf{R}\cdot\mathbf{n}_0 = \epsilon\sin\chi + 0(s^2), \tag{65}$$

and, for $\delta > 0$,

$$\lim_{\epsilon\to 0+} \int_{-\delta}^{\delta} \frac{\epsilon \sin \chi \, ds}{[(s - \epsilon \cos \chi)^2 + \epsilon^2 \sin^2 \chi]}$$

$$= \lim_{\epsilon\to 0+} \left[\tan^{-1} \left(\frac{s - \epsilon \cos \chi}{\epsilon \sin \chi} \right) \right]_{-\delta}^{\delta} = \pi \operatorname{sgn}(\sin \chi). \quad (66)$$

We first consider $\mathbf{r} \to \mathbf{r}_0 \in \Gamma'$. Then, from (52) and Fig. 3, $\sin \chi > 0$. Hence, from (64), with the help of (59), (61), (65), and (66), we obtain

$$\frac{i}{4} k_t P \int_{\Gamma} H_1^{(1)}(k_t R_0) \frac{\mathbf{R}_0 \cdot \mathbf{n}_0}{R_0} \frac{\partial E_z}{\partial n} (\rho) \, ds$$

$$+ \frac{i}{4} k_t \int_{L} H_1^{(1)}(k_t R_0) \frac{\mathbf{R}_0 \cdot \mathbf{n}_0}{R_0} \left\{ \left[\frac{\partial E_z}{\partial n} (\rho) \right]_+ + \left[\frac{\partial E_z}{\partial n} (\rho) \right]_- \right\} ds$$

$$= \frac{\partial E_z^s}{\partial n} (\mathbf{r}_0) - \frac{1}{2} \frac{\partial E_z}{\partial n} (\mathbf{r}_0) = \frac{1}{2} \frac{\partial E_z}{\partial n} (\mathbf{r}_0) - \frac{\partial E_z^i}{\partial n} (\mathbf{r}_0), \quad \mathbf{r}_0 \in \Gamma', \quad (67)$$

where $\mathbf{R}_0 = \mathbf{r}_0 - \boldsymbol{\rho}$.

We now consider $\mathbf{r} \to \mathbf{r}_0 \in L'$. If the approach is from the plus side then, from (52) and Fig. 3, $\mathbf{n}_0 = \mathbf{n}_+(\mathbf{r}_0)$ and $\sin \chi > 0$. Hence, from (64), with the help of (59), (61), (65), and (66), it follows that

$$\frac{i}{4} k_t \int_{\Gamma} H_1^{(1)}(k_t R_0) \frac{\mathbf{R}_0 \cdot \mathbf{n}_0}{R_0} \frac{\partial E_z}{\partial n} (\rho) \, ds$$

$$+ \frac{i}{4} k_t P \int_{L} H_1^{(1)}(k_t R_0) \frac{\mathbf{R}_0 \cdot \mathbf{n}_0}{R_0} \left\{ \left[\frac{\partial E_z}{\partial n} (\rho) \right]_+ + \left[\frac{\partial E_z}{\partial n} (\rho) \right]_- \right\} ds$$

$$= \left[\frac{\partial E_z^s}{\partial n} (\mathbf{r}_0) \right]_+ - \frac{1}{2} \left\{ \left[\frac{\partial E_z}{\partial n} (\mathbf{r}_0) \right]_+ + \left[\frac{\partial E_z}{\partial n} (\mathbf{r}_0) \right]_- \right\}$$

$$= \frac{1}{2} \left\{ \left[\frac{\partial E_z}{\partial n} (\mathbf{r}_0) \right]_+ - \left[\frac{\partial E_z}{\partial n} (\mathbf{r}_0) \right]_- \right\} - \left[\frac{\partial E_z^i}{\partial n} (\mathbf{r}_0) \right]_+, \quad \mathbf{r}_0 \in L'. \quad (68)$$

On the other hand, if the approach to $\mathbf{r}_0$ is from the minus side, then $\mathbf{n}_0 = -\mathbf{n}_-(\mathbf{r}_0)$ and $\sin \chi < 0$, and it follows that the left-hand side of (68) is equal to

$$- \left[\frac{\partial E_z^s}{\partial n} (\mathbf{r}_0) \right]_- + \frac{1}{2} \left\{ \left[\frac{\partial E_z}{\partial n} (\mathbf{r}_0) \right]_+ + \left[\frac{\partial E_z}{\partial n} (\mathbf{r}_0) \right]_- \right\}$$

$$= \frac{1}{2}\left\{\left[\frac{\partial E_z}{\partial n}(\mathbf{r}_0)\right]_+ - \left[\frac{\partial E_z}{\partial n}(\mathbf{r}_0)\right]_-\right\} + \left[\frac{\partial E_z^i}{\partial n}(\mathbf{r}_0)\right]_- . \qquad (69)$$

Since $[\partial E_z^i/\partial n\ (\mathbf{r}_0)]_- = -[\partial E_z^i/\partial n\ (\mathbf{r}_0)]_+$, for $\mathbf{r}_0 \in L'$, we again obtain (68).

We remark that when the integral equation (35) has been used to determine $\partial E_z/\partial n\ (\mathbf{r})$ for $\mathbf{r} \in \Gamma'$ and $[\partial E_z/\partial n\ (\mathbf{r})]_+ + [\partial E_z/\partial n\ (\mathbf{r})]_-$ for $\mathbf{r} \in L'$, then (68) may be used to calculate $[\partial E_z/\partial n\ (\mathbf{r})]_+ - [\partial E_z/\partial n\ (\mathbf{r})]_-$ for $\mathbf{r} \in L'$, and hence the separate values of $[\partial E_z/\partial n\ (\mathbf{r})]_+$ and $[\partial E_z/\partial n\ (\mathbf{r})]_-$. This is analogous to the earlier remark concerning the use of (43). If L consists of segments of a straight line, then $\mathbf{R}_0 \cdot \mathbf{n}_0 = 0$ for $\mathbf{r}_0 \in L'$, $\rho \in L'$ and $\rho \neq \mathbf{r}_0$, and the second integral in (68) vanishes. If, in addition, Γ is empty, then (68) reduces to

$$\left[\frac{\partial E_z}{\partial n}(\mathbf{r}_0)\right]_+ - \left[\frac{\partial E_z}{\partial n}(\mathbf{r}_0)\right]_- = 2\left[\frac{\partial E_z^i}{\partial n}(\mathbf{r}_0)\right]_+, \quad \mathbf{r}_0 \in L'. \qquad (70)$$

This reduction was pointed out by Noble[3] in the case of an infinitesimally thin strip.

It is of interest to note that (35), and the integral equation obtained from (55) by letting $\mathbf{r} \to \mathbf{r}_0 \in \Gamma' \cup L'$, are Fredholm equations of the first kind. On the other hand, (41) and (67) which hold for $\mathbf{r} \in \Gamma'$, and Fredholm equations of the second kind, which is usually preferable from the viewpoint of numerical solution. It is somewhat unfortunate, in this respect, that the corresponding equations which hold for $\mathbf{r} \in L'$, namely (43) and (68), do not determine $[\mathrm{H}_z(\mathbf{r})]_+ - [H_z(\mathbf{r})]_-$ and $[\partial \mathrm{E}_z/\partial n\ (\mathbf{r})]_+ + [\partial \mathrm{E}_z/\partial n\ (\mathbf{r})]_-$.

NOTE ADDED IN PROOF

In the 3-dimensional case of an open thin shell, the use of a degenerate integral equation, analogous to (68), to determine the current densities on both sides of the shell, once the total current density is known, was pointed out by Stakgold.[16]

APPENDIX

We here give an alternate derivation of the integral equation (51). Let $\mathbf{e}$ be a constant vector. Then from (16), it follows that

$$(\nabla_t^2 + k_t^2)(\mathbf{e}\cdot\nabla_t\psi^s) = 0, \quad (\nabla_t^2 + k_t^2)(\mathbf{e}\cdot\nabla_t\psi^i) = 0. \qquad (71)$$

If we apply Green's theorem, as in Section III, but this time to $\mathbf{e}\cdot\nabla_t\psi^s$ and G, then, with the help of (21) and (27), we obtain

$$\mathbf{e}\cdot\nabla_t\psi^s(\mathbf{r}) = \int_C \left\{[\mathbf{e}\cdot\nabla_t'\psi^s(\rho)]\frac{\partial G}{\partial n} - G\frac{\partial}{\partial n}[\mathbf{e}\cdot\nabla_t'\psi^s(\rho)]\right\} ds, \qquad (72)$$

where ∇_t' denotes the transverse component of the gradient with respect to the coordinates of $\boldsymbol{\rho}$. In a manner analogous to that used in Section III, we also obtain

$$\int_C \left\{ [\mathbf{e}\cdot\nabla_t'\psi^i(\rho)] \frac{\partial G}{\partial n} - G \frac{\partial}{\partial n} [\mathbf{e}\cdot\nabla_t'\psi^i(\rho)] \right\} ds = 0. \tag{73}$$

Hence, with the help of (17), we may rewrite (72) in the form

$$\mathbf{e}\cdot\nabla_t\psi^s(\mathbf{r}) = \int_C \left\{ [\mathbf{e}\cdot\nabla_t'\psi(\rho)] \frac{\partial G}{\partial n} - G \frac{\partial}{\partial n} [\mathbf{e}\cdot\nabla_t'\psi(\rho)] \right\} ds. \tag{74}$$

Now

$$\frac{\partial}{\partial n}(\mathbf{e}\cdot\nabla_t'\psi) = \mathbf{n}\cdot\nabla_t'(\mathbf{e}\cdot\nabla_t'\psi), \tag{75}$$

and[7]

$$\begin{aligned}\nabla_t'(\mathbf{e}\cdot\nabla_t'\psi) = (\mathbf{e}\cdot\nabla_t')\,(\nabla_t'\psi) &= \mathbf{e}\nabla_t'^2\psi - \nabla_t' \times (\mathbf{e} \times \nabla_t'\psi) \\ &= -\,\mathbf{e}k_t^2\psi - \nabla_t' \times (\mathbf{e} \times \nabla_t'\psi),\end{aligned} \tag{76}$$

from (16). But[7]

$$\nabla_t' \times [G(\mathbf{e} \times \nabla_t'\psi)] = \nabla_t'G \times (\mathbf{e} \times \nabla_t'\psi) + G\nabla_t' \times (\mathbf{e} \times \nabla_t'\psi), \tag{77}$$

and, from Stokes' theorem,[15]

$$\int_C \mathbf{n}\cdot\{\nabla_t' \times [G(\mathbf{e} \times \nabla_t'\psi)]\}\, ds = 0. \tag{78}$$

Hence,

$$\begin{aligned}\int_C G\mathbf{n}\cdot[\nabla_t' \times (\mathbf{e} \times \nabla_t'\psi)]ds &= -\int_C \mathbf{n}\cdot[\nabla_t'G \times (\mathbf{e} \times \nabla_t'\psi)]\,ds \\ = -\int_C (\mathbf{n} \times \nabla_t'G)\cdot\,(\mathbf{e} \times \nabla_t'\psi)\,ds &= \int_C \mathbf{e}\cdot[(\mathbf{n} \times \nabla_t'G) \times \nabla_t'\psi]\,ds.\end{aligned} \tag{79}$$

It follows, from (74) to (76) and (79), that

$$\mathbf{e}\cdot\left\{ \nabla_t\psi^s(\mathbf{r}) - \int_C \left[\frac{\partial G}{\partial n} \nabla_t'\psi + k_t^2 G\mathbf{n}\psi(\rho) + (\mathbf{n} \times \nabla_t'G) \times \nabla_t'\psi \right] ds \right\} = 0. \tag{80}$$

Since $\mathbf{e}$ is an arbitrary (constant) vector, the expression in the curly brackets in (80) must vanish. But

$$(\nabla_t'G\cdot\mathbf{n})\nabla_t'\psi + (\mathbf{n} \times \nabla_t'G) \times \nabla_t'\psi = (\nabla_t'\psi\cdot n)\nabla_t'G + (\mathbf{n} \times \nabla_t'\psi) \times \nabla_t'G, \tag{81}$$

and $\mathbf{n} \times \nabla_t' \psi = -\mathbf{i}_z \, \partial\psi/\partial s$ on C. Hence, we obtain

$$\nabla_t \psi^s(\mathbf{r}) = \int_C \left[k_t^2 \, G \mathbf{n} \psi(\rho) + (\nabla_t' G) \frac{\partial \psi}{\partial n} + (\nabla_t' G \times \mathbf{i}_z) \frac{\partial \psi}{\partial s} \right] ds. \qquad (82)$$

If we substitute for G from (19) and make use of (20) and (36), we obtain (51). The representation (82) for the gradient of the scattered field is the 2-dimensional analog of that derived by Mitzner[4] in 3 dimensions. We remark that the above derivation differs from his in that we used Green's second identity, whereas he used Green's first identity, and consequently some different transformations to reduce the result to the form corresponding to (82). Mitzner derived his result for the field inside a bounded volume, whereas we are considering scattered fields outside bounded cross-sectional curves, and have used Green's second identity so that we could make use of the radiation condition.

REFERENCES

1. J. J. Bowman, T. B. A. Senior, and P. L. E. Uslenghi, eds., *Electromagnetic and Acoustic Scattering by Simple Shapes,* Amsterdam: North-Holland, 1969, p. 90.
2. A. J. Poggio and E. K. Miller, "Integral Equation Solutions of Three-Dimensional Scattering Problems," in *Computer Techniques for Electromagnetics,* R. Mittra, ed., New York: Pergamon, 1973.
3. B. Noble, "Integral Equation Perturbation Methods in Low-Frequency Diffraction," in *Electromagnetic Waves,* R. E. Langer, ed., Madison: University of Wisconsin, 1962.
4. K. M. Mitzner, "Acoustic Scattering from an Interface between Media of Greatly Different Density," J. Math. Phys., *7,* No. 11 (November 1966), pp. 2053-2060.
5. J. L. Blue, unpublished work.
6. Ref. 1, p. 2.
7. C. E. Weatherburn, *Advanced Vector Analysis,* London: Bell, 1949, pp. 9 and 11.
8. Ref. 2, p. 167.
9. W. Magnus and F. Oberhettinger, *Formulas and Theorems for the Functions of Mathematical Physics,* New York: Chelsea, 1954, pp. 16–17.
10. Ref. 7, p. 30.
11. Ref. 9, p. 22.
12. Ref. 1, p. 6.
13. R. F. Millar, "Scattering by a Grating. II," Can. J. Phys., *39,* No. 1 (January 1961), pp. 104–118.
14. C. E. Weatherburn, *Elementary Vector Analysis,* London: Bell, 1949, pp. 84–85.
15. Ref. 7, p. 24.
16. I. Stakgold, *Boundary Value Problems of Mathematical Physics,* Vol. II, New York: MacMillan, 1968, Exercise 7.45, p. 327.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 58, No. 2, February 1979
Printed in U.S.A.

The Evolution and Special Features of Bell System Telephone Equipment Buildings

By W. PFERD

(Manuscript received June 27, 1978)

There are slightly over 20,000 telephone buildings that house the switching and transmission equipment of the Bell System telephone network. These structures provide a dedicated operational environment for the communication equipment by employing special-purpose mechanical, electrical, and structural systems. Although varying greatly in size, similiar systems appear in all modern central offices and transmission stations. The special systems are required to interconnect the cable and wire, to support and protect the equipment, to power the circuits, and to properly control the spatial environment. This paper describes the 100-year evolution of the standards that control the design of the various classes of Bell System telephone equipment buildings and the sequence of actions necessary to plan and construct a modern facility. Also included is detailed information about the more important aspects of the equipment-building systems, along with numerous photographs that illustrate the special features.

I. INTRODUCTION

Telephone company equipment buildings, known generically as wire centers, central offices, and transmission stations, are geographically placed and specifically designed and constructed to function as effective parts of the nationwide telephone network. As a result, the planning of such facilities requires different considerations than those found in conventional architectural and building design activities. The basic purpose of a telephone equipment facility, and therefore the primary objective in its design, is to provide the appropriate assembly of equipment, cable, wire, and control, operation, and support systems within a protective enclosure to satisfy the needs for local and nationwide telephone service. The enclosure, equipment, and circuits are so tightly interrelated that they are commonly identified as an equipment-building system.

This paper describes the various classes of telephone equipment buildings in the Bell System and the evolution of the design standards for the modern central office and transmission station. The sequence of events that occur in the planning and construction of a new equipment building is presented and is followed by information about the special design and construction of the electrical, mechanical, and structural portions of central offices and transmission stations.

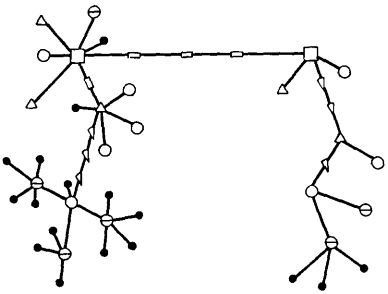
II. TELEPHONE EQUIPMENT FACILITIES

The interconnection of almost 160 million telephones across the North American continent is possible because of complex switching equipment and circuits located at nodes in the nationwide network.[1] Typically, thousands of wires in aerial and underground cables come together at each switching node location. The cables and related apparatus, such as utility poles, conduits, and manholes, are called outside plant. To minimize outside plant costs, the wires leading to a node must converge to a small geographic area which defines the most economical location from which all customers in an area can be served. At the node, the term "wire center" is consequently often used to designate the end portion of outside plant, the apparatus, the interconnecting equipment, and the support structure at that location.

More important than the terminology, however, is the function of these facilities, for they are the means by which telephones and data sets are connected to one another. It is the wire center, or central office, that provides the dial tone on the calling customer's telephone and provides the connection to the line (pair of wires) of the called party. The two telephones are connected through a maze of wiring by switching equipment located in the central office. In other cases, the calling and called customers are served by equipment in different wire centers or are connected to wire centers through a toll office that serves different geographical locations. When these locations are a considerable distance apart, the call will be routed through the nationwide network that comprises hundreds of toll-switching offices widely distributed yet interconnected with high-message-capacity transmission facilities. The network is engineered so that calls processed from one toll office to another will be routed automatically to utilize the transmission facilities as efficiently as possible. If the call is blocked due to overload at some intermediate location, alternate routes are chosen sequentially, and the call is completed through different toll offices and transmission routes. Thus, numbers of central offices or toll offices are needed to process long-distance calls, and they are located especially to serve called and calling customers and the connecting routes.

The nationwide network contains a hierarchy of five classes of offices in which the lowest (class 5) is the local office to which telephones are connected. For toll calls, several class 5 offices that are in contiguous geographic areas may each be connected to a single class-4 office or one of higher rank (lower number) in much the same manner that several thousand customers are each connected to a single local office except that toll circuts, rather than outside plant pairs of wires, are used. A class 4 or higher toll office is, therefore, the switching node and the wire center for a group of class 5 offices. In similar fashion, a group of class 4 and class 5 offices is connected to a single class 3 office, a group consisting of classes 3 to 5 is connected to a class 2 office, and a group consisting of classes 2 to 5 is connected to a class 1 office. The general scheme is shown symbolically in Fig. 1, where a connection is traced from one local office to another, through the intervening lower numbered offices in the hierarchy. The locations of the class 1, 2, and 3 offices in the Bell System are shown on the map in Fig. 2. In addition to the offices classified in the hierarchy, there are tandem offices used exclusively to switch calls between offices in the same region. In summary, there are three general types of central offices listed in five classes. Essential to the interconnection of customers are the different local, tandem, and toll offices. In the hierarchy are local class 5 offices and toll class 4 to 1 offices. Each is a vital element in the Bell System's telephone network.

In addition to the central offices, the other special-purpose structures shown in Fig. 1 between the class 1 to 3 offices enclose equipment that is associated with radio, satellite, and cable transmission facilities. Repeater, power-feed, and main equipment-building systems, known



LEGEND

SYMBOL	CLASS	NAME
□	1	REGIONAL CENTER
△	2	SECTIONAL CENTER
○	3	PRIMARY CENTER
⊖	4	TOLL CENTER
●	5	LOCAL OFFICE
▭		TRANSMISSION STATION
◁		RADIO REPEATER STATION

Fig. 1—Switching hierarchy in two regional areas.

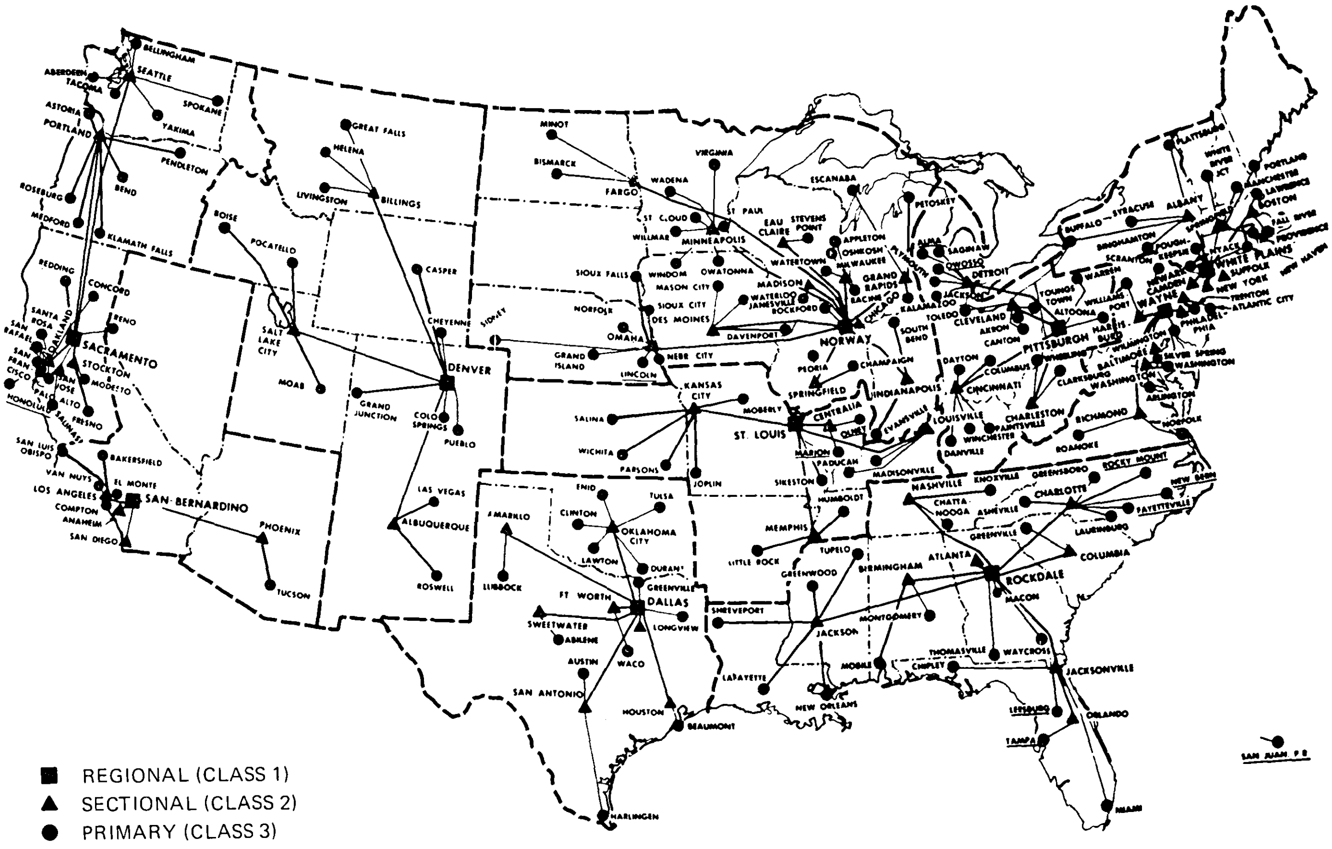


Fig. 2—Locations of regional, sectional, and primary switching centers in the toll network.

generically as transmission stations, are placed regularly along the connecting routes. Main transmission stations, located near metropolitan areas, are large in size and contain equipment, power, and support systems necessary to connect toll offices in nearby cities to the circuits of the long-distance transmission facility. Whether the facility is wire, lightguide, coaxial cable, waveguide, or point-to-point microwave radio, the transmitted signals must be amplified every few miles, the specific distance depending on the type of system. Repeater equipment installed in manholes, in cabinets at the bases of antenna towers, or in special-purpose structures provide the required amplification. Repeater transmission stations have floor-plan areas that are typically 1000 square feet. All provide power and a protective environment for the transmission equipment. The repeater equipment on the coaxial cable routes is powered through the coaxial cable from dc plants located in power-feed transmission stations. These stations are about 6500 square feet in area, are usually underground, and are spaced approximately 75 miles apart on routes which traverse the nation.

The building portions of the system provide the essential environment, power, and structural support for the installation and operation of the telephone circuits and equipment. Based on long-range plans, the building must accommodate, in various stages of growth, the installation of equipment and the interconnection of circuits and must provide control of the environment within the building. The "environment" includes the temperature, humidity, and purity of the internal space needed for proper functioning of the telephone equipment and circuits. Additionally, it includes highly specialized spatial and structural arrangements for routing miles of power and communication wire and cable to interconnect the various units of telephone equipment in three-dimensional (e.g., both interfloor and intrafloor) lattices. Also, special construction may be required to prevent malfunction of equipment by penetration of stray electromagnetic and electrical fields.

The planning for an equipment-building system consists of three stages—first, the establishment of long-range circuit forecasts; second, the determination of the means by which they will be achieved; and finally, the design and construction of the facility. Preceding any architectural design are the important and painstaking tasks of selecting the optimum geographic location for the proposed equipment-building system and making complete plans that identify the telephone equipment's physical and operational characteristics, interconnecting cable-length requirements, compatibility with future expansion of the facility, and supporting subsystems such as the reserve power and equipment-cooling machinery. The outputs of the planning stages, that is, the forecast, the location of the facility, the equipment plan, the cabling plan, and the requirements for electrical, mechanical, and structural support, provide the basis for the construction drawings and specifications that are prepared by the architect/engineer.

In summary, all the Bell System's equipment buildings can be placed in one of two categories: central office and transmission station. The central offices are local, tandem, toll, and various combinations called multi-entity offices, while the transmission stations are repeater, power-feed, and junction or main. There are slightly over 11,500 central office buildings that serve as nodes in the nationwide telephone network and close to 9000 transmission stations located along the transmission routes. Of the two general types of equipment buildings, the central offices are the most varied, ranging in size from the smallest (400 square feet) to the largest (1 million square feet) structures in the Bell System. In total, the Bell System equipment buildings contain about 230 million square feet of floor area for equipment and support systems.

III. EVOLUTION OF THE MODERN TELEPHONE EQUIPMENT BUILDING

Telephone technology has experienced a vast amount of change since its inception 100 years ago, including the way in which telephone buildings are designed and used. At first, the building was simply a place to interconnect jumper wires. Later, this awkward-to-use system was replaced with manual switchboards that had cords fitted with plugs for operator usage. By the end of the nineteenth century, automatic connection systems were in development, but it was not until the 1920s that switching machines, called entities, were introduced into the buildings. By 1930, 9 percent of the Bell System's offices were dial; by 1940, the figure had grown to 38 percent. The shortages of material and manpower resulting during World War II reduced temporarily the rate of conversion to dial, but activity was renewed in 1945, and by 1960, 94 percent of the central offices were dial. Most of the Bell System's manual central offices were retired during the change to automatic dial service. Table I contains a tabulation of the types of central offices in service between 1950 and 1977.[2] Four types of switching equipment are involved: step-by-step, panel, crossbar, and elec-

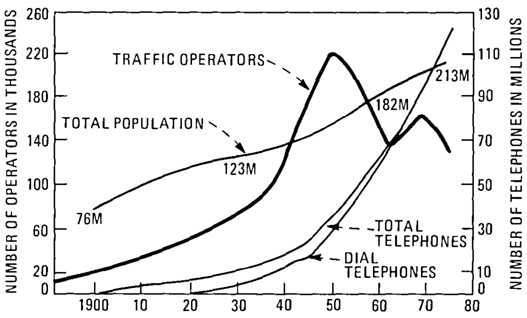
Table I—Number of central office codes

Dec. 31	Manual	Panel	S × S	X-Bar	ESS	Total*
1950	3,257	502	4,107	604		8,470
1955	1,991	512	6,087	1,161		9,751
1960	715	494	7,511	2,258		10,978
1965	94	528	8,212	4,281	1	13,121
1970	11	451	8,393	5,637	264	14,756
1975	1	144	7,911	6,549	2,183	16,788
1977	1	68	7,223	6,537	3,477	17,306

* Some buildings are multi-entity facilities containing more than one central office code. Therefore, the number of Bell System central offices exceeds the actual number of central office buildings.

tronic. During the last five years, new offices have been almost entirely of the advanced electronic type that operates automatically for the processing of calls.

As telephone usage by the general public grew through the years and the demand for increased service materialized, more complex central offices and transmission stations were needed. This was the basis for telephone company growth and modernization programs that cause the network today to have such a large number of offices and stations. During the 55 years of steady growth in usage of switching and transmission equipment, the extent of operator assistance has undergone dramatic changes which have also influenced the design of equipment building. Figure 3 shows the number of traffic operators along with the number of dial telephones and the total population for the years 1890 to 1975.[2] Because of the change from a manual to a mechanized mode of operation for toll calling that was made possible by crossbar central offices, the number of operators in the Bell System declined after 1950. By 1960, the number began to increase again in response to the increase in special toll calls needing operator assistance. After 1970, improvements in call processing made possible by the Traffic Service Position System (TSPS) caused a second decline in the number of operators. TSPS comprises electronic equipment at local and toll offices connected by circuits to operator consoles installed in administrative-type space. Typically, the equipment and the operators are located in different buildings. This arrangement offers the advantage of concentrating the operator activity for efficient call processing, and operators can be located in less expensive conventional office buildings rather than in equipment buildings. The switchboards replaced by these TSPS consoles are the last items requiring heavy personnel staffing within a central office. As a consequence of the introduction of TSPS during the past ten years, all new central office buildings are now designed primarily for equipment.



REDUCTIONS IN TRAFFIC OPERATOR REQUIREMENTS IN 1950 AND 1970 RESULTED FROM THE INTRODUCTION OF DIRECT DISTANCE DIALING AND FROM THE TRAFFIC SERVICE POSITION SYSTEM RESPECTIVELY.

Fig. 3—Bell System growth compared to total population growth.

The long-term trend toward central offices with few human occupants has been accelerated further by the introduction of automated maintenance equipment. This has occurred during the last 25 years with crossbar and electronic offices where fewer personnel are required due to machine-controlled self-diagnosis and repair procedures. In modern switching machines, checks on performance are made automatically, problems are diagnosed and reported on a teletypewriter printout or cathode ray tube for correction. In some cases, the malfunctioning parts are automatically disconnected from service and the needed replacement parts are identified along with the trouble-alert information. Similarly, computerized automatic maintenance systems are also used for remote-surveillance and testing of equipment associated with the other functions of a central office, namely, subscriber loop plant, trunks, carrier systems, and special services. At the present time, over 60 centralized maintenance systems exist to serve the equipment along with providing administrative records for the facilities and circuits.

A specific example of a modern facility that is designed for few occupants is the electronic toll office at Rego Park in Queens, New York. This multi-story structure contains about 200,000 square feet, has an ultimate capacity of approximately 100,000 toll-line terminations handling over 500,000 busy-hour calls, but will require only about 60 people for operation of the telephone equipment systems. In contrast, if the previous generation of switching equipment had been used, 425 operating personnel would be required and, although it is not technically possible to use manual cord-type switchboards of this capacity, an office handling this many calls would require approximately 2300 operators.

The trend toward increased size and full automatic operation also appears in the transmission stations. The first toll routes were limited to only a few telephone conversations (channels) per pair of wires. Analog carrier systems, introduced in 1941, could accommodate 600 channels per coaxial cable pair or 1800 channels in a cable of eight coaxial tubes with repeater stations about 20 miles apart. With innovations introduced over the years, it is now possible to accommodate over 10,000 channels on a single coaxial pair and over 100,000 channels in a cable sheath containing 22 coaxial tubes. Similarly, the channel capacity of radio relay systems has grown over the years as a result of advances in circuit and antenna technology. An important consequence of the increase in cable and radio system capacity is reduced expenditure per channel; however, this has increased the amount of equipment in transmission stations and has required more stations on a route. Funds that would have otherwise been invested in cable plant have, therefore, been used for equipment-building systems. Because of the number and the geographic locations of the transmission stations at remote sites, the use of complex equipment for monitoring and control purposes is necessary. The result is virtually unattended operation of all radio and cable transmission stations. Only main or junction stations require assigned craftspersons. Typically, less than a dozen craftspersons can operate the largest of the main stations and, thereby, the building can be designed almost entirely for equipment purposes.

The conversion from manual operation to dial service and from attended to automatic systems has also caused a change in the characteristics of the interior space used for equipment interconnection. The automatic switching and transmission equipment introduced physical and environmental requirements that were entirely different from those of the manual switching offices. Previously, the labor-intensive activities of operators demanded an office-type environment. Interestingly, the term "office" that originated during this period is still associated with the modern switching center. The change from office-type space to equipment space for growth was considered so important during the early days of dial conversion that special engineers were employed to be responsible for providing the space and structures with the features necessary for proper functioning of the switching and transmission equipment. Continuing to the present, these telephone company building engineers interact with the planning and equipment engineers to provide buildings that are designed to special standards and specifications.

During the past 100 years, Bell System telephone building standards have been established during three periods. The first standards were set before 1900 and were concerned principally with provisions for the entrance of cable into the building, shafts for running cable from floor to floor, and the long switchboards located in the operating rooms.[3] The manual service of this period required large staffs on continuous duty in the central city buildings. Additional space equal to that required for the switchboards was devoted to quarters for the operators. In some small places, the central office was also the home of the operator and his or her family. These earliest of standards resulted in offices of the types shown in Fig. 4. Many were of wood construction to harmonize with the suburban and rural areas, while those located in major cities were of multistory, fireproof office-building type design and construction.

About 1925, AT&T engineers studied the problem of standardizing equipment and buildings in anticipation of the introduction of electromechanical switching equipment. At that time, the frames on which equipment units were mounted varied in height from 9 to 14-½ feet. High ceilings for temperature control were common, and cabling was relatively simple. As a result of this second study, the Bell System adopted a standard of 11-½ feet for the height of equipment frames. These so-called high bay requirements, which at that time made best

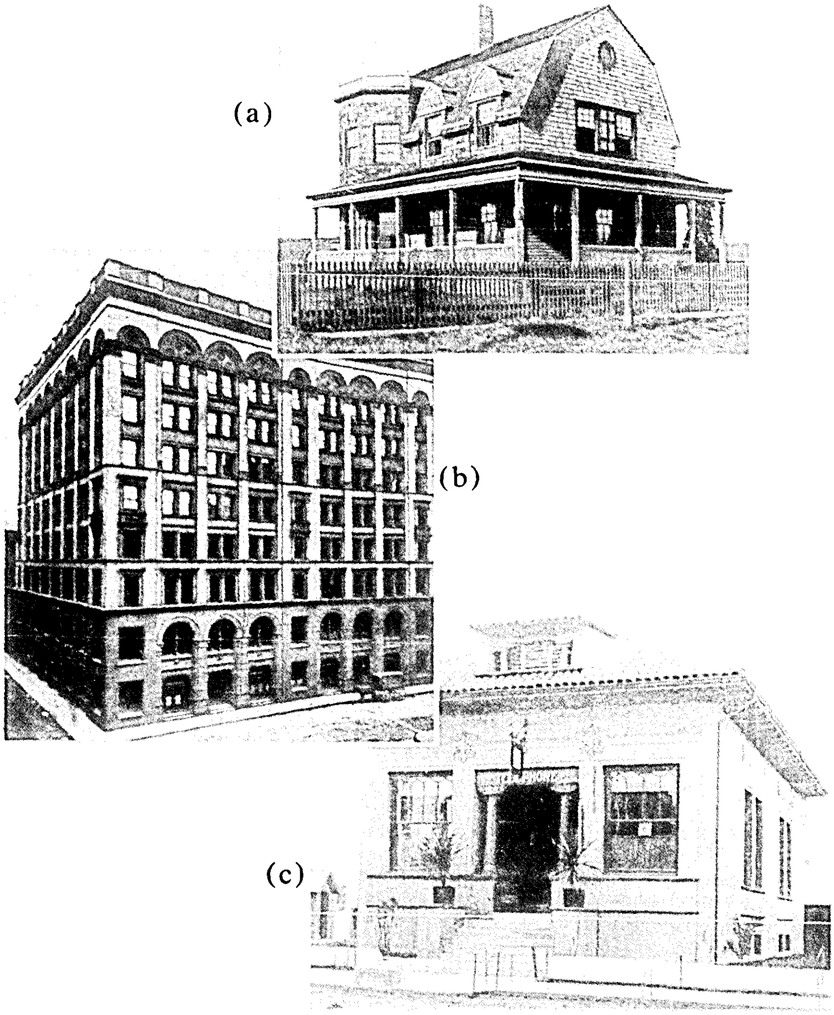


Fig. 4—Pre-1900 Bell Telephone central offices. (a) Typical country exchange with manager's dwelling above. (b) Main telephone office. (c) Local exchange.

use of building space, held until the mid-1960s and dealt exclusively with providing structural, spatial, and cabling accommodations for tall frameworks that characterize electromechanical switching systems. During the years 1925 to 1965, over 10,000 central offices and transmission stations were constructed to the high bay standards. Typical buildings of this era are shown in Fig. 5. All are fireproof, have high ceilings, and are of heavy construction to support the heavy, tall frames of equipment and large bundles of interconnecting cable. Most were designed for occupancy by craftpersons and the larger ones also provided space for operator service.

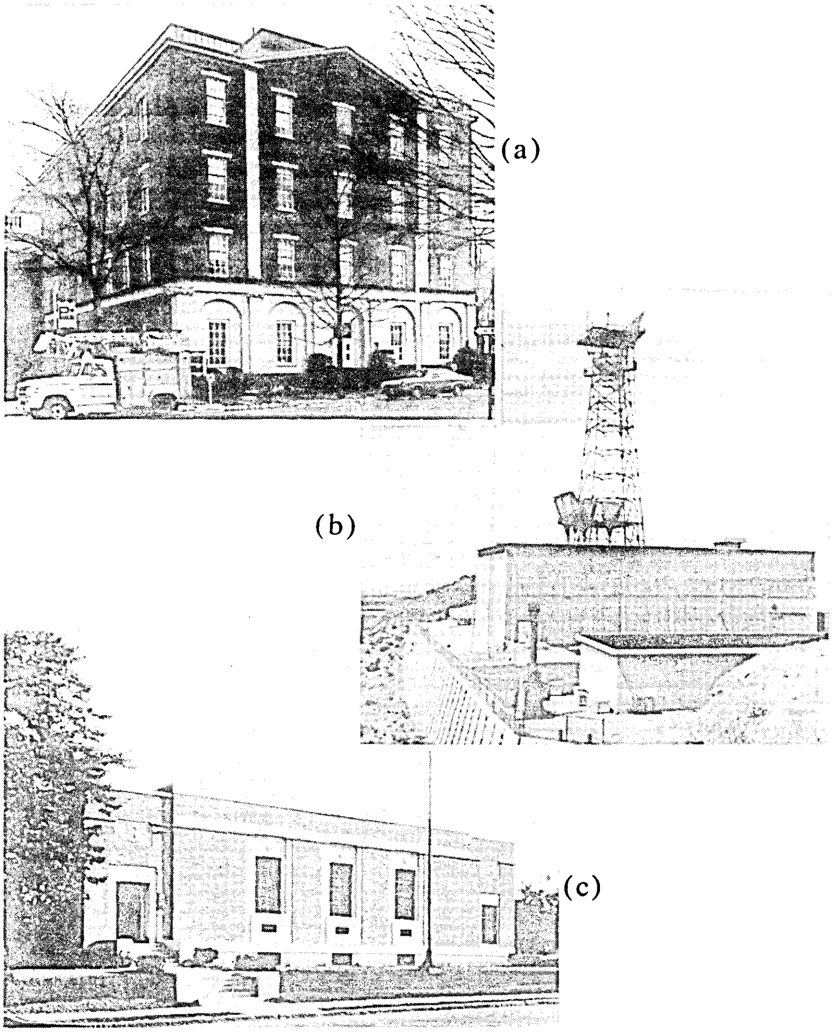


Fig. 5—Structures erected, from the 1930s to the 1960s, for tall equipment. (a) Large multi-entity office. (b) Mountaintop radio repeater station. (c) Local central office.

With the introduction of the Electronic Switching System (ESS) equipment in 1965, studies were again undertaken to assess the need for alternative standards for equipment buildings. The miniaturization of circuit elements in the electronic equipment enabled capacity and features never before attainable in switching and transmission systems; however, the individual assemblies exhibited unprecedented heat, cabling, and weight concentrations. Consequently, the third look at equipment building standards was undertaken.[4] A building and equipment task force representing all systems development areas of Bell Laboratories was convened to study and to recommend appropriate new standards. The task force devised a third group of complementary specifications for equipment and buildings that are contained in the New Equipment-Building System (NEBS) documents: (*i*) the *Equipment Design Standards,* BSP 800-610-164 for Bell System use and PUB 51001 for the general trade, provide the spatial and environmental performance requirements for all new equipment systems, (*ii*) the *Building Engineering Standards* (BSP-760-100-xxx and 760-200-xxx) specify the planning and design of buildings to adequately accommodate the modern equipment, and (*iii*) the *NEBS Catalog* lists Western Electric equipment to be installed in these facilities.

The latest in the series of Bell System building standards was adopted for use in the design of all local offices after 1972 and for all other types of central offices and transmission stations after 1974. Examples of equipment buildings constructed to NEBS standards are shown in Fig. 6. They are characterized by large windowless rooms for equipment, with associated space for elaborate environmental control and reserve power systems. Features such as these, plus the designated locations of the office at the wire center, the provisions for extensive interconnecting cable systems within the structure, and the provisions for only limited human occupancy, are typical of the modern equipment facility.

IV. NEBS STANDARDS

Equipment building standards are important not only to the performance of the facility but also to the cost of buildings since each of the physical characteristics of telephone equipment has a corresponding effect on the design of the building intended to house that equipment. Frame height and cabling space, for example, control the so-called "clear" ceiling height from floor to lowest overhead obstruction. Equipment weight determines the building's "live load," the weight that foundations, columns, and floors must support in addition to their own weight. The amount and location of heat emanating from the equipment sets the size of the cooling plant and the location of the air ducts and diffusers. But when equipment units differ greatly, the variations in their physical characteristics prevent the efficient use of space in telephone buildings and complicate the already difficult task of office planning and design. In the absence of uniform standards, space in equipment buildings must be engineered to meet the most severe requirements and to accommodate the widest possible range of conditions. The result is costly designs with dubious performance capability.

To avoid these complications, the NEBS standards include the full range of spatial and environmental conditions. The requirements cover equipment frame areas, distributing frame areas, power equipment areas, operations support systems areas, cable distribution systems,

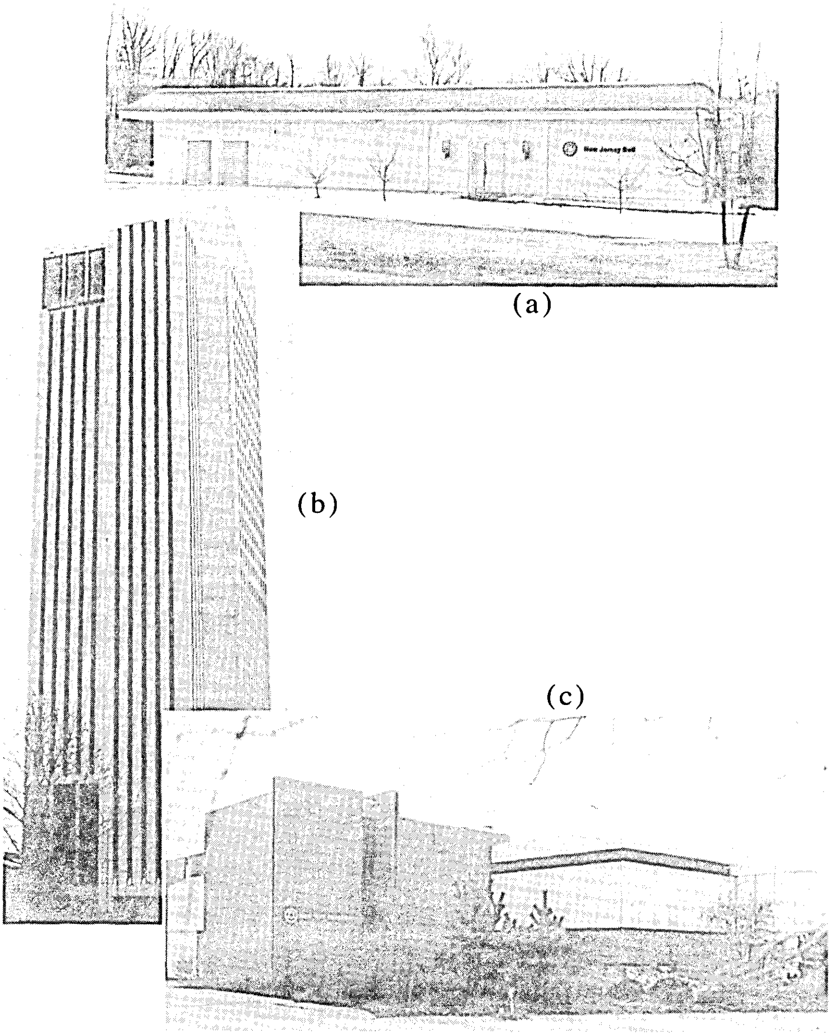


Fig. 6—NEBS central offices for 7-foot equipment. (a) Surburban switching office. (b) Metropolitan toll center. (c) Urban switching office.

and cable entrance facilities. The environmental requirements are grouped according to functional effects and include thermal, fire resistance, shock, vibration, earthquake, airborne contaminants, grounding, accoustical noise, illumination, and radio frequency interference. Standards, design requirements and planning guidelines exist in the NEBS documents for each equipment area in the building and for all the building support systems and structure.

Of principal note is the height standard of 7 feet for all new electronic equipment systems. Because of the compact electronic equipment, a standard decrease in the floor-to-floor height of the building is possible for all new Bell System central offices and leads to substantial savings. For each foot the clear ceiling height is reduced, the cost per square foot of space in a multi-story building decreases by about 4.5 percent. When the overall expense for a central office building designed to the older high bay standards is given a reference value of 1.00, the relative cost for a similiar building to house 7-ft frames is 0.92, or an 8 percent overall cost saving. The curves in Fig. 7 show that, for a 7-ft frame, costs are at the reference level when the floor-to-floor height is 15-½ ft. But these costs decrease with decreasing story height. At a floor-to-floor height of 14-½ ft, the costs are 96 percent of the reference level, and at 13-½ ft, (the NEBS standard floor-to-floor height), costs are only 92 percent of the reference level. Additional savings result from less stringent requirements for floor loading, from more compact floor plans, more accessible cable racks, and more efficient lighting, and from eliminating some supports and ladders used with taller frames.

The allocation of vertical space on a floor and the partitioning of the allowable floor load for NEBS equipment space is listed in Table II. To encourage the efficient use of the floor area in buildings, standard floor plans are available for each of the major types of equipment. The NEBS standard floor plan for 12-in. deep equipment frames is shown in Fig. 8, where five line-ups per building bay are indicated. Cable holes, miscellaneous frames, power equipment, and a process cooler can be located in a sixth line-up at the column line. In most equipment areas, the 2-½ ft maintenance aisles and 2-ft wiring aisles provide adequate space for operation, maintenance, and cabling.

The vertical space allocation is shown schematically in Fig. 9, where equipment and two identifiable cable systems are placed under the 10-ft clear ceiling. Cable pathways are designated for system cabling and

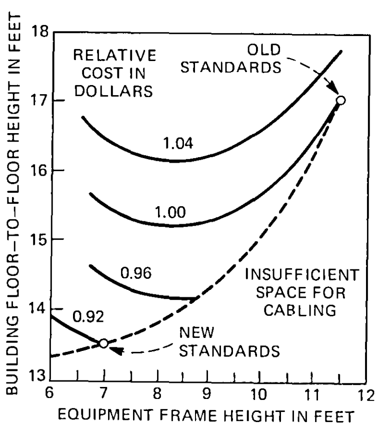


Fig. 7—Cost study for NEBS standards.

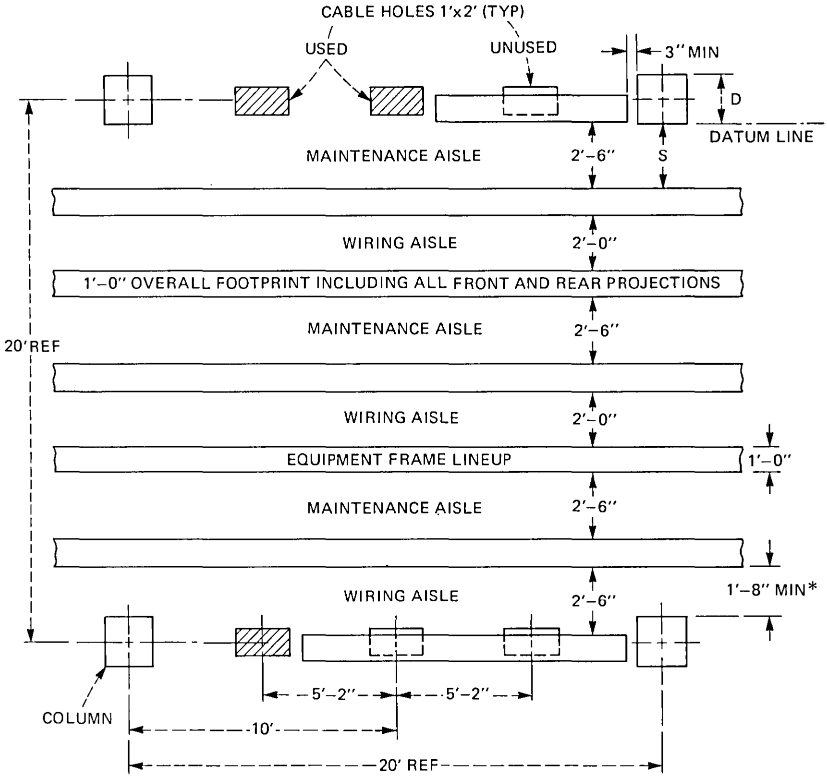
Table II—NEBS equipment space and load allocations

Equipment	Vertical Space	Floor Load
Equipment Frame Area:		
Frames	Floor to 7 ft	115 psf
Cable distribution system and installation clearances (includes allocation of 5 psf for via CDSS)	7 to 10 ft	25 psf
Power Area:		
All equipment, cable, and installation clearances	Floor to 10 ft	140 psf
Distributing Frame Area:		
All equipment, cable, and installation clearances	Floor to 9 ft	135 psf
Via CDSS	—	5 psf
Cable Entrance Area:		
All equipment, cable, and installation clearances	Floor to 10 ft	140 psf
Operations Support Systems Area:		
All equipment, cable, and installation clearances	Floor to 10 ft	140 psf
Conventional Cooling System:		
Overhead ducts and diffusers	10 to 12-½ ft	—
Modular Cooling System:		
Raised floor	Floor to 1-½ ft	10 psf
Supply, return, and drain piping	Floor to 1-½ ft	—
Process coolers	1-½ to 11-½ ft	115 psf
Suspended ceiling	11-½ to 12-½ ft	—
Transient Loads	—	10 psf

via cabling, that which joins and powers the various systems throughout the building. Pathways for three levels of cable racking occupy the 7-ft to 10-ft space with provision over the life of the equipment-building system for lights, openings for cooling air, and installer access. The application of the NEBS standards for equipment, floor plans, and cabling produces near-optimum use of space. Additionally, use of the standards simplifies building design and equipment engineering, streamlines equipment and cable installation, and allows for flexibility in growth patterns.

V. THE PLANNING OF A CENTRAL OFFICE

To achieve the special design of new equipment buildings and additions requires not only company ownership rather than leasing, but also company planning, engineering, and preparation of specifications. Figure 10 shows a typical sequence of central office planing activities. The charted durations, representative of an office of 20,000 to 30,000 telephone lines, could vary somewhat depending on circumstances associated with a particular project. A new office is an outgrowth of a continuing process called fundamental planning that is performed in each operating company by the plant-extension department. Possible patterns of growth for the residential, commercial and industrial areas of a community are studied, and predictions are made that form the basis for additions to the outside plant and the construction of a new central office. The actual location, the timing, and the expected size for the new facility are interrelated study factors. Minimized investment is sought by decisions that balance the investment



COLUMN DEPTH "D"	SPACE AT COLUMN FACE "S"
1'-10" OR LESS	2'-6"
2'-0"	2'-4"
2'-2"	2'-2"
2'-4" OR GREATER*	2'-0"

*FOR COLUMN DEPTHS GREATER THAN 2'-4" IT MAY BE NECESSARY TO OMIT SOME FRAMES IN THE EQUIPMENT LINEUP OPPOSITE COLUMNS (WIRING AISLE SIDE)

Fig. 8—NEBS standard floor plan for principal depth (12-in.) frame.

in outside plant for the area and the investment in the equipment-building system at the wire center.

It is the plant-extension department's responsibility to perform economic studies of potential alternative solutions to handle growth situations and to make recommendations that eventually will result in either new central offices or additions to existing structures. At existing central offices, telephone usage is observed and compared against the traffic-handling capabilities of the equipment and network for the area. As telephone usage increases with time and the existing circuit capacity becomes limiting, the existing central office and network facilities must be enlarged. A management decision on the construction of a new or

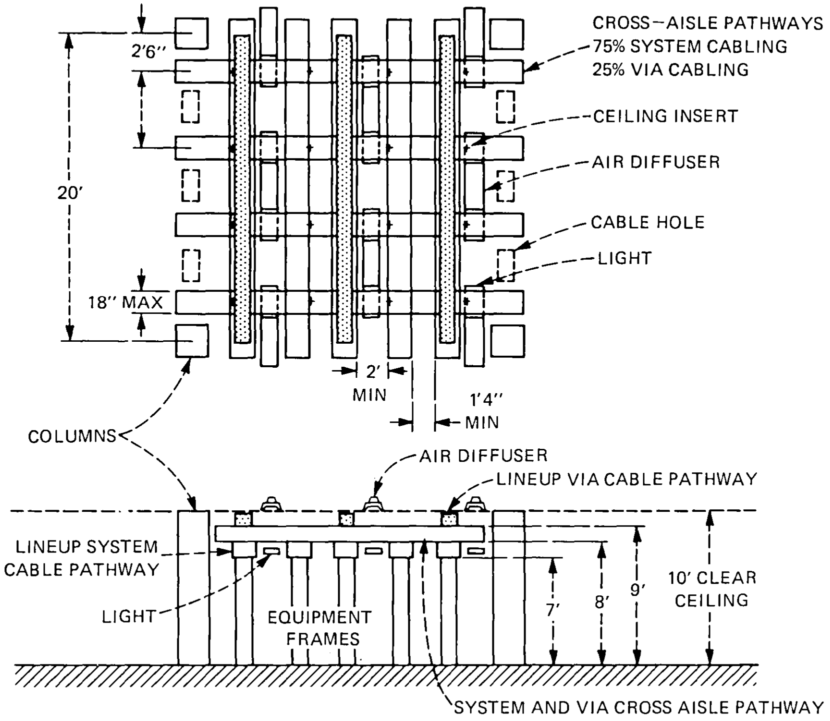


Fig. 9—NEBS cable pathways plan for 12-in. deep frame areas.

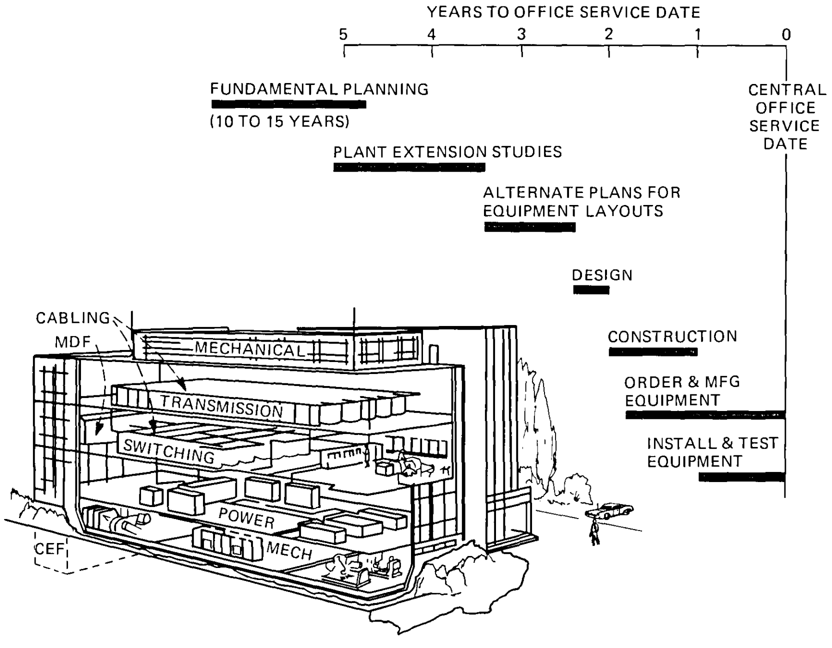


Fig. 10—Typical central office planning process. The isometric building illustration shows the types of space and equipment.

expanded facility is typically made three and sometimes five or more years before the proposed facility is put into service.

The decision to construct initiates a new-office project that will be the responsibility of other telephone company personnel, experienced not only in the equipment portion of planning but also in electrical, power, mechanical, structural, real estate, and legal aspects. The first task is to estimate equipment requirements, those for initial service and for the fully developed office. The final installation of equipment may be 5, 10, or even 20 years from the time of planning. The types of equipment that are likely to be used are local switching equipment, toll switching equipment, terminal and transmission equipment for the various toll networks that include analog and digital coaxial cable routes and analog and digital microwave radio routes, operator-service equipment, ac and dc power equipment, maintenance and administrative equipment that uses embedded minicomputers to service the telephone network, cable entrance facilities, and extensive vertical and horizontal cable-distribution systems.

The space planners must select the type and quantities of this equipment to custom-fit the service requirements of the particular central office or station, locate each of these equipment systems within the facility so that each has future growth potential without interfering with the others, and make accurate layouts from floor plan data of the individual pieces of equipment to ensure adequate spacing. Once all telephone equipment requirements are established, they become the basis for determining the capacity and physical requirements for equipment cooling, ventilation, and normal and reserve ac and dc power. The entire process requires considerable interaction with the different engineering and operating groups and may take anywhere from several months to several years. Furthermore, it is usually complicated by the fact that all planning is dependent on the telephone-usage forecast which may be amended during the planning interval and on equipment technology which is characterized by rapid changes in capability and support requirements. Finally, an architect is called upon, but only after the site of the wire center is selected and procured and study plans are completed and mutually agreed upon by the involved telephone company personnel, e.g., the planners, the equipment engineers, the building engineers, and the departments which will operate the facility.

The study plans, which show the location for all telephone, electrical, and mechanical equipment and the special structural configurations needed to accommodate this equipment, contain virtually all the basic dimensional information needed to describe the facility; that is, number of stories, floor-plan dimensions, ceiling heights, column spacing, and locations of cable-hole openings in the floors, of removable walls for future growth, of mechanical rooms and of electrical rooms. The study plans and the design standards, such as those to control construction material quality, floor-load, floor-levelness, and capacity of the standby power and air-conditioning systems, are the basic information that is passed on to the architect who then prepares the drawings and specifications to be used by the construction contractors. In addition, the study plans go to the equipment manufacturer who also receives the contract to install the equipment and interconnect it with cable.

A typical plot and floor plan of a local electronic switching system office in Liverpool, New York, is shown in Fig. 11. Planning for this central office began in 1971 and includes considerations for equipment installations to 1985. In this building, 63 percent of the gross floor area will be used for equipment, 20 percent for ac and dc electrical systems, 15 percent for mechanical systems, and 2 percent for nonequipment purposes. It is apparent from these plans that the goal of a fully integrated and interrelated equipment-building system is achieved.

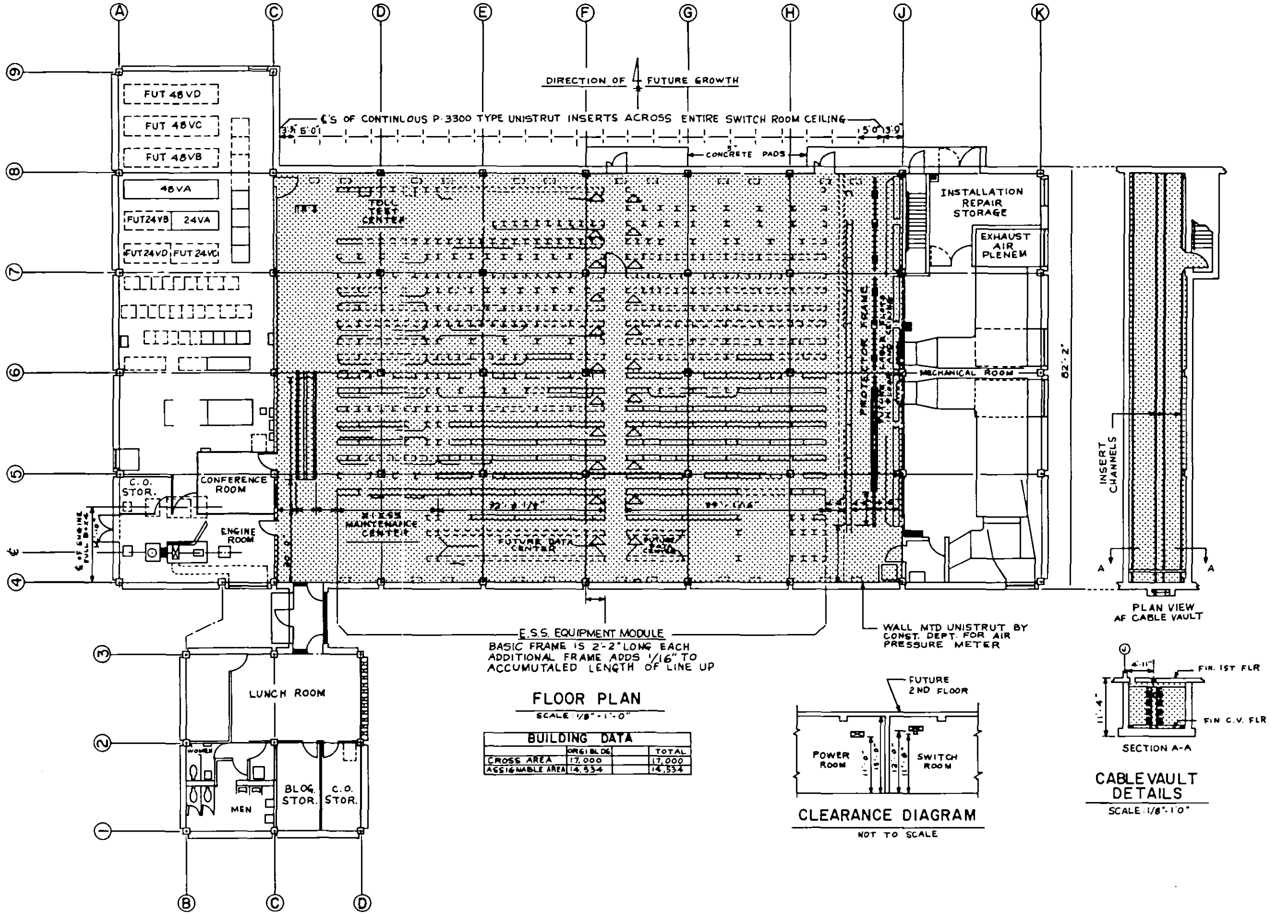
VI. SPECIAL FEATURES OF CENTRAL OFFICES AND TRANSMISSION STATIONS

Although modern central offices and transmission stations vary greatly in size, similar features appear in all of them. From the small local office and repeater station to the large multistory toll office and main station, the basic elements of the equipment-building system are repeated. The special features are required to facilitate the interconnection of internal and external cables and wires, to support and position the equipment, to power the circuits, to obtain proper environmental control, and to provide for installation, maintenance, and operation.

Detailed information about the more prominent features in the modern equipment building is given below, with illustrative photos of the larger facilities.

6.1 Cable entrance facility

As a wire center, the central office must have provision for bringing in thousands of pairs of wires. Similarly, provisions must be made at stations along coaxial cable routes for the entrance of cable and at radio stations to support the waveguide between the antenna and the equipment. A medium-sized central office cable entrance facility (CEF) is shown in Fig. 12 in a below-grade situation. The CEF is a vault-like area typically 12 to 15 ft high, 12 ft wide, and the length of the central office directly under the main distributing frame. It can be over 200 ft long. One or both of the end walls contain a conduit termination with its built-in gas-venting chamber that is employed to guard against water and hazardous gas entering the central office. As shown in Fig. 12, the terminating conduit formation provides an entrance area for



BUILDING DATA

	ORG'L BLDG.		TOTAL
CROSS AREA	17,000		17,000
ASSIGNABLE AREA	14,534		14,534

Fig. 11—Central office study plan.

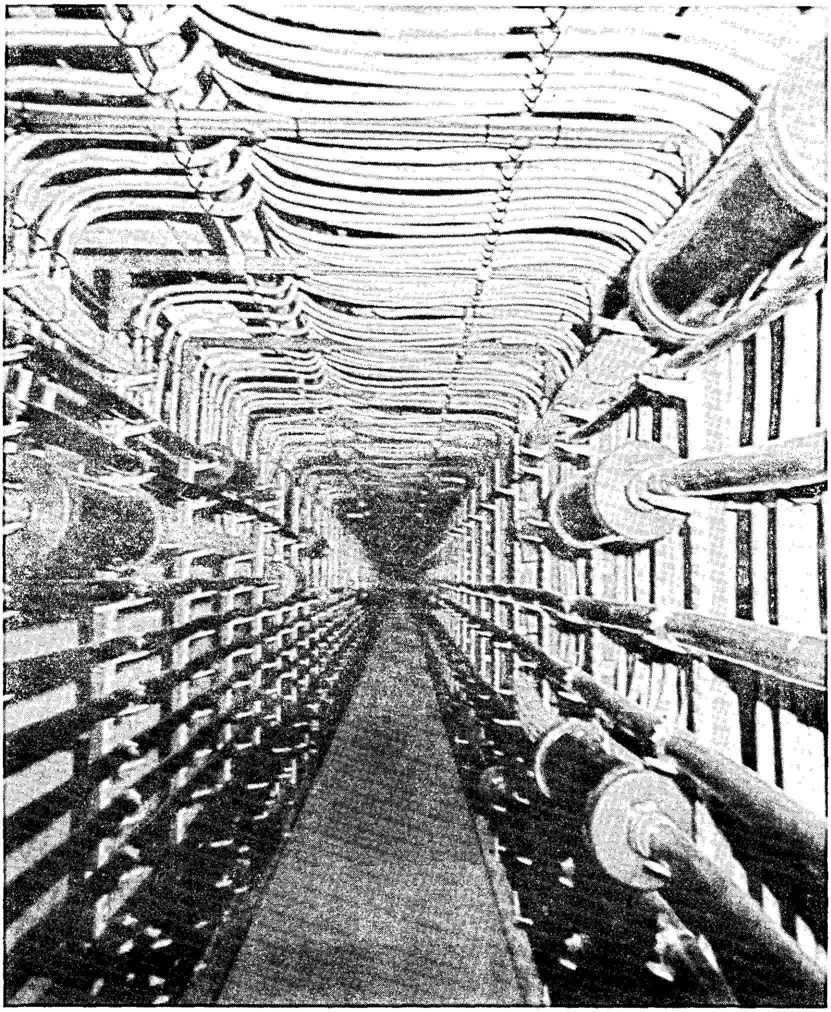


Fig. 12—A subgrade cable entrance facility with provision for cable penetration through basement walls, ceiling, and the outer wall to upper floors.

multiconductor cables, each approximately 3 inches in diameter. These cables are placed on steel support racks that are attached to the long walls, and brought to a location where they are needed upstairs on the main distributing frame. At this point, the cables are spliced to smaller terminating stub cables and routed upward through a ceiling penetration. This is the usual situation, but sometimes riser cables are directed to upper floors through conduits or shafts in the outer wall. Also, local site conditions, such as a rock ledge, high water-table level, or the projected size of the facility may preclude the use of a sub-grade structure. In such cases, the CEF is constructed paralleling the distributing frame and the outside-plant cables are brought into the office at grade level through a conduit structure built adjacent to the office.

These cable entrance facilities for underground feeder, trunk, and toll cables also provide space and structural support for the following:

(*i*) Pressurization of outside-plant cables to prevent moisture from penetrating the inner core of the cable.
(*ii*) Isolation of dc potentials for corrosion protection.
(*iii*) Grounding of cable shields for electrical protection and noise reduction.
(*iv*) Installation of splices between the feeder cable coming from the outside plant and riser, stub and bridging cables going to distributing frames.
(*v*) Routing of feeder, riser, and terminating stub cable including required cable spreading.
(*vi*) Cable-placing activities including provisions for pull-in irons, feedholes, and work platforms.

6.2 Telephone equipment areas

Approximately 60 to 85 percent of the interior space in a central office consists of telephone equipment areas. These are large, usually windowless, partitionless rooms designed to contain the equipment, the appropriate cable support systems, and the air ducts for environmental control needed for the equipment to function. The switching and transmission equipment installed in these areas are usually mounted, as shown in Fig. 13, on steel frameworks that are 1 to 2 ft deep and up to 6-½ ft wide. The frames are installed side by side in lineups usually 30 to 50 ft long and, because of the enormous size and weight, are bolted to the floor or ceiling or both with concrete-embedded anchors to prevent vibration and toppling.

The aisles between each lineup are typically between 2 to 4 ft wide and are used for access to the equipment for wiring and maintenance. Above the equipment frameworks are other steel structures that support tons of cabling which connect the equipment together and to the distributing frames.

The main distributing frame, shown in Fig. 14, is also in the equipment area. Distributing frames are the wiring interfaces between the subscriber plant and toll cables that enter through the CEF and the telephone switching and transmission equipment. In multistory offices, distributing frames and equipment are on a number of floors; therefore, vertical access for cables must be provided. Many heavy bundles of interfloor cabling, as shown in Fig. 15, are used. In the equipment area, steel frameworks are positioned between floors to support the vertical cabling and special reinforcement is employed in the floors to maintain structural integrity at points of penetration.

The dc power cabinets and batteries that provide the power conversion and uninterrupted energy source for the telephone circuits are also located in equipment areas but are usually separated in power rooms from the transmission and switching equipment. Power equipment occupies approximately 10 to 15 percent of the gross equipment area and consists of lineups of lead-acid batteries, as shown in Fig. 16, and cabinets of electrical equipment such as rectifiers, converters,

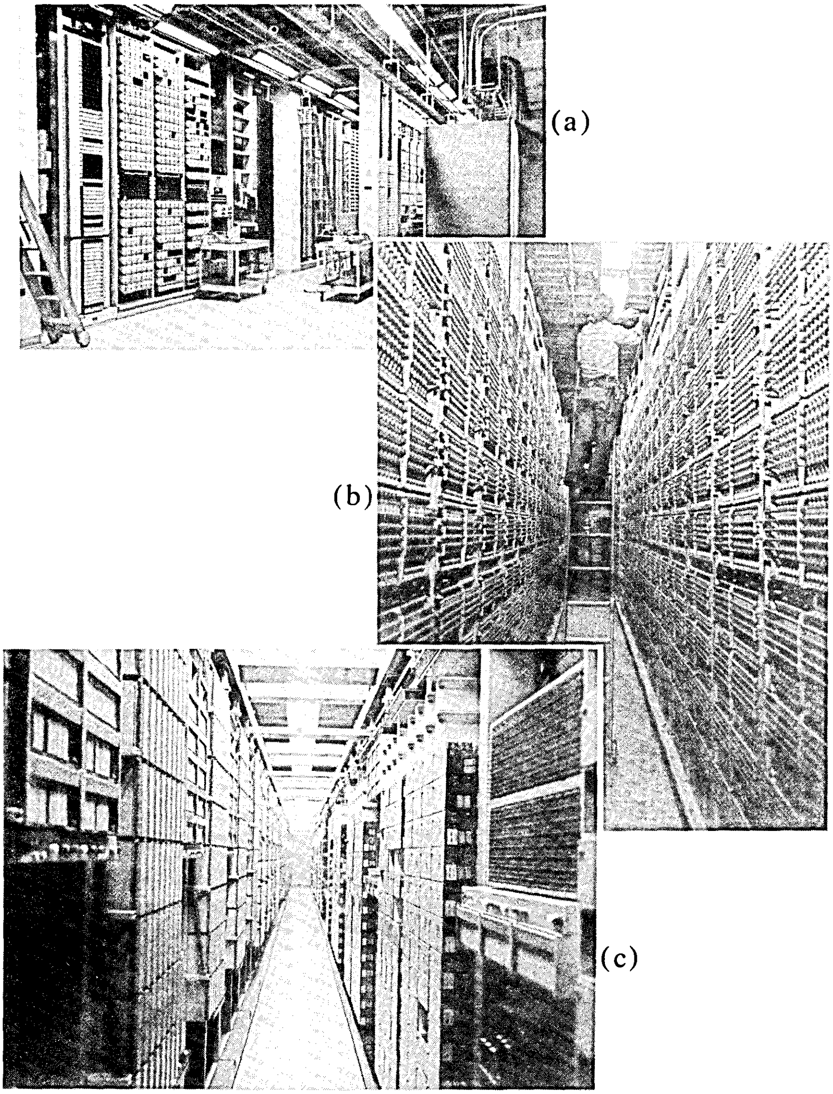


Fig. 13—(a) Toll equipment and (b) crossbar switching equipment mounted on 11-½-ft frameworks. (c) ESS equipment mounted on 7-ft frameworks.

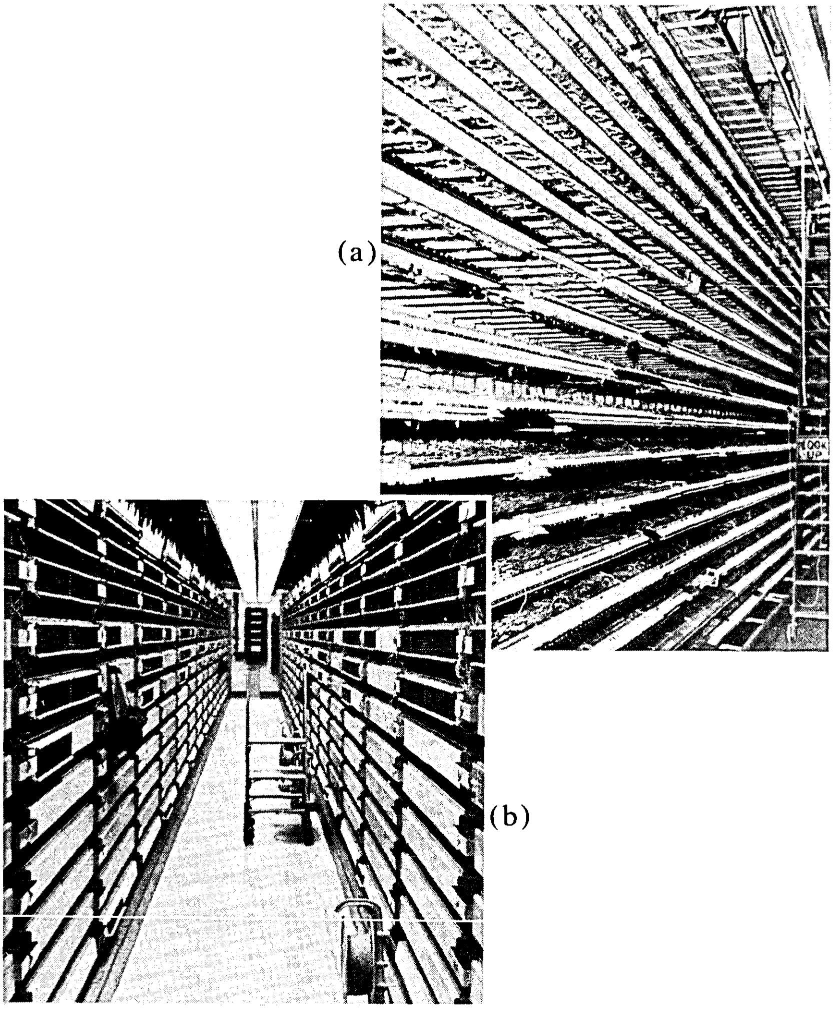


Fig. 14—The main distributing frame, the wiring interface between outside plant cables entering cable vault and the telephone equipment. (a) 12-½-ft MDF used with 11-½-ft frameworks for central offices. (b) 8-ft COSMIC MDF used with 7-ft modern electronic equipment.

panels, and service boards that control the distribution of power. The batteries and the associated overhead power distribution bus bars are assemblies typically 2-½ ft deep, 10 to 15 ft wide, up to 9-½ ft high and weigh about 10 tons per 50-kW module.

The spatial and weight characteristics of the different types of telephone equipment described above impose special design conditions on the structure. The more important structural design requirements for the modern Bell System equipment buildings are summarized in Table III, along with values used in a typical office building.

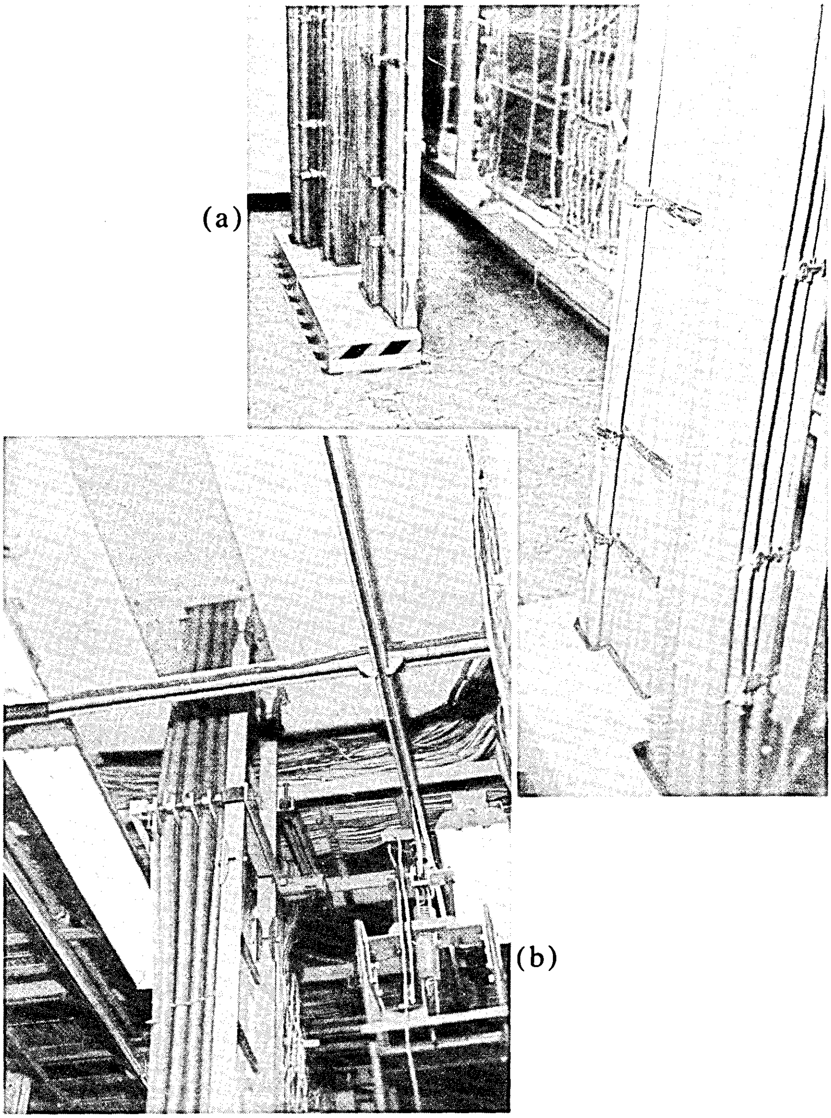


Fig. 15—Telephone equipment rooms, showing cable routed above the frames and from one floor to another.

6.3 Electrical systems

In addition to the dc power that directly serves the telephone equipment, there are a number of other special electrical features. Their design and construction are characterized by the need for extreme reliability and high capacity. The electrical systems provide means for uninterrupted service under all types of emergency condi-

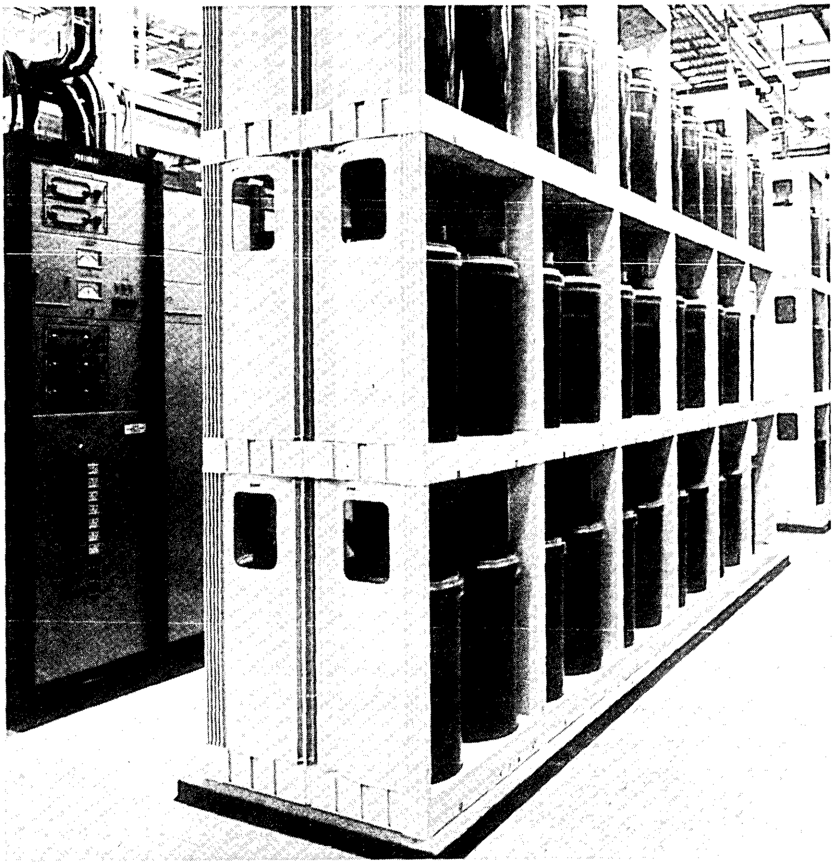


Fig. 16—Central office battery plant with dc control and distribution cabinets.

tions. Additionally, special control and protective circuits are employed. The variety and composition of the electrical systems are as follows:

(*i*) The ac power system consists of an entrance transformer for utility power, switch gear for primary distribution, motor-control centers, branch circuit protection equipment and conductors, switchboards, service cabinets, and bus-duct cable assemblies. These elements are necessary to power the central office telephone lines, terminals, and equipment under normal operating conditions. Additionally, the ac power system serves the mechanical systems that heat or cool the equipment.

(*ii*) An emergency system, shown in Fig. 17, is employed to provide power in the event of failure of the commercial source. Provisions exist to transfer all circuits to locally generated or reserve power. Typically, one or more diesel- or turbine-driven alter-

Table III—Structural characteristics

	NEBS Telephone Facilities	Office Buildings
Floor load*	150 lb/ft†	75 to 100 lb/ft†
Maximum deflection†	½ in.	1 to 2-½ in.
Floor levelness‡	±⅛″ in 8 ft	±¼″ in 10 ft
Column spacing*	20 ft	30 to 50 ft
Ceiling height§	12-½ ft	8 to 10 ft

* Floors must be able to support loads of at least 150 lb/ft^2 throughout the entire structure and higher in the areas where certain battery plants and mechanical equipment will be located. To achieve the most economical design to carry these heavy live-loads, column spacing in NEBS buildings is 20 ft rather than the 30 to 50 ft commonly found in office buildings. The net effect is that floors and foundations are massive, and there are 2 to 3 times more columns than in conventional buildings.

† A criterion related to floor strength is the maximum deflection under load. For telephone structures, it must be less than ½ in. to avoid creating additional problems involved in leveling and aligning the equipment units. For conventional buildings, the lighter floor slabs may deflect 1 to 2-½ in. under design loads, depending on local building codes.

‡ For proper structural support and to simplify installation of equipment frameworks, floor levelness from high to low points must be within 1/4 in. in 8 ft, 3/4 in. in 20 ft, and 2 in. over the entire floor area. In conventional buildings, the floor-levelness requirement is less stringent than ±1/4 in. in 10 ft.

§ The height of equipment frames and the need for clear overhead space for cabling and process cooling-air distribution assemblies require higher floor-to-ceiling heights under all obstructions than usually provided in conventional office buildings.

nators, shown in Fig. 18, are provided with the necessary ac switch gear to transfer loads from the normal ac power system to the reserve system. The heavy diesels impose extreme loads on the structure and can be a source of vibration that can damage equipment; therefore, special protective structural design features are required. Additionally, support systems for fuel, air intake, and engine exhaust are necessary. A special fuel supply and storage system is provided that permits at least three days of central office operation during commercial power-failure emergencies. The air-intake systems are large, consistent with the high capacity of the emergency diesel engines and turbines, and the structure and equipment areas must be protected from the effects of the high-temperature exhaust with special insulating assemblies. Exhaust silencers must be used with diesel and turbine plants to eliminate much of the noise that would otherwise be radiated from the engines and that would be objectionable to the building occupants or neighbors.

(iii) Extensive electrical grounding systems are placed throughout each building for the purpose of eliminating noise on lines, reducing high-speed data errors, and protecting the telephone equipment from electrical short circuits and lightning strikes. The ground circuits connect all steel in the structure and in the equipment frameworks with a connection to the earth. Certain electronic equipment requires dedicated grounding arrange-

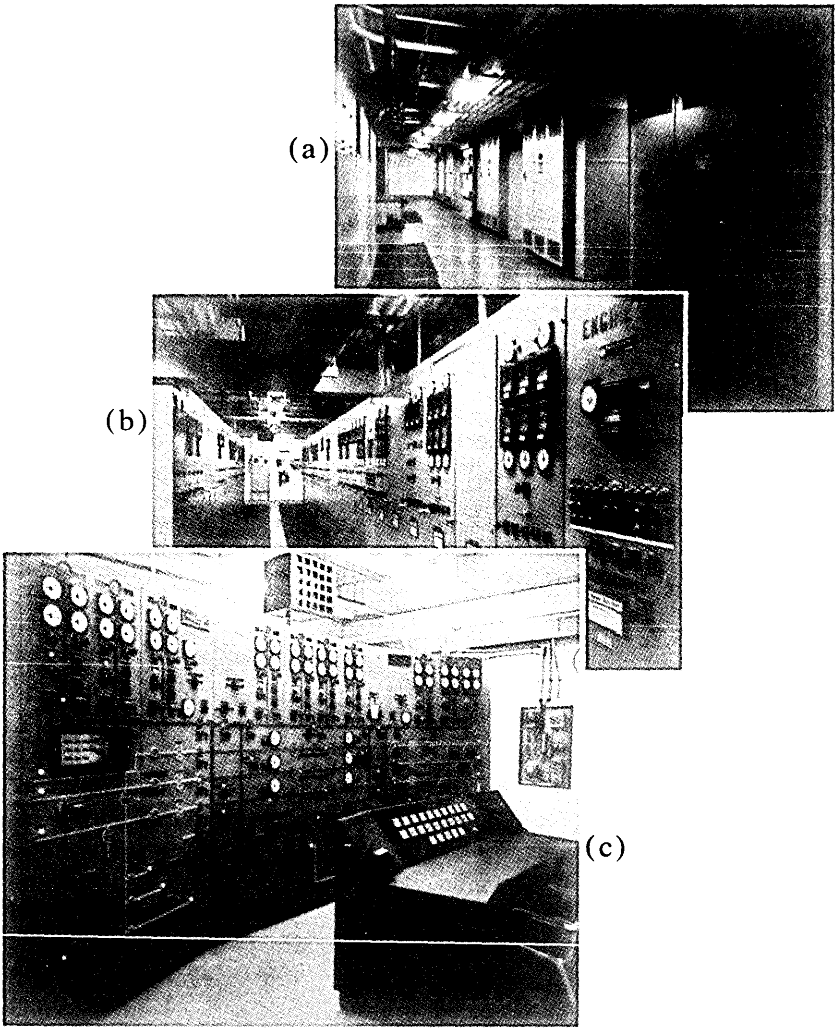


Fig. 17—AC service equipment in a large central office. (a) Stepdown power transformers reduce engine-alternator voltage to commercial power level. (b) AC switching gear transfers power from commercial to engine-alternator standby sources. (c) Mimic panel for remote control of ac equipment.

ments in addition to that provided by the building grounding system to avoid circuit malfunctions.

(*iv*) Shielding systems are provided in offices located near sources of high-intensity electromagnetic or electrical fields such as radio broadcasting stations, electrical power stations, or certain high-tension lines. Shielding prevents electromagnetic fields from penetrating and causing malfunction of electronic equipment. Figure 19 is an example of an internal shield. Wire mesh is embedded into the precast concrete wall and roof panels

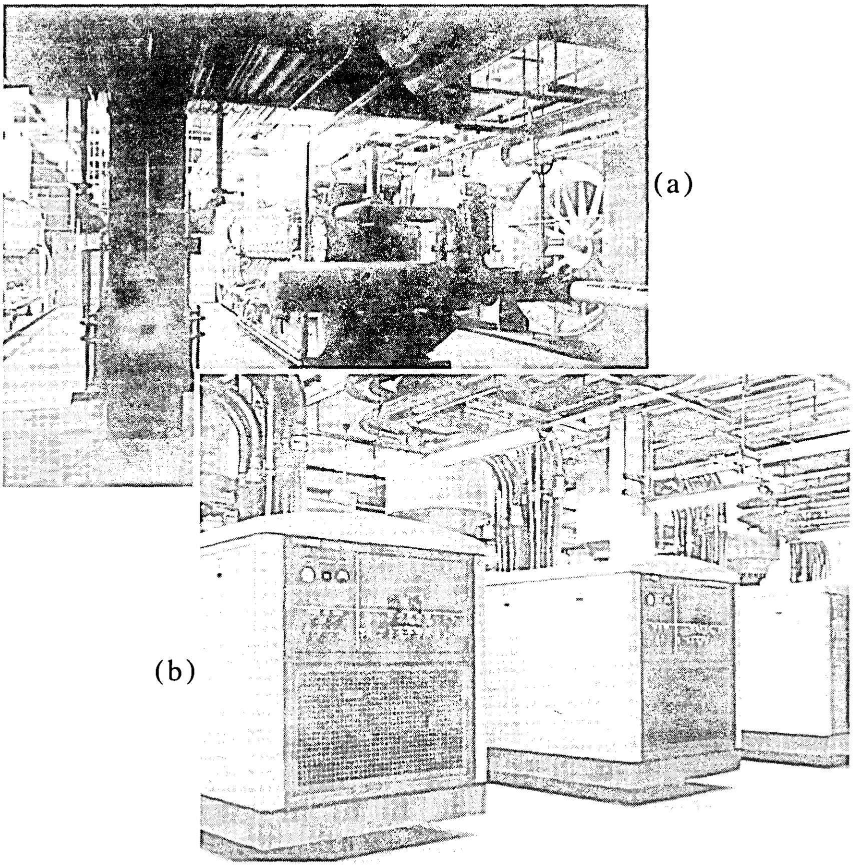


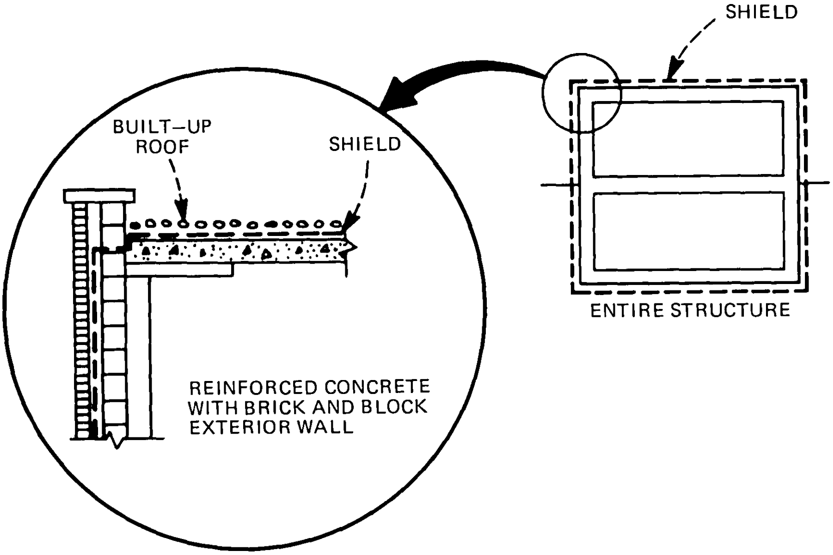
Fig. 18—Standby engine alternators. (a) 2500-kW diesels. (b) 750-kW turbines.

used in this type of construction. Shields are also obtained by welding the reinforcing bars of the structure to form a conducting cage, or by placing and joining copper or steel sheets either on the inside or outside of the structure to form a complete metal enclosure.

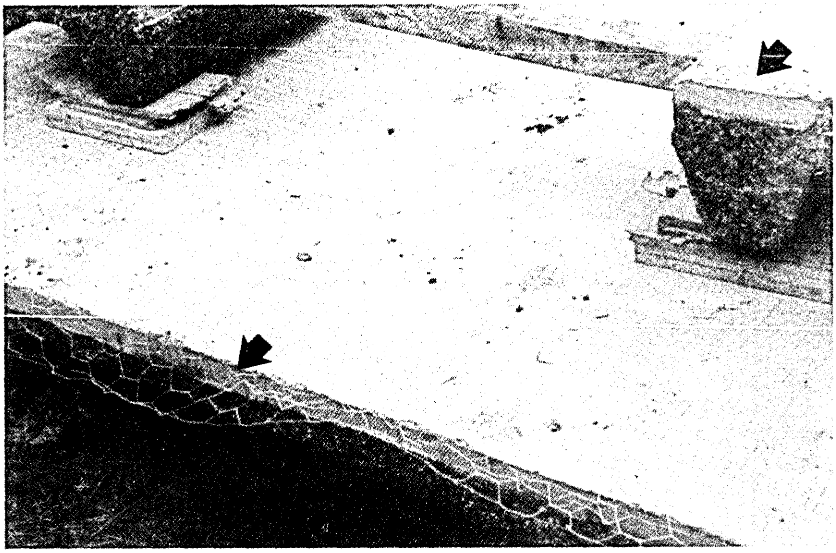
(*v*) Extensive detector, alarm and control systems, as shown in Fig. 20, are employed throughout the building for protection against circuit and service impairment due to fires. These are especially important since automatic water sprinkler systems cannot be placed in equipment rooms because of the great hazard from water to the electrically powered equipment.

6.4 Mechanical systems

Equipment buildings have mechanical systems to provide temperature control, to regulate the humidity of the surrounding air, to maintain appropriate amounts of outside air of high purity, to provide



(a)



(b)

Fig. 19—Construction methods for protection against radio frequency interference. (a) Building encased in continuous sheet-metal shield. (b) Panels precast with embedded galvanized mesh.

means for vertical access in multifloor structures, and to provide movement of air, water, and fuel. Mechanical equipment areas take from 5 to 25 percent of the gross space, depending upon the provisions for process cooling, humidification, and air filtration needed by the

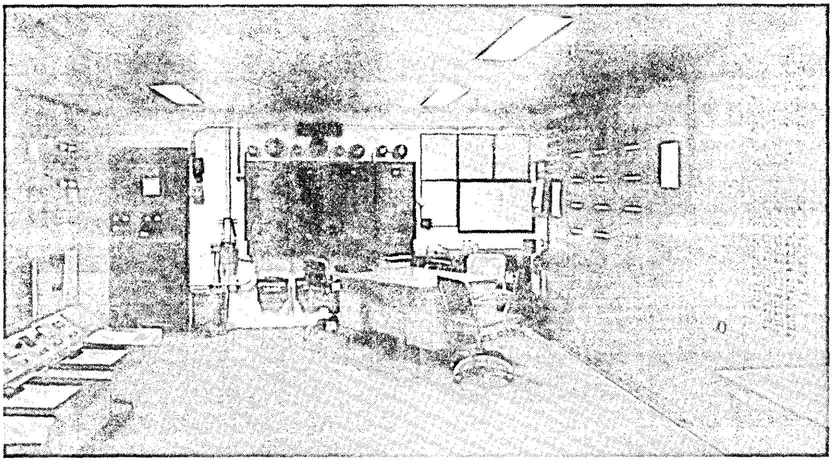


Fig. 20—Fire and mechanical equipment alarm panels. Left: Remote-control panel for equipment cooling systems. Rear: Flow alarms for fire extinguisher lines. Right: Smoke alarm panels and controls to operate dampers and evacuate smoke through ventilation system.

telephone equipment and reserve power engines. The variety of mechanical equipment systems are:

(*i*) Large refrigeration rooms are required in central offices because the cooling systems are designed to remove heat, released from telephone equipment, that can range up to 100 watts per square foot of occupied floor area, depending on the type of equipment installed. The range in heat released over one building bay of 400 square feet in local crossbar and electronic offices and for toll electronic offices is shown in Fig. 21. Recent telephone equipment developments have been aimed at miniaturizing components through the use of solid-state electronic devices. This close-packed equipment dissipates large amounts of heat and, therefore, requires exceptional amounts of cooling. For example, a 10,000-square-foot toll office will require up to 100 tons of cooling capacity when it houses electronic transmission and switching equipment. Conventional office buildings with equivalent floor area requiring cooling for human comfort would rarely be provided with more than 15 percent of this air-conditioning tonnage. Most equipment buildings have very large mechanical rooms close to the equipment areas to accommodate the high-capacity fans, chillers, chilled water and condenser water pumps, and chemical water-treatment tanks. Also, because of the high-heat dissipation, cooling is often required all year, even in northern regions, and naturally the heating plants in these structures are minimal. Figure 22 shows a portion of the chillers for the cooling system of the 323 Broadway, New York, office.

(*ii*) The ventilating fan systems are of sufficient capacity so that,

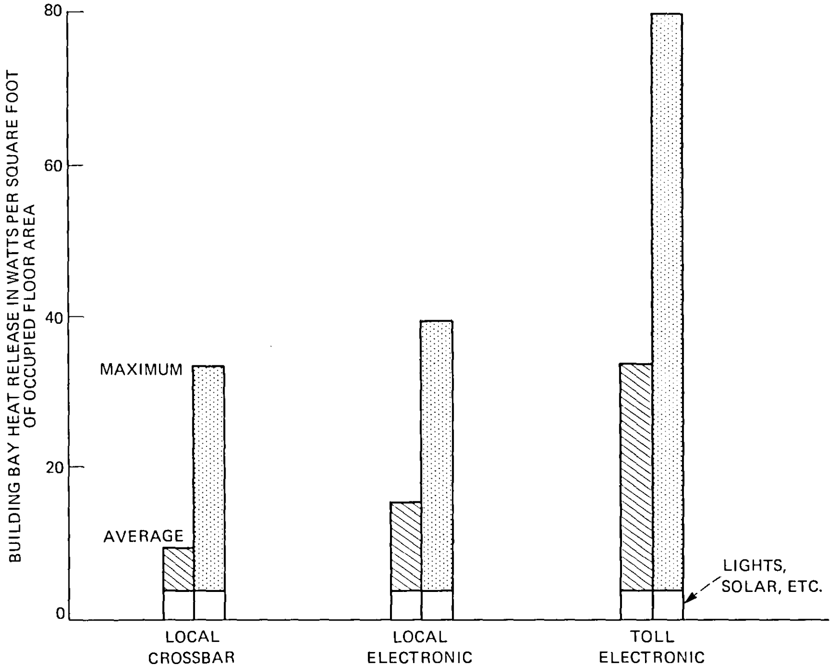


Fig. 21—Heat release in modern central offices.

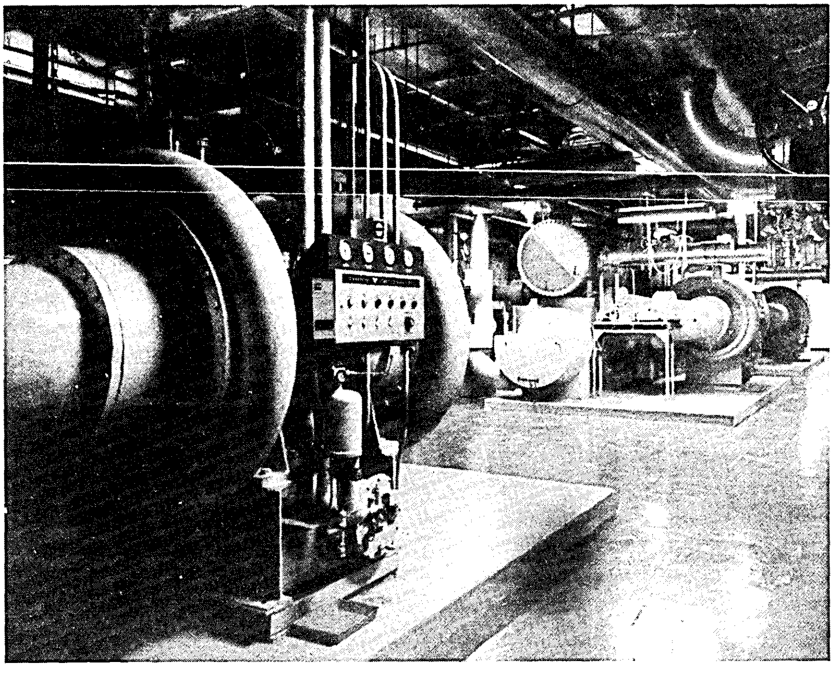


Fig. 22—Three 600-ton chillers for telephone equipment cooling.

in the event of a refrigeration-plant failure, short-term ambient operating environments for the telephone equipment can be maintained by the use of outside air while repairs are being made. The systems are characterized by elaborate air-ducting assemblies that have terminal diffusers aligned closely with the equipment lineups. The extensive overhead ducting in a new structure for this type of cooling system is shown in Fig. 23. Alternately, modular cooling systems comprised of a raised floor, plenum ceiling, and process coolers are employed to remove heat from the equipment room. In this type of system shown in Fig. 24, conditioned air is injected into the room through slots in the plenum ceiling, is heated by the equipment, and returns through slots in the raised floor to the process cooler fan-coil assemblies that remove the excess heat and return the air to the ceiling. Due to the quantity of air that is exchanged to remove the heat from the equipment space, large air-intake chambers, fans, plenums, and exhaust ducts are required. The air inlets shown in Fig. 25 illustrate the size of this air-handling equipment in a large facility. The fan room of a toll office and one of the machines used to pressurize the air distribution system are shown in Fig. 26. In addition to air for equipment cooling, special features are needed to provide for the large quantity of cooling and process air required by reserve engines and turbines that are also located within the building.

(*iii*) All the air required for process control is subject to quality standards to reduce the adverse effects of contaminants on the equipment and reserve power plants. High-capacity air filtra-

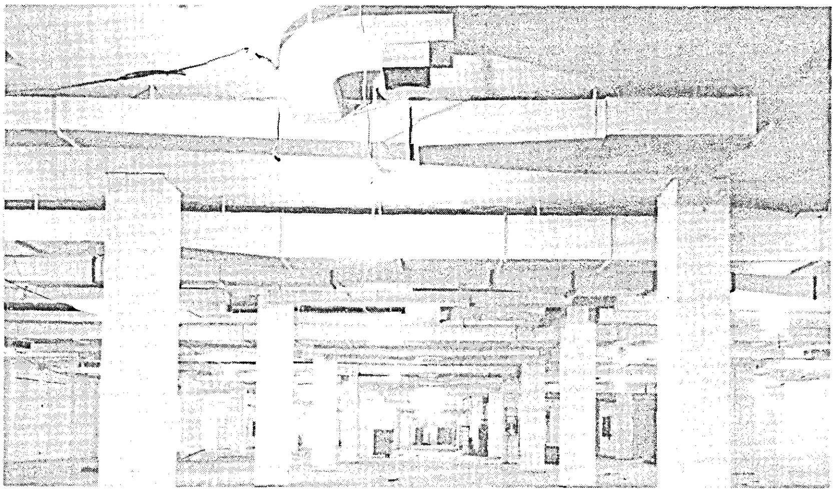


Fig. 23—Extensive overhead ducting to cool high-heat dissipating telephone equipment.

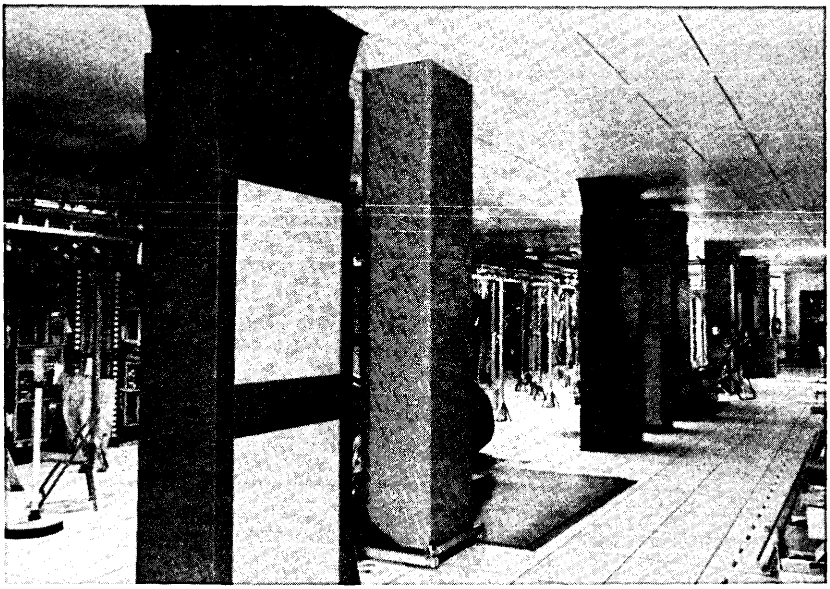


Fig. 24—Equipment room with a modular cooling system (raised floor, plenum ceiling, and process coolers in the column line).

tion systems, shown in Fig. 27, are necessary to prevent dust and products of combustion from infiltrating the building to cause electrical contact failures. Where extreme levels of air pollution occur, special high-efficiency filters are employed to remove potentially damaging material. Because of the volume of air that must be handled for ventilating the equipment rooms, large support frames are necessary for mounting the filters. These in turn require that significant space be provided in the basic structure of the facility to accommodate the air filtration plant.

(*iv*) Vertical access space within the structure is provided for the routing of all water and drain lines used to interconnect the elements of the mechanical systems that are located on different floors. Vertical access provision is also provided for the fuel lines and exhaust stacks of the engines of the reserve power plants that may be located in below-grade rooms, on intermediate floors, or on the roof of the central office. The location of the vertical runs is coordinated with the equipment plan so as not to interfere with the subsequent placement of future generations of equipment, while special provisions are made so that leaking fluids offer the least hazard to telephone equipment. Vertical access is also provided to permit the movement in the building and to upper floors of the large bulky equipment

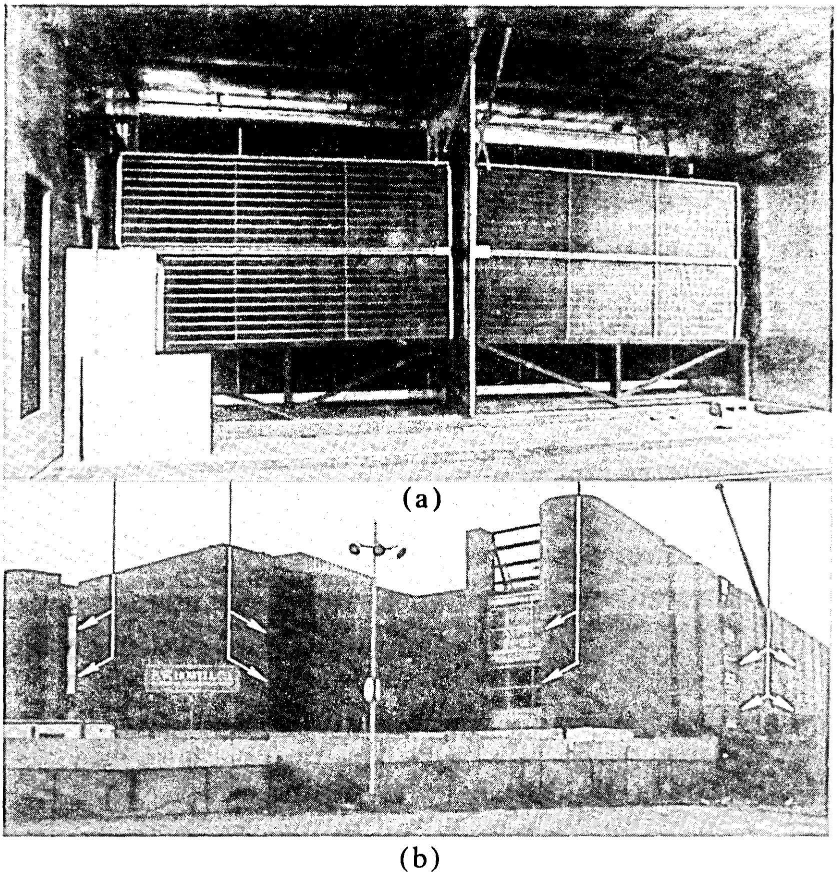


Fig. 25—Air inlets and exhausts for removing large amounts of heat released from equipment in modern central offices. (a) Interior of an inlet port. (b) Central office with 16 air inlet and exhaust chambers (10 visible).

assemblies that are added periodically to handle increases in demand for service. Large loading docks, freight elevators, or hoisting shaftsways are used to transfer and move the heavy loads, and dedicated open areas adjacent to equipment rooms are needed for the uncrating and erection of the equipment assemblies.

6.5 Special construction

Another very important characteristic of a central office or transmission station is the provision for expansion. If a horizontal addition is anticipated, the rear or a side wall must be designed for removal without interfering with the structural integrity of the roof and floors or with equipment assemblies that are operating to provide service. Also, extensions to the air distribution ducts and refrigeration machinery of the process cooling system must be accommodated. Virtually

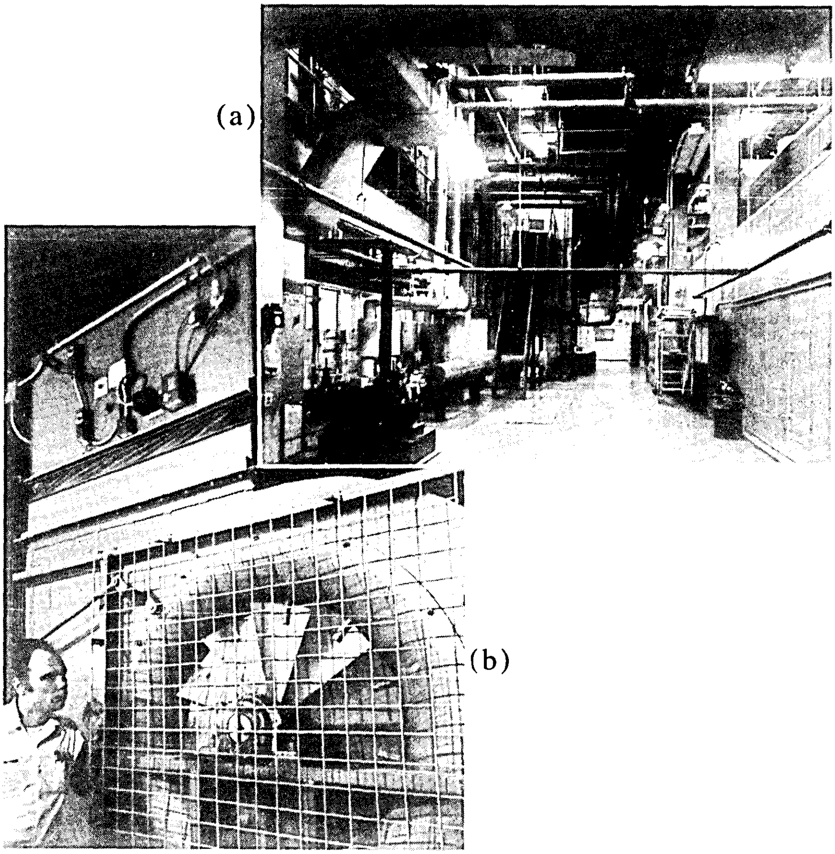


Fig. 26—Mechanical room in large central office for pressurizing chilled air to cool equipment. (a) Sheet-metal wall, portion of fan plenum that houses 6 fans. (b) One of the fans in the plenum.

every study plan for an office or station shows the planned growth directions. If vertical additions are anticipated, the footings, columns, and load-bearing walls must be adequate for the ultimate structure, and this will necessitate what appears to be greatly oversized initial construction as indicated in Fig. 28.

Special construction is also required where offices have roof-mounted microwave radio towers, as shown in Fig. 29. The typical means of transmission for many toll routes is by point-to-point microwave radio. This is particularly true when toll offices are located in metroplitan or other areas where installation of underground cables between offices is costly. Such towers must be capable of supporting several antennas and the composite assemblies weighing hundreds of tons. With the tower and antennas on top of the building, the load is carried through the building to the foundation of the structure and

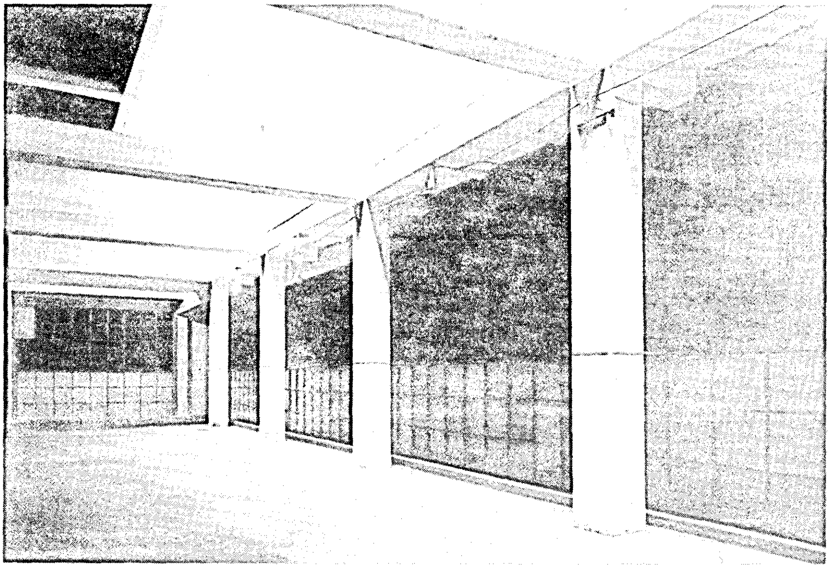


Fig. 27—High-efficiency filter banks for processing air for ventilation, for equipment cooling system make-up air, and for standby-engine air supply.

requires a massive internal support system. Additionally, vertical access for waveguide, power and personnel, and appropriate fascia must be provided until the time when the antenna support tower is enveloped by vertical additions to the wire center or transmission station.

VII. SUMMARY AND ACKNOWLEDGMENT

During the 100-year history of the telephone, three different sets of building design standards have been used in the Bell System. The earliest central offices were designed primarily for operator switchboards. In the mid-twenties, building standards were changed to accommodate tall equipment frameworks. These controlled the design of central offices and stations until the early 1970s, when the NEBS standards were adopted. During the later decades, advances in technology permitted the reassignment of most operators and craftspersons to locations remote from the switching and transmission equipment so that, today, the central office and station are designed primarily for equipment. In those few areas where operating support systems require the presence of craftspersons, the new standards permit building operating rooms to have very attractive interior design.

Modern equipment buildings have many special design features that set them apart from conventional buildings. The special features occur because of the need to provide a dedicated operational environment

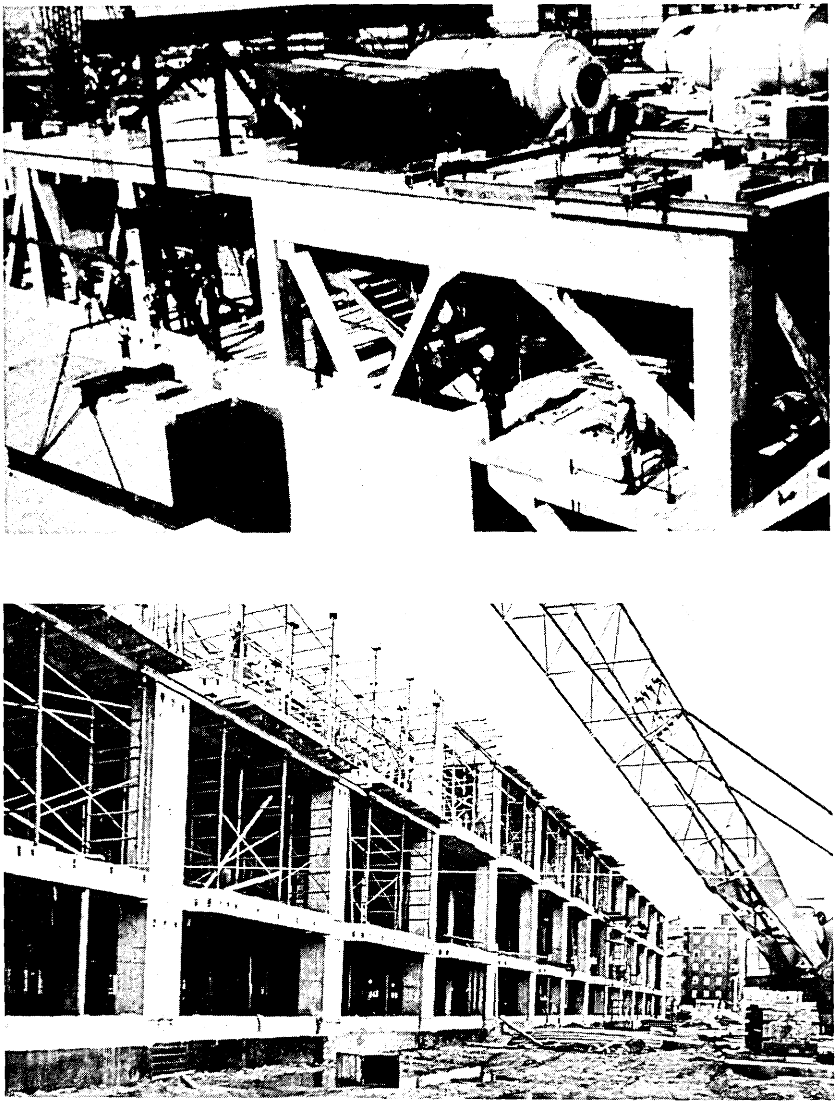


Fig. 28—Massive structural elements required for heavy equipment loads and future vertical expansion.

for the cable, wire, equipment, and apparatus that are assembled in the central office or transmission station. Once installed, the circuit elements and the mechanical, electrical, and structural elements function under the control of operations support systems as an equipment-building system that becomes an integral and vital part of the telephone network.

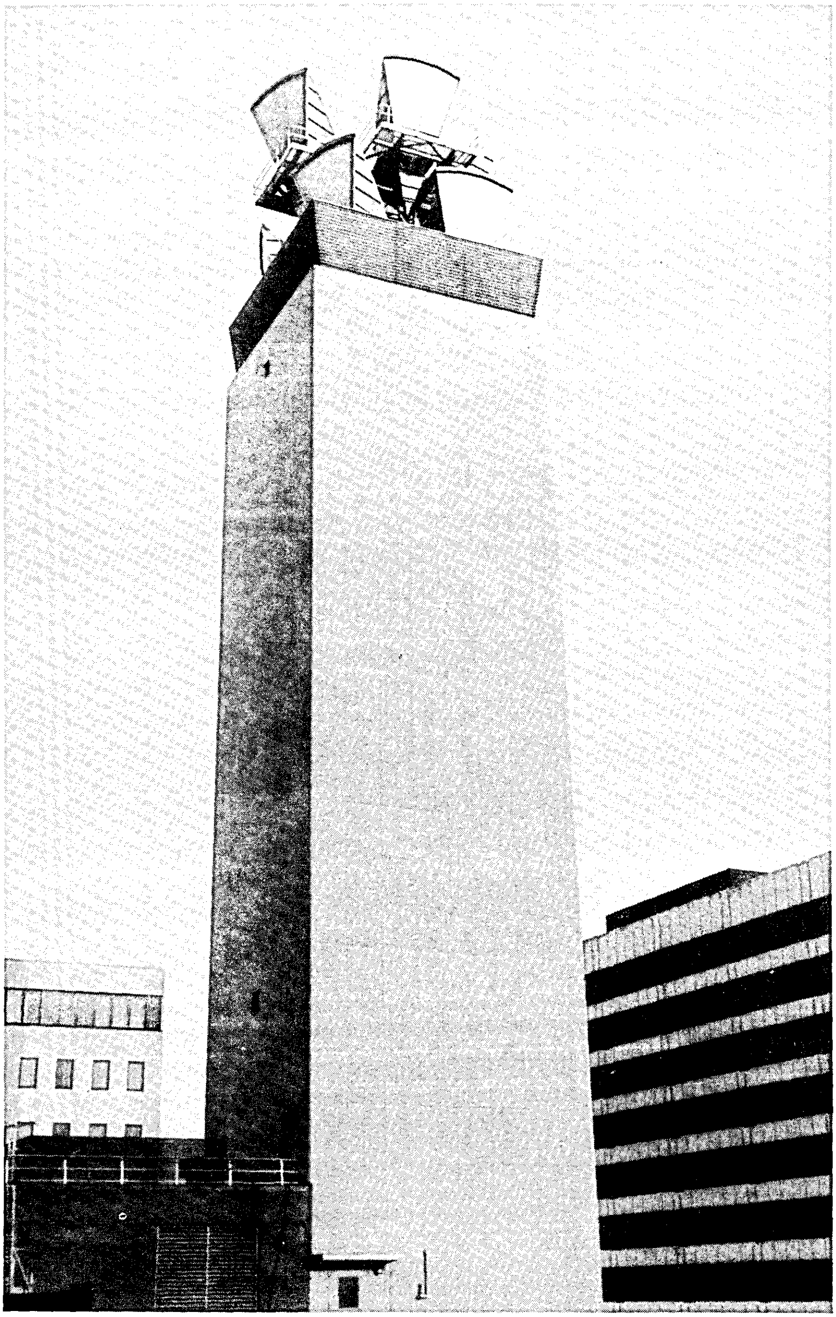


Fig. 29—Enclosed radio tower, with vertical access provisions and antenna waveguide systems, erected on an equipment building.

VIII. ACKNOWLEDGMENTS

While this paper has described in detail physical and environmental aspects, I wish to acknowledge the human dimension in the evolution of telephone equipment buildings. In this paper, I have had the opportunity to be the recorder of the work of many engineers and technicians who have worked in this field over many years. As chairman of the BTL Equipment-Building Task Force, I had the pleasure of working with many Bell Labs physical design engineers who are the originators and developers of the variety of equipment systems that are described in this paper. The development of the NEBS standards was in itself a ten-year program that owes much to the efforts of R. J. Skrabal and J. P. White, the members of their supervisory groups, the members of the Task Force, and the AT&T-WE-BTL NEBS Implementation Committee, chaired by D. M. Byrd. Many operating telephone company space planning and building engineers who are responsible for the planning, design, and construction have provided valuable information about their buildings. Finally, recognition is due L. W. Fagel for obtaining the many excellent photographs used to show the various equipment and for his study of the many special features of Bell System telephone equipment-building systems.

REFERENCES

1. *Engineering and Operations in the Bell System,* Bell Laboratories, 1977.
2. *Bell System Statistical Manual,* AT&T, Comptroller's—Accounting Division, May 1977.
3. "Housing the Telephone Exchange," Louisiana Purchase Exposition, St. Louis, 1904.
4. W. Pferd, "NEBS: Equipment Buildings of the Future," Bell Laboratories Record, December 1973.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 58, No. 2, February 1979
Printed in U.S.A.

Monolithic Electronic Devices Based on Domain Wall Motion in a Ferroelectric Crystal

By J. M. GEARY

(Manuscript received February 24, 1978)

Ferroelectric domain wall motion can provide the basis for monolithic electronic devices which are able to gate, amplify, perform digital logic, read out permanent analog messages, provide digital shift register storage, and scan optical images. No use of ferroelectric coercivity is made in these devices. Most of the functions have been experimentally demonstrated in simple form. The basic operating principles of these devices are explained, and their material requirements are discussed. The specific operation of five categories of devices is described.

I. INTRODUCTION

A variety of electronic functions can be accomplished in a novel way by taking advantage of the properties of domain wall motion in ferroelectric single crystals. Using lead germanate, an elementary gating/amplifying device and a simple analog read-only memory have been demonstrated.[1] Using gadolinium molybdate, a much more complicated analog read-only memory capable of storing four seconds of speech waveforms has been produced.[2] The theory of several other wall motion devices, including image scanners and logic gates, has been investigated. These latter devices have not been demonstrated but rely on the same principles as those which have. The purposes of this paper are (*i*) to summarize the basic concepts common to all the above devices, (*ii*) to discuss material problems and requirements as they relate to improving the devices already demonstrated and to realizing the undemonstrated ones, and (*iii*) to describe specific device configurations using presently available materials and to show configurations achievable with improved materials.

II. BASIC PRINCIPLES OF THE DEVICES

The devices discussed here employ ferroelectric crystals which permit only two anti-parallel polarization states. The crystals used are in the form of thin slabs with the polarization axis generally normal to the broad faces of the slab, as in Fig. 1. In this configuration, domain walls can be moved sideways by applying a potential between electrodes on opposite sides of the slab. Coherent wall motion of this type (as opposed to nucleation of new domains) will be assumed to be the only mode of domain switching to occur during device operation.

Depending on the type of ferroelectric material used, the domain's configuration and motion may either be arbitrary, as in Fig. 1, or rigidly anisotropic. A domain like the inner one in the figure (i.e., with downward directed P vector) will be termed a positive-favoring domain because it expands in response to a positive potential applied to the top of the crystal. Domains with upward directed P vector exhibit the opposite behavior and will be termed negative-favoring.

As the domain advances across the electrode, the sign of the spontaneous polarization, and hence the compensating conduction charge on the electrodes, must change. This requires a flow of current in the electrode leads. Presuming intimate electrode contact to the ferroelectric surface, the induced wall motion current is

$$I = 2P_s \frac{dS}{dt}, \tag{1}$$

where P_s is spontaneous polarization and S is the area of the positive-favoring domain lying under the electrode.

The monolithic electronic devices described here function by employing two ferroelectric effects: wall motion and induced currents.

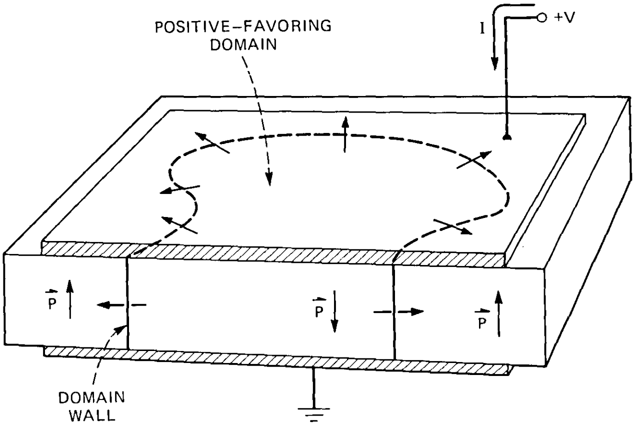


Fig. 1—Ferroelectric single crystal shown in perspective and cross section. Voltages applied to electrodes on opposite faces of the crystal slab produce an electric field parallel to the ferroelectric polarization axis. The wall of the arbitrarily shaped domain propagates outward under the influence of this field. As the wall moves, it induces a current I in the leads of the electrodes. Since the domain of the polarity shown expands in response to the positive voltage applied to the upper electrode, it is called a positive-favoring domain.

When suitable voltages are applied to appropriate configurations of electrodes, domains can be gated, shifted sequentially from place to place, made to perform scanning motions, or kept isolated from each other. Currents induced by wall motion can be used directly as outputs, digital or analog. Since the domains need no power to maintain their existence, the devices can have nonvolatile properties. Most importantly, currents due to wall motion can be fed to other parts of the crystal to control further wall motion. The specific configurations of a number of devices are presented in Section IV.

Superficially, such devices might appear to be the ferroelectric analogs of magnetic bubble devices, but in fact they differ in two important ways. The more important difference derives from the incompleteness of the analogy between electrical and magnetic phenomena. Electric field lines can terminate on electrical charges. It is the rearrangement of the terminating charges that accompanies wall motion which gives rise to the induced ferroelectric current. Since no analogous magnetic charge exists, this phenomenon has no direct analog in bubble devices.

A second distinction is that magnetic bubbles are ferromagnetic domains of a special character, ones which represent a minimum energy configuration created by the presence of an opposing field. The ferroelectric materials used, however, do not exhibit analogous "electric bubbles." Instead, they support conventional domains whose dimensions are determined by the configuration of the electrodes.

Past efforts to apply ferroelectric phenomena to electronic devices have been based on the concept of ferroelectric coercivity. These devices used the coercive strength as a threshold and were in many ways analogous to magnetic square-loop core devices such as ferrite core memories or magnetic amplifiers. Unfortunately, coercivity does not appear to be a fundamental property of ferroelectrics. Despite much experimentation, true coercivity is not observed in ferroelectric materials.[3] Ferroelectric wall motion devices, on the other hand, do not utilize threshold principles. While they have their own material requirements and constraints, they employ only properties believed to be fundamental to ferroelectricity.

III. DEVICE REALIZATION

3.1 Materials

The ideal ferroelectric material for use in wall motion devices should have a Curie temperature well above room temperature, be mechanically stable, and possess only two domain states (preferably optically distinguishable, as a diagnostic). It must switch only by rapid, consistent wall motion, not by nucleation of new domains. It is desirable that the material permit domains of arbitrary shape (as in Fig. 1).

The central materials problem of the wall motion devices is that no single material has yet been found that has all these ideal properties. The working experimental devices demonstrated to date[1,2] have been made using either gadolinium molybdate or lead germanate. Gadolinium molybdate satisfies every property above except that its domains can only assume a restricted planar shape. Lead germanate satisfies all conditions except that, in all samples tested to date, it permits nucleation to occur at all but very low fields. The properties and applicability of these materials are summarized below.

Gadolinium molybdate,[4] $Gd_2(MoO_4)_3$, is orthorhombic below its Curie temperature of 159°C and has a room temperature spontaneous polarization of 0.17 $\mu C/cm^2$. It is ferroelastic as well as ferroelectric. When a wall sweeps across the crystal, the *a* and *b* axes (which differ in lattice spacing by about 0.3 percent) are exchanged. Thus the change of electrical polarization is accompanied by a slight change in the shape of the crystal. Since the *a* and *b* axes are optically dissimilar, the domains may be easily viewed by polarized light microscopy.

To make the slightly different strain of adjacent domains compatible, the domain walls must take the form of planes stretching across the whole crystal and must be oriented at a 45-degree angle to the *a* and *b* axes. Strain compatibility also requires that the walls do not intersect. These constraints mean that the domains must stack up like parallel plates across the crystal slab. Hence, one can at best deal only with a one-dimensional array of domains.

On the other hand, it is probable that these same strain compatibility effects are also responsible for making gadolinium molybdate so resistant to the nucleation of new domains. Walls can be moved back and forth across the crystal at velocities of several m/s without danger of nucleation.[5] This is true in spite of polishing damage, saw-cut edges, inclusions, or scratches.

Recent electron microscopic investigations of wall in gadolinium molybdate indicate that, for very thin samples (around 1000Å), strain compatibility effects may no longer restrict domain shape.[6–8] The micrographs actually show numerous closed domains with right-angle bends in the walls. Under these circumstances, gadolinium molybdate would appear to satisfy every condition for an ideal material for application to the devices proposed. Neither device fabrication on samples of this type nor investigation of the possibility of closed domains in thicker samples has yet been tried.

Lead germanate,[9,10] $Pb_5Ge_3O_{11}$, is a trigonal ferroelectric below its Curie temperature of 176°C and has a room temperature spontaneous polarization of 4.6 $\mu C/cm^2$. The two permitted domain states are enantiomorphic and because of their optical activity can be observed by polarized light microscopy. Domain shape in lead germanate is observed to be quite arbitrary, and the demonstrated devices employing it have made use of this property. A small wall velocity anistropy exists, however, imparting a rounded hexagonal shape to domains which grow outward from a point and are unrestricted by electrode geometry.

The ease with which new domains nucleate in lead germanate is the primary drawback to the use of this substance in practical wall motion devices. Theoretical studies of ideal ferroelectrics show that a very large increment of energy, on the order of 10^8 kT for barium titanate, for example, is necessary to produce the nucleus of a new domain capable of further expansion.[11] Similar studies of wall motion in barium titanate[12] show that the minimum energy to initiate wall motion is on the order of 10 kT. This large ratio of nucleation and wall motion thresholds should be typical of ferroelectrics. Consequently, there appears to be no fundamental reason why a wall motion device cannot operate without causing nucleation of new domains. Unfortunately, presently available samples of lead germanate do not even approach these limits of nucleation immunity. It is hypothesized that nucleation may occur with anomalous ease due to crystal flaws, flaws in electrode coatings, or tiny unswitched domains left behind by a propagating wall.

3.2 Substrate fabrication

The devices described here require a single crystal substrate with a thickness on the order of a few microns to a few tens of microns. Both sides of the substrate must be accessible for electroding, though only one side carries patterned electrodes in most of the devices. This paper is basically concerned with what can be done with such substrates, rather than how than can be fabricated. However, we can list some techniques which could be employed.

(*i*) Hetero-epitaxy on a crystal substrate which is then selectively dissolved away so that the back surface can be electroded.

(*ii*) Hetero-epitaxy on a substrate crystal which has sufficient electrical conductivity to serve as a back-surface electrode.

(*iii*) Careful etching of saw-cut crystal wafers. Techniques of this kind[13] when applied to silicon can produce a crystal as thin as around 25 μm with a thicker rim for support.

(*iv*) Conventional polishing of saw-cut wafers. Reduction of thickness to a few mils can be practical.

3.3 The use of a resistive layer

In nearly all the devices described in this paper, domain walls are required to propagate across gaps separating adjacent electrodes. The surface polarization charge of the ferroelectric is always fully compensated for by conduction charge where an electrode is present, but the polarization charge in the interelectrode gaps has no direct source of compensating charge. Thus, when a wall attempts to cross a gap, a strong depolarizing field will be set up (see Fig. 2a). This field can hinder wall motion or stop it completely.

A layer of resistive material in intimate contact with the crystal surface can alleviate this effect by providing a source of compensating charge as shown in Fig. 2b. The layer should be conductive enough to allow rapid wall motion, but not so conductive that it causes needless power dissipation when the electrodes are at differing potentials or causes excessive loss of currents intended as outputs. The layer may be deposited before the electrodes are formed as in the figure, or it can be deposited into the gaps after the patterning of the electrodes. It may also be possible to produce a resistive layer by doping the surface of the ferroelectric crystal itself.

Experience with devices made from lead germanate shows that the use of a resistive layer is necessary to the crossing of electrode gaps by domain walls in this substance. This held true for gaps of all the widths

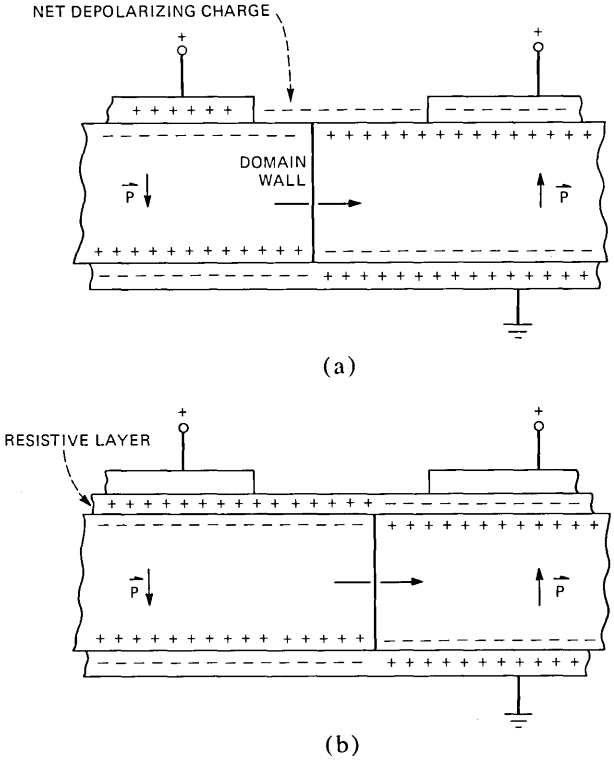


Fig. 2—(a) Cross-section view showing a domain wall attempting to cross an inter-electrode gap. The surface polarization charge of the ferroelectric changes as the wall advances, but the surface conduction charge remains fixed because the ferroelectric is a near-insulator. This creates a net depolarizing charge which can slow or stop the wall. (b) With a resistive layer in contact with the ferroelectric surface, charge compensation is maintained and the wall can cross the gap easily.

employed, including the narrowest of about 5 μm. In gadolinium molybdate, however, gaps as large as 10 μm can be crossed by the wall, on the condition that the gap extends along only a relatively small part of the length of the wall, the remaining wall length residing beneath unbroken electrodes.

IV. SPECIFIC DEVICE CONFIGURATIONS

In the remainder of this paper, the specific configurations of five device types are presented. The descriptions center on the more general versions made possible by an ideal material permitting arbitrary domain shapes. Specialization of these devices to planar domain geometry (i.e., gadolinium molybdate) are made where appropriate.

4.1 Gating device

This device permits one current to control another and is capable of amplification. Its operation illustrates many principles employed in the other devices. An experimental gating device has been demonstrated in lead germanate.[1]

The basic wall-motion gating principle is reviewed in Figs. 3a to 3d. A positive-favoring domain, permanently residing under an origin electrode, can propagate under the detector electrode if the intervening trigger electrode is positive. If the trigger is kept negative, the detector's domain state cannot change from negative-favoring to positive-favoring (in spite of a positive detector voltage) because no positive-favoring domain is available to the detector electrode. A negative confiner electrode is employed to ensure that the trigger provides the sole access of positive-favoring domain to the detector. Since the motion of a domain wall under the detector induces a current in the detector lead, the detector current is therefore gated by the trigger voltage. The device must operate in a pulsed mode: detector current ceases when the wall traverses the detector, and both trigger and detector voltage must be made negative to restore the device for a new gating cycle. The crossing of the interelectrode gaps is aided by a resistive layer if necessary. A similar device can be made using a planar-wall material by simply deleting the confiner electrode.

It can be shown that the ferroelectric gating device is capable of gain of various sorts. First note that the trigger draws a current of its own when traversed by the domain wall. By integrating eq. (1), it is clear that the total charge in the current pulses of both detector and trigger is proportional to the electrode areas. Hence by making the detector arbitrarily larger than the trigger, any desired charge gain can be obtained:

$$G_{\text{charge}} = \frac{S_d}{S_t}, \qquad (2)$$

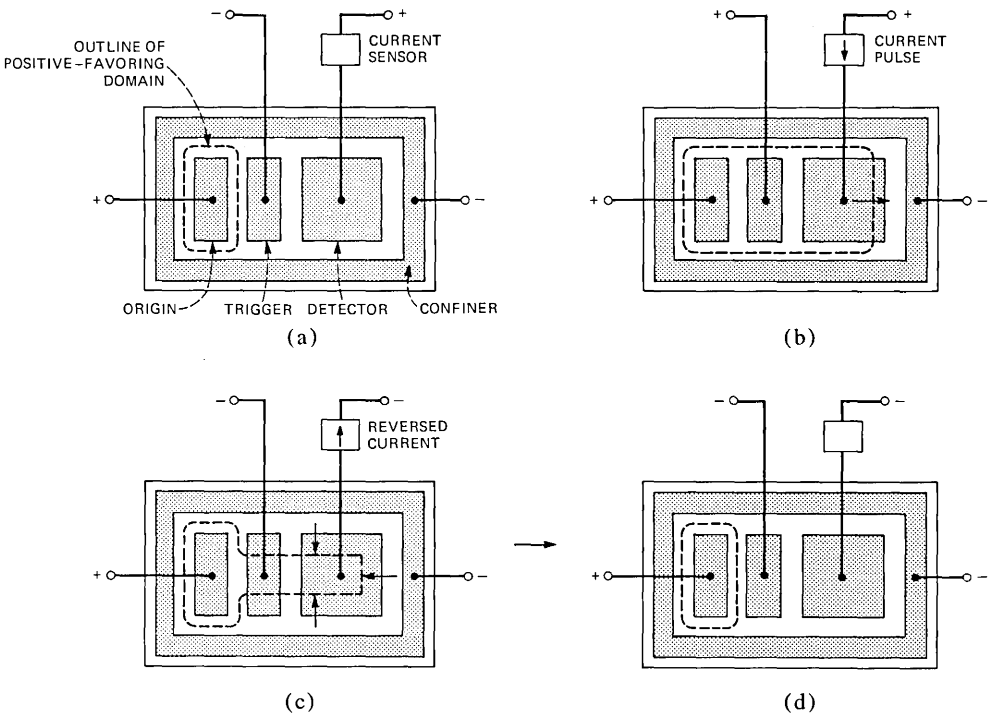


Fig. 3—Top view illustrating the operation of a ferroelectric gating device capable of gain. The trigger electrode serves to control the detector electrode's access to the positive-favoring domain beneath the origin electrode. (a) When a negative trigger voltage blocks this domain, the detector domain cannot change polarity because no domain wall is present and because nucleation cannot occur. Therefore, no current is sensed. (b) When the trigger is positive, a wall motion current is induced in the current sensor. By making trigger and detector supplies negative, the device is restored to its original state (c and d).

where S_d and S_t are detector and trigger areas. For a crystal of a given thickness, wall velocity is a function of applied voltage, $V(v)$. By applying a greater voltage to the detector than to the trigger, the detector wall velocity can be made larger than that of the trigger, yielding a current gain. Assuming the width of the detector, trigger, and origin to be identical (as in Fig. 3),

$$G_{\text{current}} = \frac{V(v_d)}{V(v_t)}, \tag{3}$$

where v_d and v_t are the detector and trigger voltages.

Suppose a resistor is placed in the detector lead to yield an output voltage, and suppose the trigger voltage is taken as the input voltage. To analyze this case in general requires the trial-and-error solution of a functional equation involving $V(v)$, an experimental curve. We will therefore consider a convenient special case in which v_{ds}, the detector supply voltage, is set so that wall velocity is the same under trigger and detector. For this case,

$$G_{\text{voltage}} = \frac{v_{ds}}{v_t} - 1 \tag{4}$$

so that gain greater than 1 is achieved when $v_{ds} > 2v_t$. The value of v_{ds} required to make the trigger and detector wall velocities equal is given by

$$v_{ds} = v_t + 2P_s WRV(v_t), \tag{5}$$

where W is the width of origin, trigger, and detector and R is the load resistance. This implies that voltage gain greater than 1 results when

$$2P_s WR > \frac{v_t}{V(v_t)}. \tag{6}$$

Hence, voltage gain exceeding 1 is always attainable simply by designing the gating device to have W and R satisfying eq. (6).

The conditions of equal wall velocity and equal electrode width imply that trigger current and detector current are also equal. If we define instantaneous power gain as the ratio of power dissipated in resistor R to the power input to the trigger lead, then

$$G_{\text{power}} = \frac{v_{ds}}{v_t} - 1 \tag{7}$$

for conditions of eq. (5). Instantaneous power gain will exceed 1 when the device is designed so that eq. (6) is also satisfied. Finally, we may define energy gain as the ratio of total energy dissipated in R to the total energy input into the trigger lead. Since we are analyzing the special case of equal trigger and detector velocities, the duration of input and output power pulses will have the same ratio as the lengths of trigger and detector electrodes measured in the direction of wall motion, ℓ_t and ℓ_d. Hence,

$$G_{\text{energy}} = \frac{\ell_d}{\ell_t}\left(\frac{v_{ds}}{v_t} - 1\right) \tag{8}$$

for conditions of eq. (5). Energy gain will exceed 1 when, for instance, the design fulfills eq. (6) and $\ell_d > \ell_t$.

The behavior of the gating device has been analyzed without taking into account the interelectrode current leakage which will result if a resistive layer is employed. This leakage can be reduced by making appropriate compromises in the design of the resistive layer. Leakage can be avoided by placing the output device (current sensor or load resistor) in the lead of the underside grounded electrode. The output will then be the trigger's plus the detector's current, yielding a greater gain. Alternatively, the detector electrode may be surrounded by a guard ring electrode connected to the detector supply, thus blocking leakage to the detector.

4.2 Analog read-only memories

The ferroelectric analog read-only memory provides a function not provided by any other monolithic technique: the direct readout of permanently stored continuous analog waveforms. A simple analog readout device made using lead germanate has been demonstrated.[1] Application to speech waveforms is especially appropriate. A gadolinium molybdate device which can read out either of two 2-second sentences has been demonstrated.[2]

Ferroelectric analog readout devices in general operate by allowing a domain wall to scan down a long detector whose shape varies in accordance with the intended output. A multiple track version of the device is shown in Fig. 4, as it would be configured for use with a material supporting arbitrarily shaped domains. To read out a track, the wall of the origin domain is gated by the trigger and is allowed to propagate down the detector. The wall is restricted to follow the outline of the detector by the negative confiner. The detector's width is a function *W(x)*, where *x* is the distance down the long axis of the detector. If the advancing wall has the form of a straight line oriented

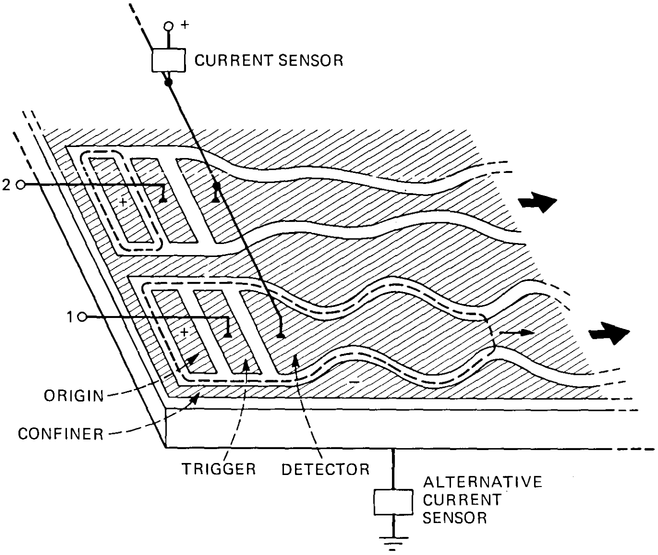


Fig. 4—Multiple-track analog readout device using one detector per track and employing arbitrarily shaped domains. Separate triggers are used to select the desired track. The output detected by the common current sensor is proportional to the width of the wall moving down the selected track.

perpendicular to its direction of advance, then the wall will sweep out surface area of the crystal at a rate of

$$\frac{dS}{dt} = v_0 W(x_0) = v_0 W(v_0 t), \tag{9}$$

where v_0 is the constant wall velocity and x_0 is the wall position. By eq. (1), this will cause a current to flow both in the detector lead and in the lead of the underside ground electrode:

$$I = 2P_s v_0 W(v_0 t). \tag{10}$$

Hence, the time variation of the output will be directly proportional to the spatial variation of the selected detector's width. If the thickness of the crystal does not vary, constant velocity can be maintained simply by applying a fixed voltage to the detector. Otherwise, a feedback circuit is employed to control the detector supply voltage in such a way as to maintain the time-averaged detector current at a desired fixed value.

Analog read-only memories with data tracks oriented parallel to each other may be made using planar wall materials such as gadolinium molybdate.[2] A differential, multiple-track version of such a device is seen in Fig. 5. The total width of the two detectors which constitute a

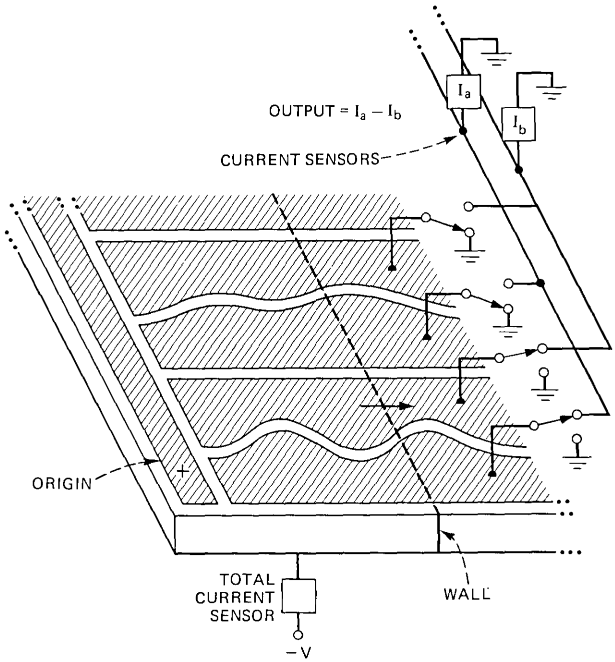


Fig. 5—Multiple-track differential analog readout device utilizing a planar wall material. External switches are used to connect the desired track to the current sensors. All unselected tracks are switched to ground.

track is constant, but the difference of their widths is equal to the desired function $W(x)$. The output may therefore be obtained as the difference of the two detector currents:

$$I_2 - I_1 = 2P_s v_0 W(v_0 t). \tag{11}$$

The sum of the two currents is directly proportional to wall velocity and can be conveniently utilized by a feedback system to regulate wall velocity.

Because the wall extends across the entire crystal, all tracks are necessarily scanned at once, and no common connection to all detectors can be used. Separate leads are therefore brought out, two per track in the differential version. The current in the lead of the desired track is selected by external circuitry, and all other leads are connected to a common potential.

4.3 Optical image scanners

In this section, optical area and line scanners are described which employ ferroelectric domain wall motion as the scanning mechanism. Though such devices have not yet been experimentally demonstrated, their operation depends on the same ferroelectric phenomena as the devices which have been demonstrated.

Figure 6 shows the structure of an optical area scanner employing a material permitting arbitrary domain shapes. The electrode geometry consists of numerous scanning electrodes interleaved with confiners. All the scanning electrodes are connected to one common terminal, and the confiners are connected to another common terminal. A ferroelectric shift register (see Section 4.4) lying along the left-hand end of the scanned area transports a single positive-favoring origin domain from one detector electrode to the next. When the detectors are made positive, the origin domain in the shift register crosses over to the adjacent scanning electrode and propagates down it. After thc domain reaches the end, the common lead is made negative, rapidly restoring the scanning electrode by a motion of the walls inward toward the axis of the scanning electrode. Repeating this process for each scanning electrode in sequence generates a domain wall scan of the whole area of the chip.

The scanning action of the domains is used to sense resistivity changes in a photoconductive layer applied to the bottom surface of the ferroelectric crystal. Figure 7 shows this layer in a cross-section view taken through one scanning electrode. Electrical contact is made to the surface of the photoconductor by a thin transparent electrode. The image to be scanned is focused on the photoconductor from below by an optical system. Induced ferroelectric currents originate specifically from the line where the moving domain wall intersects the surface of the crystal, since that is where the polarization surface charge changes its sign as the wall moves. Thus a scanning domain wall acts

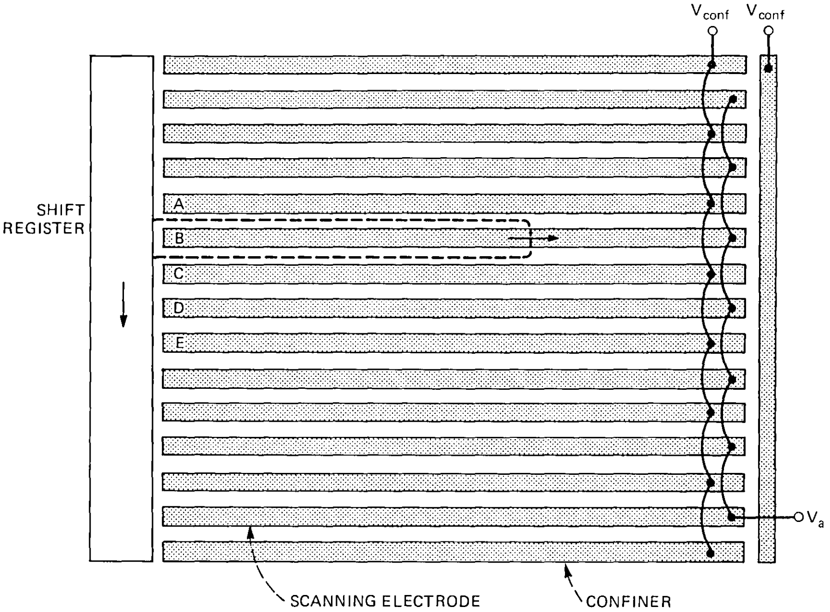


Fig. 6—Top view of electrode geometry used for area image scanner. A single positive-favoring domain held in the shift register at the left is used to initiate domain wall scans along the strip electrodes. The scanning domain is confined to one electrode by the interleaved confiner strips. When the scanning of one electrode is completed, it is restored by a brief negative pulse and the shift register is incremented to the next scanning electrode.

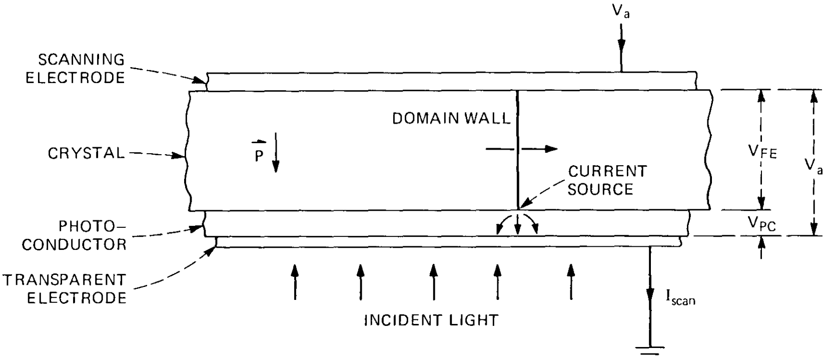


Fig. 7—Detection principle of the image scanner shown in cross section. The scanning wall acts as a moving current source. This localized current passes through a photoconductive layer whose conductivity is dependent on the intensity of the local illumination. The value of applied voltage V_a required to keep the scanning rate constant (i.e., to keep V_{FE} constant) varies in accordance with the conductivity of the photoconductive layer.

as a moving current source. The current from this source passes downward through the photoconductive layer and exits through the transparent electrode. If we apply a fixed voltage to the scanning electrode, the wall will not travel at a constant velocity, but rather will speed up when passing places where the photoconductor has high conductivity and slow down when passing places of low conductivity. The current exiting through the transparent electrode, I_{scan}, is directly proportional to the scanning wall velocity. Under these conditions, the ferroelectric scanning system yields an external current indicative of the illumination present at the position of the advancing wall.

The illumination data conveyed by I_{scan} are obtained at the price of a varying scanning rate. By regulating the wall velocity through a simple feedback system, a constant scanning rate can be obtained while still generating a signal indicative of the conductivity of the photoconductive layer. A block diagram of this system is seen in Fig. 8. A differential amplifier senses the difference between I_{scan} and a fixed reference signal. The amplifier's output is applied directly to the common scanning electrode lead. The applied voltage, V_a, is thus varied in just such a way as to keep I_{scan} equal to a desired constant. V_a can be viewed as a sum of two components: V_{FE}, the drop across the ferroelectric crystal, and V_{PC}, the drop across the photoconductive layer:

$$V_a = V_{FE} + V_{PC}. \tag{12}$$

Because wall velocity is held constant, it follows that V_{FE} is constant (assuming the ferroelectric crystal to be of constant thickness). This is so because, at a given crystal thickness, wall velocity is solely dependent on applied voltage. Consequently, any fluctuation of V_a must be due entirely to the fluctuation of V_{PC}:

$$\Delta V_a = \Delta V_{PC}. \tag{13}$$

Since I_{scan} is held constant, ΔV_{PC} will be proportional to $\Delta \rho_{PC}$, the resistivity of the photoconductive layer at the point of scanning. Thus

$$\Delta V_a \propto \Delta \rho_{PC}. \tag{14}$$

That is, the fluctuation of the feedback voltage itself constitutes the light signal output of the image scanner.

In the scanning electrode geometry shown in Fig. 6, it can be seen that the scanning electrodes and confiners are interchangeable. By alternately switching the role played by the two sets of electrodes, a doubling of the scanning density can be obtained.

A striking feature of the scanning electrode geometry is the fact that it is free of structure in the direction of scanning. Present silicon image scanners, by way of comparison, consist of two-dimensional arrays of individual light sensors. Each sensor has a structure of its own, that is, a configuration of diffusions, windows and metalized areas. In the ferroelectric scanner, each single scanning electrode performs the function of an entire row of structured light sensors. Figure 9 shows the "resolution cell," or picture element of a ferroelectric scanner.

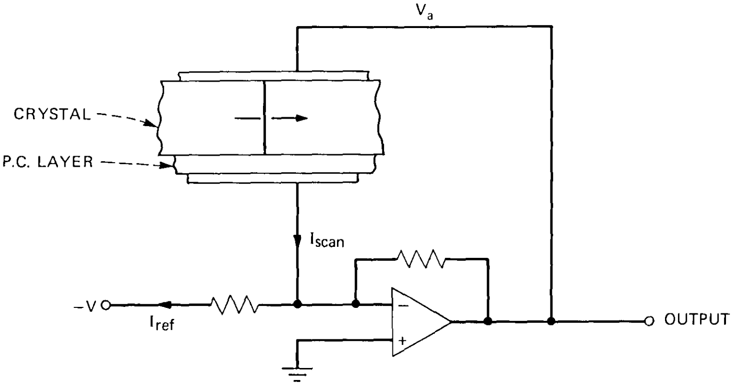


Fig. 8—Feedback circuit used to maintain constant wall velocity in the image scanner. The wall current I_{scan}, which is proportional to wall velocity, is compared with a reference current I_{ref} so as to generate a feedback signal V_a. This signal is used to drive the scanning electrodes and constitutes the output signal as well.

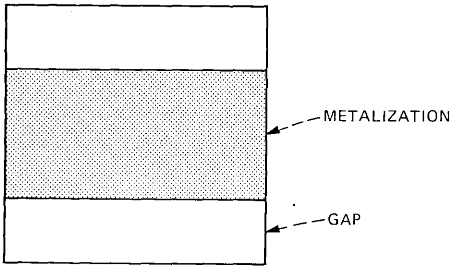


Fig. 9—Resolution cell of the area scanner for the case where scanning and confiner electrodes are exchanged on alternate scans. The cell is notable for its simplicity.

Though the resolution of the scanner in the horizontal direction is greater than in the vertical direction, the cell has been drawn conservatively as a square. The cell consists of a metal strip with two half gaps to either side. It is hard to imagine a solid-state area scanner with a simpler resolution cell than this.

A color scanner can be made by overlapping the photoconductive layer with an array of narrow parallel red, blue, and green filters oriented perpendicular to the direction of scanning (see Fig. 10). An output will be produced which represents the detected amounts of the different color components in a rapidly repeating sequence. External circuitry employing a phase-locked loop is used to sort out the different color indications into three separate parallel outputs. A narrow transparent strip interposed between each group of color filters provides a phase reference for the separation of the three color signals. Such a scheme involves a substantial reduction of resolution in the direction of scanning, but since the resolution of ferroelectric scanners should be high to begin with, the compromise is especially appropriate.

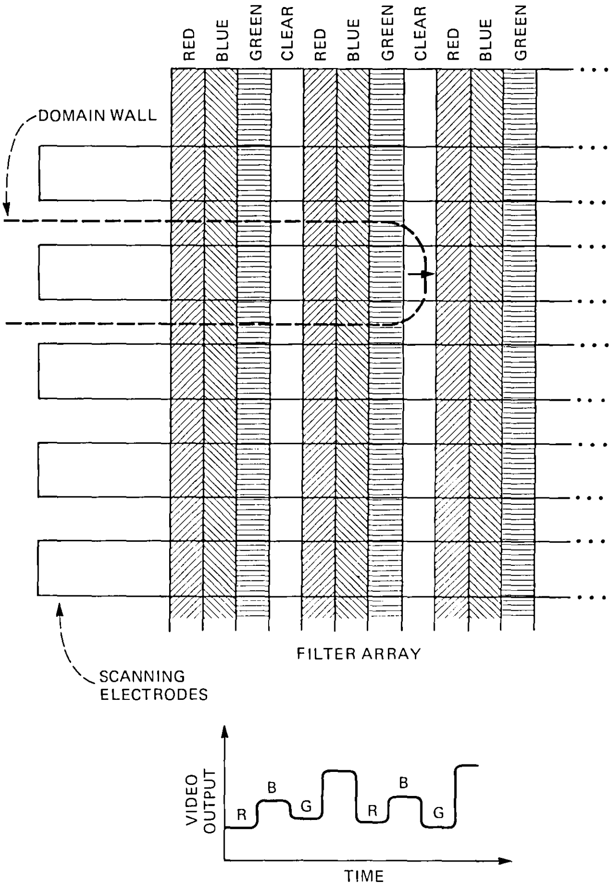


Fig. 10—Area scanner geometry with superimposed strip filter array for color applications. An example of the resulting output is seen at the bottom. External electronics are employed to separate out the three color signals into three parallel analog channels. Sync for doing this is provided by the interposed clear filters which are recognizable by their greater magnitude in the output signal. Note that no precise alignment of the filter array in position or angle is required.

A line scanner of very simple design can be made using a planar domain ferroelectric. The structure of such a device is shown in Fig. 11. The operating principle is the same as in the area scanner, but now a single domain wall extending across the width of the crystal is used for scanning. The only electrodes used are the origin, the scanning electrode, and the negative stop electrode which prevents the wall from being lost at the end of the crystal.

Ideally, one would want the crystal to be very narrow so that only a thin strip of the incident image would be scanned. Unfortunately, this leads to chip proportions which may be impractical. Instead, one can employ an optical system which itself defines the strip to be scanned and projects it on a crystal which has as narrow a shape as

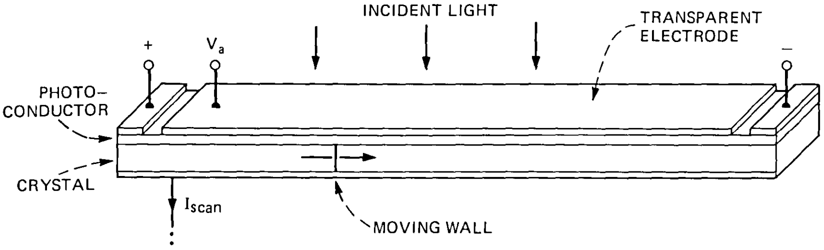


Fig. 11—Line scanner employing a planar wall ferroelectric. The principle of operation is the same as in the area scanner except that the wall extends across the full width of the crystal.

possible. This image is then defocused by a weak cylindrical lens to spread it out in a direction parallel to the domain wall. Thus the whole extent of the wall, all of which is sensitive to illumination, is illuminated by the same part of the image at a given time.

At first, it may appear that an area scanner may be obtained by dividing the top or bottom electrode of a planar wall scanner into numerous strips. This scheme will not work because the planar domain wall must scan all such strips simultaneously and would therefore be sensitive to the conductivity of the photoconductive layer associated with each of the strips at the point of scanning. This cannot be remedied by disconnecting all strips except that desired to be sensed, because the ferroelectric current from those strips would have nowhere to go, and wall motion would therefore cease. Even if the strips other than the one being sensed were connected to a voltage source (as was done in the planar wall analog readout device, Fig. 5), the wall's motion would still be influenced equally by conductivity changes anywhere along the wall.

4.4 Nonvolatile shift register memories

Static, nonvolatile, digital shift register memories can be made which employ ferroelectric domain wall motion phenomena. An experimental 4-bit shift register using lead germanate has been demonstrated.*

The basic shift register consists of a series of shifting electrodes preceded by an origin and trigger and completely surrounded by a confiner† (see Fig. 12). The shifting electrodes are connected in repeating sequence to voltage sources V_1, V_2, and V_3, as shown. By

* Unpublished experimental demonstration by the author. Origin, trigger, confiner, and shifting electrodes were all formed by small wires positioned close to the surface of a thin lead germanate crystal. The crystal was immersed in a slightly conductive fluid which established electrical contact between the wires and the crystal surface and served as a resistive layer. Domains could be introduced at will by the trigger and shifted along by the shifting electrodes. The moving domains could be observed in polarized light.

† A device for shifting domains in gadolinium molybdate was independently discovered by J. E. Geusic, T. J. Nelson, and D. P. Schinke (U.S. Patent 3,701,122). Domain detection was optical in this invention, however, and no use was made of the induced ferroelectric current.

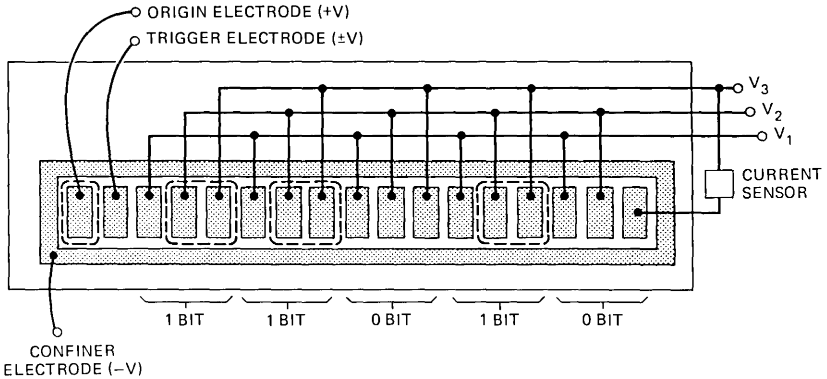


Fig. 12—Top view of a ferroelectric shift register memory. Domains introduced by the trigger are swept along by an array of 3-phase shifting electrodes. A "1" bit is represented by the presence of a positive-favoring domain in a 3-electrode cell, a "0" bit by the absence of such a domain. The arrival of a domain at the end of the register is sensed by means of the induced ferroelectric current in the last electrode.

allowing these voltages to go through the cycle of alternations shown in the waveforms of Fig. 13, positive-favoring domains introduced by the trigger can be shifted from electrode to electrode. A full shifting cycle has six stages or time periods as illustrated in the cross-section views of Fig. 13. It can be seen that each domain advances by expanding forward when the electrode ahead of it is made positive, and by pulling up from the rear when the electrode behind it changes negative. The surrounding confiner denies the shifting electrodes access to positive-favoring domains from any other part of the chip. The domains will therefore propagate down the shift register as isolated positive-favoring islands, one for each set of three shifting electrodes. The presence of such a domain represents a "1" bit and its absence of an "0" bit. The arrival of a domain at the end of the shift register is detected by the induced ferroelectric current generated in the lead of the last electrode. With suitable design, this current is adequate to drive external silicon devices directly.

Ferroelectrics generally exhibit a significant field threshold for wall motion. With supply power removed from the shift register, all electrode voltages go to zero, thus making wall motion impossible. Pyroelectric and piezoelectric effects can generate tiny voltages in the absence of supply power but, with appropriate design (e.g., finite shunt resistance between power supply terminals and ground), these will fall far below the threshold. Thus nonvolatile memory should be obtainable with ferroelectric shift registers.

4.5 Logic gates

Modifications to the prototype geometry can yield ferroelectric logic gates. Figure 14a shows an AND gate produced by interposing two

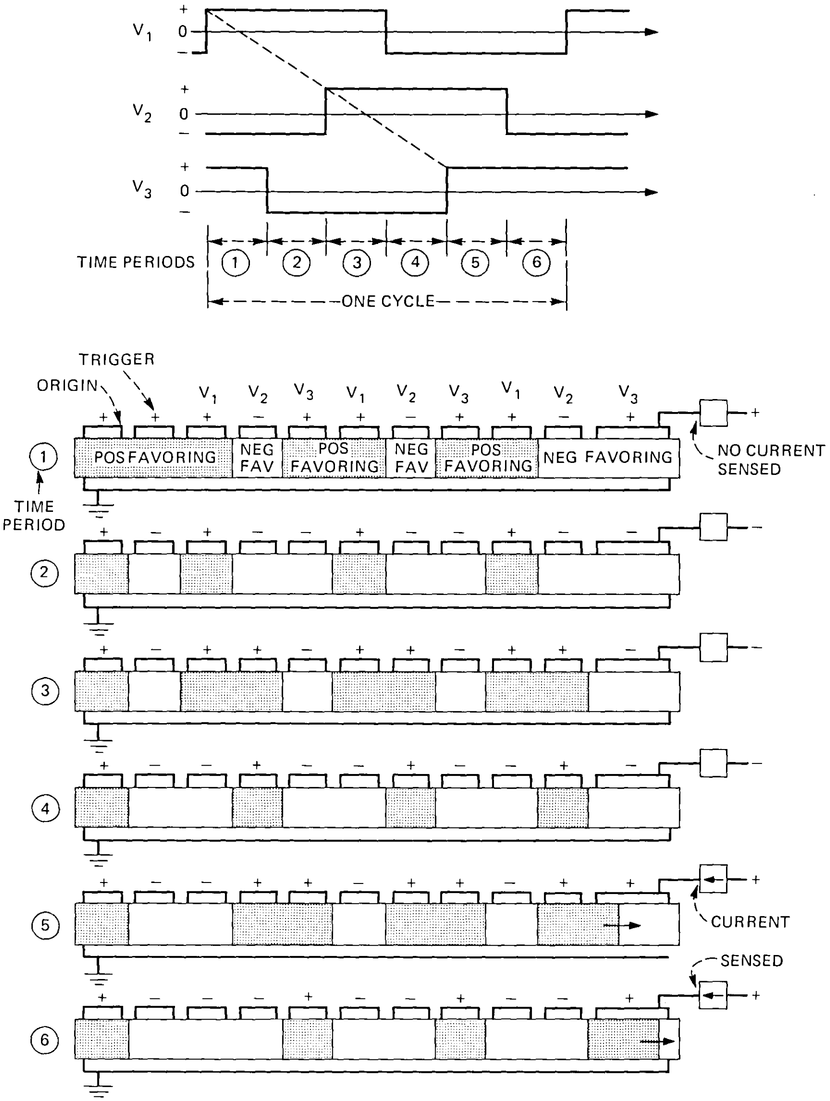


Fig. 13—Details of operation of the ferroelectric shift register. The six time periods making up a full shifting cycle are shown above, and an example illustrating the domain state of the register for each time period is seen below.

triggers between the origin and detector. A logic "1" input to the triggers is represented by a positive voltage during the first half of an operating cycle (the active period) and by a negative voltage during the second half (the restoring period) of the cycle. A logic "0" input is represented by a zero voltage throughout the full operating cycle. The detector is always supplied with a positive voltage during the active

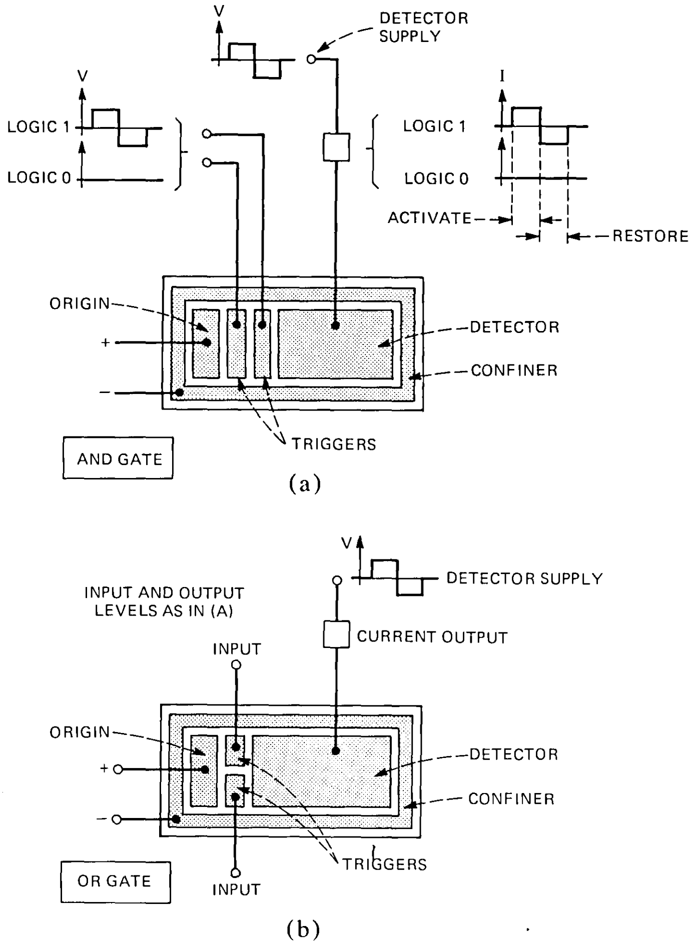


Fig. 14—Top view of simple ferroelectric logic gates. (a) AND gate made by interposing two triggers in sequence between detector and origin. Both triggers must be activated to obtain an output. (b) OR gate made by placing the triggers side by side so that either may cause an output.

period and with a negative voltage in the restoring period. Clearly, a logic "1" input to both triggers is required to produce an output current in the detector lead. A second detector current of opposite polarity will then flow during the restoring period. Should a logic "0" be applied to either trigger, no current will flow in the detector lead during either the active or restoring periods. Thus, an AND function is obtained with voltage inputs and current output. Note that the same logic level definitions used for the inputs also apply to the output: positive and negative half cycles = "1," no signals = "0."

An OR gate may be obtained if the triggers are arranged side by side,

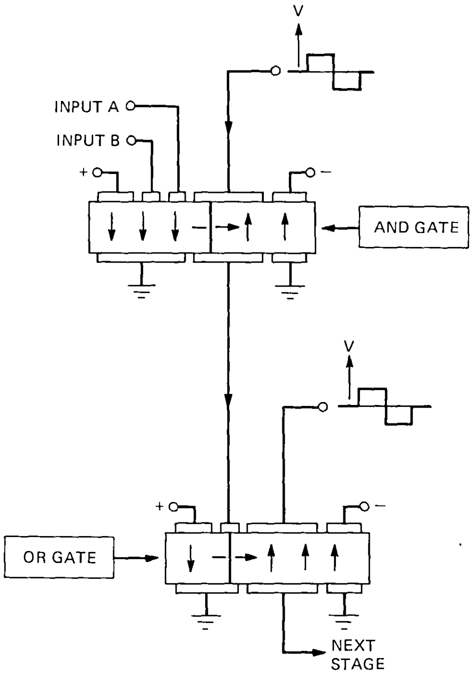


Fig. 15—Series-coupling of an AND gate output to one input of an OR gate. Both gates are seen in cross section and are schematically depicted as though they were on separate ferroelectric crystals. The wall motion current from the AND gate is fed to the OR gate trigger so that both walls must advance together. The wall motion current of the OR gate's detector can be fed to a third gate in a similar way.

as in Fig. 14b. The operation is identical to that of the AND gate except that either trigger alone may cause a 1 output. Both AND and OR gates may be made with any number of inputs in sequence or side by side.

If the above logic gates are to be of substantial utility, a means must be found to couple the output of one gate to the input of another. This can be done by series coupling. Figure 15 shows this scheme as applied to the coupling of an AND gate output to input of an OR gate. The AND gate (seen in cross section) is of conventional design except that its underside metallization has been patterned to isolate that portion of the ground electrode which lies under the detector. Wall motion currents to this electrode are fed to one trigger of an OR gate. For clarity, this is depicted as though the OR gate were on a second crystal. The series connection of detector and trigger implies that a domain wall can move under the trigger electrode if and only if a domain wall also moves under the detector electrode. If the logic condition is satisfied for the first gate, domain walls will move simultaneously in both its own detector and in the trigger of the next gate. During the restoring period, the domain walls in both gates return to their original positions, again by a series flow of current. In practice, the two gates of Fig. 15 would be on the same chip, with the OR gate turned upside down. This configuration makes it unnecessary for leads to pass through the crystal.

Series coupling may be viewed as a means of electrically transporting a domain from one part of the crystal to another. The output detector of a shift register, for example, could be series-coupled to the input trigger of the same shift register. Thus, a continuous data loop could be created without the design constraint of actually forming a physical loop.

The ability to perform logical inversion is fundamentally necessary for generation of the full range of logic functions. The output of a series-coupled ferroelectric gate can be inverted by a subtraction scheme as illustrated in Fig. 16. The figure shows a conventional AND gate connected in series with a gate with no trigger (an "inverter") which is connected to a polarity reversed version of the usual detector square wave supply. Suppose that the logic condition of the AND gate is satisfied, so that a domain wall is allowed to travel under the AND gate's detector. When viewed from a point on the series connection lead (point A in Fig. 16), this condition is symmetrical: point A is connected through mobile domain walls to both normal and inverted power supplies. Hence, the voltage at point A will be zero during both the active and restoring periods. Point A constitutes the inverted output of the first gate and is connected to a trigger input of a succeeding gate. Since point A remains at zero volts, no current flows in this connection, and the succeeding gate's trigger will not be activated.

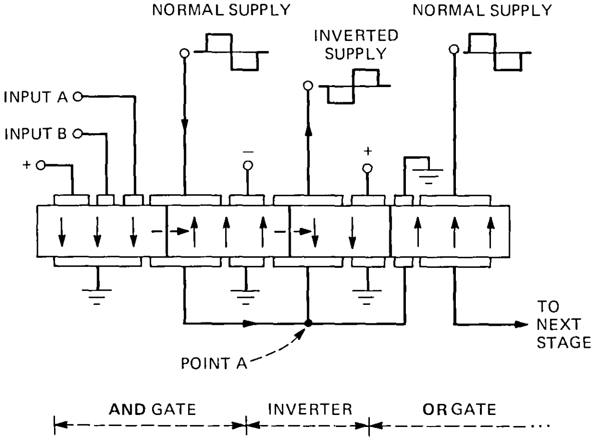


Fig. 16—Cross-section view showing an AND gate's output being inverted and series coupled to an OR gate input.

When the logic condition of the AND gate is not satisfied, no ferroelectric current will be provided to point A from the AND gate's square wave supply. However, the detector of the triggerless inverter gate always has access to a domain wall and will provide current to point A from the reversed polarity square wave supply. This current will activate the trigger of the succeeding gate. Thus the desired NAND logic function is accomplished. Since its trigger is activated from a reversed polarity source, all other aspects of the succeeding gate must be reversed in polarity also.

REFERENCES

1. J. M. Geary, Appl. Phys. Lett., *32* (1978), p. 455.
2. R. A. Lemons, J. M. Geary, L. A. Coldren, H. G. Mattes, Appl. Phys. Lett., *33* (1978), p. 373.
3. L. K. Anderson, IEEE Trans. Son. Ultrason., *19* (1972), p. 213.
4. H. J. Borchardt and P. E. Bierstadt, Appl. Phys. Lett., *8* (1966), p. 50.
5. J. R. Barkley *et al.*, Ferroelectrics (GB), *3* (1972), p. 191.
6. N. Yamamoto, K. Yagi, and G. Honjo, Phys. Stat. Sol. (a), *41* (1977), p. 523.
7. N. Yamamoto, K. Yagi, and G. Honjo, Phys. Stat. Sol. (a), *42* (1977), p. 257.
8. A. Balla and L. E. Cross, J. Mater, Sci., *12* (1977), p. 2346.
9. H. Iwasaki, Appl. Phys. Lett., *19* (1971), p. 92.
10. J. P. Dougherty, Appl. Phys. Lett., *20* (1972), p. 364.
11. R. Landauer, J. Appl. Phys., *28* (1957) p. 227.
12. R. C. Miller and G. Weinreich, Phys. Rev., *117* (1960), p. 1460.
13. H. C. Nathanson and J. Guldberg, "Topologically Structured Thin Films in Semiconductor Device Operation" in *Physics of Thin Films,* Vol. 8, New York: Academic Press, 1975.

Copyright © 1979 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 58, No. 2, February 1979
Printed in U.S.A.

Combining Echo Cancellation and Decision Feedback Equalization

By K. H. MUELLER

(Manuscript received August 2, 1978)

A system for two-wire, full-duplex data transmission is proposed. It consists of two adaptive transversal filters, one accepting the transmitted symbols and working as an echo canceller, the other accepting the received symbols and functioning as a decision feedback equalizer. A joint stochastic adjustment algorithm (updates at each baud) is analyzed, and it is shown that the sum of the mean-squared errors in the coefficients of both filters can be decoupled from its difference by selecting identical gain constants in each loop. The optimum gain equals the reciprocal of the sum of the taps of both loops. Convergence is exponential, and its time is 0.23 adjustments/dB/tap. This is completely independent of all channel parameters. Implementation of the proposed structure requires neither multipliers nor A/D converters. Promising applications are seen in channels with moderate precursor distortion, such as highpass channels (dc-restoration), two-wire PBX systems with a need for high-speed, full-duplex communication, limited distance cable channels, and, most important, two-wire digital subscriber lines for digital voice/data terminals.

I. INTRODUCTION

In a previous publication, a new approach to adaptive echo cancelling for full-duplex data transmission over two-wire facilities was presented.[1] Its novelty was that the compensation signal is synthesized directly from the data symbols, rather than from the transmitter output signal, and canceller adjustments are controlled by the receiver's estimated error signal, rather than the receiver's input signal, as has been done in previous echo cancellers.[2-5] This approach can be applied as long as the underlying modulation concept is linear; it allows for considerable economies in circuit implementation and also eliminates the double talker problem. The number of taps can be kept minimal if echo compensation is done at the baud rate, in synchronism with the receiver sampling operation. For this case, it has been shown in Ref. 1 that rapid convergence (time proportional to the number of echo taps) can be achieved, and that this convergence does not depend on channel response, echo response, timing phase, carrier phase, or the energy ratio of the echo signal to the distant received signal. Further studies dealing with this scheme are presented in Refs. 6 and 7.

On many real channels, the echo canceller alone would solve only part of the problem, since intersymbol interference (ISI) is severe and must be properly dealt with. The well-known adaptive equalizer is the proper cure for this, and during the past decade its art has been refined to a level of high sophistication. However, for two-wire full-duplex communication, one now must in general deal both with an adaptable echo canceller *and* an equalizer. Their joint adaptive adjustment will create new problems as far as updating techniques and dynamic behavior are concerned. Preliminary investigations of the behavior of a linear equalizer and an echo canceller have been carried out by Falconer and Weinstein,[8] indicating that convergence critically depends on the received signal to echo power ratio. These results have also been summarized in Ref. 6. Undoubtedly, this is a field where further studies are essential.

In this paper, a new system is proposed which combines both adaptive echo cancellation and equalization but retains the properties of rapid, channel-independent convergence under joint adjustments. As will be seen, the equalizer has to be somewhat restricted to obtain these advantages. The architecture of this system is discussed in the next section. After this, the convergence behavior is discussed, and finally simulation runs are presented which confirm the previously established analytical results.

II. SYSTEM ARCHITECTURE AND DEFINITIONS

The basic arrangement of the proposed system is shown in Fig. 1. We concentrate on a baseband system, not because of any such limitations (all linear modulation schemes can be represented in equivalent baseband), but rather to keep notation simple and concentrate on the essentials. The system contains two adaptive transversal filters; one is connected to the transmit data symbols and the other is connected to the received data symbols. The first works as an echo canceller as proposed in Ref. 1 to mitigate the effects of hybrid mismatch; the second filter is a decision feedback equalizer which compensates for intersymbol interference in the received far end signal due to linear distortion on the channel. With the structure in Fig. 1, this compensation is limited to trailing distortion components (postcursors) and we say more about this shortly. The outputs of both filters

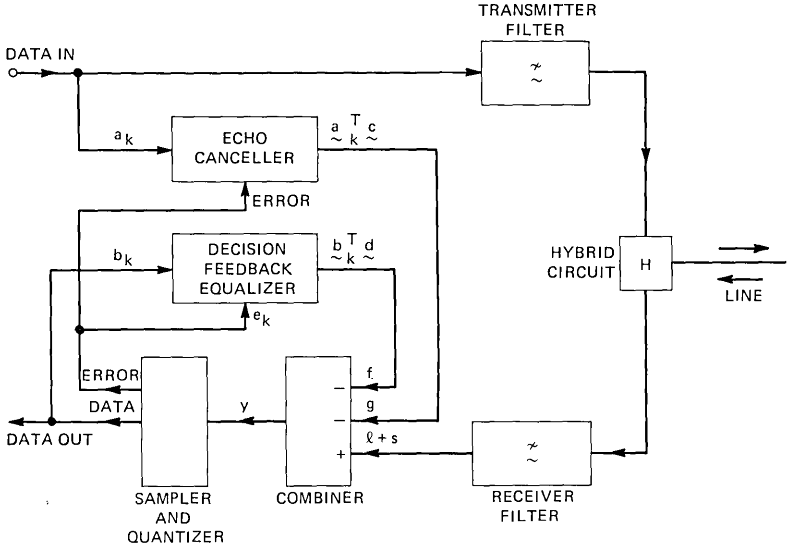


Fig. 1—Block diagram of data set with combined echo canceller and decision feedback equalizer.

are subtracted from the received signal and the resulting "cleaned-up" waveform is sampled to yield estimates $\hat{b}_k$ of the far end data b_k. Error samples e_k are generated in the usual way and are used as a common control signal to adjust both the canceller and the equalizer.

Since the equalizer has no linear taps to compensate for precursors, its abilities are somewhat limited. However, several significant advantages are also obtained: the implementation is economic since no A/D converter is required and the usually painful multiplications are replaced by simple additions (at least for binary and pseudo-ternary signals; some other codes may require multiplications where one factor is a two- or three-bit number). A further advantage is the total stability and predictable performance of this system, as is apparent from the analysis presented in Section III. Decision feedback equalization alone is well suited for highpass channels requiring dc restoration. It will also have interesting applications in transmission over cables and other channels where the distortion is predominantly of the trailing type.

In accordance with Ref. 1, let the near-end and far-end symbols be statistically and mutually independent random variables a_k and b_k. With h_k and r_k, we denote the samples of the channel response and echo response. At $t_0 + kT$, the received signal consists of a far-end component

$$\ell_k = \sum_{i=-\infty}^{\infty} b_{k-i} h_i \qquad (1)$$

and an echo signal

$$s_k = \sum_{i=-N}^{\infty} a_{k-i} r_i. \tag{2}$$

The receiver in Fig. 1 synthesizes the two compensation signals

$$g_k = \sum_{i=-N}^{M} a_{k-i} c_i = \mathbf{a}_k^T \mathbf{c} \tag{3}$$

and

$$f_k = \sum_{i=1}^{J} b_{k-i} d_i = \mathbf{b}_k^T \mathbf{d}, \tag{4}$$

where (3) is formed by an echo canceller with $L = M + N + 1$ taps $c_{-N} \cdots c_M$ and (4) is the output of the decision feedback equalizer* comprising taps $d_1 \cdots d_J$. The combined output y is

$$y_k = \ell_k + s_k - f_k - g_k + \zeta_k, \tag{5}$$

where ζ is some additive channel noise with variance σ^2. The error signal becomes

$$\begin{aligned} e_k &= y_k - b_k \\ &= \sum_{i=1}^{J} b_{k-i}(h_i - d_i) + \sum_{i=-N}^{M} a_{k-i}(r_i - c_i) + w_k, \end{aligned} \tag{6}$$

where

$$\begin{aligned} w_k = \zeta_k &+ \sum_{i=-\infty}^{-1} b_{k-i} h_i + \sum_{i=J+1}^{\infty} b_{k-i} h_i \\ &+ \sum_{i=M+1}^{\infty} a_{k-i} r_i + b_k(h_0 - 1). \end{aligned} \tag{7}$$

One recognizes that w_k is the remaining error component after optimum settings for both the canceller and the equalizer have been obtained. These optimum settings are, of course, given by

$$c_i = r_i \quad i = -N \cdots M \tag{8}$$

$$d_i = h_i \quad i = 1 \cdots J, \tag{9}$$

where all echo and intersymbol interference components within the reach of the adaptive structures are fully cancelled. It is further clear that some kind of automatic gain control should be used to force h_0

* We make the usual assumption that correct decisions are entered into the equalizer.

= 1 to minimize the variance of (7). In accordance with (8) and (9), we introduce error vectors

$$\phi = \mathbf{c} - \mathbf{r} \tag{10}$$

$$\psi = \mathbf{d} - \mathbf{h} \tag{11}$$

for the canceller and equalizer coefficients. The error can now be expressed in the simple form

$$e_k = w_k - \mathbf{b}_k^T\psi - \mathbf{a}_k^T\phi. \tag{12}$$

It will be our goal to adaptively minimize the mean-square error (MSE) which is given as

$$E\{e_k^2\} = \psi^{\mathrm{T}}\psi + \phi^{\mathrm{T}}\phi + R, \tag{13}$$

where R denotes that part of the MSE which cannot be further reduced with the canceller/equalizer combination, i.e.,

$$R = E\{w_k^2\} = \sum_{i=-\infty}^{-1} h_i^2 + \sum_{i=J+1}^{\infty} h_i^2 + \sum_{i=M+1}^{\infty} r_i^2 + \sigma^2 + (h_0 - 1)^2, \tag{14}$$

and the first two terms in (13) represent the excess error due to misadjustment. We now investigate how this excess error can be minimized.

III. JOINT STOCHASTIC ADJUSTMENTS

A joint, stochastic updating algorithm of the form

$$\mathbf{c}_{n+1} = \mathbf{c}_n + \gamma e_n \mathbf{a}_h \tag{15}$$

$$\mathbf{d}_{n+1} = \mathbf{d}_n + \beta e_n \mathbf{b}_n \tag{16}$$

with constant step sizes γ and β is proposed; i.e., no averaging is used. Together with (12) one obtains the coupled recursions

$$\boldsymbol{\phi}_{k+1} = (I - \gamma \mathbf{a}_k \mathbf{a}_k^T)\boldsymbol{\phi}_k - \gamma \mathbf{a}_k \mathbf{b}_k^T \boldsymbol{\psi}_k + \gamma w_k \mathbf{a}_k \tag{17}$$

$$\boldsymbol{\psi}_{k+1} = (I - \beta \mathbf{b}_k \mathbf{b}_k^T)\, \boldsymbol{\psi}_k - \gamma \mathbf{b}_k \mathbf{a}_k^T \boldsymbol{\sigma}_k + \beta w_k \mathbf{b}_k, \tag{18}$$

which demonstrate that adjustments in the two loops are not independent of each other. Both ϕ_k and ψ_k and therefore the excess error ϵ are of course influenced by the past history of the data symbols. Defining

$$q_k = E\{\phi_k^T\phi_k\} \tag{19}$$

$$p_k = E\{\psi_k^T\psi_k\} \tag{20}$$

and applying the independence assumptions discussed in Ref. 1, after some manipulations one obtains

$$q_{k+1} + p_{k+1} = (1 - 2\gamma + \gamma^2 L + \beta^2 J)q_k + (1 - 2\beta + \beta^2 J + \gamma^2 L)p_k + (\gamma^2 L + \beta^2 J)R \quad (21)$$

or, after introducing

$$\epsilon_k = p_k + q_k \quad (22)$$

$$\delta_k = p_k - q_k, \quad (23)$$

this can be written as

$$\epsilon_{k+1} = (1 - \beta - \gamma + \beta^2 J + \gamma^2 L)\epsilon_k + (\gamma - \beta)\delta_k + R(\beta^2 J + \gamma^2 L). \quad (24)$$

Our only interest is to minimize the combined MSE stemming from both echoes and intersymbol interference, i.e., ϵ; we are not concerned with the behavior of either component alone. A simple recursion which depends only on the excess error ϵ alone results if we set

$$\gamma = \beta, \quad (25)$$

i.e., if equal gains are used in the echo canceller and the decision feedback equalizer loop. Note, however, that $\gamma = \beta$ will *not* eliminate the coupling between the two loops; but then this has never been our concern since our objective is to minimize the total error without regard to the convergence behavior of its components. An illustration of what this practically means will be presented in the next section.

With the difference term δ_k disappearing, the recursion (24) can easily be solved,

$$\epsilon_k = \epsilon_\infty + [1 - 2\beta + \beta^2(L + J)]^k(\epsilon_0 - \epsilon_\infty), \quad (26)$$

where ϵ_0 is the initial mean-square excess error, and

$$\epsilon_\infty = \frac{\beta(J + L)R}{2 - \beta(J + L)} \quad (27)$$

is the steady-state mean-square excess error (tap fluctuation noise) in the tracking mode. During the training mode, ϵ converges exponentially until it reaches ϵ_∞. Fastest convergence will occur if

$$\beta = \beta_{\text{opt}} = \frac{1}{J + L}, \quad (28)$$

in which case

$$\epsilon_\infty(\beta_{\text{opt}}) = R. \quad (29)$$

A comparison with Ref. 1 shows that the results regarding the echo canceller alone and those for the combination of the echo canceller and the decision feedback equalizer are related; one need only replace the number of echo canceller taps with the sum of the taps of both the canceller and the equalizer.

Convergence with $\beta = \beta_{\text{opt}}$ is governed by

$$\epsilon_k = R + \left(1 - \frac{1}{J + L}\right)^k (\epsilon_0 - R) \tag{30}$$

and during the training phase the excess mean-square error is thus reduced at an average rate of

$$\frac{4.343}{J + L}\ \text{dB/adjustment}, \tag{31}$$

and convergence time is

$$0.230\ \text{adjustments/dB/tap}. \tag{32}$$

Both the above results assume $J + L \gg 1$ to linearize the logarithm in (30).

IV. SIMULATION RESULTS

The simplicity of the result obtained in the previous section speaks for itself. Most important, convergence is only determined by the sum of the taps; in particular, the ratio of echo signal power to received signal power is immaterial. Nothing could be more desirable for an actual system which is subjected to a wide variety of channel conditions.

As an example, consider a system with 127 echo canceller taps and 31 decision feedback taps. Samples of the echo response and the channel response have been taken at random. The combined mean-square error is about 100 times stronger than the received signal and consists mainly of echo noise; the ratio of echo/signal/ISI power being 100/1/1. Convergence of the total excess error ϵ is shown in Fig. 2. Note that convergence in the simulated system is somewhat faster than predicted by theory. This is because (30) was obtained as an average over *all* possible data sequences, whereas the simulation made use of maximum length pseudorandom sequences which have ideal spectral properties (and are also efficient to store). In the analysis, averaging includes such nonconverging patterns as steady mark or space. The analytical results of this and all similarly related problems tend therefore to be on the pessimistic side as far as they relate to the real world where pseudorandom training patterns are commonly used.

Our objective has always been to minimize the total noise power from all sources without caring about the reduction of the individual components. However, it is interesting to observe how each of the components p_k and q_k behaves separately during the joint training. To make this case more clear, an equalizer with $J = L = 8$ is selected, and

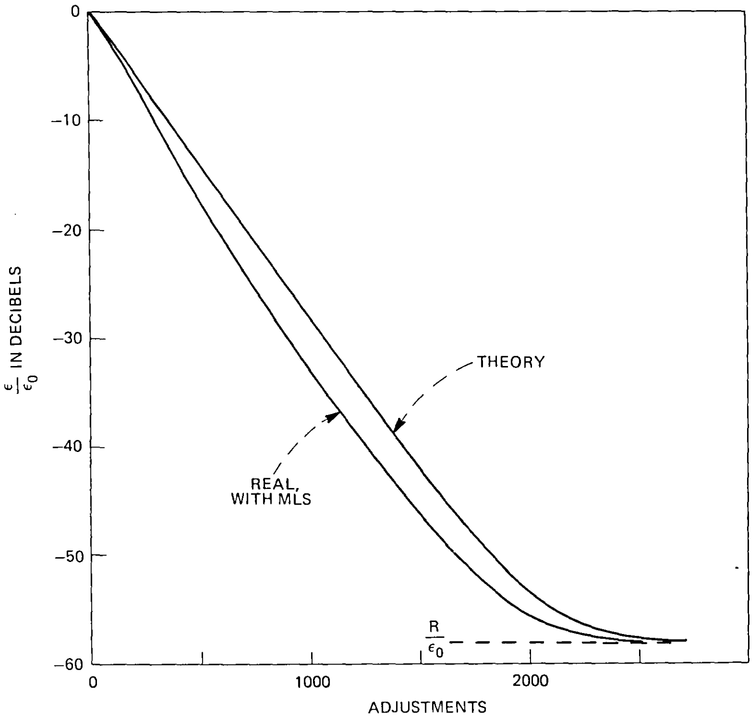


Fig. 2—Reduction of excess error in a system with 127 canceller taps and 31 decision feedback taps.

we consider various ratios of echo/signal/ISI power. The number of taps in this example may be in the order of what one would consider for two-wire full-duplex baseband transmission over limited distance cable facilities. Figure 3 depicts what is happening to the individual components p_k and q_k for echo-to-ISI power ratios r of 100, 1, and 0.01. The number of adjustments is written as a parameter along each curve. In the case where impairments due to echo noise and ISI are equal, they are reduced at about the same rate. However, if one component initially dominates, this component is reduced first, and this may actually perturb the tapsettings for the other (weaker) component in such a way as to increase its contribution temporarily until both components have been reduced to about a comparable level. From there on, they are jointly reduced at the same rate.

V. CONCLUSIONS AND SUMMARY

A combined structure incorporating an echo canceller and a decision feedback equalizer has been proposed. The structure has some appealing symmetries which can probably be exploited to realize efficient signal processing, in particular for the special case analyzed in this

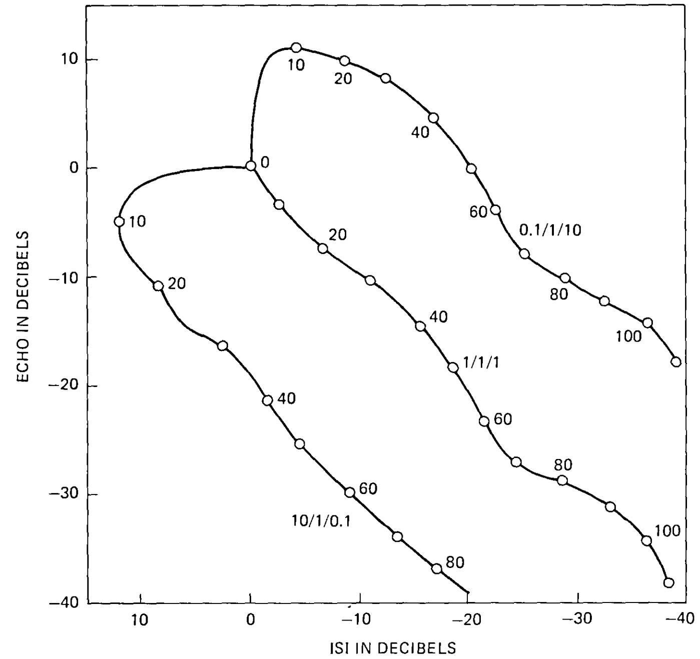


Fig. 3—Convergence of p and q components versus number of adjustments (bauds) for three ratios of echo/signal/ISI power.

paper where it has been assumed that the two communicating stations are mutually synchronized to a common master clock,* that taps are spaced at symbol intervals T, and that adjustments occur at each baud (no averaging). It has been shown that a direct, linear, first-order recursion can be obtained for the *total* excess error stemming from both the canceller and the decision feedback equalizer, provided that equal gain factors are selected for both loops. The optimum gain (in the sense of fastest convergence) equals the reciprocal of the sum of the number of taps in the two loops. Convergence is exponential, and the number of bauds required to obtain a certain improvement is 0.23 baud/dB/tap. Using the optimum gain results in a 3-dB steady-state mean-square error degradation, but this could easily be reduced (at the expense of tracking ability) to a negligible amount via gearshifting. The arrangements of either only a decision feedback equalizer (no linear taps) or only an echo canceller alone are both contained in our results; simply set either $L = 0$ or $J = 0$.

The absence of multiplications and A/D converters in both the canceller and the decision feedback equalizer will make implementation attractive. However, despite all the mentioned advantages, both

* This would, for example, be required in DDS extension service.

in regard to economics and convergence properties, it must be realized that there are many channels where compensation of the postcursor ISI is not sufficient. The equalizer will then require linear taps (the canceller, where the bulk of the taps will be concentrated for voiceband data applications, will fortunately never require linear taps). The inclusion of only a few linear taps drastically changes the joint convergence behavior, and more investigation is needed to determine economic architectures and algorithms which, under these conditions, would essentially retain the independence characteristics of the structure shown in Fig. 1.

REFERENCES

1. K. H. Mueller, "A New Digital Echo Canceller For Two-Wire Full-Duplex Data Transmission," IEEE Trans. Commun., *COM-24,* No. 9 (September 1976), pp. 956-962.
2. F. K. Becker and H. R. Rudin, "Application of Automatic, Transversal Filters to the Problem of Echo Suppression," B.S.T.J., *45,* No. 10 (December 1966), pp. 1847-1850.
3. M. M. Sondhi and A. J. Presti, "A Self Adaptive Echo Canceller," B.S.T.J., *45,* No. 10 (December 1966), pp. 1851-1854.
4. V. G. Koll and S. B. Weinstein, "Simultaneous Two-Way Data Transmission Over a Two-Wire Circuit," IEEE Trans. Commun., *COM-21,* No. 2 (February 1973), pp. 143-147.
5. N. Demytko and K. S. English, "Echo Cancellation on Time-Variant Circuits," Proc. IEEE, *65,* No. 3 (March 1977), pp. 444-453.
6. D. D. Falconer, K. H. Mueller, and S. B. Weinstein, "Echo Cancellation Techniques for Full-Duplex Data Transmission on Two-Wire Lines," Conference Record, NTC 1976, pp. 8.3-1 to 8.3-7.
7. S. B. Weinstein, "A Passband Data Driven Echo Canceller for Full-Duplex Transmission on Two-Wire Circuits," IEEE Trans. Commun., *COM-25,* No. 7 (July 1977), pp. 654-666.
8. D. D. Falconer and S. B. Weinstein, "High-Speed Two-Way Data Communication on a Two-Way Channel," unpublished work.

The Bell System Technical Journal
Vol. 58, No. 2, February 1979
Printed in U.S.A.

Imaging Reflector Arrangements to Form a Scanning Beam Using a Small Array

By C. DRAGONE and M. J. GANS

(Manuscript received May 23, 1978)

To obtain the performance of a large aperture phased array, a small phased array is combined with a large main reflector and an imaging arrangement of smaller reflectors to form a large image of the small array over the main reflector. An electronically scanable antenna with a large aperture is thus obtained, using a small array. An attractive feature of the imaging arrangement is that the main reflector need not be fabricated accurately, since small imperfections can be corrected efficiently by the array. As an application, a 4.2-m diameter antenna is discussed for a 12–14 GHz satellite with a field of view of 3 degrees by 6 degrees required for coverage of the continental United States.

I. INTRODUCTION

Use of a phased array in a satellite of large aperture diameter is proposed in Ref. 1 to form a narrow beam to communicate with ground stations in the United States. A large array in this case is not attractive, because of its weight and the loss and complexity of the long interconnections required by the large spacing between the array elements. Thus, we here propose the use of a small array combined with several reflectors as shown in Figs. 1 to 3. The reflectors are arranged so that a magnified image of the array S_0 is formed over the aperture of the main reflector. The magnification M relating the diameters D_0 and D_1 of the main reflector and the array, respectively, is chosen much greater than unity, i.e.,

$$M = \frac{D_0}{D_1} \gg 1, \tag{1}$$

so that the array is much smaller than the main reflector.

The main reflector S_0 in Figs. 1 and 2 may be difficult to fabricate accurately because of its large diameter. However, it is pointed out in

Section II that small imperfections are easily corrected because S_0 and the array are conjugate elements.* Another important property of the arrangements described here is that the transformation relating the field over the array aperture to the field over the main reflector aperture is essentially frequency independent, and therefore it can be approximated by its asymptotic behavior at high frequency. That is, the transformation can be determined accurately using the laws of geometric optics, as pointed out in Section II.

Use of several reflectors combined with a small array is not new. In particular, the Gregorian configuration of two confocal paraboloids shown in Fig. 1 is discussed in Refs. 2 and 3. In Ref. 3, the performance of this arrangement is analyzed by representing the field over the array aperture in terms of plane waves, and by then determining separately the transformation for each plane wave. Here, however, we shall see that our condition (3) allows the analysis to be carried out entirely using the laws of geometrical optics, as already pointed out. We first consider the arrangement of Fig. 1.

II. ANALYSIS

In Fig. 1, the first paraboloid, S_0, transforms a plane wave, propagating in the direction of the paraboloid axis, into a spherical wave converging toward the focus F. This spherical wave is then transformed into a plane wave, by the second paraboloid S_1, which is large enough to intercept all incident rays. After the second reflection, the reflected rays illuminate the array plane $\sum_1$. Since the illuminated area corresponds to the projection of the first paraboloid, its diameter D_1 is determined by D_0, and from Fig. 1,

$$M = \frac{D_0}{D_1} = \frac{f_0}{f_1}, \tag{2}$$

where f_1 and f_0 are the axial focal lengths of the two paraboloids. Thus, by choosing

$$\frac{f_0}{f_1} \gg 1,$$

a small array diameter D_1 is sufficient to intercept all the incident rays. Notice on $\sum_1$ the center of illumination is determined by the ray corresponding to the center C_0 of the paraboloid. The center C_1 of the array must therefore be placed on this ray, which will be called the *central ray.*

* Conjugate elements in an optical system have the property that the rays originating from a point of one element are transformed, by the optical system, into rays which pass through a corresponding point of the other element. Two such corresponding points are called conjugate points.

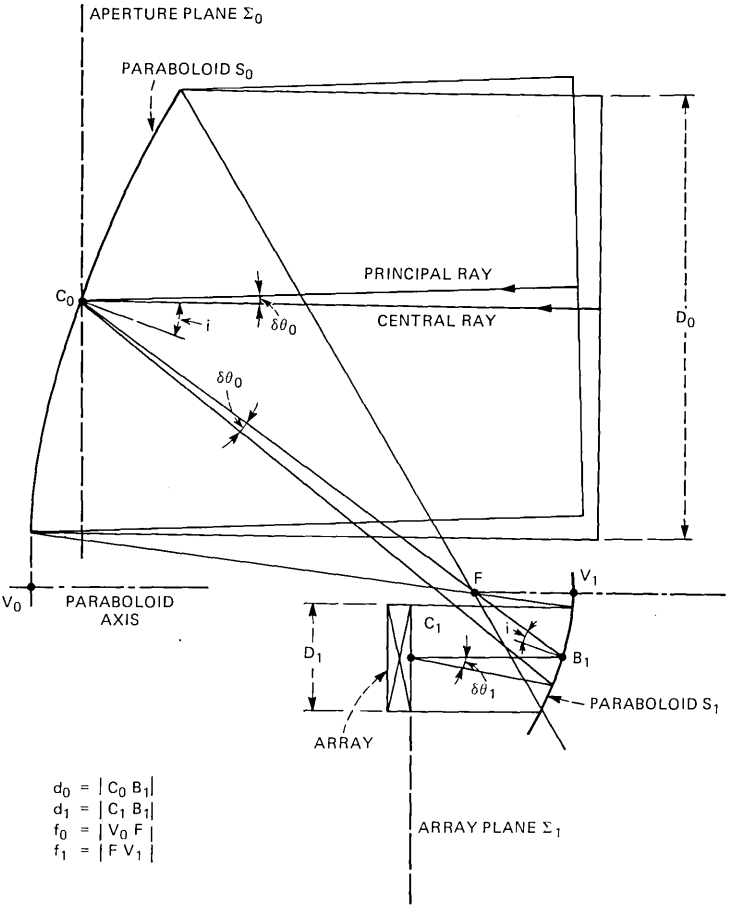


Fig. 1—A Gregorian arrangement of two confocal paraboloids magnifying a small array. The main reflector S_0 and the array are conjugate elements.

Now suppose in Fig. 1 the direction of the incident wave is changed so that the ray incident at C_0 makes a small angle $\delta\theta_0$ with respect to the central ray. The center of illumination will then vary with $\delta\theta_0$, unless C_0 and C_1 are conjugate points, as in Fig. 1. In this case, for small $\delta\theta_0$, all rays reflected at C_0 pass through C_1 after the second reflection. We thus conclude that, for maximum efficiency of illumination, the following condition must be satisfied:

> The center of the main reflector and the center of the array must be conjugate points. (3)

When this condition is satisfied, the field in the vicinity of C_1 is the image of the field in the vicinity of C_0. More precisely, let $\sum_0$ and $\sum_1$ be the two planes orthogonal to the central ray, through C_0 and C_1. Then $\sum_0$ and $\sum_1$ are conjugate planes, in the vicinity of C_0 and C_1, and therefore the field E_1 on $\sum_1$ is the image of the field E_0 on $\sum_0$.

The consequences of this basic condition are now discussed. The transformation which relates the input field E_0 to the output field E_1 in Fig. 1 involves several reflections and, because of diffraction, the field propagating from one reflector to the next *cannot* be determined accurately using the laws of geometric optics. Thus, suppose Fresnel's diffraction formula is used to determine the transformation from one reflector to the other, or from the reflector to the array in Fig. 1. The details of such calculations are given in Ref. 4, where the general transformation between the input and output fields E_0 and E_1 of an optical system is derived. It is shown in Ref. 4 that, when the input and output planes are conjugate planes of the optical system, the output field E_1 is simply the image of the input field E_0, and it can be calculated using the laws of geometric optics. This result is quite remarkable for, in general, the laws of geometric optics give correctly only the field in the output plane, not the field inside the optical system. An important consequence of this result, which was used in Ref. 5 to obtain frequency independence in the far-field of a satellite antenna, is now pointed out.

Suppose in Fig. 1 the surface of the main reflector is not perfect, but it contains a small imperfection $\delta\ell$ causing a phase error $\delta\psi_0 \simeq 2k\delta\ell$ after reflection. Then, if P_0 is the location of the imperfection on the paraboloid, and P_1 is the corresponding point on the array plane, one has that E_1 will contain at P_1 a phase error $\delta\psi_1$ approximately equal to $\delta\psi_0$,

$$\delta\psi_1 \simeq 2k\delta\ell.$$

Since $\delta\ell$ is independent of frequency, the reflector deformation can be corrected by a frequency independent change in the time delay of the array element corresponding to P_1. This would not be true if C_0 and C_1 were not conjugate points.

Consider also the effect of a surface deformation on the subreflector in Fig. 1. Now, unless the subreflector and the array are conjugate elements, the resulting perturbation of E_1 will not be independent of $\delta\theta_0$, but its location will vary with $\delta\theta_0$. Furthermore, because of diffraction, E_1 will be perturbed both in amplitude and phase, and the perturbations will in general vary with frequency. Thus, surface deformations on the subreflector cannot be easily corrected.

2.1 Location of C_1

The location of C_1 is now determined. The ray reflected in Fig. 1 at C_0 for $\delta\theta_0 \neq 0$ will be called the principal ray. Let $\delta\theta_1$ be the angle this ray makes with the central ray at C_1. Then, since M is the magnification of the two conjugate planes $\sum_0$ and $\sum_1$, the angles $\delta\theta_0$ and $\delta\theta_1$ must satisfy the well-known relation

$$\delta\theta_1 = M\delta\theta_0.$$

Now, from Fig. 1,

$$d_1\delta\theta_1 = d_0\delta\theta_0, \tag{4}$$

d_1 and d_0 being the distances of C_1 and C_0, respectively, from the center B_1 of the subreflector. One can show that

$$d_0 = \frac{f_1 + f_0}{\cos^2 i}, \tag{5}$$

i being the angle of incidence at C_0 (or B_1) for the central ray. From the above relations one obtains

$$d_1 = \frac{f_1 + f_0}{\cos^2 i}\frac{1}{M} \tag{6}$$

or

$$d_1 = \frac{f_1}{\cos^2 i}\ \frac{M+1}{M} = |FB_1|\frac{M+1}{M}. \tag{7}$$

2.2 Use of an additional reflector to increase the distance d_1

In Fig. 1, the array is relatively close to the subreflector S_1, and this may be a disadvantage for some applications. In the application discussed in Section III, for instance, a greater distance d_1 will be needed to place a grid between the array and the subreflector for polarization or frequency diplexing. In this case, it is advantageous to use three reflectors S_0, S'_0, and S_1 arranged as shown in Fig. 2.

To determine the distance $d_1 = |C_1B_1|$ between the array and the last reflector, which is a paraboloid, it is convenient to introduce the parameters ℓ, ξ_1, ξ_2, M_0 defined by

$$\ell = |C_0F'|$$

$$\frac{\ell}{\xi_1} = |B_0F'|$$

$$\frac{M_0\ell}{\xi_1} = |FB_0|$$

$$\frac{\ell}{\xi_2} = |B_1F|.$$

To determine the location of C_1 for small $\delta\theta_0$, consider in Fig. 2 the two rays reflected by the main paraboloid at C_0. One of the two rays is the central ray. Notice that the hyperboloid subreflector forms a

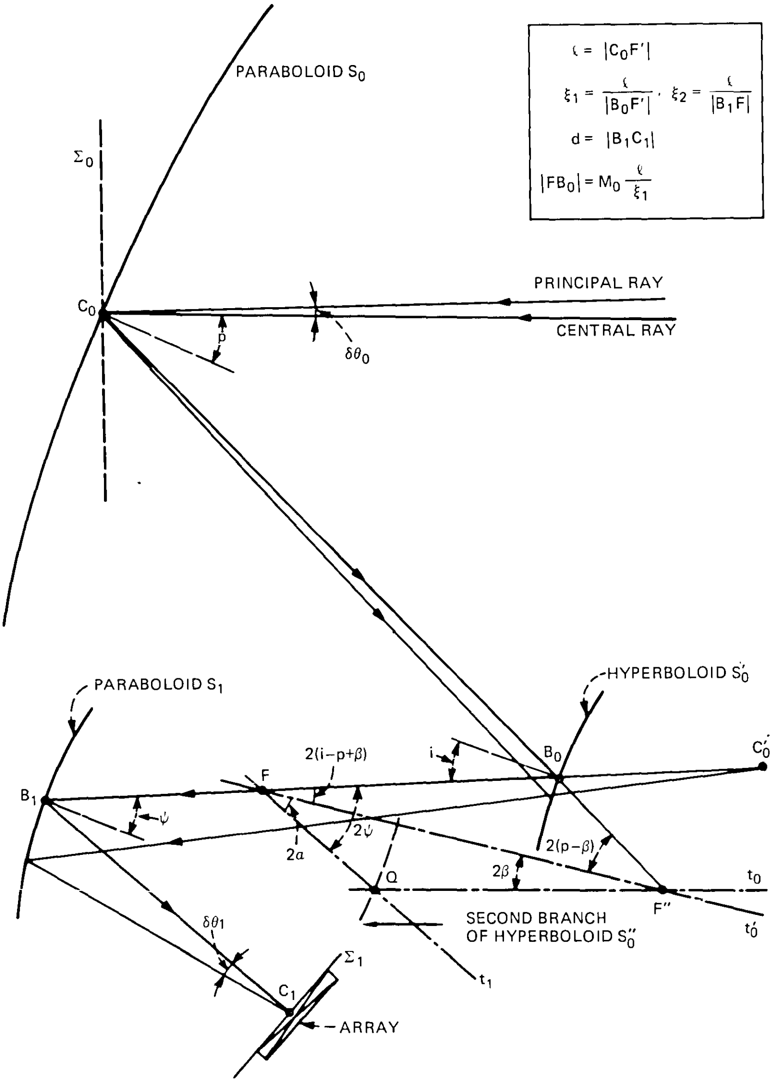


Fig. 2—Imaging arrangement of three reflectors.

virtual image C_0' of C_0. The last paraboloid subreflector transforms this virtual image into a real image C_1, where both rays meet after reflection by S_1

To determine the location of C_0', one has to find the paraxial focal length of the hyperboloid reflector. Taking into account that F' and F are conjugate points, whose distances from S_0' are ℓ/ξ_1 and $M_0\ell/\xi_1$, respectively, the focal length in question is*

$$\frac{\ell}{\xi_1}\,\frac{M_0}{M_0 - 1}.$$

Thus, since the distance of C_0 from B_0 is

$$\ell\,\frac{\xi_1 - 1}{\xi_1},$$

using the lens equation one finds that the distance of C_0' from B_0 is

$$\ell\, M_0\,\frac{\xi_1 - 1}{\xi_1}\,\frac{1}{1 + \xi_1(M_0 - 1)}.$$

The location of C_1 is next determined. The paraxial focal length of the last reflector S_1 is ℓ/ξ_2, and the distance of C_0' from B_1 is

$$\frac{\ell}{\xi_2} + \frac{\ell M_0^2}{1 + \xi_1(M_0 - 1)}.$$

Therefore, using once more the lens equation, one finds for the distance of C_1 from B_1

$$d_1 = \frac{\ell}{\xi_2}\left[1 + \frac{1}{M_0^2}\,\frac{1 + \xi_1(M_0 - 1)}{\xi_2}\right]. \tag{8}$$

One can verify that

$$M = \frac{D_0}{D_1} = M_0\xi_2, \tag{9}$$

which allows ξ_2 in eq. (8) to be expressed in terms of M and M_0, giving the result

$$d_1 = \frac{\ell}{M}\left\{M_0 + \frac{1}{M}[1 + \xi_1(M_0 - 1)]\right\}. \tag{10}$$

This expression, which for $M_0 = 1$ can be shown to coincide with eq. (6), is a monotonic function of M_0. Thus, by choosing $M_0 >> 1$, as in Fig. 2, a distance d_1 appreciably greater than that of Fig. 1 is obtained.

An important difference between the two arrangements of Figs. 1 and 2 is that the various surfaces of revolution of the reflectors in Fig. 2 are not centered around the same axis, as in Fig. 1. In fact, in Fig. 2 the axis t_0' of the hyperboloid is tilted by the angles 2β and 2α, with respect to the axes t_0 and t_1 of the two paraboloids. This difference is now explained.

* According to the lens equation (Ref. 4), the inverse of the focal length must equal the sum of the inverses of the distances from the conjugate points to the reflector.

2.3 Orientation of the axes t_0, t'_0, t_1

For some applications, it is important that everywhere on the array plane the polarization of E_1 coincide with that of E_0. For $\delta\theta_0 = 0$, one can show this condition is satisfied in Fig. 1, provided the reflectors are centered around the same axis. In Fig. 2, on the other hand, either of the two angles α, β may be chosen arbitrarily provided the other angle satisfies the condition[6,7]

$$\tan\alpha = m\tan\beta, \tag{11}$$

where m is related to the eccentricity e of the hyperboloid through the relation

$$m = \frac{e+1}{e-1}. \tag{12}$$

It can be calculated once p, β, i are known, using the relation

$$m = \frac{\tan(p-\beta)}{\tan(i-p+\beta)}, \tag{13}$$

where the angles α, β, i, and p are as shown in Fig. 2.

Equation (11) has the following geometric significance. In Fig. 2, t_0 represents the axis of the main reflector, t'_0 the axis of the subreflector, and t_1 the axis of the imaging reflector. The reflector S'_0 is derived from one of the two branches of an hyperboloid. If S''_0 denotes the other branch, then it is shown in Ref. 7 that the point of intersection Q of the two axes t'_0 and t_1 must be a point of S''_0, as shown in Fig. 2. Then, since 2α and 2β can be interpreted as the angles the two focal radii FQ and $F'Q$ make with the axis t'_0 one obtains eq. (11).

From the triangle FB_0F' in Fig. 2, taking into account that $|FB_0| = M_0|B_0F'|$, one has

$$\sin(2p-2\beta) = M_0\sin 2(i-p+\beta), \tag{14}$$

and, therefore, β can be considered a function of i, p, M_0. Notice from Fig. 2 that the angle of incidence ψ for the central ray on the last reflector is given by

$$2\psi = 2(\alpha+i-p+\beta). \tag{15}$$

This angle ψ can be shown to increase as i is decreased. In the application to be discussed next, a relatively large value for ψ is desirable to allow a frequency diplexer to be used as shown in Fig. 3.

III. AN APPLICATION

As an application, consider the design of an antenna which must transmit at 12 GHz and receive at 14 GHz in a 4.2 m diameter satellite in synchronous orbit at 105° W longitude. Assume a field of view of 3 degrees by 6 degrees (which corresponds approximately to the conti-

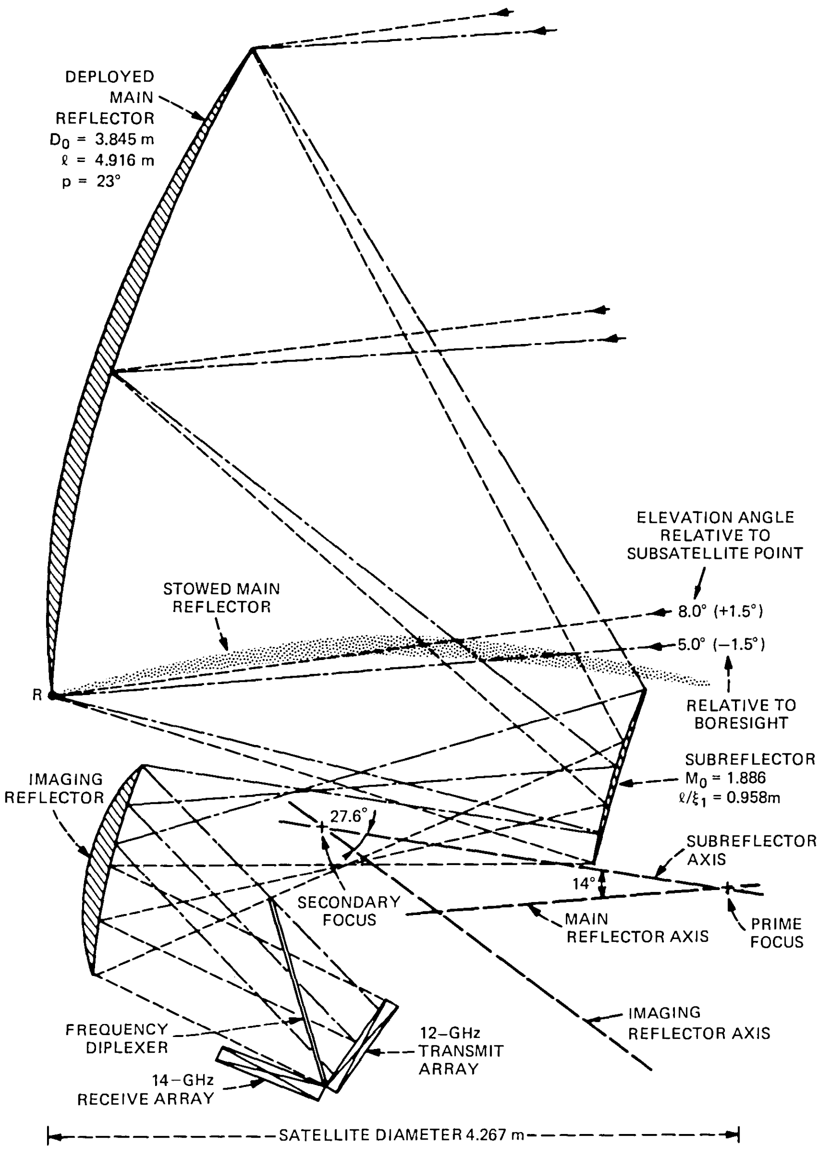


Fig. 3—Imaging satellite antenna with 12/14 GHz frequency diplexing and overall magnification of 7.

nental United States) is required, and suppose separate arrays must be employed for transmission and reception. An arrangement suitable for this purpose is shown in Fig. 3, using a quasi-optical diplexer*

* See, for example Refs. 8 and 9. In Fig. 3, the design requirements are more stringent than in Refs. 8 and 9 because of the wide range of incident angles experienced by the diplexer.

between the two arrays and the last reflector. Figure 4 shows a detail of the feed arrays and diplexer, and Fig. 5 shows a front view of the antenna. The values of M, M_0, etc. are listed in Fig. 3. They were chosen taking into account the requirement that the arrangement should be free of blockage, it should be efficient and, of course, the array should be reasonably small. It is assumed that in Fig. 3 the main reflector can be rotated around R, so that it can be initially stowed in the satellite horizontally as indicated in Fig. 3. Then, once in orbit, it will be rotated into its final position shown in Fig. 3. The size of the main reflector can be increased and still fit in the satellite diameter either by using an elliptical shape or by not stowing the reflector in a completely horizontal position.

Figures 3 and 4 show the paths of the marginal rays in the plane of symmetry for $\delta\theta_0 = \pm 1.5$ degrees. Also shown in Fig. 4 are the rays for $\delta\theta_0 = 0$.

In Section II, the angle $\delta\theta_0$ was assumed to be very small, in which case the illumination over the array aperture can be considered inde-

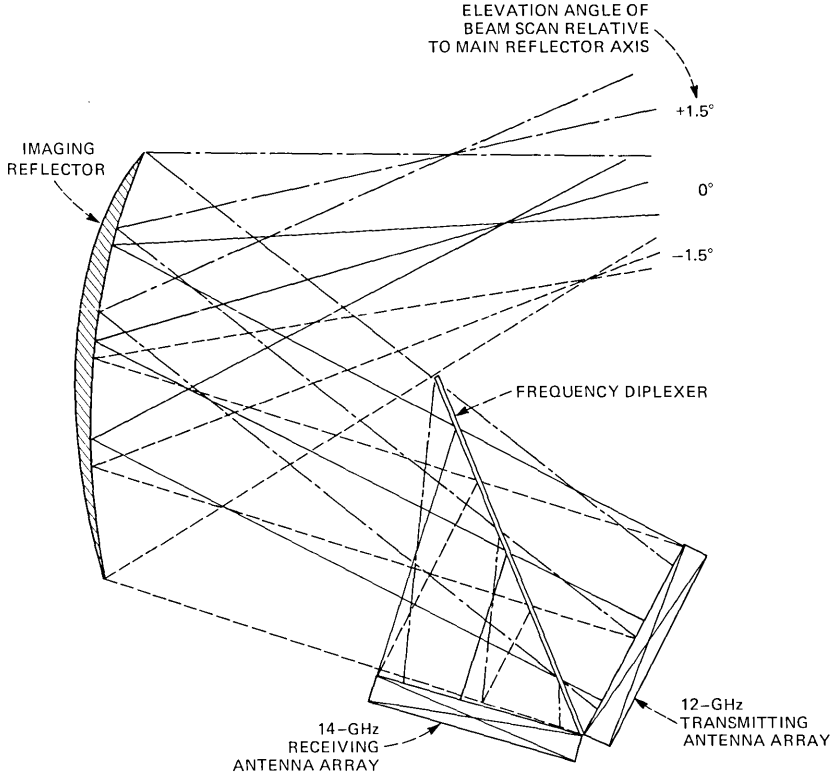


Fig. 4—Detail of feed arrays and diplexer.

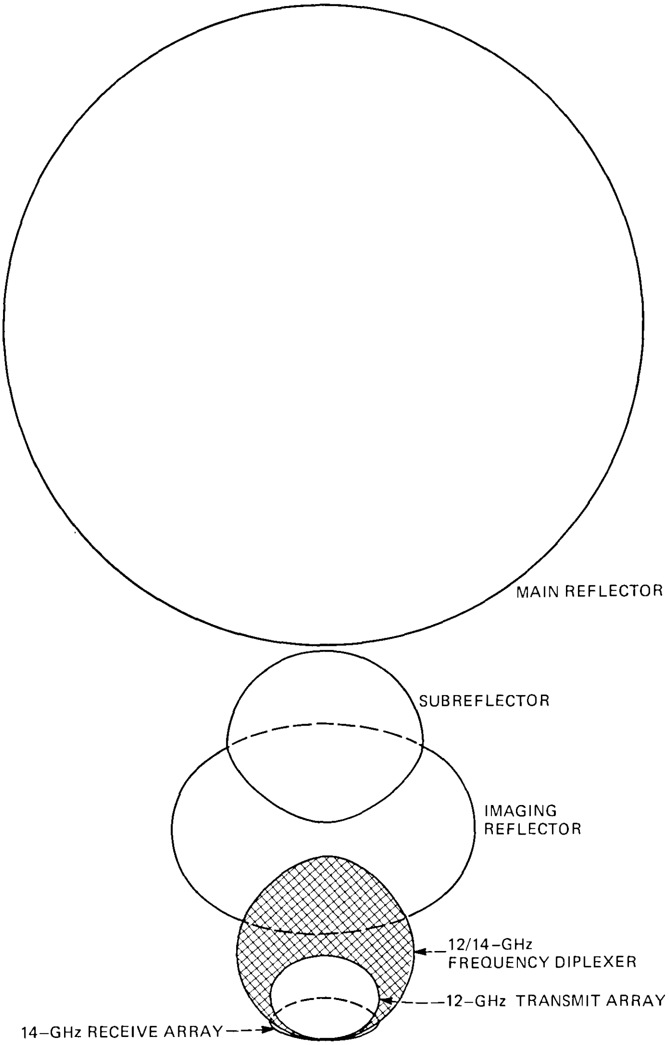


Fig. 5—Front view of imaging satellite antenna.

pendent of $\delta\theta_0$. In the application considered here, however, the angle of incidence $\delta\theta_1$ assumes relatively large values ($\delta\theta_1 = 21$ degrees, for $\delta\theta_0 = 3$ degrees) because of the large magnification $M = 7$. Thus, there is appreciable variation in illumination over the array aperture, and this causes a loss in gain which is now discussed. It is assumed the spacing of the array elements is very small, so that any desired phase distribution over $\sum_1$ can be produced by the array excitation. Then, if g^2 denotes the power distribution on $\sum_1$ due to the array excitation, the efficiency η of illumination is given by the familiar expression

$$\eta = \frac{\left(\int\int_{-\infty}^{\infty} fg\,dx\,dy\right)^2}{\int\int_{-\infty}^{\infty} f^2\,dx\,dy \int\int_{-\infty}^{\infty} g^2\,dx\,dy}, \tag{16}$$

where $f^2\,dx\,dy$ is the power incident on the element of area $dx\,dy$, caused on $\sum_1$ by a plane wave incident on $\sum_0$. Figure 6 shows for different scan angles the geometric optics array illuminations and the corresponding losses in gain given by eq. (16) for uniform array excitation. Also shown are the losses for a tapered excitation of −10 dB at the edge of the array. Two cases, A and B, are shown in Fig. 6. In case A, the array is centered at C_1 with diameter given by D_0/M. In this case, the scan loss is zero for $\delta\theta_0 = 0$, but it becomes relatively high at the edge of the field of view. In case B, the scan losses for $\delta\theta_0 = \pm 3$ degrees were minimized by increasing the array size and slightly offsetting the array center as shown in Fig. 6 (case B). All losses for −10 dB taper in Fig. 6 are normalized with respect to the value (−0.45 dB taper loss) given by eq. 16 for $\delta\theta_0 = 0$ in case A. This sacrifice in antenna directivity is often made to obtain the sidelobe reduction provided by a −10 dB edge taper.

Curves of scan loss for −10 dB taper in case B are shown in Fig. 7. The positive values near the east and west coasts are due to the above normalization.

IV. CONCLUSIONS

In a conventional reflector antenna, a relatively small feed is usually placed at the focus of a reflector arrangement which then transforms the spherical wave radiated by the feed into a plane wave. In such an antenna, only in part is the power radiated by the feed intercepted by the aperture of the main reflector. Thus, to minimize the loss due to spillover, the edge illumination is usually chosen appreciably lower (−10 dB or less) than the illumination at the center of the aperture. The loss due to spillover is then typically −0.5 dB (in addition to the taper loss mentioned above, giving a total of about −0.9 dB). On the other hand, by using the imaging reflectors and the properly sized and located feed array, less loss is obtained over most of the United States, as shown in Fig. 7. For example, Fig. 7 shows that the loss due to vignetting (i.e., spillover) is largest near the center of the country at a value of −0.6 dB. Losses suffered in the feed array itself are a function of the size and number of feed elements and have not been included.

The imaging arrangements discussed here are particularly useful

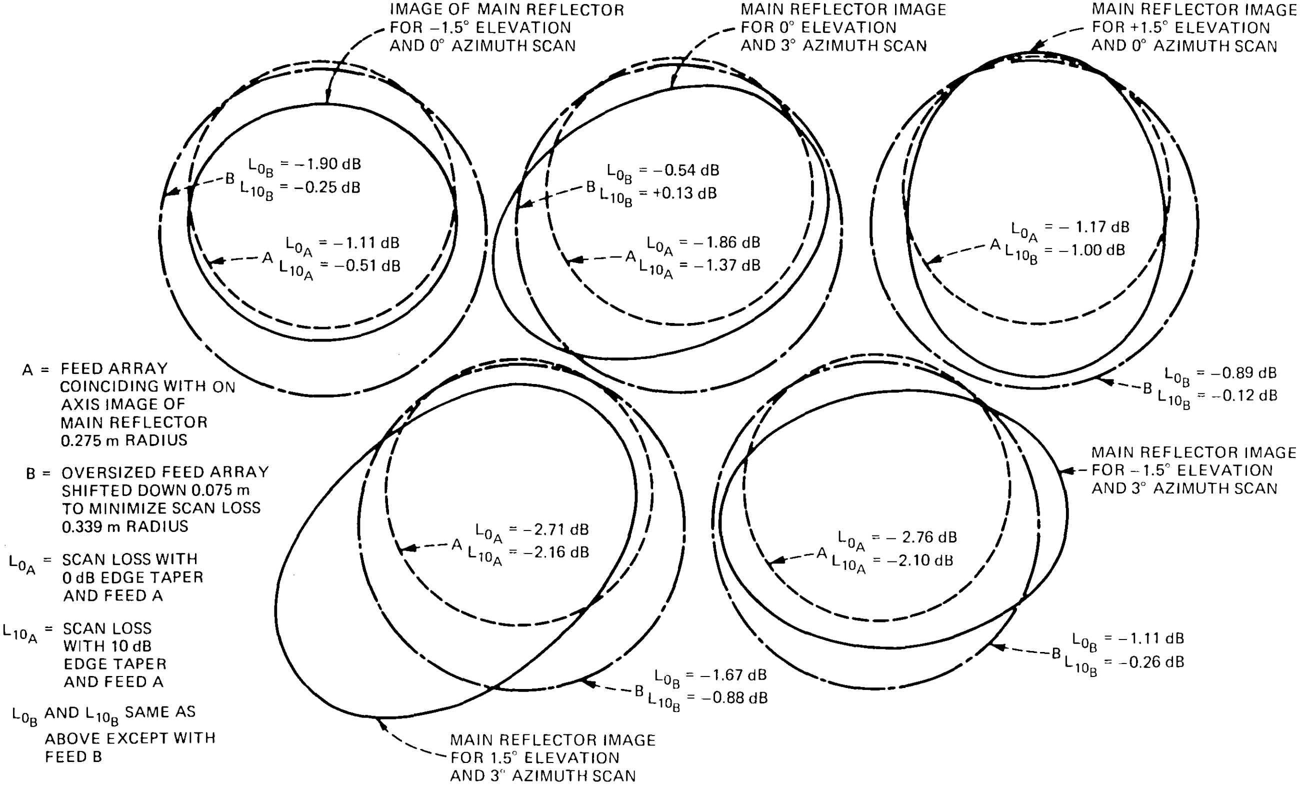


IMAGE OF MAIN REFLECTOR FOR −1.5° ELEVATION AND 0° AZIMUTH SCAN
MAIN REFLECTOR IMAGE FOR 0° ELEVATION AND 3° AZIMUTH SCAN
MAIN REFLECTOR IMAGE FOR +1.5° ELEVATION AND 0° AZIMUTH SCAN
B L_{0_B} = −1.90 dB L_{10_B} = −0.25 dB
A L_{0_A} = −1.11 dB L_{10_A} = −0.51 dB
B L_{0_B} = −0.54 dB L_{10_B} = +0.13 dB
A L_{0_A} = −1.86 dB L_{10_A} = −1.37 dB
A L_{0_A} = − 1.17 dB L_{10_B} = −1.00 dB
B L_{0_B} = −0.89 dB L_{10_B} = −0.12 dB
A = FEED ARRAY COINCIDING WITH ON AXIS IMAGE OF MAIN REFLECTOR 0.275 m RADIUS
B = OVERSIZED FEED ARRAY SHIFTED DOWN 0.075 m TO MINIMIZE SCAN LOSS 0.339 m RADIUS
L_{0_A} = SCAN LOSS WITH 0 dB EDGE TAPER AND FEED A
L_{10_A} = SCAN LOSS WITH 10 dB EDGE TAPER AND FEED A
L_{0_B} AND L_{10_B} SAME AS ABOVE EXCEPT WITH FEED B
MAIN REFLECTOR IMAGE FOR −1.5° ELEVATION AND 3° AZIMUTH SCAN
A L_{0_A} = −2.71 dB L_{10_A} = −2.16 dB
A L_{0_A} = − 2.76 dB L_{10_A} = −2.10 dB
B L_{0_B} = −1.67 dB L_{10_B} = −0.88 dB
B L_{0_B} = −1.11 dB L_{10_B} = −0.26 dB
MAIN REFLECTOR IMAGE FOR 1.5° ELEVATION AND 3° AZIMUTH SCAN

Fig. 6—Vignetting at the array plane for various scan angles.

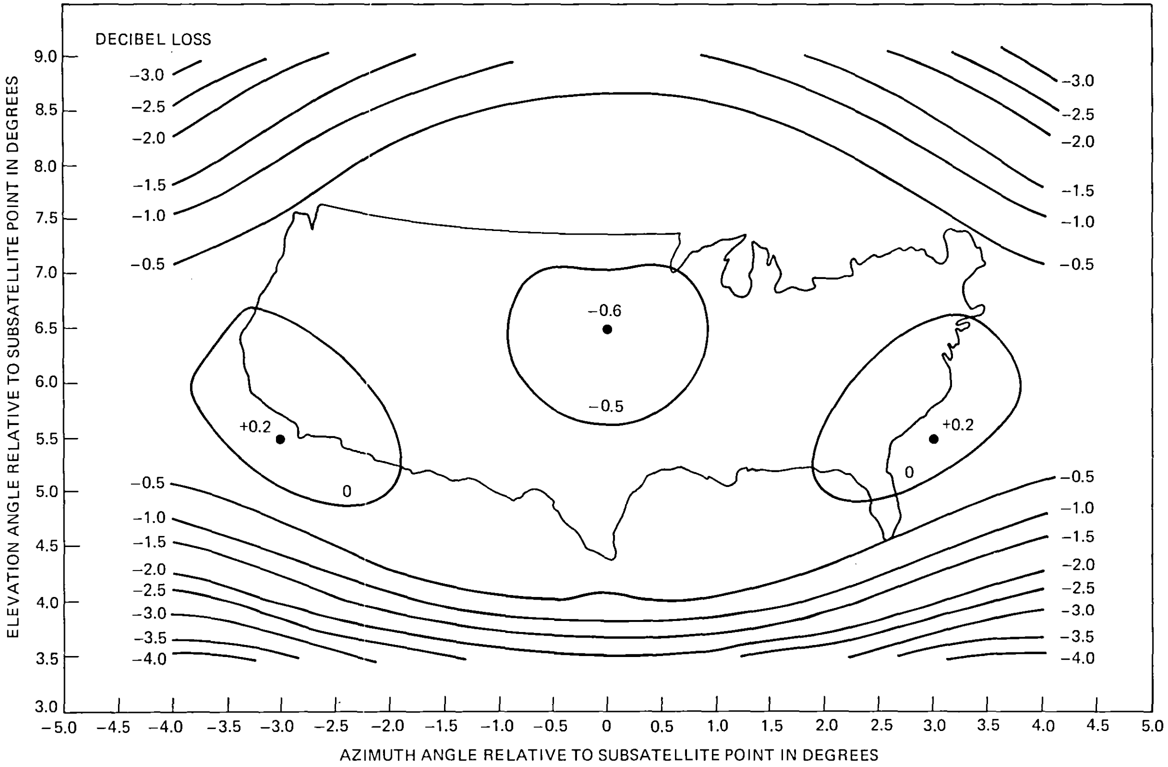


Fig. 7—Contours of scan loss due to imperfect illumination over the array aperture.

when the aperture diameter D_0 is large. The requirements on surface accuracy for the main reflector are greatly reduced because of the ability of the array to correct efficiently for small surface imperfections or reflector displacements. In order for this to work, the array and the reflector must be conjugate elements; i.e., condition (3) is required. This further assures that the transformation relating E_1 to E_0 is essentially frequency independent.

REFERENCES

1. D. O. Reudink and Y. S. Yeh, "A Rapid Scan Area-Coverage Communication Satellite," B.S.T.J., *56,* No. 8 (October 1977), pp. 1549–1561.
2. W. D. Fitzgerald, "Limited Electronic Scanning with an Offset Feed Near-Field Gregorian System," M.I.T. Lincoln Laboratory, Tech. Rep. 486, September 24, 1971, DDC AD-736029.
3. Ming H. Chen and G. N. Tsandoulas, "A Dual-Reflector Optical Feed for Wide-Band Phased Arrays," IEEE Trans. Ant. Prop., *AP-22,* No. 4 (July 1974), pp. 541–545.
4. J. W. Goodman, *Introduction to Fourier Optics,* New York: McGraw-Hill, 1968.
5. C. Dragone, "An Improved Antenna for Microwave Radio Systems Consisting of Two Cylindrical Reflectors and a Corrugated Horn," B.S.T.J., *53,* No. 7 (September 1974), pp. 1351–1377.
6. Hirokawa Tanaka and M. Mizusawa, "Elimination of Cross-Polarization in Offset Dual-Reflector Antennas," Electronics and Communications in Japan, *58-B,* No. 12, 1975.
7. C. Dragone, "Offset Multireflector Antennas with Perfect Pattern Symmetry and Polarization Discrimination," B.S.T.J., *57,* No. 7 (September 1978), pp. 2663–2684.
8. A. A. M. Saleh and R. A. Semplak, "A Qausi-Optical Polarization Independent Diplexer for Use in Beam Feed Systems of Millimeter-Wave Antennas," IEEE Trans. Ant. Prop., *AP-24,* No. 6 (November 1976), pp. 780–784.
9. J. A. Arnaud and F. A. Pelow, "Resonant-Grid Quasi-Optical Diplexers," Elec. Lett., *9,* No. 25 (December 13, 1973), pp. 589–590.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 58, No. 2, February 1979
Printed in U.S.A.

Properties of Kruithof's Projection Method

By R. S. KRUPP

(Manuscript received July 24, 1978)

In 1937, J. Kruithof introduced a scheme for projecting from measured point-to-point teletraffic data to some future values, based upon estimates of total originating and terminating traffic only. This study seeks to give a unified picture of Kruithof's projection method and its generalizations, with some practical details and recommendations for implementation. The main text deals with existence and convergence testing, treatment of ill-conditioned or slowly convergent cases, and various extensions of the basic method. An appendix includes proofs of existence, uniqueness, convergence, and continuity.

I. INTRODUCTION

J. Kruithof's method[1] for projecting from measured point-to-point teletraffic data q_{ij} to some future values p_{ij} is based upon estimates of total originating and terminating traffic only. While the original publication (in Flemish) did not receive as much attention as it may deserve, the idea was good enough that it has been independently reinvented numerous times in the intervening years. Related techniques have turned up in economics, statistics, biophysics, pattern recognition, and vehicular traffic studies, for instance. Such repetition largely seems due to a scientific "Babel" effect: workers in different technical disciplines can no longer read each other's work and recognize the same problem in a new context.

While Kruithof showed that his method had certain properties which are clearly desirable in a projection scheme, he did not investigate the underlying mathematical problems. Subsequent workers, such as Bear,[2] Kullback,[3] Sinkhorn,[4,5] Theil,[6] and particularly Csiszar,[7] have thrown much light on these matters. This paper seeks to give a unified picture of Kruithof's method and its many generalizations, with some practical details and recommendations for implementation. Much of the more intricate mathematics is relegated to an appendix, including proofs of existence, uniqueness, convergence, and continuity. These are cited as needed in the main text, which contains information on existence and convergence testing, treatment of ill-conditioned and slowly convergent cases, and various extensions of the basic method. The final comments propound a rationale for Kruithof projection in the context of Bell System planning.

II. KRUITHOF'S BASIC METHOD

In 1937, J. Kruithof[1] proposed a technique for predicting the point-to-point traffic p_{ij} in a given year from a number of originating points $i = 1, 2 \cdots M$ to a number of terminating points $j = 1, 2 \cdots N$. The units could be calls, trunks, or erlangs, for instance, as long as it makes sense to add up various entries in the matrix $\mathbf{p} = [p_{ij}]$. It is assumed that the corresponding traffic matrix $\mathbf{q} = [q_{ij}]$ is known for some other year, while total traffic b_i at each point i and d_j at each j have already been estimated by some external means to yield:

$$b_i = \sum_j p_{ij} \tag{1}$$

$$d_j = \sum_i p_{ij}. \tag{2}$$

Then Kruithof's formula for projecting $\mathbf{p}$ from $\mathbf{q}$ is:

$$p_{ij} = q_{ij} E_i F_j. \tag{3}$$

The "growth factors" E_i and F_j for the originating and terminating points are implicitly defined, and must be computed by solving (1)–(3) simultaneously.

Kruithof recommended that (1)–(3) be solved by starting with the estimate $\mathbf{p} = \mathbf{q}$, then alternately normalizing the rows of $\mathbf{p}$ to satisfy (1) and the columns to satisfy (2), until it stops changing. In practice, this scheme suffers from a tendency to accumulate roundoff error. A mathematically equivalent procedure with better numerical properties would be to substitute (3) into (1) and (2), solving for E_i and F_j to obtain:

$$E_i = b_i / \sum_j q_{ij} F_j \tag{4}$$

$$F_j = d_j / \sum_i q_{ij} E_i. \tag{5}$$

Starting from an arbitrary estimate, such as $F_j = 1$ for all j, (4) and (5) may be evaluated alternately until $\mathbf{p}$ converges.

Various questions arise naturally in connection with this projection method:

(*i*) Under what conditions do (1)–(3) possess a solution $\mathbf{p}^*$?
(*ii*) Can there be more than one solution for $\mathbf{p}^*$?
(*iii*) How does $\mathbf{p}^*$ vary with the estimates of total traffic?
(*iv*) Does the iteration converge, and if so, to what?

(v) When should we stop iterating a given case?
(vi) Are there valid generalizations of this scheme?

Many of these points are treated at considerable length in the appendix, which discusses a generalized problem: find a probability distribution **P** which satisfies arbitrary linear constraints, such as (1)–(2), and is related to a given distribution **Q** by a product formula, such as (3). To make the connection in the present case, our first step is to divide **p** by the sum $\hat{p}$ of all its elements, reducing it to a joint probability $\mathbf{P} \equiv \mathbf{p}/\hat{p}$ that a call, trunk, or other increment of traffic is from i to j. Now (1) and (2) yield relations:

$$B_i \equiv b_i/\hat{p} = \sum_j P_{ij} \qquad (1')$$

$$D_j \equiv d_j/\hat{p} = \sum_i P_{ij}, \qquad (2')$$

with B_i and D_j the marginal distributions of traffic on i and j. Similarly, **q** is divided by the sum $\hat{q}$ of its elements to get the joint distribution $\mathbf{Q} \equiv \mathbf{q}/\hat{q}$. Let the events $\{e\}$ in the appendix be the set of pairs $e = (i, j)$ and the constraints $\{c\}$ corresponding to (29) be (1′)–(2′) for all points i and j. Then taking $E_i = V_i\hat{p}/\hat{q}$ and $F_j = W_j$ puts the projection formula (3) in exactly the product form (32):

$$P_{ij} = Q_{ij}E_iF_j\hat{q}/\hat{p} = Q_{ij}V_iW_j, \qquad (3')$$

so that we have a case of the general Kruithof problem defined in the appendix.

Now the existence conditions from the appendix show that there is a solution $\mathbf{p}^*$ of the form (3) if and only if (1) and (2) have some solution p_{ij} that vanishes for each q_{ij} that vanishes and is positive whenever q_{ij} is positive. The uniqueness results show there is at most one solution $\mathbf{p}^*$. The solution is continuous in all b_i and d_j whenever it exists. The iteration on (4) and (5) can be recognized as an example of a relaxation procedure, as discussed and analyzed in the appendix. If (1)–(2) possess a solution p_{ij} that is zero for each q_{ij} that is zero, the iteration will converge to some **p** with these same properties. The limit may not be of the form (3) though, since p_{ij} can also vanish for some $q_{ij} > 0$. When a solution $\mathbf{p}^*$ to (1)–(3) exists, however, the iteration converges to it uniquely. This explains the resistance of (4)–(5) to roundoff effects, since such perturbations die out in the process of converging. In the special case that q_{ij} is symmetric and $b_i = d_i$ for all i, uniqueness shows that p_{ij} is also symmetric, since $\mathbf{p}^*$ and its transpose both satisfy (1)–(3). We should note that only $\mathbf{p}^*$ is unique, not the factors E_i and F_j in (3). For instance, replacing them by E_i/a and aF_j for all i, j, and any $a > 0$ will produce the same $\mathbf{p}^*$. The extent of this nonuniqueness is characterized later.

Kruithof pointed out two desirable properties possessed by the projection scheme. The first, called "reversibility," says that after projecting traffic from **q** at time 1 to **p** at time 2, we can turn around and project backward to time 1, recovering **q** exactly. The second property, "divisibility," says projecting from **q** to **p** and then from **p** to **r** at time 3 yields just the same result as projecting directly from **q** to **r**. Thus the projection scheme is not noisy, in the sense of irretrievably losing information along the way about the initial traffic. Rather, **p** depends only on **q** and the row and column sums, not on the path followed over time. Kruithof also described a third desirable property, "separability," which did not hold for his basic method. The idea is to be able to merge or split a collection of points i or j, without affecting the projected traffic for any other points. By careful generalization of Kruithof's method, a property similar to this can be introduced.

III. NETWORK FLOW CONSIDERATIONS

An important aspect of Kruithof's method may be visualized by means of a flow on the simple directed graph in Fig. 1. The nodes $i = 1, 2 \cdots M$ and $j = 1, 2 \cdots N$ represent originating and terminating points. The edge joining node i to node j carries traffic p_{ij}. Interpreting (1)-(2) as conservation laws, the remaining edges to source s and sink t carry total traffic quantities b_i and d_j, while the net flow from s to t is the sum $\hat{p}$ of all the flows p_{ij}. When an element q_{ij} vanishes in **q**, the corresponding edge from i to j is deleted in the network, so that flow p_{ij} is automatically zero. For the flow **p** to have the form (3), all remaining edges must have nonzero flow p_{ij} on them. Conversely, if a flow p_{ij} can be constructed that satisfies (1)-(2) and does not vanish on any edge of the network, then it fulfills the existence conditions, so that (1)-(3) have a solution $\mathbf{p}^*$.

The labeling method of Ford and Fulkerson[8] immediately springs to mind as a means of constructing a flow p_{ij} to satisfy (1)-(2). This is a simple, efficient, easily programmed algorithm that maximizes the net flow from s to t. We just assign maximum capacities of b_i for edges from s to i, infinity for edges from i to j, and d_j for edges from j to t. If the maximum flow obtained is less than:

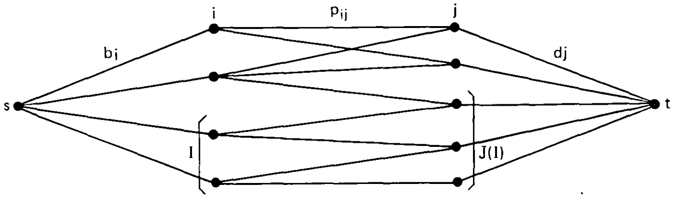


Fig. 1—Network model.

$$\hat{p} = \sum_i b_i = \sum_j d_j, \tag{6}$$

then no such solution of (1)–(2) exists. The only trouble is that the flow obtained may not be positive on all edges, as required to establish that (1)–(3) have a solution. This can be cured by assigning a sufficiently small lower bound $p_{ij} \geqslant h > 0$ to flow on the edges from i to j. An initial flow of h on each such edge will then be feasible.

In practice, the total traffic quantities b_i and d_j are integers or can be scaled up and rounded off to an approximate integer problem without loss of credibility. The labeling method may then be modified to carry h as an infinitesimal quantity, while preserving pure integer arithmetic for efficiency and to avoid roundoff error. Indeed, since one wants to draw a yes-or-no conclusion about existence of a solution, roundoff introduces an unwelcome uncertainty. In the modified scheme, all capacities and flow variables are carried as pairs (m,n) of integers, representing the expression $m + nh$. In the steps of the basic labeling algorithm,[8] one adds and subtracts various capacities and flows, attaching labels to nodes according to whether or not the result exceeds zero. In the modification, the obvious rules apply for adding and subtracting quantities $m + nh$; it remains, however, to specify an interpretation for inequalities involving such quantities. Two classes of inequalities must be defined:

(*i*) $m + nh > 0(1)$ means $m \geqslant 1$.
(*ii*) $m + nh > 0(h)$ means $m \geqslant 1$ or $m = 0$ but $n \geqslant 1$.

In designating h an infinitesimal, we are really promising to choose it as small as necessary so that $|nh| < 1$ for all n that arise.

The solution proceeds in two phases. First the standard labeling method is used, but with all inequality tests to be taken as $> 0(1)$. Thus, in each iteration, an augmenting path of capacity $> 0(1)$ is found. The net flow increases by at least one unit (plus or minus some nh), so that this phase terminates after a finite number of iterations. At this point, the net flow must be $\hat{p} - \hat{n}h$ for some $\hat{n} \geqslant 0$, or else no flow solution of (1)–(2) exists; in effect, we have the maximum flow for the special case $h = 0$. The second phase repeats the standard labeling method, but with all inequality tests to be taken as $> 0(h)$. Now each augmenting path has capacity $> 0(h)$, increasing net flow by at least h units and, again, the iterations terminate in a finite number of steps. If the maximum flow reaches $\hat{p}$, then a solution $\mathbf{p}$ of (1)–(2) has been constructed that fulfills the existence conditions; otherwise, no such solution exists. The two-phase algorithm was programmed, for the network associated with an arbitrary $M \times N$ matrix $\mathbf{q}$, in about eighty lines of Fortran. In tests, it was able to settle the question of existence of solutions to specific Kruithof problems with gratifying rapidity. The same technique can be used to maximize flow on any network under any mixture of $>$ and $\geqslant$ capacities.

Still more can be gleaned from the network model above. Observe that in (6) the sum of the b_i must equal the sum of the d_j in order to conserve flow. This requirement is just a simple example of a class of necessary conditions arising from (1)–(3). More generally, let I be any subset of originating nodes i and $J(I)$ be the subset of terminating nodes j with edges to nodes in I. That is, node j is in $J(I)$ if q_{ij} is positive for some node i in I. Then the flow b_i to node i in I can only pass to nodes in $J(I)$, so that it is included in the flows d_j out of these nodes. Thus the total flow $Y(I)$ from nodes in $J(I)$ cannot be less than the total flow $X(I)$ to all the nodes in I:

$$X(I) \equiv \sum_{I} b_i \leqslant Y(I) \equiv \sum_{J(I)} d_j \tag{7}$$

for each proper subset I (that is, I not empty and not containing all i). Conversely, these necessary conditions guarantee that every cut has a capacity of at least $\hat{p}$, so that (6)–(7) are also sufficient conditions for existence of a flow p_{ij} that satisfies (1)–(2).

Accounting for eq. (3), the requirement of positive flow on each edge allows conditions (7) to be strengthened. Indeed, flow $Y(I)$ from $J(I)$ includes the flow $X(I)$ into I plus any additional flow to $J(I)$ from nodes i that are not in I. If any edge joins a node outside I to $J(I)$, then its flow is positive and (7) becomes a strict inequality:

$$X(I) < Y(I). \tag{8}$$

If equality holds in (7), the rest of the network has no connection to I and $J(I)$, except through s and t. Such a disconnected situation represents two or more independent Kruithof problems, which ought to be treated separately from the outset. Indeed, the rows and columns of q_{ij} can then be renumbered so that it is partitioned into two or more uncoupled blocks. To simplify the statement of later results, we assume that the problem does not decompose in this way, so that the network is connected and (8) holds for every proper subset I of the nodes i. In complete analogy to the case of (6)–(7), necessary conditions (6)–(8) are also sufficient for existence of a positive flow satisfying (1)–(2), and hence for existence of $\mathbf{p}^*$. To see this, we first reduce all network flows and capacities by the initial feasible flow h used in the modified labeling method. For sufficiently small $h > 0$, conditions (7)–(8) now apply to the reduced b_i and d_j as well. But these again show enough capacity on every cut that a flow $p_{ij} - h \geqslant 0$ exists satisfying (1)–(2).

Conditions (6)–(8) can also be used in a direct proof that $\mathbf{p}^*$ exists, independently of results in the appendix. The idea is to seek a stationary point of the quotient Num/Den of two multinomials in the growth factors $F_j \geqslant 0$. The numerator is positive on the interior and vanishes at boundaries (that is, where some $F_j = 0$):

$$Num \equiv \prod_j F_j^{d_j}, \tag{9}$$

while the denominator is also positive on the interior:

$$Den \equiv \prod_i \left[\sum_j q_{ij} F_j \right]^{b_i}. \tag{10}$$

Setting a derivative of Num/Den with respect to F_j to zero yields the same result as substituting (4) into (5). Thus, if the values F_j make Num/Den stationary, they will also yield a solution $\mathbf{p}^*$ of (1)–(3). Note that Num and Den are both homogeneous of order $\hat{p}$ from (6), so that their quotient is positive and constant along interior rays (that is, along aF_j for all $a > 0$). Since the rays form a compact set and Num/Den is positive and continuously differentiable on the interior, it is enough to show that the quotient goes to zero on the boundary to deduce that it achieves a (stationary) interior maximum. Now suppose that Den vanishes at some boundary point $\mathbf{F}$ and let I be the set of nodes i for which the factor $\sum q_{ij}F_j$ in (10) is zero. Then F_j must vanish if q_{ij} is positive for some i in I, and hence for every j in $J(I)$. If Z measures distance from an interior point $\mathbf{F}'$ to boundary point $\mathbf{F}$, then the numerator will vanish as $\mathbf{F}'$ approaches $\mathbf{F}$ at least as fast as $Z^{Y(I)}$, while the denominator goes like $Z^{X(I)}$, from (7). Thus Num/Den approaches zero at each boundary point $\mathbf{F}$, from (8), completing the proof. The enterprising reader may find it instructive to ferret out the connection between the preceding proof of existence and the more general proofs in the appendix.

One immediate consequence is that the Kruithof problem always has a solution $\mathbf{p}^*$ if $\mathbf{q}$ is strictly positive and (6) holds. Indeed, each node i has an edge to every node j, so that $Y(I) = \hat{p}$ for every proper I and (8) follows. To verify this case more directly, one can construct the positive flow $p_{ij} = b_i d_j / \hat{p}$, which satisfies (1)–(2). Another useful example is a square matrix q_{ij} with zeros only on the diagonal. Then $Y(I) = \hat{p}$ for every I with two or more nodes, so that only the unit sets i must be checked. Conditions (8) now reduce to the requirement $b_i + d_i < \hat{p}$ for every node i. An important application involves choosing all b_i and d_j equal to one, so that $M = \hat{p} = N$ and $\mathbf{q}$ must be reduced to a doubly stochastic matrix $\mathbf{p}^*$. But (7) says that $X(I)$ and $Y(I)$ are just the sizes $|I|$ and $|J(I)|$ of I and $J(I)$, respectively. A result from matching theory, the "marriage theorem,"[9] now shows that the conditions $|I| \leqslant |J(I)|$ from (7) are equivalent to $\mathbf{q}$ having some positive principal diagonal, and that $|I| < |J(I)|$ from (8) imply that each positive element of $\mathbf{q}$ lies on a positive principal diagonal. These are the conditions cited by Sinkhorn and Knopp[5] for the doubly stochastic case. In a typical case, the modified labeling method will ordinarily be easier to apply than (6)–(8) in testing for existence of a solution.

The growth factors $\mathbf{E}$ and $\mathbf{F}$ are essentially unique, except for the possibility of scaling them along rays as described earlier. Indeed, let $\hat{\mathbf{E}}$ and $\hat{\mathbf{F}}$ satisfy $p_{ij} = q_{ij}E_iF_j = q_{ij}\hat{E}_i\hat{F}_j$ for all i and j. Then $\hat{E}_i/E_i = F_j/\hat{F}_j = a$ whenever q_{ij} is not zero, and the factor a is independent of i and j since the network is connected. When the problem decomposes into independent Kruithof subproblems, the network breaks into separate connected components, each with a separate scaling factor a. The case that $\mathbf{p}$ and $\mathbf{q}$ are symmetric implies that $\mathbf{E} = a\mathbf{F}$, so that the growth factors may be scaled to satisfy $\mathbf{E} = \mathbf{F}$ uniquely.

IV. CONVERGENCE CONSIDERATIONS

In iterating (4) and (5) to solve for $\mathbf{p}^*$, a good convergence test can be based on the norm consisting of the sum of the absolute values of the differences between the left and right sides of (1) and (2):

$$g = \sum_i | b_i - E_i \sum_j q_{ij}F_j | + \sum_j | d_j - F_j \sum_i q_{ij}E_i |. \tag{11}$$

Clearly, g is the net error in traffic units and goes to zero as $\mathbf{p}$ goes to the solution $\mathbf{p}^*$; we can show that, in fact, g does so monotonically. Indeed, g is continuous, convex, and piecewise linear in $\mathbf{E}$, with E_i-derivatives of the form:

$$-\sum_j q_{ij}F_j\left[\operatorname{sgn}\left(b_i - E_i \sum_{j'} q_{ij'}F_{j'}\right) + \operatorname{sgn}\left(d_j - F_j \sum_{i'} q_{i'j}E_{i'}\right)\right],$$

and this expression always takes the sign of $E_i \sum q_{ij}F_j - b_i$ or else vanishes. Thus g can only decrease or remain constant as E_i is increased or decreased to satisfy (4) and g achieves its minimum, for any fixed value of $\mathbf{F}$, when each E_i is given by (4). A similar discussion holds with respect to the F_j-derivative of g. Accordingly, when g becomes small during iteration, it will remain so, and it is appropriate to stop.

During the process of iterating to compute $\mathbf{E}$ and $\mathbf{F}$, we can accumulate a value of g with very little additional effort. Specifically, when all the F_j have a new value, the second term of (11) vanishes. Proceeding to update the E_i, we compute $\sum q_{ij}F_j$ for use in (4). With two more additions and one multiplication, we get the corresponding contribution to g in the first term of (11). By the time all new E_i are computed, a value of g for $\mathbf{F}$ and the old $\mathbf{E}$ is available. Clearly, the iteration may be interpreted as a relaxation scheme to minimize g by cyclically minimizing over the E_i and F_j.

Since nonexistence of a solution $\mathbf{p}^*$ when (6) holds can only accompany zero elements in $\mathbf{q}$, one might attempt to force a solution by substituting a small positive value for each zero. In general, this is a terrible idea; Sinkhorn,[4] for instance, shows some examples of pathological behavior associated with such schemes. The iteration process will seek to get significant flow on some edges with small q_{ij} by using very large growth factors. The solution **p*** will not greatly resemble **q** and the iteration will ordinarily converge very slowly. Indeed, slow convergence is a possible warning that the problem is ill-conditioned in some way. For such cases, it is prudent to check the "existence margin" by setting successively larger elements of **q** to zero and running the existence test until it fails—just the reverse of adding small positive terms. The labeling method is sufficiently economical of computing time that it may be repeated often in preference to performing a great many iterations.

In general, ill-conditioning occurs when some of the total traffic values b_i and d_j are not particularly consistent with one another. That is, when some collection of small elements q_{ij} are set to zero, (1)–(2) no longer have any solution for which p_{ij} vanishes if and only if q_{ij} does. The limit as these elements approach zero may not exist, or it may depend on the specific way in which they vanish. In short, **p*** is not necessarily continuous in the elements of **q** at the boundaries of the feasible region. On the other hand, **p*** is provably continuous in the row and column sums, or any other constraint levels, so that adjusting them to achieve consistency at the boundary is a stable procedure. Thus, the proper way to treat ill-conditioning is by readjusting some values of b_i and d_j to make them more consistent, though it may not be obvious how to do this. We see later that there is an easy way to extend the Kruithof method so that it automatically allocates traffic among the rows or columns of prespecified aggregations in a consistent and reasonable way.

It is also possible for a quite reasonable problem to converge at a very slow rate. For example, a problem may decompose into multiple independent subproblems, whose network components are only connected by way of s and t. Now, each of the subproblems may be well-behaved, so that the overall iteration process converges rapidly. Nevertheless, when a few small positive values of q_{ij} are introduced to couple the subproblems, convergence will be rapid at first and then become rather slow, as a rule. Moreover, the solution to which the problem converges is a sensible one that differs only slightly from the decoupled case. This model could apply to two or more countries, for example, with much more traffic internally than across their borders.

What causes the above difficulty is the difference in degree of uniqueness between coupled and decoupled cases. For n subproblems, n independent arbitrary scaling factors a will appear in the general solution. The actual values they assume will be determined by the initial values assigned to **E** or **F**. This effect appears as some arbitrariness in the relative sizes of the E_i for those portions of **E** associated with the various subproblems. For the coupled case, only a single overall scaling factor is appropriate; the portions of **E** from the different subproblems must now be scaled in a correct ratio to each other. In the rapidly convergent phase, each subproblem is solved to within its scale factor; the slow phase corresponds to a process of adjusting scale factors to account for small coupling terms.

The slowly convergent phase may be shortened appreciably by means of standard numerical schemes for acceleration of convergence. Wynn's algorithm,[10] in particular, has been employed successfully to project the values of the E_i for successive steps, in order to estimate their limit. The strategy is to calculate the norm g for the projected E and, when this is much less than the current error g (one-fiftieth, say), restart the iteration from the projected value. Additional computational effort and program steps for a convergence acceleration option in Kruithof's method are minor. Of course, acceleration techniques cannot rescue a truly ill-conditioned case, where the existence margin is small. All this is not meant to imply that slow convergence is the rule with the Kruithof method. In fact, some rather large examples, involving several hundred rows and columns, have been solved quite readily.

V. VARIOUS EXTENSIONS

An immediate generalization of the basic Kruithof scheme would be to stratify the traffic data in more than two dimensions. Besides the originating and terminating points i and j, other indices k, l, m $\cdots$ might specify time of day, week, or year, type of traffic (business or residential, for instance), and so on. A three-dimensional case of the general Kruithof problem then might take the specific form:

$$p_{ijk} = q_{ijk}E_iF_jG_k \tag{12}$$

$$b_i = \sum_{j,k} p_{ijk} = E_i \sum_{j,k} q_{ijk}F_jG_k \tag{13}$$

$$d_j = \sum_{i,k} p_{ijk} = F_j \sum_{i,k} q_{ijk}E_iG_k \tag{14}$$

$$f_k = \sum_{i,j} p_{ijk} = G_k \sum_{i,j} q_{ijk}E_iF_j. \tag{15}$$

Reduction to the standard case in the appendix proceeds as before. We define events $e = (i, j, k)$ and constraints c corresponding to each value of i, j, and k, while $\mathbf{p}$ and $\mathbf{q}$ are normalized to probabilities by dividing by the sums of their elements.

The proofs of existence, uniqueness, and convergence in the appendix still hold for this case. In particular, a solution $\mathbf{p}^*$ of (12)-(15) exists if and only if (13)-(15) have a solution p_{ijk} that is positive or zero accordingly as q_{ijk} is positive or zero. The norm g consisting of the sum of absolute values of differences between left and right sides in (13)-(15) is still net error in traffic units, though it no longer decreases monotonically with each new value of E_i, F_j, or G_k. Nevertheless, g

goes to zero as **p** goes to **p*** and it is still a reasonable indicator of convergence. When **q** is strictly positive, the interior solution $p_{ijk} = b_i d_j f_k/\hat{p}^2$ of (13)–(15) demonstrates that **p*** also exists, provided that the b_i, d_j, and f_k each add up to $\hat{p}$. A more general existence test would involve solving the linear programming problem described in the appendix, with appropriate precautions against being misled by the effects of any roundoff error. Network flow models no longer apply, and integer solutions need not occur in the case of integer constraints.

Another three-dimensional example of the general problem might be as follows:

$$p_{ijk} = q_{ijk}E_{ij}F_{jk}G_{ik} \tag{16}$$

$$b_{ij} = \sum_k p_{ijk} = E_{ij} \sum_k q_{ijk}F_{jk}G_{ik} \tag{17}$$

$$d_{jk} = \sum_i p_{ijk} = F_{jk} \sum_i q_{ijk}E_{ij}G_{ik} \tag{18}$$

$$f_{ik} = \sum_j p_{ijk} = G_{ik} \sum_j q_{ijk}E_{ij}F_{jk}, \tag{19}$$

and the same sort of discussion applies to this case as to (12)–(15). For four or more dimensions, the reader should have no difficulty creating a great many extensions of this general class, such as $p_{ijkl} = q_{ijkl}B_{ij}C_{jk}D_{kl}E_{il}F_{ik}G_{jl}$. The rule is to multiply $q_{ijkl\ldots}$ by one factor $E_{ij\ldots}$ for each constraint in which $p_{ijkl\ldots}$ appears, and then solve each constraint for its factor by dividing into the constraint level $b_{ij\ldots}$. In higher dimensions, the number of elements in **p** and **q** grows much faster than the numbers of constraints and multipliers. The numbers of the latter thus remain reasonable, if only to stay within storage and computational limits on the former. The corresponding linear program to test for existence will therefore have a basis of reasonable size, as well.

Another class of extensions involves specifying less about total traffic quantities in the two-dimensional case (or any other dimension, using the previous generalization). That is, various sets I or J of originating or terminating points i or j may be lumped together, with only their total traffic $\tilde{\mathbf{b}}$ and $\tilde{\mathbf{d}}$ to be supplied externally. Now (1)–(2) yield the following constraints:

$$\tilde{b}_I \equiv \sum_I b_i = \sum_I \sum_j p_{ij} \tag{20}$$

$$\tilde{d}_J \equiv \sum_J d_j = \sum_J \sum_i p_{ij}. \tag{21}$$

Results on the general Kruithof problem in the appendix show that $E_i = \tilde{E}_I$ and $F_j = \tilde{F}_J$ in this case, so that (3) becomes:

$$p_{ij} = q_{ij}E_iF_j = q_{ij}\tilde{E}_I\tilde{F}_J, \tag{22}$$

assuming each i belongs to one I and each j to one J.

To solve for $\tilde{\mathbf{E}}$ and $\tilde{\mathbf{F}}$, we can collapse the problem to a simpler form: add together all rows i in I and all columns j in J to work with reduced matrices $\tilde{\mathbf{p}}$ and $\tilde{\mathbf{q}}$, as follows:

$$\tilde{p}_{IJ} \equiv \sum_I \sum_J p_{ij} \qquad \tilde{q}_{IJ} \equiv \sum_I \sum_J q_{ij}. \tag{23}$$

Now (20)-(22) are reduced to exactly the same form as (1)-(3):

$$\tilde{b}_I = \sum_J \tilde{p}_{IJ} \qquad \tilde{d}_J = \sum_I \tilde{p}_{IJ} \qquad \tilde{p}_{IJ} = \tilde{q}_{IJ} \tilde{E}_I \tilde{F}_J, \tag{24}$$

but with many fewer constraints to be satisfied. Note that this procedure of aggregating traffic nodes may absorb some elements of $\mathbf{q}$ that were zero or small, in order to ameliorate ill-conditioning. Indeed, the corresponding disaggregation formula (22), to be used after we have solved (24), simply allocates traffic p_{ij} to the element q_{ij} in proportion to its relative contribution to $\tilde{q}_{IJ}$ in (23). This in turn automatically shares out $\tilde{b}_I$ and $\tilde{d}_J$ among their b_i and d_j in a consistent manner, as mentioned earlier. As another of its virtues, aggregation is a smoothing process that can cut down the effects of errors in the predictions of total traffic by reducing the number of independent parameters. The reduction in manual and computational effort is also an evident advantage. To organize the computation in an efficient manner, we would start with two tables, $I(i)$ and $J(j)$, that assign each point i or j to its appropriate aggregate. Then we run through the pairs (i, j), adding each q_{ij} into its correct $\tilde{q}_{I(i)J(j)}$. After we solve (24) for $\tilde{\mathbf{E}}$ and $\tilde{\mathbf{F}}$, the answer is just $p_{ij} = q_{ij} \tilde{E}_{I(i)} \tilde{F}_{J(j)}$ from (22). Existence testing can be performed directly on $\tilde{\mathbf{q}}$, $\tilde{\mathbf{b}}$, and $\tilde{\mathbf{d}}$ with the labeling method.

The scheme above illustrates a sense in which a "separability" property can be introduced, similar to what Kruithof sought. There is no real need to require that different I or J be disjoint. If overlap is permitted, then q_{ij} would be multiplied by $\tilde{E}_I$ for each I that contains i, and by $\tilde{F}_J$ for each J containing j. Collections of rows or columns can now be aggregated only if they all lie in the same sets I or J. Since some p_{ij} may now be counted twice, the constraints can no longer be interpreted as conservation laws for a flow, and the existence test becomes a linear program, as described in the appendix.

Another way of specifying less in the Kruithof problem is to leave some rows or columns unconstrained. At such i and j, we can specify the growth factors E_i and F_j arbitrarily; Bear[2] has considered choosing these multipliers to be one. One advantage of not constraining is that any element q_{ij} for which row i and column j are not constrained plays no part in the solution process and may be set to zero for convenience. For computational efficiency, we multiply each of these rows by its fixed growth factor and add together all such rows to form a single new row; similarly, all unconstrained columns combine. Of course, fewer specifications may be introduced in higher-dimensional schemes as well.

Anyone can tailor their own ad hoc constraints to account for additional knowledge of the future. For instance, it may be known that certain items of point-to-point traffic p_{ij} are growing considerably faster than the other, more typical items in their rows and columns. Then a constraint may be created to fix the projected value of the sum of all such items. This produces one more growth factor to be multiplied into q_{ij} for these selected items only. Similar treatment may be given to a class of items that have slower than normal growth, that decrease, or that just vanish. The general Kruithof problem defined in the appendix includes all such cases, as well as any other linear constraints that may need to be introduced.

VI. SUMMARY AND COMMENTS

In this study, we have sought to give a unified view of Kruithof's teletraffic projection method, including theoretical aspects as well as practical details for its implementation. The mathematics in the appendix treats the theory of a general Kruithof problem. Necessary and sufficient conditions for existence and convergence of its solution are derived, along with proofs of uniqueness and continuity. The main text considers special cases of this problem that are of particular interest. Schemes for existence and convergence testing and for handling slow convergence are discussed. We conclude by trying to place this projection scheme in the context of the Bell System planning function.

An early and important step in the Bell System planning process is that of predicting future demand for the various services offered. By their nature, such projections can be quite uncertain, since they will include cumulative effects of several years' fluctuations in the United States and world economy, for instance. Indeed, analysis of time series of typical traffic data[11] indicates that about five percent per year of random error remains in even the best projections, and must be regarded as inherently unpredictable. Nevertheless, a strategy is available to cope with such uncertainties, for the purposes of planning.

The fundamental assumption required in this strategy is that traffic increases monotonically with time. First a plan can be generated, based upon some "best guess" of the demand profile over time. From year to year, the time scale of the plan can then be corrected to match up the originally projected demand with actual values or better estimates, based upon more recent data. In effect, a parameter such as total traffic is thus used as a new independent variable in the plan, while time is a dependent variable that absorbs much of the economic fluctuations and other error. However, this leads us to view the overall process of planning as a system, rather than a collection of independent modules, one of which is projection. We see that a "sliding time scale" approach to planning now places a premium not so much on absolute accuracy of traffic predictions over time as on "relative accuracy" or consistency and uniformity among the various traffic quantities that are projected.

A typical plan may be based upon many thousands or tens of thousands of projected traffic items. We wish to predict these quantities such that a single readjustment of the plan time scale (based on some total traffic measure, for instance) can do a reasonable job of correcting for the error in each item. This requires that projection be done in such a way that the prediction errors in individual traffic items will tend to be highly correlated. Thus, a scheme which projected individual time series for each separate traffic quantity, for example, might give the best absolute accuracy for each item, but still be unsuitable for planning purposes because the noise components in these time series would tend to be independent. At the opposite extreme, initial measurements of all traffic quantities might simply be increased by a single overall growth factor for each year under study. This would produce very strong correlations but fails to take into account detailed knowledge of growth patterns, say, for separate portions of a study area.

The general Kruithof method offers many middle roads. Any collection of average or overall traffic quantities **b** may be predicted externally (from time series, for example, or market surveys). As shown in the appendix, the remaining items can then be projected to be consistent with whatever is given. Inserting a great many external predictions introduces more detailed knowledge, but reduces the correlation. Supplying fewer external specifications yields stronger correlations, at some loss in accuracy; various tradeoffs are possible.

All the schemes of Kruithof type act to minimize the net information change in the projection, subject to those external constraints being enforced. This gives them the remarkable properties called "reversibility" and "divisibility" by Kruithof. Essentially, all that is lost of the original data **q** in projecting it to future values **p** is whatever is inherent in the externally provided average quantities **b**. Supplying new values **b**′ for these quantities will thus allow us to continue the projection from **p** to another year or recover the base data **q** exactly. Effectively, Kruithof's method is able to resolve **q** and **p** into a part which determines some arbitrarily chosen system of average quantities **b** that are to be changed and an "orthogonal" part that does not change. This latter part is essentially the equivalence class $C(\mathbf{Q})$ discussed in the appendix.

Kruithof's original proposal, and much of the subsequent work on the subject, is concerned with projecting two-dimensional arrays p_{ij} of traffic data. In this case, the most natural overall quantities to be specified externally are the sums of rows i and columns j in **p**. For example, supplying two hundred parameters for a 100 × 100 matrix would suffice to project a total of ten thousand items. An option is to aggregate rows and columns, perhaps in collections of average size four, so that only fifty external sums are needed. This is a tradeoff that increases correlations but may decrease accuracy. Meanwhile, it can alleviate possible ill-conditioning, reduces manual and computational effort, and still projects ten thousand items. Examples of data organized in three or more dimensions are also known in Bell System planning. Kruithof's method generalizes to such cases without any particular difficulty.

VII. ACKNOWLEDGMENTS

Kruithof's method was first brought to my attention by Pierre L. Bastien; many of the ideas here originated in discussions with him. Various users of Kruithof's method in Bell System applications have pointed out interesting properties and extensions. Several colleagues were kind enough to read and comment upon draft copies; P. Kazakos was particularly helpful in supplying references.

APPENDIX

In this appendix, we define the general case of the Kruithof problem and derive various properties. In particular, necessary and sufficient conditions for existence and convergence of solutions are developed, and uniqueness and continuity are proved. To begin with, consider two probability distributions, P_e and Q_e, over the same finite set of disjoint events $\{e\}$, so that:

$$\sum_e P_e = 1 \qquad P_e \geqslant 0 \tag{25}$$

$$\sum_e Q_e = 1 \qquad Q_e \geqslant 0. \tag{26}$$

The information content of a probability, in appropriate units, is minus its logarithm. Thus the change in information from Q_e to P_e is just log P_e − log Q_e = $\log(P_e/Q_e)$. The average of this information change is defined as:

$$K(\mathbf{P}, \mathbf{Q}) \equiv \sum_e P_e \log(P_e/Q_e), \tag{27}$$

which will be interpreted as a measure of how close distribution **P** is to distribution **Q**. A simple example is the case that all probabilities Q_e are equal; now K reduces to a linear function of the entropy of distribution **P**. Thus entropy measures departure from the equiprobable case (corresponding to classical equilibrium). The expression (27) goes by several names in the literature; for instance, Kullback distance, I-divergence, relative entropy, discrimination information, and Gibbs free energy.

As a distance measure, K has two desirable properties: it is not negative and vanishes only when $\mathbf{P} = \mathbf{Q}$. Two other properties, symmetry and the triangle inequality, are lacking; Csiszar[7] points out, however, that laws analogous to the parallelogram identity and Pythagoras' theorem do hold, with K playing the role of squared length. To show that K is nonnegative, first note that $F(X) \equiv \log(1/X)$ is strictly convex, since $F'' = 1/X^2 > 0$. Defining $X_e \equiv Q_e/P_e$ and using convexity in (27) yields the inequality:

$$K = \sum_e P_e F(X_e) \geqslant F\left(\sum_e P_e X_e\right) = F(1) = 0, \tag{28}$$

with equality only if all $P_e F(X_e)$ vanish, in which case $\mathbf{P} = \mathbf{Q}$ from (25)-(26).

When Q_e is positive and P_e approaches zero, $P_e \log(P_e/Q_e)$ goes to a zero limit; to ensure continuity, we will define it to vanish at $P_e = 0$, even for the case $Q_e = 0$. If some $Q_e = 0$ for positive P_e, then K is infinite. Confining our attention to $\mathbf{P}$ that are not infinitely far from $\mathbf{Q}$, P_e must vanish whenever Q_e does, so that e is an event of probability zero. With no loss of generality, such e may be excluded from the set $\{e\}$ of events for now, so that all Q_e are positive.

Consider the problem of minimizing $K(\mathbf{P}, \mathbf{Q})$ for fixed $\mathbf{Q}$, over all distributions $\mathbf{P}$ subject to (25) and a finite set $\{c\}$ of arbitrary linear constraints having the general form:

$$\sum_e P_e A_{ec} = B_c. \tag{29}$$

This amounts to finding the distribution $\mathbf{P}$ that is closest to $\mathbf{Q}$ on the intersection (denoted $S(\mathbf{B})$, or just S) of the positive orthant $\mathbf{P} \geqslant 0$ and the hyperplanes (29), which prescribe that certain averages over $\mathbf{P}$ take on the values B_c. (To simplify the notation, assume that the equality in (25) is designated $\hat{c}$ and is included among the constraints c.) Observe that S is a compact convex polytope, so that the continuous function K achieves its minimum value on S, whenever S is not empty. Further, this minimum occurs at a unique point $\mathbf{P}^*$ in S, and there are no other local minima of K, since it is strictly convex. To see this, note that the matrix of second partial derivatives of K with respect to $\mathbf{P}$ is diagonal and positive definite. We assume from now on that S contains more than one point.

Suppose that S contains an interior point $\mathbf{P}^{Int}$ (that is, no P_e^{Int} vanishes); then $\mathbf{P}^*$ is also an interior point. Indeed, consider any boundary point $\mathbf{P}^{Bdy}$ and the line $\mathbf{P}^{Bdy}\mathbf{P}^{Int}$ joining it to $\mathbf{P}^{Int}$. The gradient of K has components $1 + \log(P_e/Q_e)$ that become arbitrarily large negative on some neighborhood of $\mathbf{P}^{Bdy}$ for those P_e^{Bdy} that vanish, while all other components remain bounded. Thus, a segment of $\mathbf{P}^{Bdy}\mathbf{P}^{Int}$ containing $\mathbf{P}^{Bdy}$ can be found along which K decreases toward the interior, and $\mathbf{P}^{Bdy}$ cannot be a minimum point of K. Since the gradient of K is continuous on the interior of S, an interior $\mathbf{P}^*$ is a stationary point of K and, by strict convexity, there are no other stationary points of K on S.

When $\mathbf{P}^*$ is an interior point, the minimization problem can be solved by using stationarity. Specifically, adjoin constraints (29) to K with multipliers v_c to form the Lagrangian function $L(\mathbf{P}, \mathbf{v})$, as follows:

$$L \equiv \sum_e P_e \left[\log(P_e/Q_e) - \sum_c A_{ec} v_c \right] + \sum_c B_c v_c. \tag{30}$$

Now L is strictly convex on the positive orthant $\mathbf{P} \geqslant 0$ and becomes infinite if any P_e does, while its gradient is negative infinite on the boundaries. Thus it achieves a unique minimum over $\mathbf{P} \geqslant 0$ at some interior stationary point $\hat{\mathbf{P}}(\mathbf{v}) > 0$ for any fixed values of v_c. This point is found by requiring the P_e-derivative of L to vanish for each e, yielding the following necessary conditions:

$$w_e \equiv \log(\hat{P}_e/Q_e) = \sum_c A_{ec} v_c - 1. \tag{31}$$

Setting $V_c \equiv \exp(v_c)$ in (31), except for the constraint $\hat{c}$ corresponding to (25) with $V_{\hat{c}} \equiv \exp(v_{\hat{c}} - 1)$, now yields

$$\hat{P}_e = Q_e \exp(w_e) = Q_e \prod_c V_c^{A_{ec}}, \tag{32}$$

with all V_c strictly positive, so that $\hat{\mathbf{P}}(\mathbf{v})$ cannot be on the boundary.

Suppose some v_c are found such that $\hat{\mathbf{P}}(\mathbf{v})$ satisfies the constraints (29) and thus lies in S. Then since $L = K$ on S, $\hat{\mathbf{P}}(\mathbf{v})$ is a stationary point of K on S, and hence is the unique minimum point $\mathbf{P}^*$. Conversely, the linear program of minimizing dK for all small variations $d\mathbf{P}$ that satisfy (29) has (31) as its dual constraints. Since the primal problem has $d\mathbf{P} = 0$ as an optimum at $\mathbf{P}^*$, the dual is feasible there, sc that $\mathbf{v}$ can be found to satisfy (31) at $\mathbf{P}^*$. The point of all this reasoning is that a solution to (29) can be found with the specific product form (32) if and only if S possesses an interior. When such a solution exists, it is also the unique minimum point $\mathbf{P}^*$ of K over S. We can now define the Kruithof problem, in general, as that of finding the factors V_c in (32) so as to satisfy the constraints (29).

Kruithof's "reversibility" property amounts to symmetry of the "closeness" relation: whenever $\mathbf{P}$ is closest to $\mathbf{Q}$, then $\mathbf{Q}$ is closest to $\mathbf{P}$ in the same sense. That is, we define some new constraint levels:

$$\hat{B}_c \equiv \sum_e Q_e A_{ec} \tag{33}$$

and seek a distribution $\mathbf{R}$ to minimize:

$$\hat{K} \equiv K(\mathbf{R}, \mathbf{P}) = \sum_e R_e \log(R_e/P_e) \tag{34}$$

over the set $S(\hat{\mathbf{B}})$ defined by linear constraints:

$$\sum_e R_e A_{ec} = \hat{B}_c \qquad R_e \geq 0. \tag{35}$$

Defining $\hat{V}_c \equiv 1/V_c$, we can write Q_e from (32) in the product form:

$$Q_e = P_e \prod_c \hat{V}_c^{A_{ec}}. \tag{36}$$

From (26), (33), and (36), $\mathbf{Q}$ is an interior point of $S(\hat{\mathbf{B}})$ having the product form (32), so that $\mathbf{R} = \mathbf{Q}$ is the unique minimum solution for $\mathbf{R}$, and $\mathbf{Q}$ is closest to $\mathbf{P}$. Kruithof's "divisibility" property is just transitivity of the closeness relation: when $\mathbf{P}$ is closest to $\mathbf{Q}$ and $\mathbf{R}$ is closest to $\mathbf{P}$, then $\mathbf{R}$ is closest to $\mathbf{Q}$. Specifically, we choose an arbitrary constraint vector $\hat{\mathbf{B}}$ in (35), such that $S(\hat{\mathbf{B}})$ has an interior point, and seek $\mathbf{R}$ to minimize $\hat{K}$ in (34) over $S(\hat{\mathbf{B}})$. The solution for $\mathbf{R}$ is the right side of (36) for some $\hat{\mathbf{V}}$, and substituting for $\mathbf{P}$ from (32) yields:

$$R_e = Q_e \prod_c (\hat{V}_c V_c)^{A_{ec}}, \tag{37}$$

which again has the product form (32). Thus $\mathbf{R}$ minimizes $K(\mathbf{R}, \mathbf{Q})$ over $S(\hat{\mathbf{B}})$ and is closest to $\mathbf{Q}$ in this sense because it is closest to $\mathbf{P}$.

Symmetry and transitivity show that "closest to" is an equivalence relation determined by the particular matrix $[A_{ec}]$. This relation partitions the set of positive distributions over $\{e\}$ into equivalence classes. Each class $C(\mathbf{Q})$ can be generated from any one of its members $\mathbf{Q}$ by using (32): choose all positive values of the V_c for $c \neq \hat{c}$ and scale the resulting values $\hat{\mathbf{P}}$ as necessary to meet the normalization condition (25). Since the column of $[A_{ec}]$ corresponding to $\hat{c}$ is all ones, $V_{\hat{c}}$ appears linearly in (32), and the scaling above represents a particular choice of that variable. Uniqueness says that each constraint vector $\mathbf{B}$ is achieved at most once in each class, while the existence condition asserts that $\mathbf{B}$ is achieved in every class if it is achieved in one class. A natural mapping $\mathbf{B} = f(\mathbf{P})$, namely the linear mapping (29), takes any $C(\mathbf{Q})$ into the set E of constraint vectors $\mathbf{B}$ for which $S(\mathbf{B})$ has an interior. Clearly, f is continuous and is one-to-one and onto E from the existence and uniqueness results. We will see later that f is one-to-one on the closure of $C(\mathbf{Q})$, which is compact from (25). (However the closure is not necessarily an equivalence class.) It follows that f is a homeomorphism of $C(\mathbf{Q})$ and E. In particular, $\mathbf{P}^*$ is a uniformly continuous function of the constraint vector $\mathbf{B}$ on E and its closure.

A useful result follows from the linear relation between $\mathbf{v}$ and $\mathbf{w}$ in (31). Specifically, let $\mathbf{R}$ be any solution of (25) and (29), multiply R_e by w_e, and sum over e, using (29) and (31) to obtain

$$\sum_e R_e w_e = \sum_e R_e \left[\sum_c A_{ec} v_c - 1 \right] = \sum_c B_c v_c - 1 \tag{38}$$

by exchanging the order of summation. It is easy to show from (31) that the left side of (38) is just $K(\mathbf{R}, \mathbf{Q}) - K(\mathbf{R}, \hat{\mathbf{P}})$, while the right side does not depend on the particular choice of $\mathbf{R}$ in $S(\mathbf{B})$, so that

$$K(\mathbf{R}, \mathbf{Q}) + K(\mathbf{P}, \hat{\mathbf{P}}) = K(\mathbf{R}, \hat{\mathbf{P}}) + K(\mathbf{P}, \mathbf{Q})$$

for all $\mathbf{R}$, $\mathbf{P}$ in S and all $\hat{\mathbf{P}}$, $\mathbf{Q}$ in C. But now the choice $\mathbf{P} = \mathbf{P}^* = \hat{\mathbf{P}}$ yields an example of the Pythagorean theorem noted by Csiszar:

$$K(\mathbf{R}, \mathbf{Q}) = K(\mathbf{R}, \mathbf{P}^*) + K(\mathbf{P}^*, \mathbf{Q}). \tag{39}$$

Roughly, it says that the distance from $\mathbf{R}$ to $\mathbf{Q}$ breaks into a component within S from $\mathbf{R}$ to $C(\mathbf{Q})$ at $\mathbf{P}^*$ and an "orthogonal" component from $\mathbf{P}^*$ to $\mathbf{Q}$ within $C(\mathbf{Q})$.

Now we will investigate the nonlinear (Wolfe) dual problem to minimization of K on S. Let $H(\mathbf{v})$ be the minimum of $L(\mathbf{P}, \mathbf{v})$ over the positive orthant $\mathbf{P} \geqslant 0$, so that

$$H(\mathbf{v}) = L(\hat{\mathbf{P}}(\mathbf{v}), \mathbf{v}) \leqslant L(\mathbf{P}^*, \mathbf{v}) = K(\mathbf{P}^*, \mathbf{Q}) \leqslant K(\mathbf{R}, \mathbf{Q}), \tag{40}$$

which shows that $H(\mathbf{v})$ is bounded above if S is not empty. Substituting (31), (32), and (38) into (30) allows us to express H as a function of each of $\mathbf{v}$, $\mathbf{V}$, $\mathbf{w}$ and $\hat{\mathbf{P}}$ as follows:

$$\begin{aligned} H &= \sum_c B_c v_c - \sum_e Q_e \exp\left(\sum_c A_{ec} v_c - 1\right) \\ &= 1 + \sum_c B_c \log(V_c) - \sum_e Q_e \prod_c V_c^{A_{ec}} \\ &= 1 + \sum_e [R_e w_e - Q_e \exp(w_e)] = 1 + \sum_e [R_e \log(\hat{P}_e/Q_e) - \hat{P}_e]. \end{aligned} \tag{41}$$

Direct differentiation of $H(\hat{\mathbf{P}})$ shows that it achieves a unique maximum over the positive orthant at $\hat{\mathbf{P}} = \mathbf{R}$ where $H = K(\mathbf{R}, \mathbf{Q})$. (Break H into a linear part for those components of $\mathbf{R}$ that vanish and a strictly concave part.) Indeed, if $H(\hat{\mathbf{P}})$ goes to $K(\mathbf{R}, \mathbf{Q})$ on $\hat{\mathbf{P}} \geqslant 0$, we can conclude that $\hat{\mathbf{P}}$ approaches $\mathbf{R}$. Now the difference between $K(\mathbf{R}, \mathbf{Q})$ and $H(\hat{\mathbf{P}})$ achieves a unique minimum of zero at $\hat{\mathbf{P}} = \mathbf{R}$:

$$K - H = \sum_e [R_e \log(R_e/\hat{P}_e) + \hat{P}_e] - 1 = K(\mathbf{R}, \hat{\mathbf{P}}) + \sum_e \hat{P}_e - 1. \tag{42}$$

Whenever this expression vanishes for some $\hat{\mathbf{P}}$ of the form (32), we have $K(\mathbf{P}^*, \mathbf{Q}) = K(\mathbf{R}, \mathbf{Q})$ from (40), and thus $\mathbf{P}^* = \mathbf{R}$ by uniqueness of the minimum.

Suppose that values of $\hat{\mathbf{P}}(\mathbf{v})$ of the form (32) approach some limit $\mathbf{R}$ on the closure of $C(\mathbf{Q})$. Such $\mathbf{R}$ satisfies (25) and thus (29) for the constraint vector $\mathbf{B} = f(\mathbf{R})$. Now $K - H$ becomes arbitrarily small, and we conclude that $\mathbf{R}$ is the $\mathbf{P}^*$ that minimizes K. This shows that the closure of $C(\mathbf{Q})$ consists of points $\mathbf{P}^*$ that minimize $K(\mathbf{P}, \mathbf{Q})$. The uniqueness of the minimum then says that the mapping f is one-to-one on the closure, as promised earlier. Choosing $\mathbf{R} = \mathbf{P}^*$ in (41) shows that $H(\mathbf{P}^*) = K(\mathbf{P}^*, \mathbf{Q})$ even when $H(\hat{\mathbf{P}})$ is extended to the closure of $C(\mathbf{Q})$ by continuity. Finally, the result (39) again follows by setting $\hat{\mathbf{P}} = \mathbf{P}^*$ in (42).

The nonlinear dual problem consists of maximizing H over all $\mathbf{v}$, and thus over positive $\mathbf{V}$, or over the linear affine space of values $\mathbf{w}$ generated by (31). By its construction, H is concave in $\mathbf{v}$, while differentiation shows that it is strictly concave in each individual v_c. Setting the V_c-derivatives of H to zero yields necessary conditions for a stationary maximum:

$$B_c = \sum_e Q_e A_{ec} \prod_{c'} V_{c'}^{A_{ec'}}. \tag{43}$$

These are the same relations that would be obtained by substituting (32) into (29), namely, the general Kruithof problem. One possible scheme for solving eqs. (43) would be by relaxation: choose some variable V_c and adjust it to maximize H with all other variables $V_{c'}$ held fixed; then choose some other variable and repeat, cycling through all V_c infinitely often. (When H is bounded above by $K(\mathbf{R}, \mathbf{Q})$ in (40), a value of V_c to maximize H and satisfy (43) always exists uniquely, because $-H$ is strictly convex in each v_c and is arbitrarily large for large v_c.) If the values of $\hat{\mathbf{P}}(\mathbf{v})$ for the iterates $\mathbf{v}$ approach the limiting value $\mathbf{P}^*$, then the relaxation procedure represents a means of computing $\mathbf{P}^*$. More generally, any collection of constraints c could be solved simultaneously in (43), followed by another collection, and so on, so that each c appears infinitely often. Simultaneous solution of several constraints can be harder than the scheme of treating one variable at a time, however. Another possibility might be to solve (43) approximately for V_c, so that H increases at each step, but not necessarily to its exact maximum in V_c. The general relaxation procedure is just an attempt to maximize H over all variables by doing a few at a time. Such relaxation schemes are known[12] to converge to the maximum of a concave function under very general conditions.

In practice, A_{ec} will generally be a zero-one matrix, so that the powers of V_c in (32) do not become a nuisance. In such a case, the constraints (29) have a simple interpretation, since they assign probability B_c to the event c that is the disjoint union of those e for which $A_{ec} = 1$. Pursuing this view, the sum in (29) is taken over all events e included in c (denoted $e \subset c$) and the product in (32) is taken over all events c that include event e (denoted $c \supset e$). Substituting (32) into (29) now yields

$$B_c = \sum_{e \subset c} Q_e \prod_{c' \supset e} V_{c'} = V_c \sum_{e \subset c} Q_e \prod_{c \neq c' \supset e} V_{c'}, \tag{44}$$

as the form taken by (43) in this case. The relaxation iteration step now reduces to dividing B_c by the sum on the right in (44), in order to calculate the new value of V_c that maximizes H. Under certain weak restrictions on the sequence of variables chosen for iteration, it can be shown that H increases to the limit $K(\mathbf{P}^*, \mathbf{Q})$ and that $\hat{\mathbf{P}}(\mathbf{v})$ converges to $\mathbf{P}^*$. Specifically, assume that the iteration scheme includes an infinity of intervals of length M, for some sufficiently large M, in each of which H is maximized over every variable V_c at least once. Then the relaxation iteration converges if and only if S is nonempty; the limit is $\mathbf{P}^*$, which minimizes K over S and lies on the closure of $C(\mathbf{Q})$.

Indeed, with S nonempty, the consecutive values of H are nondecreasing but bounded above in (40). Thus H approaches some limit H^*, and the successive increases dH must eventually go to zero. Direct computation from (41) and (44) now yields the relation:

$$dH = B_c[dv_c - 1 + \exp(-dv_c)], \tag{45}$$

where dv_c is the corresponding change in v_c. Differentiation shows this expression to be strictly convex in dv_c (except in the trivial case $B_c = 0$), vanishing only at its minimum, namely at $dv_c = 0$. Thus each dv_c also goes to zero as H approaches H^*, so that $d\mathbf{w}$ goes to zero from (31). Consider the values of $\hat{\mathbf{P}}(\mathbf{v})$ that are obtained each time the constraint $\hat{c}$ corresponding to (25) is satisfied in one of the postulated intervals of length M. Since these values are confined to the simplex defined by (25), they have a subsequence that converges to some limit point $\mathbf{R}$. Now each member of the subsequence differs from a solution of constraint c by at most M changes, each of order $d\mathbf{w}$. It follows that the limit $\mathbf{R}$ will satisfy every constraint c. But then, from the discussion after (42), H goes to $K(\mathbf{R}, \mathbf{Q})$ on the subsequence and $\mathbf{R}$ is $\mathbf{P}^*$. Finally, H increases to its maximum over $\hat{\mathbf{P}} \geqslant 0$ for all iterates $\hat{\mathbf{P}}$, so that they converge to $\mathbf{P}^*$. Conversely, whenever the $\hat{\mathbf{P}}$ converge, the limit $\mathbf{R}$ satisfies all constraints, so that S is nonempty. Csiszar proves convergence for cylic iteration on collections of constraints, if each collection contains $\hat{c}$, by an elegant application of (39). In general, such cases do not include single-variable relaxation schemes, such as those treated above.

The artificial restriction that no Q_e may vanish can now be dropped, since (32) shows that P_e is zero whenever Q_e is in any case. The definition of S and its interior must be modified to account for all such conditions, of course. That is, (25) and (29) are supplemented by requirements that P_e vanish whenever Q_e does, while $P_e = 0$ makes $\mathbf{P}$ a boundary point of S only if Q_e is positive. Thus the Kruithof solution exists if and only if some $\mathbf{P}$ satisfies the constraints and vanishes for exactly the same events that $\mathbf{Q}$ does.

The question of whether the existence condition is met for a particular constraint vector $\mathbf{B}$ can be resolved, in principle, by constructing such an interior $\mathbf{P}$ with standard linear programming techniques.

Indeed, consider the problem of maximizing $h \geqslant 0$ subject to linear constraints:

$$\sum_{Q_e \neq 0} (T_e + h) A_{ec} + u_c = B_c \qquad T_e \geqslant 0 \qquad u_c \geqslant 0, \tag{46}$$

where each equality c has been written such that $B_c \geqslant 0$. The slack variables form the initial basis $u_c = B_c$, and a phase one procedure minimizes their sum. If not all u_c are forced to zero at optimum, then $S(\mathbf{B})$ is empty. Otherwise, a point in S has been constructed, so that the relaxation iteration will converge. In this case, all the u_c are dropped and phase two proceeds to maximize h. As soon as some step causes h to exceed zero, $P_e = T_e + h$ is the desired interior point. If the optimum still has $h = 0$, then $S(\mathbf{B})$ has no interior, and the iterative solution will not take the product form (32).

REFERENCES

1. J. Kruithof, "Telefoonverkeersrekening," De Ingenieur, *52*, No. 8(1937), pp. E15–E25.
2. D. Bear, *Principles of Telecommunication-Traffic Engineering*, London: Peter Peregrinus, 1976, Ch. 13.
3. Solomon Kullback, *Information Theory and Statistics*, New York: Wiley, 1959.
4. Richard Sinkhorn, "A Relationship Between Arbitrary Positive Matrices and Doubly Stochastic Matrices," Ann. Math. Statist., *35*, No. 2(1964), pp. 876–879.
5. Richard Sinkhorn and Paul Knopp, "Concerning Nonnegative Matrices and Doubly Stochastic Matrices," Pacific J. of Math., *21*, No. 2(1967), pp. 343–348.
6. Henri Theil, *Statistical Decomposition Analysis*, Netherlands: North-Holland, 1972, Ch. 3.
7. I. Csiszar, "I-Divergence Geometry of Probability Distributions and Minimization Problems," Ann. Prob., *3*, No. 1(1975), pp. 146–158.
8. L. R. Ford, Jr. and D. R. Fulkerson, *Flows in Networks*, Princeton: Princeton University Press, 1962, Ch. 1.
9. Oystein Ore, *Graphs and Their Uses*, New York: Random House, 1963, Ch. 4.
10. Peter Wynn, "On a Device for Computing the $e_m(S_n)$ Transformation," Math. Comp., *10*, No. 54(1956), pp. 91-96.
11. William A. Gale, private communication and unpublished work.
12. S. Schechter, "Minimization of a Convex Function by Relaxation," Chapter 7 in *Integer and Nonlinear Programming*, Netherlands: North-Holland, 1970, J. Abadie, ed.

Contributors to This Issue

Richard R. Anderson, B.S.M.E., 1949, Northwestern University; M.S.E.E., 1960, Stevens Institute of Technology; Bell Laboratories, 1949—. Mr. Anderson first engaged in research on electronic switching systems for telephone central offices. In 1956 he joined the data transmission exploratory development department and made several prototype magnetic-tape transports for storing digital data. He has conducted theoretical studies of data transmission systems by computer simulation. Member, AAAS, Sigma Xi, Tau Beta Pi.

Morton Antler, B.A., 1948, New York University (University College); Ph.D. (Chemistry), 1953, Cornell University; Ethyl Corp., 1953-1958; Borg-Warner, 1958-1959; IBM, 1959-1963; Burndy Corp., 1963-1970; Bell Laboratories, 1970—. Mr. Antler's research interests include inorganic and surface chemistry, corrosion, electrodeposition technology, tribology, and electrical contact science. Since 1959 he has been involved in studies of the properties of electric contact materials, particularly as they relate to connector applications. Currently he is studying the wear behavior of experimental contact materials and the influence of environment on contact performance. Precious Metal Plating Awards (1968 and 1971), American Electroplaters Society; the Alfred E. Hunt Memorial Award (1971), American Society of Lubrication Engineers; Special Recognition Award (1975), Electronic Connector Study Group, Inc. Member, American Chemical Society, American Electroplaters Society, American Society of Lubrication Engineers, American Society for Testing and Materials, Sigma Xi; Associate Director of the Annual Holm Conference on Electrical Contacts of the Illinois Institute of Technology; U.S. Representative to the Advisory Group for the International Conferences on Electrical Contact Phenomena.

Corrado Dragone, Laurea in E.E., 1961, Padua University (Italy); Libera Docenza, 1968, Ministero della Pubblica Istruzione (Italy); Bell Laboratories, 1961—. Mr. Dragone has been engaged in experimental and theoretical work on microwave antennas and solid-state power sources. He is currently concerned with problems involving electromagnetic wave propagation and microwave antennas.

Michael Drozdowicz, Owens Technical Institute, 1968; Bell Laboratories, 1969—. In the Connector Technology Department, Mr. Drozdowicz has worked on the development of wear, porosity, and environmental tests for the evaluation of connector contact materials.

Gerard J. Foschini, B.S.E.E., 1961, Newark College of Engineering; M.E.E., 1963, New York University; Ph.D. (Mathematics), 1967, Stevens Institute of Technology; Bell Laboratories, 1961—. Mr. Foschini initially worked on real-time program design. Since 1965, he has mainly been engaged in analytical work concerning the transmission of signals. Currently, he is working in the area of data communication theory. Member, IEEE, Sigma Xi, Mathematical Association of America, American Men of Science, New York Academy of Sciences.

Richard L. Franks, B.S.E.E., 1963, University of Washington; M.S., 1969, Ph.D. (Electrical Engineering), 1970, University of California, Berkeley; Bell Laboratories, 1970—. Mr. Franks taught Mathematics and Physics at the U.S. Navy Nuclear Power School from 1963 to 1967. His M.S. and Ph.D. theses were on autonomous oscillations in nonlinear systems. He began work at Bell Laboratories on modeling and analysis of large-scale systems related to telephone traffic. After becoming a supervisor in 1973, that work expanded to include fault detection systems. He currently is Head of the Network Management Department.

Michael J. Gans, B.S. (E.E.), 1957, Notre Dame University; M.S., 1961, Ph.D. (E.E.) 1965, University of California, Berkeley; Bell Laboratories, 1966—. At Bell Laboratories, Mr. Gans has been engaged in research on antennas for mobile radio and satellite communications.

J. M. Geary, B.S.E.E., 1968, University of Maryland; M.S.E.E., 1969 and Ph.D.(E.E.), 1973, Carnegie-Mellon University; Bell Laboratories, 1973—. Mr. Geary has worked on optical investigation and computer simulation of plasmas, magnetic and optical data entry devices, infrared optical receiver design, plasma panel displays, and psychoacoustics of inharmonic sound. He is presently engaged in work on magnetic bubble detectors and further ferroelectric devices.

Richard D. Gitlin, B.E.E., 1964, City College of New York; M.S., 1965, and D. Eng. Sc., 1969, Columbia University; Bell Laboratories 1969—. Mr. Gitlin is supervisor of the Data Techniques Group in the Advanced Data Communications Department. He is a member of the Communication Theory Committee of the IEEE Communications Society and is editor for Communication Theory of the IEEE Transactions on Communications. Senior Member, IEEE; Member, Sigma Xi, Eta Kappa Nu, Tau Beta Pi.

B. Gopinath, M.Sc. (Math.), 1964, University of Bombay; Ph.D. (E.E.), 1968, Stanford University; research associate, Stanford University, 1967-1968; Alexander von Humboldt research fellow, University of Göttingen, 1971-1972; Bell Laboratories, 1968—. Mr. Gopinath is engaged in applied mathematics research in the Mathematics and Statistics Research Center.

Harry Heffes, B.E.E., 1962, City College of New York; M.E.E., 1964, Ph.D., 1968, New York University; Bell Laboratories, 1962—. Mr. Heffes has previously worked in the areas of control and filtering theory. More recently, he has been concerned with modeling and analysis of teletraffic systems. He has been Adjunct Associate Professor of Electrical Engineering at New York University. Member, Tau Beta Pi, Eta Kappa Nu, American Men of Science, ORSA.

Jack M. Holtzman, B.E.E., 1958, City College of New York; M.S., 1960, University of California (Los Angeles); Ph.D., 1967, Polytechnic Institute of Brooklyn; Hughes Aircraft Company, 1958-1963; Bell Laboratories, 1963—. Mr. Holtzman has worked in systems and control theory. More recently, he has been working on problems in traffic theory and computer communications networks. He is currently Head, Teletraffic Theory and Applications Department. Member, ORSA.

Sheldon Horing, B.E.E., 1957, City College of New York; M.E.E., 1959, New York University; Ph.D. (E.E.), 1962, Brooklyn Polytechnic Institute; Bell Laboratories, 1957-1960, 1962—. Mr. Horing completed the communications development training program in 1960. He was first engaged in the design and development of an optical electromechanical control system. After spending two years on the faculty at Brooklyn Polytechnic Institute, he returned to Bell Laboratories where he joined the Mathematical Analysis and Consulting Group and engaged in research and consulting in control theory and related areas, as well as in studies of defense systems. He is currently Head of the

Performance Analysis Department, which is engaged in traffic studies of computer-based systems. Member, Sigma Xi, Tau Beta Pi, Eta Kappa Nu.

Roy Stephen Krupp, S.B., (Mathematics, Physics), 1960, Massachusetts Institute of Technology; M.I.T. Aerophysics Laboratory, 1960-65; S.M., 1967 and Ph.D., 1970 (Aeronautics and Astronautics), Massachusetts Institute of Technology; Bell Laboratories, 1970—. A member of the Toll Switching Systems Studies Department, Mr. Krupp has worked at traffic problems, modeling the toll network, and on studies of switching networks. His general interests include combinatorics, fluid mechanics, and various branches of applied mathematics.

Victor B. Lawrence, B.Sc, 1968, D.I.C., 1970, Ph.D., 1972, London University; General Electric Company of Great Britain, Hirst Research Center, 1973; Staff of Kumasi University, 1974; Bell Laboratories, 1974—. Mr. Lawrence's technical experience has been in the field of digital signal processing and data communications.

E. J. Messerli, B.A.Sc. (E.E.) 1965, University of British Columbia; M.S. (E.E.) 1966, Ph.D. (E.E. and C.S.) 1968, University of California, Berkeley; E.E. & C.S. faculty, Berkeley, 1968-69; Bell Laboratories, 1969—. Mr. Messerli has been primarily involved in systems analysis and network planning. His work includes studies on the demand assignment of capacity for a domestic satellite system, on the impact of faulty trunks on customers and the network, and on the worth of more accurate data for trunk provisioning. He is currently supervisor of a group concerned with planning for the measurement of new services. Member, IEEE, ORSA.

Kent V. Mina, B.S.E.E., 1958, Princeton University; M.S.E.E., 1960, New York University; Bell Laboratories, 1958—. Mr. Mina is currently a member of the Electronic and Computer Systems Research Laboratory and has been working in the field of digital communication circuits.

John A. Morrison, B.Sc., 1952, King's College, University of London; Sc.M., 1954, and Ph.D., 1956, Brown University; Bell Laboratories, 1956—. Mr. Morrison has done research in various areas of applied mathematics and mathematical physics. He has recently been interested in queuing problems associated with data communications networks. He was a Visiting Professor of Mechanics at Lehigh University during the fall semester 1968. Member, American Mathematical Society, SIAM, IEEE, Sigma Xi.

Kurt H. Mueller, Diploma (E.E.) 1961, and Ph.D., 1967, Swiss Federal Institute of Technology; Bell Laboratories, 1969–1972, 1973—. Mr. Mueller has worked on a variety of problems in coding, modulation, signal processing, and echo cancellation for high-speed data communication. During 1972–73, he was on leave at the Swiss Federal Institute of Technology where he lectured on information theory and was a member of the executive body of the European Informatics Network. He is currently employed by Gretag, AG in Zurich, Switzerland. Senior Member, IEEE.

William Pferd, B.S.M.E., Rutgers University; M.S.M.E., New Jersey Institute of Technology; 1st Lt., U.S. Army Air Force Intelligence, 1942–1945; Bell Laboratories, 1947—. During his Bell Laboratories career, Mr. Pferd has worked in the areas of customer products, systems engineering, and computer application. He participated in the development of ringers, rotary dials, card dialers, connectors, automatic sets and public telephones. While at the Indianapolis Laboratory, his major responsibility was the development of the "single slot" coin telephone. On returning to the Whippany Laboratory, he led engineering studies in criteria for hardened facilities; equipment standards, layout and assembly; central office planning and design and building investment tax credit. He is currently Head of the Office Planning Department within the Building and Energy Systems Laboratory, responsible for the development of computer systems engineering and systems to facilitate the planning, design, and operation of Bell System buildings. Recipient, Industrial Achievement Award, Copper and Brass Research Association (for the "insulation crushing" clip connector); Meritorious Service Award, National Automatic Merchandising Association (for the new "clad" coinage). Senior Member, IEEE; Member, ACM.

Stephen B. Weinstein, B.S.E.E., 1960, Massachusetts Institute of Technology, M.S.E.E., 1962, University of Michigan, Ph.D. (E.E.), 1966, University of California at Berkeley; Philips Research Laboratories, Eindhoven, Netherlands, 1967–1968; Bell Laboratories, 1968—. Mr. Weinstein's technical interests include data communications, data communications security, and microcomputer systems. Senior member, IEEE.

Papers by Bell Laboratories Authors

BIOLOGY

Critical Phenomena in Fluids. P. C. Hohenberg, Proc. NATO Sch. Microsc. Struct. Dyn. Liquid (1978), pp. 333-366.

Electron Energy Loss Spectroscopy and its Application to Biology. D. C. Joy and D. M. Maher, Proc. 13th Annual Conference Microbeam Analysis Society (June 1978), pp. 15A-15J.

Physics of SEM for Biologists. D. C. Joy and C. M. Maruszewski, Annual J. Scan. Electron Microsc., *11* (June 1978), pp. 379-390.

CHEMISTRY

Calculated γ-Effects on the ^{13}C NMR Spectra of 3,5,7,9,11,13,15-Heptamethylheptadecane Stereoisomers and Their Implications for the Conformational Characteristics of Polypropylene. A. E. Tonelli, Macromolecules, *11* (1978), pp. 565-567.

Chain Dynamics of Poly(but-l-ene) from Carbon-13 Nuclear Magnetic Relaxation Measurements. F. C. Schilling, R. E. Cais, and F. A. Bovey, Macromolecules, *11* (March/April 1978), pp. 325-328.

Compatibility & Physical Properties of Blends of Poly(vinyl chloride) and PB3041, a Polymeric Plasticizer. H. E. Bair, D. Williams, T. K. Kwei, and F. J. Padden, Jr., ACS Polymer Preprints, *19*, No. 1 (March 1978), pp. 143-148.

Contamination-Free Adjustment of pH During Trace Analysis. J. E. Riley, Jr., Anal. Chem., *50*, No. 3 (March 1978), pp. 541-543.

Contribution of the Conformational Specific Heat of Polymer Chains to the Specific Heat Difference Between Liquid and Glass. R.-J. Roe and A. E. Tonelli, Macromolecules, *11* (1978), pp. 114-117.

Copolymerization of Acrylamide with Sulfur Dioxide. Determination of the Effect of Copolymerization Temperature on the Monomer Sequence Distribution by Carbon-13 NMR. R. E. Cais and G. Stuk, Polymer, *19*, No. 2 (1978), pp. 179-187.

The Dimensions of the Alternation Polyesteramide 6NT6. A. E. Tonelli, Eur. Polym. J., *14* (1978), pp. 305.

Examination of Aluminum Copper Films during Anodic Oxidation—Part 1. Corrosion Studies. H. H. Strehblow and C. J. Doherty, Electrochem. Soc., *125*, No. 1 (1978), pp. 30-33.

Functional-Group Selectivities in the Reduction of Poly(vinyl chloride) with Metallic Hydrides. W. H. Starnes, Jr., F. C. Schilling, I. M. Plitz, R. L. Hartless, and F. A. Bovey, Polym. Preprints, *19*, No. 2 (September 1978), pp. 579-584.

Hydrogen Cyanide Production During Reduction of Nitric Oxide over Platinum Catalysts: Transient Effects. R. J. H. Voorhoeve, C. K. N. Patel, L. E. Trimble, and R. J. Kerl, Science, *200* (May 19, 1978), pp. 761-763.

The Iodine Coordination Polyhedron in Hydrated Sodium Periodate Determined From Second Harmonic Generation. G. R. Crane and J. G. Bergman. Inorganic Chemistry, *17*, No. 6 (1978), pp. 1613-1617.

Naphtho (1,8-CD; 4,5-CD) BIS(1,2,6 Thiadiazine. A Compound of Ambiguous Aromatic Character. P. C. Haddon, M. L. Kaplan, and J. H. Marshall, J. Amer. Chem. Soc., *100*, (1978), pp. 1235-1239.

Para-Phenyleneditetrathiafulvalene. M. L. Kaplan, R. C. Haddon, and F. Wudl, J. Chem. Soc. Chem. Commun. (1977), pp. 388.

Piezoelectric Langbeinite-type $K_2Cd_2(SO_4)_3$ Structure at Four Temperatures Below and One Above the 430°K Ferroelastic-Paraelastic Transition. S. C. Abrahams, F. Lissalde, and J. L. Bernstein, J. Chem. Phys., *68* (Feb. 15, 1978), pp. 1926-1935.

The Reaction of Poly(vinyl chloride) with Thiols. W. H. Starnes, Jr., I. M. Plitz, D. C. Hische, D. J. Freed, F. C. Schilling, and M. L. Schilling, ACS Polymer Preprints, *19*, No. 1 (March 1978), pp. 623–628.

Role of Singlet Oxygen in the Degradation of Polymers. M. L. Kaplan and A. M. Trozzolo, Singlet Oxygen, ed. H. Wasserman and R. W. Murray, New York: Academic Press, 1978.

The Structure of α-Amantin in Dimethylsulfoxide Solution. A. E. Tonelli and D. J. Patel, Biopolymers, *17*, No. 8 (1978), pp. 1973–1986.

Transducer Applications of Piezoelectric Polymers. G. R. Crane and A. A. Comparini, Proc., 35th Annual Tech. Conference, Society of Plastics Engineers, Montreal, Canada (April 25–28, 1977), pp. 380–382.

ELECTRICAL AND ELECTRONIC ENGINEERING

Electrical Properties of Binary Amorphous Alloys. J. J. Hauser and J. Tauc, Phys. Rev. B, *17* (April 15, 1978), pp. 3371–3379.

Electron Beam Induced Current. H. J. Leamy, L. C. Kimerling, and S. D. Ferros, Scanning Electron Microscopy, *1* (1978), pp. 717–725.

Low-Loss, Single-Mode Fibers With Different Boron(2) Oxygen(3)-Silicon Oxygen(2) Compositions. G. W. Tasker, W. G. French, J. R. Simpson, P. Kaiser, and H. M. Presby, Appl. Opt., *17*, No. 11 (June 1978), pp. 1836–1842.

Material Dispersion in Lightguide Glasses. J. W. Fleming, Electron Lett., *14* (May 1978), pp. 326–328.

MATERIALS SCIENCE AND ENGINEERING

Calculations of the Spin Polarization of Field Emitted Electrons from Monocrystalline Iron. J.-N. Chazalvield, N. V. Smith, and Y. Yafet, Conf. Proceedings, Transition Metals, 1977, Institute of Physics Conference *Series 39*, Bristol and London (1978), pp. 272–276.

Hydrogen Evolution from Plasma-Deposited Amorphous Silicon. K. J. Matysik, C. J. Mogab, and B. G. Bagley, J. Vac. Sci. Technol., *15* (March/April, 1978), pp. 302–304.

Mechanical Properties in Bending at Elevated Temperature of High Strength Copper Alloy Flat Spring Materials. A. Fox, and E. O. Fuchs, J. Test. Eval., *6*, No. 3 (May 1978), pp. 211–220.

Microstructure and Magnetism in Amorphous Rare-Earth Transition Metals: I Microstructure. H. J. Leamy, and A. G. Dirks, J. Appl. Phys. *49* (June 1978), pp. 3430–3438.

Microstructure and Magnetism in Amorphous Rare Earth-Transition Metal Thin Films. A. G. Dirks, and H. J. Leamy, J. Appl. Phys., *49* (1978), pp. 1735–1737.

Microstructures in Amorphous Nb_3Ge Films. G. C. Chi, and R. J. Schute, Materials Science and Engineering, *34*, No. 2 (1978), pp. 161–163.

Modelling the Contrasting Semimetallic Characters of TiS_2 and $TiSe_2$. J. A. Wilson, Phys. Stat. Sol., *86* (March 1, 1978), pp. 11–36.

The Optical Properties of Amorphous Metallic Alloys. E. Hauser, R. J. Zirke, J. Tauc, J. J. Hauser, and S. R. Nagel, Phys. Rev. Letters, *40* (June 26, 1978), pp. 1733–1736.

PHYSICS

Absolute Sign of Piezoelectric d_{33} and pyroelectric p_3 Coefficients in $LiClO_4$ $3H_2O$. R. Liminga, S. Chominilpan, and S. C. Abrahams, J. Appl. Crystallogr., *11* (April 1, 1978), pp. 128–131.

Advances in X-Ray Analysis. F. McMurdie, C. S. Barrett, and S. C. Abrahams, Appl. Opt., *17*, No. 4, p. 500.

Amorphous Nickel Films Getter-Sputtered at 25°K. J. J. Hauser, Phys. Rev. B, *17* (February 15, 1978), pp. 1908–1912.

Angle-Resolved Photoemission Study of Chlorine Chemisorbed on Cleaved Silicon (111). P. K. Larsen, N. V. Smith, and H. H. Farrell, Extended Abstracts of the Vth Int'l. Conf. on Vac. Ultraviolet Radiat. Phys., Montpellier, *2* (September 1977), pp. 241–243.

Calcium Orthovanadate, A New High Temperature Ferroelectric. A. M. Glass, S. C. Abrahams, A. A. Ballman, and G. Loiacono, Ferroelectrics, *17* (April 1978), pp. 579–582.

High-temperature Spin Dynamics in an Amorphous Ferromagnet. J. A. Tarvin, G. Shirane, R. J. Birgeneau, and H. S. Chen, Phys. Rev. *17* (1978), pp. 241–248.

A Method for Evaluating Viscosities of Metallic Glasses from the Rates of Thermal Transformations. H. S. Chen, J. Non. Cryst. Solids, *27* (1978), pp. 257–263.

A Coulometric Analysis of Iron (II) in Ferrites Using Chlorine. P. K. Gallagher, Amer. Ceram. Soc. Bull., *57* (June 1978), pp. 576–578.

Critical Phenomena and the Superfluid Transition in ^{4}He. G. Ahlers, Quantum Liquids, ed. J. Ruvalds and T. Regge, Amsterdam: North-Holland, 1978, pp. 1–26.

Determination of the E(k) Relation for a Surface State on Au(111). Z. Hussain, N. V. Smith, Phys. Letts., *66A* (June 26, 1978), pp. 492–494.

Effect of Crystallographic Factors on Spherulitic Morphology, as Evidenced in Unidirectionally Solidified Polyamides. A. J. Lovinger, Bull. Amer. Phys. Soc., *23*, No. 3 (March 1978), pp. 420.

Effect of Drawing Tension on Residual Stresses in Clad Glass Fibers. C. R. Kurkjian and U. C. Faek, J. Ceram. Soc., *61* (March–April, 1978), pp. 176–177.

The Electron Energy-Loss Spectrum and Requirements for its Processing. D. C. Joy, D. M. Maher, and P. Mochel, Proc. 13th Ann. Conference Microbeam, *13* (June 1978), pp. L9A–L9F.

Evolution of Turbulence from the Rayleigh-Benard Instability. G. Ahlers, and R. P. Behringer, Phys. Rev. Lett., *40* (March 13, 1978), pp. 712–716.

Failure of Small, Thin-Film Conductors Due to High Current-Density Pulses. E. Kinsbron, C. M. Melliar Smith, A. T. English, and S. T. Chynoweth, Proceeding of the 16th International Reliability Physics Symposium, *16* (July 1978), pp. 248–254.

Getter-Sputtering at Low Temperature (20°K). J. J. Hauser, Appl. Phys. Lett., *32* (February 1, 1978), pp. 125–127.

Light Scattering Investigation of the Ferroelectric Transition in Lead Germanate. K. B. Lyons, P. A. Fleury, Phys. Rev. B, *17* (March 15, 1978), pp. 2403–2419.

Nonlinear Optical Susceptibilities of Triphenylbenzene, Resorcinol and Meta-nitroaniline. J. G. Bergman, and G. R. Crane, J. Chem. Phys., *66*, No. 8 (1977), pp. 3803–3804.

Periodic Regrowth Phenomena Produced by Laser Annealing of Ion Implanted Silicon. H. J. Leamy, G. A. Rozgonyi, T. T. Sheng, and G. A. Celler, Appl. Phys. Lett., *32* (May 1978), pp. 535–537.

Phonon Echoes and Self-Induced Transparency in a Glass. J. E. Graebner, and B. Golding, Proc. Int. Conf. Lattice Dyn., Paris, Sept., 1977 (1978), pp. 464–466.

Plasma Energies for A15 Compounds. L. F. Mattheiss and L. R. Testardi, Phys. Rev. B, *17* (June 15, 1978), pp. 4640–4643.

Point Light Source Detection Characteristics of a SEC Vidicon Digital TV Camera. A. B. Dargis, Rev. Sci. Instrum., *49*, No. 3 (March 1978), pp. 308–313.

Pulse Non-Linearity Measurements on Thin Conducting Films. A. T. English, G. L. Miller, D. A. H. Robinson, L. V. Dodd, and T. Chynoweth, J. Appl. Phys., *49*, No. 2, pp. 717–722.

Short Range Order in Metallic Glasses. T. M. Hayes, J. W. Allen, J. Tauc, B. C. Giessen, and J. J. Hauser, Phys. Rev. Lett., *40* (May 8, 1978), pp. 1282–1285.

Specific Heat of Dilute Solutions of ^{3}He in ^{4}He and the ^{3}He-Quasiparticle Excitation Spectrum. D. S. Greywall, Phys. Rev. Lett., *41* (July 17, 1978), pp. 177–180.

Surface Energy Bands and Atomic Position of Cl Chemisorbed on Cleaved Si(111). P. K. Larsen, N. V. Smith, M. Schluter, H. H. Farrell, K. M. Ho, and M. L. Cohen, Phys. Rev. B *17* (March 15, 1978), pp. 2612–2619.

Copyright © 1979 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 58, No. 2, February 1979
Printed in U.S.A.

B.S.T.J. BRIEF

Effects of Sandstorms on Microwave Propagation

By T. S. CHU

(Manuscript received August 29, 1978)

Low rainfall volume suggests the promise of long paths using higher microwave frequencies for radio communication in desert areas. The pursuit of this promise gives rise to the need for understanding the effects of sandstorms on microwave propagation. First, a distinction should be made between large sand grains and fine sand dust.[1] Sand grains of greater than about 0.2-mm diameter are driven by the wind as a low-flying cloud with a height of less than about 2 meters above the ground. This limited height is expected to be lower than most antenna heights of a microwave station. On the other hand, dust-like sand particles can rise in dense clouds to a height of one kilometer or more. This latter type of sandstorm, which is essentially a misnomer for dust storm, may lie in the terrestrial and earth-space paths of microwave radio; hence, path attenuation data are required. Precise calculation is hampered by the uncertainty about the dielectric constant and the size distribution of sand particles. However, useful analysis and frequency dependence of the sandstorm effects can be obtained without precise knowledge of these parameters.

The relation between microwave attenuation and optical visibility will be of interest because visibility provides a convenient measure of dust density. The visibility is inversely proportional to the optical attenuation coefficient. A proportionality constant of 15 dB* will be assumed for the visibility distance in the following calculations. Sand

* This constant is simply 10 $\log_{10}$ of the measured median 0.031 of normalized difference in luminance between the sky and a mark located at the visibility distance (Ref. 2).

particles will be assumed as spheres of 0.01- to 0.1-mm radius with a dielectric constant in the range of 2.5 $(1 - j\,0.01)$ to 10 $(1 - j\,0.01)$. The assumed dielectric constant of 2.5 is that of dry soil.[3] The loss tangent of 0.01 and the other assumed dielectric constant of 10 are believed to be probable upper limits for sand particles in a desert environment. The Rayleigh approximation is valid at centimeter wavelengths, whereas the very-large-sphere approximation can be used at optical wavelengths.

The attenuation coefficient of a sandstorm is simply the sum of extinction cross sections $C(a)$ of sand spheres

$$\alpha = \int_0^\infty N\,(a)\;C\,(a)\;da, \tag{1}$$

where $N\,(a)\,da$ is the number density within the range of radii $(a, a + da)$. Assuming a single sand radius a in meters, this attenuation coefficient can be written as[4,5]

$$\alpha = \frac{3.25\;SQ_{\text{ext}}}{a}\;\text{dB/m}, \tag{2}$$

where $Q_{\text{ext}} = C/\pi a^2$ is the normalized extinction cross section, and $S = (4/3)\pi a^3 N$ is the fraction of sand in the atmospheric volume. Since $Q_{\text{ext}} = 2$ at optical wavelengths, the number of sand particles per cubic meter becomes

$$N = \frac{\alpha_0 a}{6.5\left(\frac{4}{3}\,\pi a^3\right)}, \tag{3}$$

where α_0 is the optical attenuation coefficient in dB/m.

The effective refractive index of a scattering medium is[5]

$$\tilde{m} = 1 - iS(0)\;2\pi N k^{-3}, \tag{4}$$

where k is the free space phase constant and $S(0)$ is the forward scattering function. Within the Rayleigh approximation, we have[4]

$$S(0) = ik^3\left(\frac{\epsilon - 1}{\epsilon + 2}\right)a^3 + \frac{2}{3}\,k^6\left(\frac{\epsilon - 1}{\epsilon + 2}\right)^2 a^6, \tag{5}$$

where ϵ and a are, respectively, the dielectric constant and the radius of the spherical scatterer. The second term in eq. (5) is negligible at centimeter wavelengths for the sand particle sizes under consideration. Substituting eqs. (3) and (5) into eq. (4) gives the phase shift and attenuation coefficient for centimeter waves.

$$k(\text{Re}\,\tilde{m} - 1) = \frac{3k}{13}\,\alpha_0 a\left[\text{Re}\left(\frac{\epsilon - 1}{\epsilon + 2}\right)\right]\left(\frac{180}{\pi}\right)\text{DEG}/m \tag{6}$$

$$k(\mathrm{Im}\ \tilde{m}) = \frac{3k}{13}\alpha_0 a\left[\mathrm{Im}\left(\frac{\epsilon - 1}{\epsilon + 2}\right)\right](8.68)\ \mathrm{dB/m}. \qquad (7)$$

For a given visibility, the above equations show a linear dependence on the particle radius. For two particle sizes, eqs. (6) and (7) have been plotted for 11 GHz vs visibility and optical attenuation in Figs. 1 and 2. It is seen that, for a relatively poor visibility of 0.1 km, the calculated attenuation for this uniform sandstorm is less than 0.03 dB/km, whereas the calculated phase shift is in the range 1.5 to 35 DEG/km. Some beam displacement or broadening could take place if there were strong density gradient in the sandstorm. Significant attenuation at 11 GHz will certainly occur for a very poor visibility of 10 meters or less.

It is of interest to compare our calculations with recently published 10-GHz measurements[6] on dust using an open resonator. Substituting eq. (3) into eqs. (6) and (7), the refractive index and the loss tangent of a dust medium can be obtained

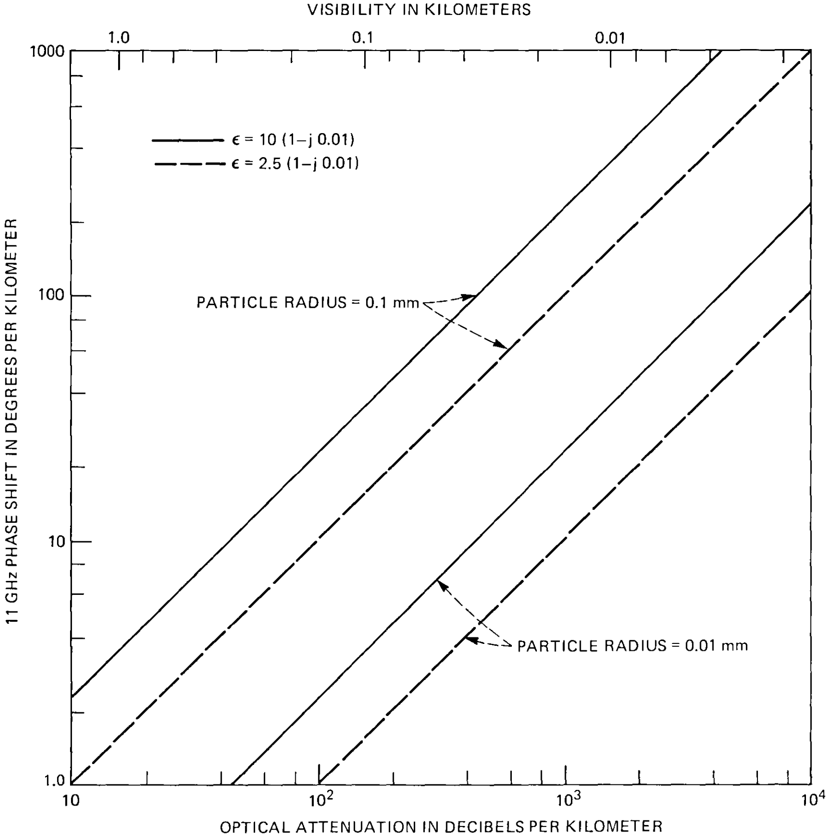


Fig. 1—Calculated 11-GHz phase shift by uniform sandstorm.

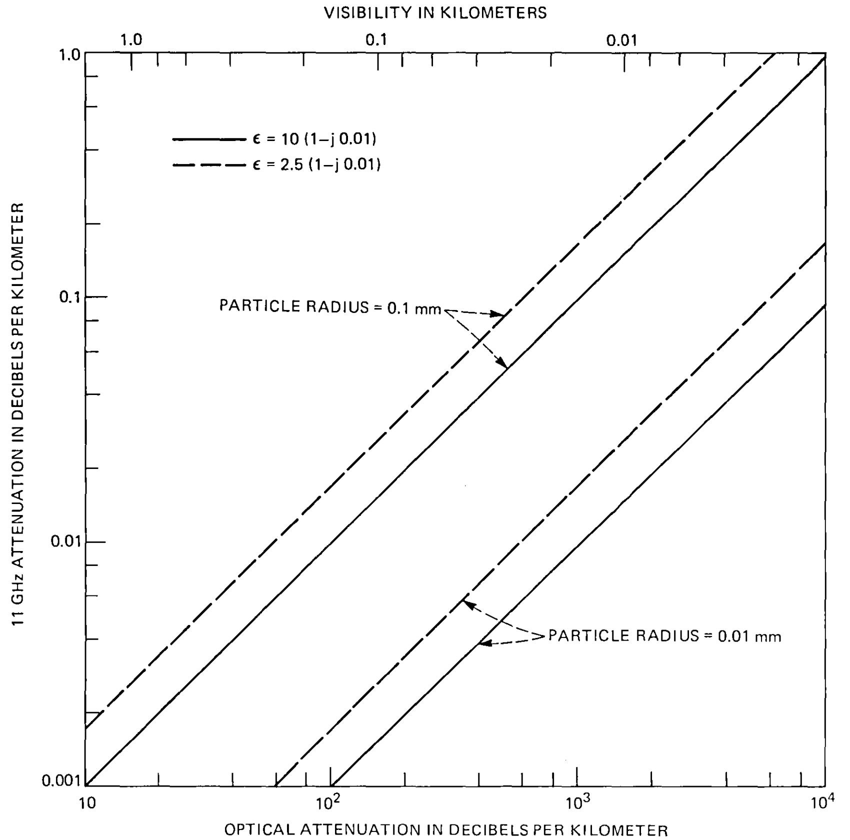


Fig. 2—Calculated 11-GHz attenuation by uniform sandstorm.

$$\mathrm{Re}\,\tilde{m} - 1 = 0.579\ W\left[\mathrm{Re}\left(\frac{\epsilon - 1}{\epsilon + 2}\right)\right] 10^{-3} \tag{8}$$

$$\tan\delta = 1.157\ W\left[\mathrm{Im}\left(\frac{\epsilon - 1}{\epsilon + 2}\right)\right] 10^{-3}, \tag{9}$$

where W is the weight in Kg/m^3 and a specific gravity of 2.6 is assumed. For a given W, the refractive index and the loss tangent are independent of the particle size. Equations (8) and (9) have been plotted along with the measured data of Ref. 6 in Figs. 3 and 4, respectively. The measured refractive indices of both sand and clay dust lie within the range of calculated values. The measured loss tangent for sand dust agrees with the calculated values within the limits of measuring error, whereas that for the clay dust is higher by an order of magnitude. The moisture content of the clay dust in Ref.

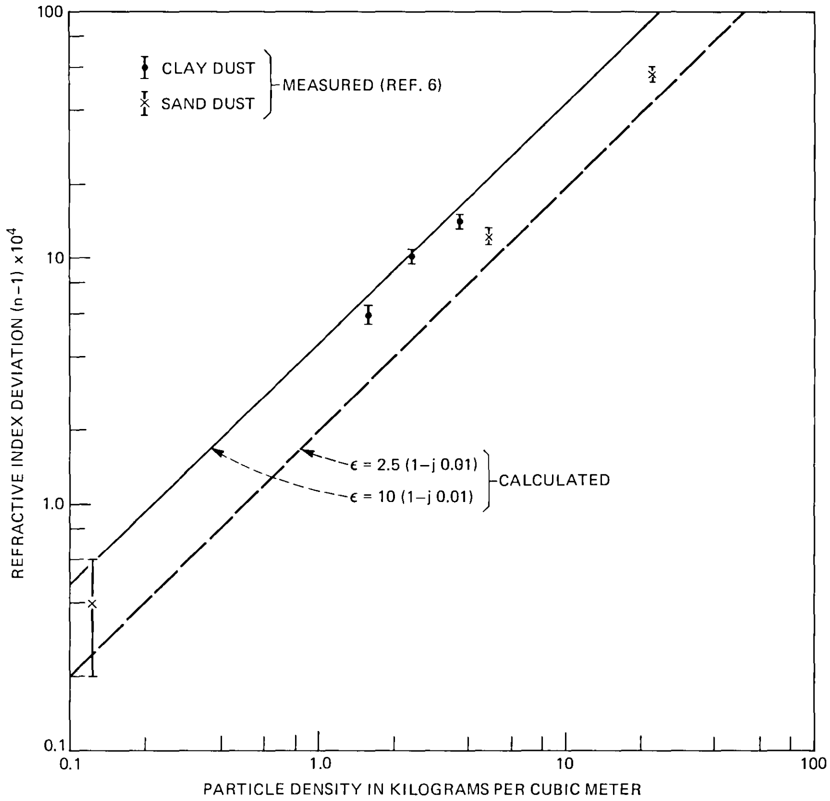


Fig. 3—Comparison between measured and calculated refractive indices for uniform dust precipitation.

6 is unknown. One notes that most particle densities ($\gtrsim$ 1 Kg/m^3) used in the aforesaid measurement are so high that their optical visibilities are less than one meter.

An important result of our calculation is the linear dependence on frequency in eqs. (6) and (7). This property implies that if effects of a sandstorm at 4 and 6 GHz are negligibly small, then at 11 GHz they will also be small. One notes the sharp contrast between the above prediction and the rain attenuation which increases very rapidly from 6 to 11 GHz. Large rain drops have diameters of several millimeters; furthermore, liquid water has a much larger dielectric constant and much larger loss tangent.

Since the particle size and density of a sandstorm are larger near the ground than at a greater height, there appears to be an incentive for using large antenna heights. It also follows that satellite microwave communication is expected to encounter less sandstorm effects than terrestrial microwave networks.

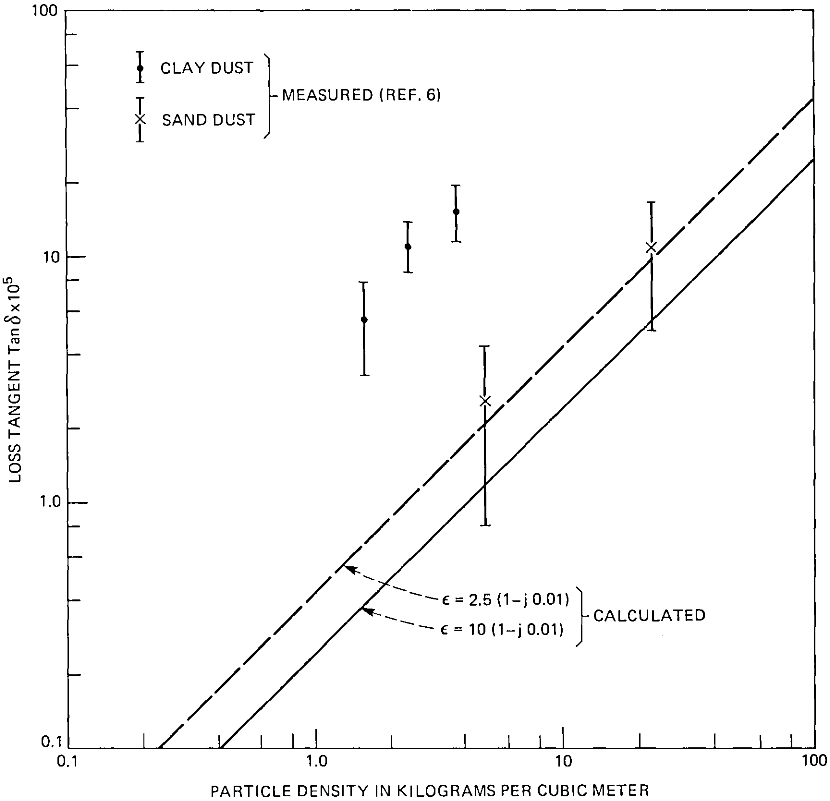


Fig. 4—Comparison between measured and calculated loss tangents for uniform dust precipitation.

ACKNOWLEDGMENTS

The author is indebted to D. C. Hogg, A. J. Giger, and A. A. M. Saleh for helpful suggestions. He also wishes to thank S. H. Lin and W. T. Barnett for reviewing the manuscript.

NOTE ADDED IN PROOF

A very recent paper[7] indicated an upper limit of 0.15 mm for measured radii of particles collected during sandstorms at Khartoum, Sudan.

REFERENCES

1. R. A. Bagnold, "The Physics of Blown Sand and Desert Dunes," New York: William Morrow, 1943, pp. 6, 10.
2. W. E. K. Middleton, "Vision Through the Atmosphere," Toronto: Univ. of Toronto Press, 1952, p. 220.

3. ITT, "Reference Data for Radio Engineers," Sixth Edition, Indianapolis: Howard W. Sams, 1975.
4. T. S. Chu and D. C. Hogg, "Effects of Precipitation on Propagation at 0.63, 3.5, and 10.6 Microns," B.S.T.J., *47,* No. 5 (May–June 1968), pp. 723–759.
5. H. C. Van DeHulst, "Light Scattering by Small Particles," New York: John Wiley, 1957.
6. L. Y. Ahmed and L. J. Auchterlonie, "Microwave Measurements on Dust, Using an Open Resonator," Elec. Lett., *12,* No. 17 (August 19, 1976), pp. 445–446.
7. S. I. Ghobrial, I. A. Ali, and H. M. Hussien, "Microwave Attenuation in Sandstorms," International Symposium on Antennas amd Propagation, Sendai, Japan, Aug. 29–31, 1978.

THE BELL SYSTEM TECHNICAL JOURNAL is abstracted or indexed by *Abstract Journal in Earthquake Engineering, Applied Mechanics Review, Applied Science & Technology Index, Chemical Abstracts, Computer Abstracts, Computer & Control Abstracts, Current Contents/Engineering, Technology & Applied Sciences, Current Contents/Physical & Chemical Sciences, Current Index to Statistics, Current Papers in Electrical & Electronic Engineering, Current Papers on Computers & Control, Electrical & Electronic Abstracts, Electronics & Communications Abstracts Journal, The Engineering Index, International Aerospace Abstracts, Journal of Current Laser Abstracts, Language and Language Behavior Abstracts, Mathematical Reviews, Metals Abstracts, Science Abstracts, Science Citation Index,* and *Solid State Abstracts Journal.* Reproductions of the Journal by years are available in microform from University Microfilms, 300 N. Zeeb Road, Ann Arbor, Michigan 48106.

Bell System