THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

Volume 54 April 1975 Number 4

Analysis of Field-Aided, Charge-Coupled Device Transfer

By J. McKENNA and N. L. SCHRYER

(Manuscript received August 7, 1974)

We study the numerical solution of a nonlinear, partial-differential equation that describes charge transport in a model of a charge-coupled device (CCD). This model differs from previous models in that field-aiding of the transfer is taken into account. Although a derivation of the transport equation is given, the main emphasis in the paper is on the numerical techniques involved, and no actual numbers are presented. Actual numerical results based on the techniques developed here can be found in several recent design studies. The equation, which is parabolic, has one space dimension and one time dimension. Galerkin's method, with standard chapeau functions, is used to discretize in space. This results in a very stiff system of nonlinear, ordinary, differential equations. To solve these equations, we use a first-order backward Euler scheme coupled with extrapolation. A number of alternative schemes were tried and found to be inadequate.

I. INTRODUCTION

In this paper, we study the numerical solution of a nonlinear, partial-differential equation that describes charge transport in a model of a charge-coupled device (CCD). The emphasis is on the numerical techniques involved, although a derivation of the equation is given. The reader is referred to other papers where the solutions are used in device theory and design.[1,2] We briefly summarize the physical background of the equation first.

A knowledge of the dynamics of charge transfer in a CCD is, of course, central to a complete understanding of its operation. A calculation of the motion of charge in a CCD, starting from the coupled, nonlinear Poisson and charge-conservation equations and taking into account the full geometry of the device, has so far proved impossible. However, Strain and Schryer[3] and, independently, Kim and Lenzlinger[4] developed and studied an approximate, one-dimensional model of charge transfer in a CCD. The original analysis considered motion owing only to diffusion and the mutual repulsion of the charge carriers. Field-aided transfer was ignored. Since these original studies, a number of other authors have studied the effects of field-aiding.[5–8] In Refs. 5, 6, and 8, as in the original papers,[3,4] an infinite sink for the charge at one end of a cell is assumed. The assumption of an infinite sink rules out charge "bunching," which in certain situations is an important effect (for an example of this, see Ref. 1, Fig. 8). In Ref. 7, the assumption of an infinite sink is not made. In this paper, we extend the original work[3,4] to include field-aiding and more realistic boundary conditions. Our model can describe both surface[9] CCDs and buried-channel[10] CCDs (BCCDs). We do not include the effects of surface traps, since the main application[1] was to BCCDs. We feel the numerical scheme described here has advantages over that used in Ref. 7, where essentially the same model as ours was used to study surface CCDs, with the effect of traps included. Calculations using our methods show that BCCDs, which can be fabricated with present technology, should be extraordinarily fast and efficient and have reasonable charge-carrying capabilities. Transfer times of 1.8 ns are predicted for a two-phase device having 10-μm-wide electrodes.[1] Slower but similar results are obtained for surface devices.

Strain and Schryer[3] solved, by the method of finite differences, a transport equation quite similar to the one we study here. However, their method of solution proved inadequate when applied to our equation. It is possible to obtain solutions of the transport equation as follows. We use Galerkin's method[11] with standard chapeau functions in space. We treat the time behavior by polynomial extrapolation to the limit of the results of a first-order, fully implicit (nonlinear), finite difference scheme. Although the equation only roughly models the true physical situation, an accurate knowledge of the solution as it varies over many orders of magnitude is necessary if it is to be of any use. This requirement makes the numerical solution of the equation difficult. Many other schemes were tried, and the above method is the only one we found that could solve the problem.

The equation of charge transport is derived in Section II, although some more complex details are given in Appendix A. The technique

for numerically solving the equation of charge motion is given in Section III, with some details in Appendix B. Questions of existence and accuracy are discussed in Section IV, along with the use of polynomial extrapolation. An outline of the theory of extrapolation is given in Appendix C. The method by which initial solutions are obtained is the subject of Section V. Finally, in Section VI we discuss several other schemes by which we tried to solve the equation of charge motion and which failed.

II. DERIVATION OF THE TRANSPORT EQUATION

We refer the reader to the literature for a discussion of the principles of operation of either surface CCDs[9] or BCCDs.[10] Basically, however, both are devices that move packets of charge from under one electrode to under another electrode by suitably changing the voltage on the electrodes.

As in Ref. 3, we assume that the charge can be described by a charge density $q(x, t)$. Here, x is the distance under the plates (see Fig. 1) and t is the time. Then, as we show in Appendix A, the component of the electric field along the direction of motion of the charge, which is due to the mutual repulsion of the charge, is

$$E_x^q = -Sq_x. \tag{1}$$

The elastance S is assumed to be a constant independent of x and t. In all that follows, we use subscripts to denote differentiation; thus, $q_x = \partial q(x, t)/\partial x$, etc. Equation (1) holds for both surface and buried

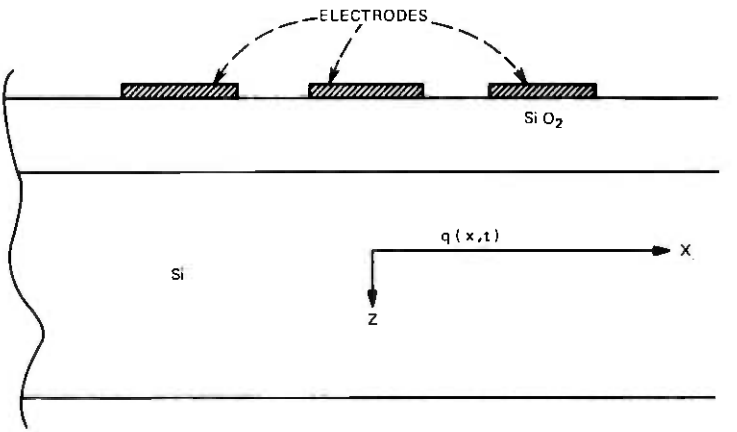

Fig. 1—Schematic of a CCD showing relation to device of x-coordinate in transport equation.

channel devices, although the values of S are different in each case. Expressions for S are given in Appendix A in terms of the physical parameters of the devices.

Let $\varphi(x, t)$ be the given driving potential due to the voltages applied to the electrodes. For a surface CCD, φ is the electric potential at the oxide semiconductor interface, while, for a BCCD, φ is the potential at the potential minimum of the buried channel. In most applications, we have approximated φ by the potential in the CCD in the absence of any mobile charge.[12,13]

The total field along the direction of motion is

$$E_x = -Sq_x - \varphi_x. \tag{2}$$

The current density is[9]

$$J(x, t) = q\mu E_x - Dq_x, \tag{3}$$

where D is the diffusion constant and μ is the mobility, which we also assume to be constant. If we substitute (2) into (3) and make use of the Einstein relation $D = (kT/e)\mu = \alpha\mu$, then

$$J(x, t) = -\mu[(\alpha + Sq)q_x + q\varphi_x]. \tag{4}$$

If we substitute (4) into the charge-conservation equation,[14]

$$q_t + J_x = 0, \tag{5}$$

we get the desired transport equation,

$$q_t = \mu[(\alpha + Sq)q_x + q\varphi_x]_x. \tag{6}$$

The appropriate solution of (6) satisfies an arbitrarily given initial distribution of charge $q(x, 0)$ and the boundary conditions $J(0, t) = J(L, t) = 0$. The boundary conditions state that there is no charge flow into or out of the device at either end. L is the length of the device.

It is convenient to write (6) in terms of dimensionless quantities, as in Ref. 3. Let

$$\tau = t/(L^2/\mu v_0), \quad y = x/L, \quad w = Sq/v_0, \quad \Phi = \varphi/v_0, \quad \beta = \alpha/v_0, \tag{7}$$

where v_0 is a reference voltage. Then (6) becomes

$$w_\tau = [(w + \beta)w_y + w\Phi_y]_y. \tag{8}$$

As it turns out, there seems to be no natural voltage unit in the problem (Ref. 3), so we typically pick $v_0 = 1$ volt.

Physically, the quantity of interest is the total charge present between any two points $0 \leqq y_1 < y_2 \leqq 1$. This suggests that, instead of $w(y, \tau)$, we consider

$$Q(y, \tau) = \int_0^y w(\xi, \tau)d\xi. \tag{9}$$

If we integrate eq. (8) with respect to y from 0 to y and make use of the boundary condition $J(0, t) = 0$, we get

$$Q_\tau = (Q_y + \beta)Q_{yy} + Q_y\Phi_y. \tag{10}$$

Since the right-hand side of (10) is just proportional to $J(y, \tau)$, we see that $Q_\tau(1, \tau) = 0$. From this last remark and (9), it follows that the correct boundary conditions on $Q(y, \tau)$ are

$$Q(0, \tau) = 0, \qquad Q(1, \tau) = Q_T = \text{const.} \tag{11}$$

The appropriate initial condition is determined from $w(y, 0)$ by setting $\tau = 0$ in (9). The transport problem we wish to solve is, thus, eq. (10), subject to boundary conditions (11) and given initial conditions. This is a much simpler problem than attempting to solve (8) for the charge density.

III. SOLUTION OF THE TRANSPORT EQUATION

We simplify the notation slightly by setting

$$\psi(y, \tau) = \Phi_y(y, \tau), \tag{12}$$

and note that (10) can be written

$$-\beta Q_{yy} - \frac{1}{2}\frac{\partial}{\partial y}(Q_y)^2 - \psi Q_y + Q_\tau = 0. \tag{13}$$

If we multiply both sides of (13) by a continuous, piece-wise differentiable function $f(y)$ which satisfies $f(0) = f(1) = 0$, integrate the result from 0 to 1, and integrate the terms containing second derivatives by parts, we obtain (letting $f' = df/dy$)

$$\int_0^1 \{[\beta Q_y + \tfrac{1}{2}(Q_y)^2]f'(y) + [-\psi Q_y + Q_\tau]f(y)\}dy = 0. \tag{14}$$

Equation (14) is the starting point for the application of Galerkin's method, because any twice-differentiable function $Q(y, \tau)$ that satisfies (14) for all continuous, piece-wise differentiable $f(y)$ satisfying $f(0) = f(1) = 0$ must also be a solution of (13).

We now discretize in space by introducing a net $\{y_1, y_2, \cdots, y_N\}$ on $[0, 1]$ and a set of standard chapeau functions $f_j(y)$, $1 \leqq j \leqq N$, as pictured in Fig. 2 and defined in Appendix B. In all that follows, the

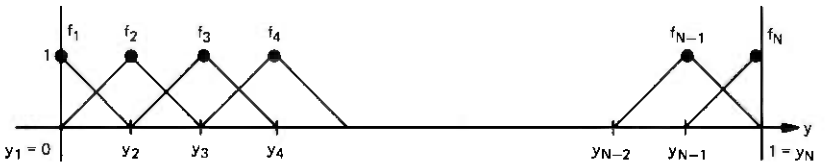

Fig. 2—Discretization of the space interval and the corresponding chapeau functions.

net $\{y_1, \cdots, y_N\}$ is assumed to be given and fixed. In terms of the basic chapeau functions, we define approximations to the solution and external field:

$$\bar{Q}(y, \tau) = \sum_{j=2}^{N-1} Q_j(\tau) f_j(y) + Q_T f_N(y), \tag{15}$$

$$\bar{\psi}(y, \tau) = \sum_{j=1}^{N} \psi_j(\tau) f_j(y). \tag{16}$$

Note that $\bar{Q}(y, \tau)$ has been constructed to satisfy the boundary conditions, $\bar{Q}(0, \tau) = 0$, $\bar{Q}(1, \tau) = Q_T$. The functions $Q_j(\tau)$ are yet to be determined, but we require that they satisfy the initial conditions

$$Q_j(0) = Q(y_j, 0). \tag{17}$$

Because of (17), $\bar{Q}(y, \tau)$ satisfies the correct initial conditions at the mesh points: $\bar{Q}(y_j, 0) = Q(y_j, 0)$. We define $\psi_j(\tau) = \psi(y_j, \tau)$, so that $\bar{\psi}(y_j, \tau) = \psi(y_j, \tau)$.

To determine the $N - 2$ functions $Q_j(\tau)$, we require that $\bar{Q}(y, \tau)$ satisfy (14) for each of the $N - 2$ choices of $f(y)$, $f(y) = f_j(y)$, $2 \leqq j \leqq N - 1$, with $\psi(y, \tau)$ replaced by $\bar{\psi}(y, \tau)$. This yields a system of $N - 2$ first-order, nonlinear, ordinary differential equations for the $Q_j(\tau)$. This technique has a robust history and has been applied, not only to many problems of the same type as (13), but to other types of problems as well. The idea is quite simple: Let the approximate solution be a linear combination of the functions f_j, $2 \leqq j \leqq N - 1$ and then make the left-hand side of (13) orthogonal to each of these functions. In geometrical terms, this means making the left-hand side of (13) orthogonal to the span of $f_2, \cdots, f_{N-1}$, denoted by $\langle f_2, \cdots, f_{N-1} \rangle$, in $\mathcal{L}^2[0, 1]$ in the usual inner product: $(f, g) = \int_0^1 f(y) g(y) dy$. Then, crudely speaking, as more points y_j are chosen, $\langle f_2, \cdots, f_{N-1} \rangle$ spans more of $\mathcal{L}^2[0, 1]$ and the left-hand side of (13) must go to zero as $N \to \infty$, so long as it remains orthogonal to $\langle f_2, \cdots, f_{N-1} \rangle$.

If we carry out the substitution of (15) and (16) into (14), with $f(y) = f_i(y)$, $2 \leqq i \leqq N - 1$, the $N - 2$ equations result:

$$\begin{aligned} \sum_{j=2}^{N-1} Q_j \left[\beta \int_0^1 f_j' f_i' dy - \sum_{k=1}^{N} \psi_k \int_0^1 f_j' f_k f_i' dy + Q_T \int_0^1 f_j' f_i' f_N' dy \right] \\ + \tfrac{1}{2} \sum_{j,k=2}^{N-1} Q_j Q_k \int_0^1 f_j' f_k' f_i' dy + \sum_{j=2}^{N-1} \dot{Q}_j(\tau) \int_0^1 f_j f_i dy \\ = - \left[\frac{Q_T^2}{2} \int_0^1 f_i' (f_N')^2 dy + Q_T \left(\beta \int_0^1 f_i' f_N' dy \right. \right. \\ \left. \left. - \sum_{k=1}^{N} \psi_k \int_0^1 f_k f_N' f_i' dy \right) \right]. \end{aligned} \tag{18}$$

In (18) $\dot{Q}_j(\tau) = dQ_j/d\tau$. The values of the integrals appearing in (18) are listed in Appendix B. If we define $h_i = y_i - y_{i-1}$, $2 \leqq i \leqq N$, and substitute into (18) the values of the integrals, we get

$$\begin{aligned}(h_i\dot{Q}_{i-1} + 2(h_i + h_{i+1})\dot{Q}_i + h_{i+1}\dot{Q}_{i+1})/6 \\ + Q_{i-1}[-\beta/h_i + (\psi_{i-1} + 2\psi_i)/6] + Q_i[\beta(1/h_i + 1/h_{i+1}) \\ + (\psi_{i+1} - \psi_{i-1})/6 + \delta_{i,N-1}Q_T/(h_N)^2] + Q_{i+1}[-\beta/h_{i+1} \\ - (\psi_{i+1} + 2\psi_i)/6] + \tfrac{1}{2}\{(Q_i - Q_{i-1})^2/h_i^2 - (Q_{i+1} - Q_i)^2/h_{i+1}^2\} \\ = \delta_{i,N-1}[Q_T^2/(2h_N^2) + Q_T(\beta/h_N + \psi_{N-1}/3 + \psi_N/6)], \quad (19)\end{aligned}$$

where $\delta_{i,N-1}$ is the Kronecker delta function. These equations hold for $2 \leqq i \leqq N - 1$ if we let $Q_1(\tau) = Q_N(\tau) \equiv 0$. The nonlinear ordinary differential-equation initial-value problem given by (17) and (19) represents the spatial discretization of (13) and must now be solved for the $Q_j(\tau)$, $2 \leqq j \leqq N - 1$.

We use a fully implicit finite difference scheme in time (backward Euler). Let

$$Q_j^n = Q_j(n\Delta\tau) \quad (20)$$

for some choice of $\Delta\tau > 0$. We then let $\dot{Q}_j(n\Delta\tau)$ be approximated by $(Q_j^{n+1} - Q_j^n)/\Delta\tau$ and set $Q_j = Q_j^{n+1}$, $\psi_j = \psi_j^{n+1}$ in (19). On rearranging, we obtain the fully implicit, first-order, finite-difference scheme for solving (19) in time:

$$\begin{aligned}T_{i1}^{n+1}Q_{i-1}^{n+1} + T_{i2}^{n+1}Q_i^{n+1} + T_{i3}^{n+1}Q_{i+1}^{n+1} + A_i(Q_i^{n+1} - Q_{i-1}^{n+1})^2 \\ - A_{i+1}(Q_{i+1}^{n+1} - Q_i^{n+1})^2 = R_i^{n+1}, \quad (21)\end{aligned}$$

where

$$T_{i1}^{n+1} = -\beta/h_i + (\psi_{i-1}^{n+1} + 2\psi_i^{n+1})/6 + h_i/(6\Delta\tau), \quad (22a)$$

$$\begin{aligned}T_{i2}^{n+1} = \beta(1/h_i + 1/h_{i+1}) + (\psi_{i+1}^{n+1} - \psi_{i-1}^{n+1})/6 + (h_i + h_{i+1})/(3\Delta\tau) \\ + \delta_{i,N-1}Q_T/(h_N)^2, \quad (22b)\end{aligned}$$

$$T_{i3}^{n+1} = -\beta/h_{i+1} - (\psi_{i+1}^{n+1} + 2\psi_i^{n+1})/6 + h_{i+1}/(6\Delta\tau), \quad (22c)$$

$$A_i = 1/(2h_i^2), \quad (22d)$$

$$\begin{aligned}R_i^{n+1} = \delta_{i,N-1}[Q_T^2/(2h_N^2) + Q_T(\beta/h_N + \psi_{N-1}^{n+1}/3 + \psi_N^{n+1}/6)] \\ + (h_iQ_{i-1}^n + 2(h_i + h_{i+1})Q_i^n + h_{i+1}Q_{i+1}^n)/(6\Delta\tau). \quad (22e)\end{aligned}$$

Equations (21) hold for $2 \leqq i \leqq N - 1$, $n = 0, 1, 2, \cdots$, with the assumption that $Q_1^n = Q_N^n = 0$, $n = 0, 1, 2, \cdots$ and with the initial conditions $Q_i^0 = Q(y_i, 0)$, $2 \leqq i \leqq N - 1$.

We now find the solution of the nonlinear system of eqs. (21) for fixed n by an iterative Newton method. We drop the superscript n denoting the time step, and for fixed n denote by $Q_i(m)$, $2 \leqq i \leqq N - 1$, the mth iterate of the solution of (21). To obtain $Q_i(m + 1)$ from $Q_i(m)$,

we set $Q_i(m+1) = Q_i(m) + r_i(m)$, substitute this into (21), and linearize the resulting equations for the $r_i(m)$:

$$\begin{aligned}
&\{T_{i1} - 2A_i[Q_i(m) - Q_{i-1}(m)]\}r_{i-1}(m) \\
&\quad + \{T_{i3} - 2A_{i+1}[Q_{i+1}(m) - Q_i(m)]\}r_{i+1}(m) \\
&\quad + \{T_{i2} + 2A_i[Q_i(m) - Q_{i-1}(m)] \\
&\qquad\qquad + 2A_{i+1}[Q_{i+1}(m) - Q_i(m)]\}r_i(m) \\
&= R_i - \{T_{i1}Q_{i-1}(m) + T_{i2}Q_i(m) + T_{i3}Q_{i+1}(m) \\
&\qquad + A_i[Q_i(m) - Q_{i-1}(m)]^2 - A_{i+1}[Q_{i+1}(m) - Q_i(m)]^2\}. \qquad (23)
\end{aligned}$$

These equations hold for $2 \leqq i \leqq N-1$ with $r_1 = r_N = 0$. This is a tridiagonal system of linear equations. Reference 15 contains a concise analysis and very efficient method of solution for such a system of tridiagonal equations.

In practice, the initial estimate of the solution Q_i^{n+1} to (21) is taken to be Q_i^n from the previous time step. So, if $\Delta\tau$ is chosen sufficiently small, the Newton sequence generated by (23) should converge and do so quickly.

What we have described so far is a method for discretizing (9) and (10) in space and time, giving the nonlinear system of eqs. (21), and we have proposed an iterative scheme, given in (23), for solving (21) at each time step. In the next section, we study the feasibility and accuracy of the method.

IV. EXISTENCE AND ACCURACY

We shall show that iteration (23) can be carried out as long as the following conditions are satisfied:

$$0 \leqq Q_2^n \leqq Q_3^n \leqq \cdots \leqq Q_{N-2}^n \leqq Q_{N-1}^n \leqq Q_T, \quad n = 0, 1, 2, \cdots, \qquad (24)$$

$$\sup_{[y_{i-1}, y_i] \times [0, \infty]} |\psi(y, \tau)| \leqq \frac{2\beta}{h_i}, \qquad 2 \leqq i \leqq N. \qquad (25)$$

These conditions are sufficient to ensure the existence of a solution of eqs. (23) for each n. We have not proved it, but in practice they also seem to be necessary. These conditions do not show that the iteration (23) must converge, merely that it is well defined. In fact, if the initial estimate of the solution of (21) is too far off, then in practice the Newton sequence given by (23) may well not converge, and it is necessary to choose $\Delta\tau$ smaller so that Q_i^n provides a better estimate of Q_i^{n+1}.

The monotonicity condition (24) on Q_i^n is merely a necessary consequence of the definition (9) of Q_i^n, since $w(\xi, \tau) \geqq 0$ by definition. The mesh restriction (25), however, is apparently new and fundamental. In practice, if (25) is violated, even at only one point and by a "small"

amount, the solution produced, if any, is highly erratic and non-monotone, and may even be negative.

We now prove that conditions (24) and (25) imply that the matrix of eqs. (23) is strictly diagonally dominant.[16] From this, we can conclude that the matrix has an inverse,[16] so the equations have a solution. From (22a) to (22c) and (25), we see that

$$T_{i1} + T_{i2} + T_{i3} = (h_i + h_{i+1})/(2\Delta\tau) > 0, \tag{26}$$

and

$$\begin{aligned} T_{i2} &\geqq (h_i + h_{i+1})/(3\Delta\tau) > 0; \\ T_{i1} \leqq h_i/(6\Delta\tau), &\qquad T_{i3} \leqq h_{i+1}/(6\Delta\tau). \end{aligned} \tag{27}$$

Because of the monotonicity property (24) and the fact that $T_{i2} > 0$, it is easy to show that

$$\Delta T_i = T_{i2} - |T_{i1}| - |T_{i3}| > 0 \tag{28}$$

implies the diagonal dominance of (23):

$$\begin{aligned} |T_{i2} + 2A_i[Q_i(m) - Q_{i-1}(m)] + 2A_{i+1}[Q_{i+1}(m) - Q_i(m)]| \\ > |T_{i1} - 2A_i[Q_i(m) - Q_{i-1}(m)]| \\ + |T_{i3} - 2A_{i+1}[Q_{i+1}(m) - Q_i(m)]|. \end{aligned} \tag{29}$$

To show that (28) is true, we consider the four possible sign combinations of T_{i1} and T_{i3} and use (26) and (27):

(*i*) $T_{1i} > 0$, $T_{i3} > 0$.

$$\begin{aligned} \Delta T_i = T_{i2} - T_{i1} - T_{i3} &= (h_i + h_{i+1})/(2\Delta\tau) \\ &\quad - 2(T_{i1} + T_{i3}) \geqq (h_i + h_{i+1})/(6\Delta\tau) > 0. \end{aligned}$$

(*ii*) $T_{i1} > 0$, $T_{i3} < 0$.

$$\begin{aligned} \Delta T_i = T_{i2} - T_{i1} + T_{i3} &= (h_i + h_{i+1})/(2\Delta\tau) \\ &\quad - 2T_{i1} \geqq \frac{h_i}{6\Delta\tau} + \frac{h_{i+1}}{2\Delta\tau} > 0. \end{aligned}$$

(*iii*) $T_{i1} < 0$, $T_{i3} > 0$.

$$\begin{aligned} \Delta T_i = T_{i2} + T_{i1} - T_{i3} &= (h_i + h_{i+1})/(2\Delta\tau) \\ &\quad - 2T_{i3} \geqq \frac{h_i}{2\Delta\tau} + \frac{h_{i+1}}{6\Delta\tau} > 0. \end{aligned}$$

(*iv*) $T_{i1} < 0$, $T_{i3} < 0$.

$$\Delta T_i = T_{i2} + T_{i1} + T_{i3} = (h_i + h_{i+1})/(2\Delta\tau) > 0.$$

This completes the proof of the diagonal dominance of (23).

We now discuss the accuracy of the spatial and time discretizations. It is well known (see Ref. 11) that the Galerkin procedure, using chapeau functions, is accurate to $0(h^2)$, where $h = \max_i h_i$ and $0(h^2)/h^2$ represents roughly an upper bound on $Q_{yy}(y, \tau)$ over $[0, 1] \times [0, \infty)$.

We shall not go into the proof of such results here. Rather, a heuristic but useful analysis of the error is presented. The $0(h^2)$ accuracy, basically, comes from the fact that replacing $Q(y, \tau)$ by its interpolant,

$$\sum_{i=1}^{N-1} Q(y_i, \tau) f_i(y) + Q_T f_N(y),$$

results in such a $0(h^2)$ error by using Taylor's theorem on each of the intervals $[y_i, y_{i+1}]$, $i = 1, \cdots, N-1$. A similar statement can be made about $\psi(y, \tau)$ and its interpolant. For the sake of clarity, assume that the mesh is uniform with $h_i \equiv h$, $i = 2, \cdots, N$. Then standard finite difference arguments show that (18) is a spatial finite difference approximation to a function $Q^*(y, \tau)$ obeying

$$Q^*_\tau = Q^*_{yy}(\beta + Q^*_y) - \psi Q^*_y + 0(h^2), \tag{30}$$

where 0 involves terms of the form Q^*_{yy} and its higher-order derivatives, $\partial^{m+n}/\partial y^m \partial \tau^n$. Then, intuitively speaking, since $Q(y, \tau)$ solves (30) to within $0(h^2)$ and $Q^*(y, 0) - Q(y, 0) = 0(h^2)$, we must have $Q^*(y, \tau) - Q(y, \tau) = 0(h^2)$.

Even though (30) is based on the assumption that the spatial mesh is uniform, it shows clearly that the h_i must be small in any region where any of the derivatives $(\partial^{m+n}/\partial y^m \partial \tau^n) Q_{yy}$ are large. Physically, such regions are precisely those regions where the field $\psi(y, \tau)$ is large. This makes restriction (25) quite reasonable, since (25) requires a smaller spatial mesh where the field ψ is large. In fact, we can estimate the number of points N_v, required by (25), using a variable mesh, in a potential rise of v volts: (25) requires that ψ change by no more than $2\beta \cong 1/20$ (at room temperature) over any mesh interval. Then, for example, a potential rise of 5 volts will have $\cong$100 points y_i modeling it. So (25) itself forces a fairly accurate representation of ψ and hence, indirectly, of Q.

However, the time mesh is another matter altogether. The time difference scheme is only first-order accurate and the local time behavior of Q near large values of ψ is rather bad. Thus, application of (21) to (23) alone to solve the problem gives rather poor results. For this reason, we have used polynomial extrapolation to the limit of the results of the first-order scheme (23). A brief discussion of the extrapolation process is given in Appendix C. Ironically, polynomial extrapolation was used because rational extrapolation converged so quickly to the solution that it led to very large $\Delta\tau$ choices (see Ref. 17 for the $\Delta\tau$ monitoring mechanism) which, in turn, led to iteration (23) not converging or taking a very long time doing it. So, even though

polynomial extrapolation is "slower" than rational, it is "better" for our purpose here.

V. CALCULATION OF $Q(y, \infty)$

In most cases of interest, the initial condition for (10) is chosen as an equilibrium solution $Q(y, \infty)$ corresponding to a time-independent potential $\Phi(y)$. It is convenient in these cases to solve for the corresponding $w(y, \infty) = w(y)$ and then integrate to get $Q(y, \infty)$.

Setting $w_\tau = 0$ in (8) yields $0 = [(w + \beta)w_y + w\Phi_y]_y$, which, when integrated twice from 0 to y with the aid of the boundary condition $J(0, \infty) = 0$, yields

$$F(w) = w + \beta \ln \frac{w}{C_0} + \Phi(y) = 0 \tag{31}$$

for some constant C_0. Let y_0 be any point in $[0, 1]$ such that $w(y_0) > 0$. Then

$$C_0 = w(y_0) \exp\left(\frac{\Phi(y_0) + w(y_0)}{\beta}\right). \tag{32}$$

Thus, given $\Phi(y)$ and a single value of $w(y_0) > 0$, the entire equilibrium distribution $w(y)$ is determined. Note that $w(y) > 0$ whenever $\Phi(y)$ is finite.

To find $w(y)$ from (31) we use Newton's method. An initial guess at the solution $w^{(0)}(y) > 0$ is made. The solution is then iterated, the $(n + 1)$th iterate being related to the nth by

$$w^{(n+1)}(y)\left\{1 + \frac{\beta}{w^{(n)}(y)}\right\} = \Phi(y) + \beta\left\{1 - \ln\left(\frac{w^{(n)}(y)}{C_0}\right)\right\}. \tag{33}$$

Since $F'(w) = 1 + \beta/w > 0$ and $F''(w) = -\beta/w^2 < 0$, we see that $F(w)$ is a concave, monotone-increasing function for $w > 0$. Thus, the Newton sequence generated by (33) will converge to the solution (31) no matter what initial $w^{(0)}(y) > 0$ is chosen.

Once the $w(y_i)$, y_i in the Galerkin net $\{y_1, \cdots, y_N\}$ are found using (33), $Q(y_i)$ may be found by the trapezoidal rule for integration. This is consistent with the representation of Q by the chapeau functions, $\tilde{Q}$, since the trapezoidal rule is exact for chapeau functions.

VI. ATTEMPTS THAT FAILED

The first attempt at solving (10) was via the finite difference scheme of Ref. 3. It was impractical because the spatial mesh restriction (25) appeared there, also, forcing the spatial mesh to be very small in some regions, although it could be quite large in others. Since any nonuniformity of mesh size in a central finite difference scheme leads to

only first-order accuracy, we were then left with a very fine mesh over the entire interval [0, 1]. This required tens of thousands of points in the spatial mesh, far too many to be practical.

After going to Galerkin's method in space, which has second-order accuracy even with a nonuniform mesh, the solution of (19) posed another problem: It is an extremely "stiff" system of ordinary differential equations, with the coefficients A_i ranging typically from 10^4 to 10^{10}. This is a reflection of the locally quick time and spatial changes in $Q(y, \tau)$ when ψ is large, this fact being transmitted to the h_i by (25). For this reason, any attempt to linearize (19) between time steps for a finite difference scheme in time led to failure—the solution is nowhere near linear over reasonable time intervals when ψ is large. The symptom of this problem, in practice, was that the $\Delta\tau$ required in the polynomial or rational extrapolation process for these linearized schemes was extraordinarily small, requiring in one case more than 10^{10} time steps to cope with a single 5-volt potential swing.

Once a nonlinear approach to the solution of (19) was recognized as probably the only route left, the most obvious "accurate" scheme to use is a fully nonlinear Crank-Nicholson solution of (19). A small digression on this scheme in a simple case is useful here. For the linear system of ordinary differential equations,

$$\mathbf{u}' = \mathbf{A}\mathbf{u}, \tag{34}$$

where $\mathbf{u}$ is a vector and $\mathbf{A}$ a matrix, the Crank-Nicholson approximation to the true solution, $\mathbf{u} = e^{\mathbf{A}\tau}\mathbf{u}_0$, is

$$\mathbf{u}(n\Delta\tau) \cong (\mathbf{I} + \tfrac{1}{2}\mathbf{A}\Delta\tau)^n(\mathbf{I} - \tfrac{1}{2}\mathbf{A}\Delta\tau)^{-n}\mathbf{u}_0.$$

This is based on the approximation[18]

$$e^{\mathbf{A}\Delta\tau} \cong (\mathbf{I} + \tfrac{1}{2}\mathbf{A}\Delta\tau)(\mathbf{I} - \tfrac{1}{2}\mathbf{A}\Delta\tau)^{-1}. \tag{35}$$

Letting $\mathbf{u}(n\Delta\tau) = (u_1^n, \cdots, u_N^n)^T$, this corresponds to the standard finite difference formulation of the Crank-Nicholson scheme:

$$(u_j^{n+1} - u_j^n)/\Delta\tau = \tfrac{1}{2}(\mathbf{A}\mathbf{u}^{n+1} + \mathbf{A}\mathbf{u}^n)_j, \quad 1 \leqq j \leqq N.$$

A nonlinear generalization of the above scheme for (19) would have an error of the form $C(\Delta\tau)^2$; however, C is very large. This is most easily seen by considering (35) for real $A\Delta\tau$ very large (positive or negative). That relation then states that $e^{A\Delta\tau} \cong -1$, which is an exceedingly bad approximation. For a "stiff" system, (34) [or (19)], one that has a wide spread in its eigenvalues for A, the above reasoning indicates that the Crank-Nicholson scheme would give very poor results unless $\Delta\tau$ is very small. In practice, as before, the symptom of

this problem was very small $\Delta\tau$ choices by the extrapolation routines—the same problem that would have required 10^{10} time steps in a linear scheme would have required "only" 10^8 with Crank-Nicholson. (In this matter, see also Ref. 19.)

In all, more than 12 different schemes were programmed and tested on this problem, (9) and (10), with the result that only the one described in Sections II to V is effective for the wide range of ψ distributions required to model both surface and buried-channel CCDs.

VII. ACKNOWLEDGMENTS

We wish to thank R. J. Strain for helpful conversations during the formulation of the problem studied in this paper. We are indebted to J. A. Morrison for suggesting that we study the equation for $Q(y, \tau)$ rather than the equation for $w(y, \tau)$.

APPENDIX A

In this appendix, we derive eq. (1), the fundamental equation of the Strain-Schryer model, for the case of a BCCD. We choose coordinates as shown in Fig. 3; the x-axis is parallel to the oxide-semiconductor interface and the z-axis is directed into the semiconductor. The potential in the oxide is $\varphi_0(x, z)$ and the potential in the semiconductor is $\varphi_1(x, z)$. The permittivity of the oxide is ϵ_{ox}, that of the semiconductor ϵ_s, and the thickness of the oxide is δ.

In the special case where all the properties of the BCCD are independent of x, the potential in the presence of the inserted charge q has been calculated by Kent[20] and Schryer.[21] They showed that the value of the potential at its minimum in the buried channel is approximately a linear function of the charge q, $\varphi_1 = S_0 q + V_0$, for all values

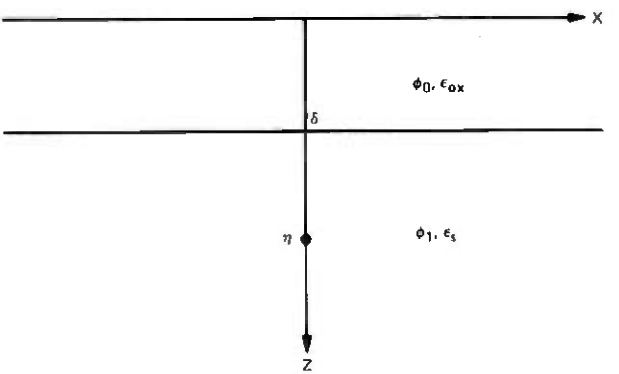

Fig. 3—Coordinate system involved in calculating the potential of a line charge.

of q in the operating range of the device. The elastance S_0 and V_0 are independent of q but depend on the oxide thickness δ and the semiconductor doping, and V_0 also depends strongly on the electrode potential.

In the general case, the Strain-Schryer model assumes that the field in the x-direction in the channel can be approximated by the sum of two terms. The first term is obtained from the above expression for φ_1 by assuming q a function of x and differentiating,

$$E_{x1} = -S_0 \frac{\partial q}{\partial x}. \tag{36}$$

The second term takes into account the field at x resulting from the charge at other points x' in the channel. Because of the metallic electrodes, the charge at x' will induce image charges that will tend to shield the field at x. For this purpose, we first calculate the potential of a unit line charge located at $x = 0$, $z = \eta > \delta$ in the semiconductor. The plane $z = 0$ is assumed to be a perfect conductor at zero potential, and the oxide and semiconductor are assumed uniform. We can write down a solution of Laplace's equation in the form

$$\varphi_0(x, z; \eta) = \int_{-\infty}^{\infty} r(\alpha) \frac{\sinh|\alpha| z}{|\alpha|} e^{i\alpha x} d\alpha, \quad 0 \leqq z \leqq \delta, \tag{37}$$

$$\varphi_1(x, z; \eta) = -\frac{1}{4\pi\epsilon_s} \Psi(x, z; \eta) + \int_{-\infty}^{\infty} s(\alpha) e^{-|\alpha|(z-\delta)} e^{i\alpha x} d\alpha, \quad \delta \leqq z < \infty, \tag{38}$$

where

$$\Psi(x, z; \eta) = \ln\{x^2 + (z - \eta)^2\} - \ln\{x^2 + (z + \eta)^2\}. \tag{39}$$

The function $\Psi(x, z; \eta)$ has the correct singular behavior at $x = 0$, $z = \eta$ and is harmonic everywhere else in $-\infty < x < \infty$, $\delta \leqq z < \infty$. The boundary condition $\varphi_0(x, 0; \eta) = 0$ is satisfied, and the unknown functions $r(\alpha)$ and $s(\alpha)$ must be chosen so that the boundary conditions

$$\varphi_0(x, \delta; \eta) = \varphi_1(x, \delta; \eta), \quad \epsilon_{ox} \frac{\partial \varphi_0}{\partial z}(x, \delta; \eta) = \epsilon_s \frac{\partial \varphi_1}{\partial z}(x, \delta; \eta) \tag{40}$$

are satisfied. It is straightforward to show that

$$\Psi(x, \delta; \eta) = -2 \int_{-\infty}^{\infty} e^{-\eta|\alpha|} \frac{\sinh|\alpha|\delta}{|\alpha|} e^{i\alpha x} d\alpha, \tag{41}$$

$$\frac{\partial \Psi}{\partial z}(x, \delta; \eta) = -2 \int_{-\infty}^{\infty} e^{-\eta|\alpha|} \cosh|\alpha|\delta e^{i\alpha x} d\alpha. \tag{42}$$

If we substitute (37), (38), (41), and (42) into (40), and Fourier-

transform with respect to x, we obtain two linear equations for $r(\alpha)$ and $s(\alpha)$. The solution of these two equations yields

$$r(\alpha) = \frac{1}{\pi} e^{-(\eta-\delta)|\alpha|}\{(\epsilon_{ox} + \epsilon_s)e^{|\alpha|\delta} + (\epsilon_{ox} - \epsilon_s)e^{-|\alpha|\delta}\}^{-1}, \tag{43}$$

$$s(\alpha) = \frac{\sinh|\alpha|\delta}{|\alpha|}\left\{r(\alpha) - \frac{1}{2\pi\epsilon_s} e^{-\eta|\alpha|}\right\}. \tag{44}$$

On substituting (43) and (44) into (37) and (38), we obtain the desired result. If we expand $r(\alpha)$ and $s(\alpha)$ in powers of $e^{-|\alpha|\delta}$, the Fourier integrals can be evaluated, and we can express the potential as the potential resulting from an infinite array of image charges. Since this result is not needed, we do not give it here.

In the buried-channel case, we need the potential resulting from a two-dimensional charge distribution. Let the density of this distribution be $\rho(\xi, \eta)$. Then $q(x) = \int \rho(x, \eta)d\eta$ is the charge appearing in eq. (36). Since the potential resulting from the image charges induced by a line charge at (ξ, η) in the semiconductor is $\varphi(x - \xi, z; \eta)$, we can now write down the second term of the field in the channel as

$$E_{x2} = -\int\int \frac{\partial \varphi_1}{\partial x}(x - \xi, z; \eta)\rho(\xi, \eta)d\xi d\eta. \tag{45}$$

From (38) and (39),

$$\frac{\partial \varphi_1}{\partial x}(x - \xi, z; \eta) = -\frac{1}{2\pi\epsilon_s}\left[\frac{x - \xi}{(x - \xi)^2 + (z - \eta)^2} - \frac{(x - \xi)}{(x - \xi)^2 + (z + \eta)^2}\right] + i\int_{-\infty}^{\infty} \alpha s(\alpha)e^{-|\alpha|(z-\delta)}e^{i\alpha(x-\xi)}d\alpha. \tag{46}$$

Since $(\partial\varphi_1/\partial x)(x - \xi, z; \eta)$ is singular at $\xi = x$, $\eta = z$, the main contribution to the integral in (45) occurs at this point. We expand $\rho(\xi, \eta)$ in a Taylor series about x, keep only the linear terms in the expansion, and extend the limits of the ξ integral from $-\infty$ to ∞. Since $(\partial\varphi_1/\partial x)(x - \xi, z; \eta)$ is an odd function of $x - \xi$, the term involving $\rho(x, \eta)$ vanishes. A straightforward calculation shows that the remaining term is

$$-\frac{1}{2\epsilon_s}\frac{\partial}{\partial x}\int (z + \eta - |z - \eta|)\rho(x, \eta)d\eta - \delta\left(\frac{1}{\epsilon_{ox}} - \frac{1}{\epsilon_x}\right)\frac{\partial q}{\partial x}. \tag{47}$$

The first integral can be transformed by the mean value theorem: $\int (z + \eta - |z - \eta|)\rho(x, \eta)d\eta = (z + \bar{\eta} - |z - \bar{\eta}|)q(x)$, where $\bar{\eta}$ is a point in the interval of integration. In many cases, it is reasonable to

replace the factor $z + \bar{\eta} - |z - \bar{\eta}|$ by a constant $2l$, independent of z. For such cases, we have

$$E_{x2} = -\left\{l/\epsilon_s + \delta\left(\frac{1}{\epsilon_{ox}} - \frac{1}{\epsilon_s}\right)\right\}\frac{\partial q}{\partial x}. \tag{48}$$

If we combine (36) and (48) we obtain (1), where

$$S = S_0 + (l - \delta)/\epsilon_s + \delta/\epsilon_{ox}. \tag{49}$$

Here S_0 must be obtained from a one-dimensional charge-insertion calculation,[20,21] and l must be estimated from the above formulas.

It should be noted that, if we let $\rho(\xi, \eta) = \rho(\xi)D(\eta - \delta)$ in the previous derivation, where $D(x)$ is the Dirac delta function, and set $y = \delta$, we should get the result of Ref. 3 for a surface device. However, in this case, (47) yields δ/ϵ_{ox} for the correction term, while in Ref. 3 the correction term is $2\delta/(\epsilon_s + \epsilon_{ox})$ [eq. (4)]. This is because, in Ref. 3, in the expansion of the field in terms of image charges, only the first image was taken into account.

APPENDIX B

In this appendix, we list several results concerning the chapeau functions $f_j(y)$:

$$\begin{aligned} f_j(y) &= 0, \qquad 0 \leqq y \leqq y_{j-1}, \\ &= (y - y_{j-1})/h_j, \qquad y_{j-1} \leqq y \leqq y_j, \\ &= (y_{j+1} - y)/h_{j+1}, \qquad y_j \leqq y \leqq y_{j+1}, \\ &= 0, \qquad y_{j+1} \leqq y \leqq 1, \end{aligned} \tag{50}$$

where $h_j = y_j - y_{j-1}$.

We list here a number of elementary integrals that are needed in obtaining eqs. (19) from eqs. (18).

$$\int_0^1 (f_j')^2 dy = 1/h_j + 1/h_{j+1}, \tag{51}$$

$$\int_0^1 f_j' f_{j+1}' dy = -1/h_{j+1}, \tag{52}$$

$$\int_0^1 (f_j)^2 dy = (h_{j+1} + h_j)/3, \tag{53}$$

$$\int_0^1 f_j f_{j+1} dy = h_{j+1}/6, \tag{54}$$

$$\int_0^1 (f_j')^3 dy = (h_j)^{-2} - (h_{j+1})^{-2}, \tag{55}$$

$$\int_0^1 (f_j')^2 f_{j+1}' dy = (h_{j+1})^{-2}, \tag{56}$$

$$\int_0^1 (f_j')^2 f_{j-1}' dy = -(h_j)^{-2}, \tag{57}$$

$$\int_0^1 (f_j)^2 f_j' dy = 0, \tag{58}$$

$$\int_0^1 (f_{j+1})^2 f_j' dy = -\tfrac{1}{3}, \tag{59}$$

$$\int_0^1 f_j f_j' f_{j+1} dy = -\tfrac{1}{6}, \tag{60}$$

$$\int_0^1 (f_j)^2 f_{j+1}' dy = \tfrac{1}{3}, \tag{61}$$

$$\int_0^1 f_j f_{j+1}' f_{j+1} dy = \tfrac{1}{6}. \tag{62}$$

In all these expressions, $f_j' = df_j/dy$.

APPENDIX C

In this appendix, we give a brief description of the extrapolation method for solving eqs. (19) in time. We used a linearized, backward Euler method for solving (19) in time. It is first-order accurate. That is, by using a time step of Δt to go from t_0 to $t_1 = t_0 + m\Delta t$, the resulting error at t_1 is $0(\Delta t)$. See either Ref. 22 or Ref. 23 for the proof of such results.

However, much much more is known about these methods. In fact, Stetter[24] has shown, in a very general setting, that processes such as the above backward Euler technique give rise to expansions of the form

$$\mathbf{T}(\Delta t) = \mathbf{T}(0) + \sum_{j=1}^{\infty} \tau_j (\Delta t)^j, \tag{63}$$

where, for our problem, $\mathbf{T}(\Delta t)$ is the value of the vector $(Q_1^m, \cdots, Q_N^m)^T$, which is the value of our approximate solution at $t_1 = t_0 + m\Delta t$, and the τ_j are vectors that depend only upon t_0 and t_1. Thus, as $\Delta t = (t_1 - t_0)/m$ goes to zero or, equivalently, as m goes to infinity, $\mathbf{T}(\Delta t)$ not only converges, with error $0(\Delta t)$, to the true solution at t_1, namely, $\mathbf{T}(0)$, but each component of $\mathbf{T}(\Delta t)$ looks more and more like a polynomial in Δt. The process of extrapolation consists of simply computing several values, $\mathbf{T}(\Delta t)$, $\mathbf{T}(\Delta t/2)$, $\cdots$, $\mathbf{T}(\Delta t/p)$, and then passing a polynomial of degree $p - 1$ through these data points corresponding to each component. The value of these interpolating polynomials at the origin is the solution $\mathbf{T}(0)$, plus terms of order $(\Delta t)^p$. Here p is called the level of extrapolation.

By using polynomial extrapolation to the limit of the result of the first-order scheme (21), we generate a process that has an error of $0[(\Delta t)^p]$ when p levels of extrapolation are used. This extrapolation process is very well described in Ref. 25, and its application to the numerical solution of ordinary differential equations is also very well described in Ref. 17. It must be stressed that the underlying process, Gragg's modified midpoint rule, which Bulirsch and Stoer extrapolate in Ref. 17, is *not* the one we are proposing to extrapolate here. That rule is second-order accurate and is actually unstable if the equations being solved are stiff. The first-order, linearized, backward Euler method we use here is highly stable under extrapolation, even for very stiff systems like (19). So Ref. 17 should be read with an eye to using extrapolation in solving ordinary differential equations and not to those peculiarities that Bulirsch and Stoer introduce to take special advantage of the nice properties of Gragg's modified midpoint rule. The same technique we have used here to solve (13) was used in Ref. 26 to solve a similar system. It is of interest that, for both these problems, polynomial extrapolation was found to be 15 to 20 percent faster than rational extrapolation. This is in contrast to the finding in Ref. 17 that rational extrapolating is usually the faster of the two, at least when extrapolating Gragg's modified midpoint rule.

REFERENCES

1. J. McKenna, N. L. Schryer, and R. H. Walden, "Design Considerations for a Two-Phase Buried-Channel Change-Coupled Device," B.S.T.J., *53*, No. 8 (October 1974), pp. 1581–1597.
2. J. A. Cooper Jr. and A. M. Mohsen, "Design Considerations and Performance Predictions for Two Phase Offset Charge Coupled Devices with Particular Reference to Their Use in Mass Memory," unpublished work.
3. R. J. Strain and N. L. Schryer, "A Nonlinear Diffusion Analysis of Charge-Coupled-Device Transfer," B.S.T.J., *50*, No. 6 (July–August 1971), pp. 1721–1740.
4. C.-K. Kim and M. Lenzlinger, "Charge Transfer in Charge-Coupled Devices," J. Appl. Phys., *42*, No. 9 (August 1971), pp. 3586–3594.
5. J. E. Carnes, W. F. Kosnocky, and E. G. Ramberg, "Free Charge Transfer in Charge-Coupled Devices," IEEE Trans. on Elec. Dev., *ED-19*, No. 6 (June 1972), pp. 798–808.
6. H.-S. Lee and L. Heller, "Charge Control Method of Coupled Device Transfer Analysis," IEEE Trans. on Elec. Dev., *ED-19*, No. 12 (December 1972), pp. 1270–1279.
7. A. M. Mohsen, T. C. McGill, and A. M. Carver, "Charge-Transfer in Overlapping Gate Charge-Coupled Devices," IEEE J. Solid State Circuits, *SC-8*, No. 3 (June 1973), pp. 191–207.
8. Y. Daimon, A. M. Mohsen, and T. C. McGill, "Final Stage of the Charge-Transfer Process in Charge-Coupled Devices," IEEE Trans. on Elec. Dev., *ED-21*, No. 4 (April 1974), pp. 266–272.
9. W. S. Boyle and G. E. Smith, "Charge Coupled Semiconductor Devices," B.S.T.J., *49*, No. 4 (April 1970), pp. 587–593.
10. R. H. Walden, R. H. Krambeck, R. J. Strain, J. McKenna, N. L. Schryer, and G. E. Smith, "The Buried Channel Charge Coupled Device," B.S.T.J., *51*, No. 7 (September 1972), pp. 1635–1640. Also presented at the Device Research Conference, Ann Arbor, Mich., June 1971.

11. P. G. Ciarlet, M. H. Schultz, and R. S. Varga, "Numerical Methods of High-Order Accuracy for Nonlinear Boundary Value Problems," Numer. Math., *9*, No. 5 (April 1967), pp. 394–430.
12. J. McKenna and N. L. Schryer, "The Potential in a Charge Coupled Device With No Mobile Minority Carriers and Zero Plate Separation," B.S.T.J., *52*, No. 5 (May-June 1973), pp. 669–696.
13. J. McKenna and N. L. Schryer, "The Potential in a Charge Coupled Device With No Mobile Minority Carriers," B.S.T.J., *52*, No. 10 (December 1973), pp. 1765–1793.
14. S. M. Sze, *Physics of Semiconductor Devices*, New York: John Wiley, 1969, p. 66.
15. R. S. Varga, *Matrix Iterative Analysis*, Englewood Cliffs, N. J.: Prentice-Hall, 1962, p. 195.
16. Ref. 15, p. 23.
17. R. Bulirsch and J. Stoer, "Numerical Treatment of Ordinary Differential Equations by Extrapolation Methods," Numer. Math., *8*, No. 1 (March 1966), pp. 1–13.
18. Ref. 15, p. 263.
19. J. L. Blue and H. K. Gummel, "Rational Approximations to Matrix Exponential for Systems of Stiff Differential Equations," J. Comput. Phys., *5*, No. 1 (February 1970), pp. 70–83.
20. W. H. Kent, "Charge Distribution in Buried-Channel Charge-Coupled Devices," B.S.T.J., *52*, No. 6 (July-August 1973), pp. 1009–1024.
21. N. L. Schryer, "Solution of a One-Dimensional Model of Charge-Coupled Device Operation," unpublished work.
22. R. D. Richtmeyer and K. W. Morton, "Difference Methods for Initial-Value Problems," Second Edition, New York: John Wiley 1967.
23. I. W. Sandberg and H. Schichman, "Numerical Integration of Systems of Stiff Nonlinear Differential Equations," B.S.T.J., *47*, No. 4 (April 1968), pp. 511–527.
24. H. J. Stetter, "Asymptotic Expansions for the Error of Discretization Algorithms for Nonlinear Functional Equations," Numer. Math., *7*, No. 1 (February 1965), pp. 18–31.
25. R. Burlirsch and J. Stoer, "Fehlerabschätzungen und Extrapolation mit Rationalen Funktionen bei Verfahren vom Richardson-Typus," Numer. Math., *6*, No. 5 (December 1964), pp. 413–427.
26. N. L. Schryer and L. R. Walker, "On the Motion of 180° Domain Walls in Uniform dc Fields," J. Appl. Phys., *45*, No. 12 (December 1974), pp. 5406–5421.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 4, April 1975
Printed in U.S.A.

Aluminum Oxide/Silicon Dioxide, Double-Insulator, MOS Structure

By J. T. CLEMENS, E. F. LABUDA, and C. N. BERGLUND

(Manuscript received May 15, 1974)

A double-insulator structure consisting of 500 Å of vapor-deposited Al_2O_3 and 1000 Å of thermally grown SiO_2 is used as the gate dielectric in a beam-lead-compatible, p-channel, MOSFET, silicon-integrated-circuit technology. The Al_2O_3 layer, in addition to serving as a sodium barrier and thereby providing a self-passivated technology, results in a positive flatband voltage shift when compared to an SiO_2 structure. The mechanism for this flatband voltage shift is the subject of this paper.

The major experimental results obtained are (i) a negative charge exists near the Al_2O_3/SiO_2 interface, its magnitude being independent of the Al_2O_3 thickness but inversely proportional to the SiO_2 thickness, (ii) the magnitude of the SiO_2/Si interface charge is inversely proportional to the SiO_2 thickness, and (iii) a potential jump of about 1.25 volts in flatband voltage is associated with the addition of the Al_2O_3 layer.

A physical model is proposed which assumes the existence of a constant voltage drop across the SiO_2 layer during the Al_2O_3 deposition and a corresponding charge buildup at the SiO_2/Al_2O_3 interface.

I. INTRODUCTION

The threshold voltage of an insulated-gate, field-effect transistor is directly dependent upon the properties of the gate insulator. A double-dielectric gate structure consisting of nominally 500 Å of vapor-deposited Al_2O_3 and 1000 Å of thermally grown SiO_2 is the basis of a beam-lead-compatible, p-channel, MOSFET, silicon-integrated-circuit technology.[1–3] The Al_2O_3 layer serves two functions. First, it is a diffusion barrier for light ions, such as sodium, and thus provides a self-passivated technology. Second, the Al_2O_3 layer shifts the threshold voltage of the MOS transistor in the positive direction (due to a flatband voltage shift). For example, for a ⟨100⟩ oriented, n-type, 10-ohm-cm, silicon substrate, a flatband voltage of 0.0 volt is obtained with the dual-dielectric structure and a titanium metal gate, characteristic of the beam-lead metallization system, whereas with just an SiO_2 structure the flatband voltage is −0.8 volt. The more positive

flatband voltage capability provided by the Al_2O_3 layer implies that MOSFET integrated circuits can be fabricated which have low power-dissipation properties and which are more easily interfaced with bipolar circuits.

Many techniques have been used for the deposition of Al_2O_3 films intended for application in an integrated-circuit technology. All of our considerations are restricted to Al_2O_3 films deposited at 900°C from an $AlCl_3$ source, the technique reported by Tung and Caffrey.[1] A brief review is given in Ref. 4 of the other Al_2O_3 deposition techniques that have been reported and the electrical characteristics of the films obtained.

The electrical properties of the Al_2O_3/SiO_2, dual-dielectric, gate insulator are quantitatively described in this paper; and, in particular, the mechanisms are delineated which cause the positive shift in flatband voltage. The experimental approach was to do a parametric study of the flatband voltage of the dual-dielectric MOS structure, the two parameters of interest being the thicknesses of the Al_2O_3 and SiO_2 layers.

Two major conclusions were obtained from the parametric study. First, a net negative charge exists near the Al_2O_3/SiO_2 interface, and the magnitude of the charge is independent of Al_2O_3 thickness over the range studied, but inversely proportional to SiO_2 thickness. Second, the magnitude of the normal interface charge associated with the SiO_2/Si interface has a component that is inversely proportional to SiO_2 thickness.

A model explaining the origin of the negative charge at the Al_2O_3/SiO_2 interface was developed, based on the assumption that the electrical conductivity of Al_2O_3 at high temperatures (>300°C) is much greater than that of SiO_2. As a result, during the high-temperature deposition of Al_2O_3, an electric field exists in the SiO_2 due to the Si/Al_2O_3 contact potential difference, and the negative charge at the Al_2O_3/SiO_2 interface terminates this field.

Recently, Aboaf, Kerr, and Bassous also reported the existence of a negative charge at the Al_2O_3/SiO_2 interface with the magnitude of the charge being independent of the Al_2O_3 thickness and inversely proportional to the SiO_2 thickness.[4] This is consistent with the insulator interface charge origin model we proposed[5] and implies that this model has general applicability in dual-dielectric structures, since they used three Al_2O_3 deposition techniques, all different from the technique used to obtain the Al_2O_3 films studied in this paper.

The organization of the paper is as follows. A simple flatband theory for dielectric structures is presented in Section II and space-charge formation in insulators is discussed in Section III. A theoretical

discussion of the Al_2O_3/SiO_2 structure is given in Section IV and the experimental results are presented in Section V. A summary is given in Section VI.

II. FLATBAND CALCULATIONS

The flatband voltage of an MOS structure is defined as that voltage which must be applied to the metal electrode to produce a zero space charge or flatband condition in the semiconductor, and it is determined by the net charge density existing in the insulator system and the various interfacial barrier energies. To calculate the flatband voltage, the voltage across the insulators under flatband conditions is calculated from the net charge density using Gauss' law, and to this is added the voltage contributed by the various barrier energies.

Consider the band diagrams of the metal/SiO_2/Si and metal/Al_2O_3/SiO_2/Si systems shown in Figs. 1 and 2, respectively. The various barrier energies for these systems are defined in the figures. We shall assume that at the SiO_2/Si interface in both structures there is an interface charge layer Q_{ss}. The net charge density in the bulk of the SiO_2 is assumed to be zero and the charge density in the Al_2O_3 is denoted by $\rho_A(x)$. Using S.I. to denote the single-insulator system and D.I. the double-insulator system, it follows that the voltages across the insulators, V_i, due to the charge densities can be written as:

$$V_i(\text{S.I.}) = (Q_{ss}/\epsilon_{ox})T_{ox} \tag{1}$$

$$\begin{aligned} V_i(\text{D.I.}) &= Q_{ss}\left[\frac{T_{ox}}{\epsilon_{ox}} + \frac{T_A}{\epsilon_A}\right] + \int_{T_{ox}}^{T_A+T_{ox}} \frac{dx}{\epsilon_A}\int_0^x \rho_A(x')dx' \\ &= Q_{ss}\left[\frac{T_{ox}}{\epsilon_{ox}} + \frac{T_A}{\epsilon_A}\right] + \int_{T_{ox}}^{T_{ox}+T_A} \frac{T_{ox} + T_A - x}{\epsilon_A}\rho_A(x)dx, \end{aligned} \tag{2}$$

where $(\epsilon_{ox}, \epsilon_A)$, (T_{ox}, T_A) = the dielectric constants and thicknesses of the SiO_2 and Al_2O_3, respectively. In particular, $\epsilon_{ox} = 3.9$; $\epsilon_A = 9.0$.

The applied voltage difference between the metal and the semiconductor, the flatband voltage V_{FB}, for both structures can be written as:

$$V_{FB}(\text{S.I.}) = -(Q_{ss}/\epsilon_{ox})T_{ox} + (\phi_{m,ox} - \phi_B - \phi_f) \tag{3}$$

$$V_{FB}(\text{D.I.}) = -V_i(\text{D.I.}) + (\phi_{m,A} + \phi_{ii} - \phi_B - \phi_f). \tag{4}$$

The insulator-insulator barrier ϕ_{ii} is assumed to be positive if it is as shown in Fig. 2.

Consider the contact potential terms in eqs. (3) and (4). If we let W and χ with the appropriate subscripts denote the vacuum work functions and electron affinities, respectively, of the various materials, and, if we assume that the energy band matching between two different

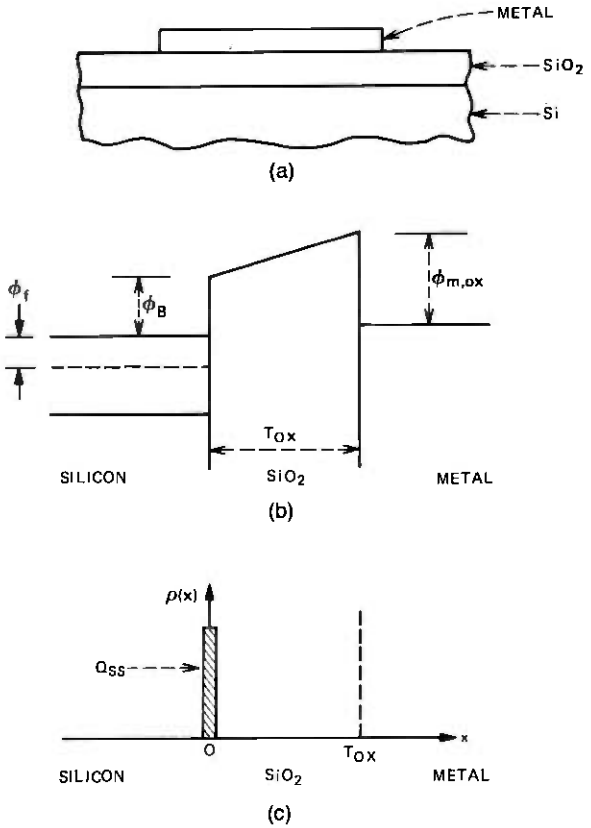

Fig. 1—(a) Cross-sectional view of the metal/SiO_2/Si capacitor structure. (b) Band diagram associated with the structure. (c) Plot of assumed charge density present in the structure.

materials is determined entirely by the difference in vacuum work functions, that is,

$$\phi_{m,ox} = W_m - \chi_{ox} \quad \text{and} \quad \phi_B = \chi_s - \chi_{ox}, \tag{5a}$$

then

$$\phi_{m,ox} - \phi_B - \phi_f = (W_m - \chi_{ox}) - (\chi_s - \chi_{ox}) - \phi_f = W_m - \chi_s - \phi_f, \tag{5b}$$

$$\phi_{m,A} + \phi_{ii} - \phi_B - \phi_f = (W_m - \chi_A) + (\chi_A - \chi_{ox}) - (\chi_s - \chi_{ox}) - \phi_f = W_m - \chi_s - \phi_f, \tag{6}$$

or

$$\phi_{m,ox} - \phi_B - \phi_f = \phi_{m,A} + \phi_{ii} - \phi_B - \phi_f = W_m - \chi_s - \phi_f. \tag{7}$$

Note that under this assumption, the contact potential terms are independent of the electron affinities of the insulators and dependent

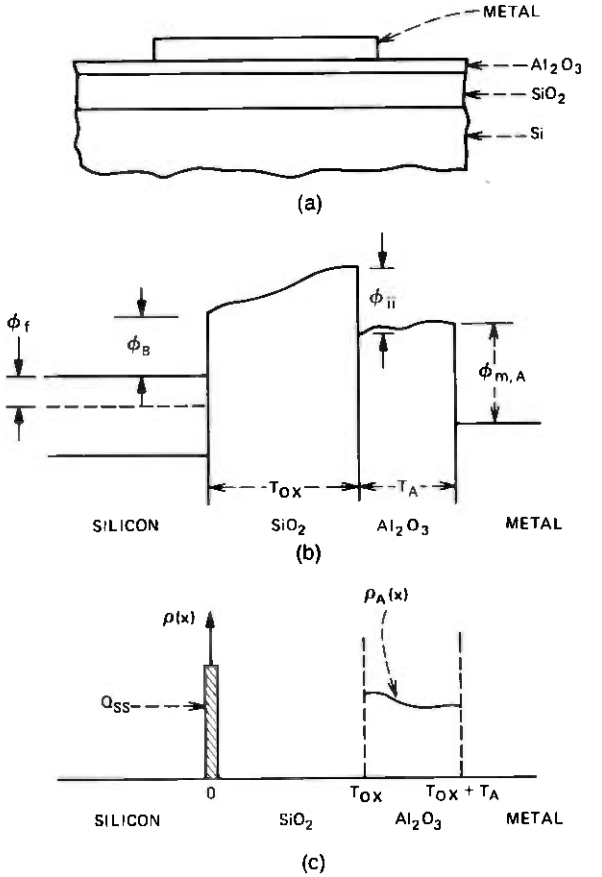

Fig. 2—(a) Cross-sectional view of the metal/Al_2O_3/SiO_2/Si capacitor structure. (b) Band diagram associated with the structure. (c) Plot of assumed charge density present in the structure.

only on the difference between the metal and semiconductor work functions. Thus, if the work-function assumption is valid, an MOS system involving a given metal and a semiconductor will have a constant flatband voltage after correction for Q_{ss} independent of the number or nature of the insulators, unless space charge exists in the insulators. In other words, if the assumption is valid, a measured V_{FB} that changes when the insulators are changed implies space charge exists in the insulators.

III. SPACE-CHARGE FORMATION IN INSULATORS

In the previous section, the flatband voltage of a dual-dielectric MOS structure was calculated assuming a given distribution of space charge. The purpose of this section is to discuss one possible model

for the origin and spatial location of space charge in insulators, namely, the one we feel represents the most likely explanation for much of the space charge in the SiO_2/Al_2O_3 system. In this discussion, we first summarize the proposal for space-charge formation suggested by Simmons for a metal-insulator-metal (MIM) system.[6] Then this model is extended and applied to double-insulator systems, in particular those using SiO_2 and Al_2O_3.

Suppose we form the MIM system shown in Fig. 3a at sufficiently low temperatures that no charge transport occurs within the insulator and no charge exchange occurs between the insulator and the metallic contacts in a time period comparable to the experimental observation time. In this case, no space charge can form in the insulator because thermal equilibrium will not exist and the potential versus position will look as shown in Fig. 3b, where the insulator is represented essentially as a wideband insulator with conduction and valence band edges. Now assume that the system is heated to a sufficiently high temperature that charge transport can occur within the insulator and charge exchange can occur between the insulator and the metallic contacts in a time period that is short compared to experimental observation. In this case, thermal equilibrium will be established and there will be two extreme possibilities between which the system will equilibrate: either the characteristic length corresponding to a space-charge region at thermal equilibrium in the insulator, the electrostatic screening, or Debye length will be large compared to the insulator thickness, and the potential versus position will be virtually identical to that shown in Fig. 3b; or the Debye length in the insulator will be small compared to the insulator thickness, and such space-charge regions as shown in Fig. 3c will form near the two metal interfaces. In the latter alternative, a well-defined "Fermi level" will exist in the bulk of the insulator, as shown in Figure 3c, which will coincide in energy with the Fermi level in the metallic contacts in much the same way that Fermi levels coincide in a conventional Schottky barrier on a semiconductor. Clearly an MIM system in which the Debye length is short compared to the insulator thickness at any given time will always lie somewhere between the extremes indicated in Figs. 3b and 3c, depending on thermal history, so that in such an insulator, space-charge regions will always exist in the vicinity of the metallic contacts. The magnitude of the charge will depend on the difference in the work function of the metal and the insulator and on the degree of thermal equilibrium which has been established.

In the above discussion, it has been assumed that there is no net voltage difference across the MIM structure. Very similar arguments can be presented for the case where a finite voltage exists between

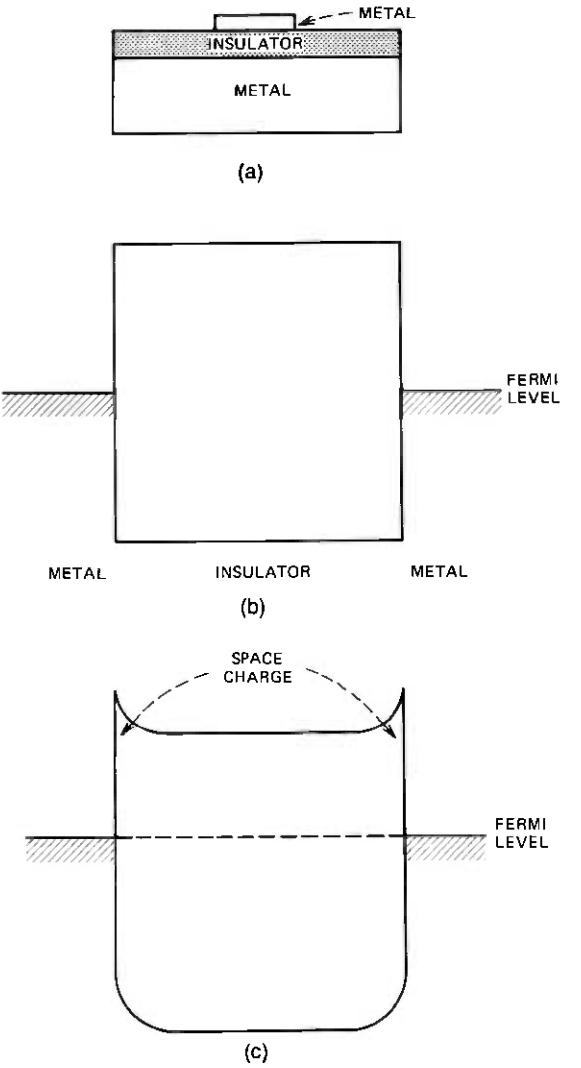

Fig. 3—(a) Cross-sectional view of a metal/insulator/metal capacitor structure. (b) Band diagram of the sturcture if the insulator Debye length is assumed to be much greater than the thickness of the insulator. (c) Band diagram of the structure if the insulator Debye length is assumed to be smaller than the thickness of the insulator.

the two metal contacts; the voltage is either applied or is due to differences in the two involved metal/insulator potential barrier heights. Assuming an insulator which forms space-charge regions that are narrow compared to the insulator thickness, the initial

potential diagram is shown dotted in Fig. 4a and the final steady-state situation is shown as the solid lines. The only electric field required in the bulk of the insulator under steady-state conditions is the ohmic field associated with any current injected from the electrodes, typically a negligible field.

Now suppose that the voltage is reduced as shown at $t = 0$ in Fig. 4b. Initially, the same space-charge that exists under applied voltage

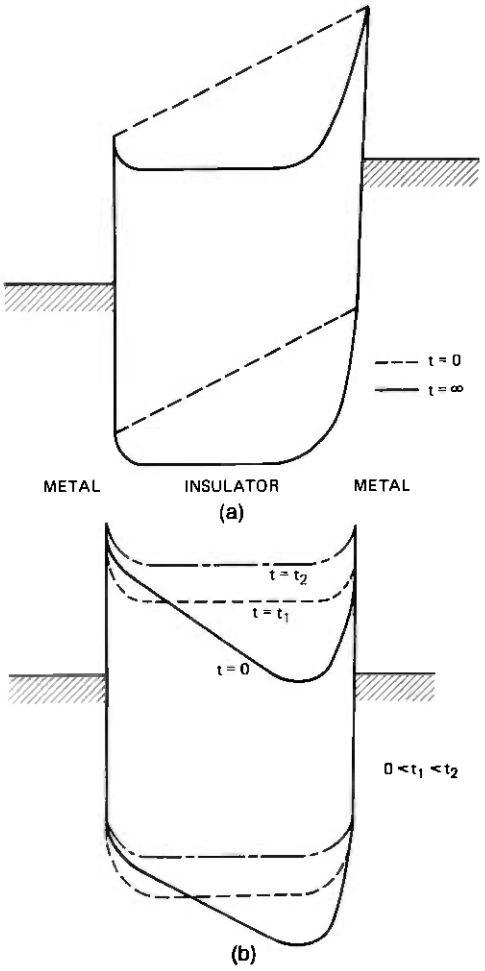

Fig. 4—(a) Band diagram of a metal/insulator/metal structure, whose insulator Debye length is less than the insulator thickness, depicting the immediate and equilibrium band structure in response to the application of an external voltage. (b) Band diagram of the same structure depicting the time response of the bands after the applied voltage is reduced to zero.

will remain. After a time t_1 at elevated temperature, there will be a redistribution of charge as shown, the total amount of charge remaining fixed. Finally, if the temperature is further raised or after an additional time t_2, carriers will be injected from one or both electrodes to bring the system to the equilibrium of Fig. 3c.

The above argument for space-charge formation in insulators is an especially powerful one because it invokes well-known concepts. It simply applies the concepts of steady state, thermal equilibrium, and Fermi level to insulators and shows that the important features of space-charge layers in insulators can be described in terms of only one parameter, the work function or Fermi level position in the insulator at thermal equilibrium. (A more detailed discussion is available in the work of Simmons.[6]) With this as a starting point, it is possible to discuss a wide variety of charging phenomena in insulating thin films in an intuitively understandable way; and it should provide a basis for more quantitative analyses of a number of such effects. In the following section, the concepts discussed here are applied to the silicon/SiO_2/Al_2O_3/metal structure with emphasis on the space-charge region that builds up near the insulator/insulator interface during deposition of the Al_2O_3.

IV. THE SiO_2/Al_2O_3 SYSTEM MODEL

The concepts of the preceding section can be applied to the double-layer structure shown in Fig. 2 in which the first layer is SiO_2 and the second is Al_2O_3. Previous experimental results have indicated that the density of trap levels in thermally grown SiO_2 is sufficiently low that the Debye length should be much larger than the typical SiO_2 film thickness of a few thousand Angstroms.[7] This supports the assumption made previously that no finite charge density exists in the bulk of the SiO_2 film. On the other hand, the trap density in deposited Al_2O_3 films has been found to be much larger so that it is reasonable to assume that the Debye length is small compared to the Al_2O_3 film thicknesses that we will consider.[8–10] The Al_2O_3 films of interest are deposited at 900°C. The double-insulator system is assumed to be at an elevated temperature during the deposition for a sufficient length of time that thermal equilibrium will be established, and under these conditions, the potential diagram versus position will be like that shown in Fig. 5. Since the Fermi level in the silicon must coincide with that in the Al_2O_3 at thermal equilibrium, a contact potential difference will exist that will result in (*i*) an electric field in the SiO_2, (*ii*) a space-charge region in the Si at the Si/SiO_2 interface, (*iii*) a space-charge region in the Al_2O_3 at the SiO_2/Al_2O_3 interface, and (*iv*) zero electric field in the bulk of the Al_2O_3 film. In addition, depending on the Al_2O_3 surface boundary

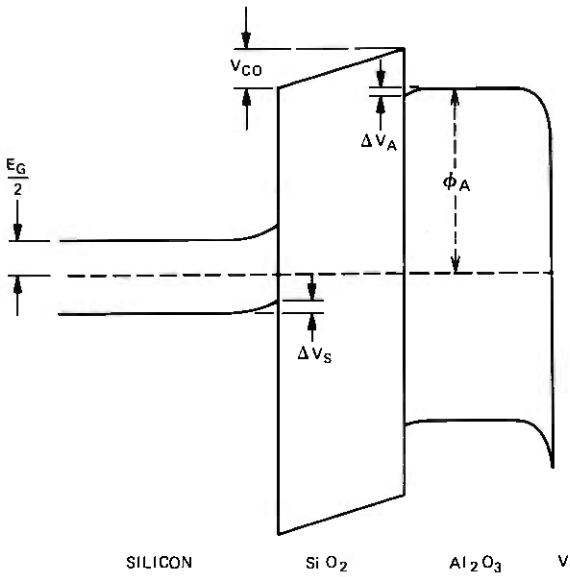

Fig. 5—Band diagram of the $Al_2O_3/SiO_2/Si$ system in equilibrium at high temperature.

conditions, a space-charge region may exist near the outer Al_2O_3 surface.

The important parameter in the SiO_2/Al_2O_3 system is the total potential difference V_c that must build up to align the Fermi levels in the silicon and the Al_2O_3. This potential difference will be made up of a potential V_{co} across the SiO_2 and the drops in potential due to the band-bending regions in the silicon near the Si/SiO_2 interface and in the Al_2O_3 near the SiO_2/Al_2O_3 interface, ΔV_s and ΔV_A, respectively (see Fig. 5). That is,

$$V_c = V_{co} + \Delta V_s + \Delta V_A, \tag{8a}$$

where the system equilibrium condition is given by

$$(E_G/2) + \phi_B + V_c - \phi_{ii} - \phi_A = 0. \tag{8b}$$

If it is assumed that the work function argument applies to the SiO_2/Al_2O_3 system, then

$$V_c = \chi'_A + \phi_A - \chi'_s - (E_G/2), \tag{9}$$

where χ'_A is the Al_2O_3 electron affinity at the deposition temperature; ϕ_A is the energy separation between the Al_2O_3 conduction band and the Fermi level in the bulk; χ'_s is the silicon electron affinity; and the $E_G/2$ term represents the assumption that the silicon is intrinsic at the elevated temperature of interest.

Since the electric field in the SiO_2 is determined directly by the potential drop V_{co} and the oxide thickness T_{ox}, the net charge present in the narrow space-charge region at the Al_2O_3/SiO_2 interface is given by

$$Q_{ii} = (\epsilon_{ox}/T_{ox})\{V_c - \Delta V_s - \Delta V_A\}. \tag{10}$$

Over a wide range of SiO_2 thickness, ΔV_s and ΔV_A will be negligible compared to V_c. For example, even assuming that V_{co} corresponds to an electric field in the SiO_2 of 10^6 volts/cm, ΔV_s is less than 0.1 volt at 900°C.* Thus, eq. (10) can be approximated by

$$Q_{ii} = (\epsilon_{ox}/T_{ox})V_c \approx \epsilon_{ox}/T_{ox}\{\chi_A' + \phi_A - \chi_s' - (E_G/2)\}. \tag{11}$$

From eqs. (10) and (11), several interesting properties of Q_{ii} are apparent. First, its magnitude is relatively independent of the quality and reproducibility of the Al_2O_3 provided only that $\phi_A + \chi_A'$ is reproducible and ΔV_A is negligible. The band-bending ΔV_A may vary markedly from sample to sample depending on the trap density, but as long as the space-charge region is relatively narrow so that ΔV_A is small compared to V_c, this variation will have no significant effect. Second, the magnitude of Q_{ii} varies inversely with the SiO_2 thickness T_{ox} and is independent of Al_2O_3 thickness T_A. This effect provides a straightforward and unique prediction of the model that can easily be tested experimentally.

If the system is now cooled to room temperature, the conductivity of the Al_2O_3 will be reduced to the point where the space-charge regions will not move or change under application of an electric field for long periods of time, and these regions will be effectively frozen in. We must consider the two possible space-charge regions in the Al_2O_3, one at the Al_2O_3/SiO_2 interface, and the other at the outer surface. The contributions V_{ii} and V_m, respectively, of these charge layers to the flatband voltage is given by (see eq. 2)

$$V_{ii} = -\int_{T_{ox}}^{T_{ox}+T_A} \left[\frac{T_{ox} + T_A - x}{\epsilon_A}\right] \rho_{ii}(x)dx = -\frac{T_A}{\epsilon_A} Q_{ii} \tag{12}$$

and

$$V_m = -\int_{T_{ox}}^{T_{ox}+T_A} \left[\frac{T_{ox} + T_A - x}{\epsilon_A}\right] \rho_m(x)dx, \tag{13}$$

where $\rho_{ii}(x)$ and $\rho_m(x)$ are the net charge densities at the SiO_2/Al_2O_3 and Al_2O_3/metal interfaces, respectively. In the limit where $\rho_{ii}(x)$ is located in a plane at the interface, an effective interface charge density Q_{ii} is defined by eq. (12). The term V_m is a constant independent

*This can be shown from an integration of Poisson's equation and using the fact that at 900°C the intrinsic charge density in the silicon is approximately 10^{19} cm^{-3}.

of T_{ox} and T_A if $\rho_m(x)$ is a function only of the distance $(T_{ox} + T_A - x)$ between the charge and the metal. Combining (2), (11), (12), and (13) gives for the flatband voltage of the double-insulator structure

$$V_{FB}(\text{D.I.}) = (\phi_{m,A} + \phi_{ii} - \phi_s) + V_m - Q_{ss}\left[\frac{T_{ox}}{\epsilon_{ox}} + \frac{T_A}{\epsilon_A}\right] - V_c\frac{\epsilon_{ox}T_A}{\epsilon_A T_{ox}}, \quad (14)$$

and, for completeness, the flatband voltage of the single insulator [eq. (3)] is

$$V_{FB}(\text{S.I.}) = (\phi_{m,ox} - \phi_s) - \frac{Q_{ss}}{\epsilon_{ox}}T_{ox}, \quad (15)$$

where $\phi_s = \phi_B + \phi_f$.

V. EXPERIMENTAL RESULTS

5.1 *Preliminary remarks*

Dual-dielectric MIS capacitor structures with various insulator thicknesses were fabricated on n- and p-type silicon substrates with resistivities of approximately 10 ohm-cm. For the n-type substrates both ⟨100⟩ and ⟨111⟩ orientations were investigated. The SiO_2 was thermally grown at 1100°C using oxygen bubbled through 80°C water as the ambient. The Al_2O_3 was vapor deposited on the SiO_2 at 900°C from an $AlCl_3$ source. The details of the Al_2O_3 deposition process are given in Ref. 1. The insulator thicknesses, T_{ox} and T_A, were varied by varying the growth and deposition times of the SiO_2 and the Al_2O_3, respectively.

The deposition of the Al_2O_3 duplicated exactly the procedure used for fabricating integrated circuits. As such, a layer of SiO_2 was also deposited on top of the Al_2O_3, which in the fabrication of integrated circuits is used as an etch mask for defining patterns in the Al_2O_3 film. For our samples, this layer of SiO_2 was chemically stripped prior to any measurements or any further processing.

One feature that may be important in this study is the method of formation of the metal electrodes. Depending on the deposition technique used, the samples may be heated for a sufficient time during the metal deposition to form a space-charge region at the Al_2O_3/metal interface. However, if this induced space charge is reproducible and constant from sample to sample and is spatially constrained to a region very near the interface, it will only influence the flatband voltage via the constant voltage term V_m in (14). Experimentally, we shall attempt to assure the reproducibility of this possible space-charge effect by measuring the MIS structures at room temperature with a

mercury electrode.[11] Several samples were also investigated with thermally evaporated titanium-aluminum electrodes.

Initially, we attempted to vary the Al_2O_3 thickness by etching in discrete steps rather than by varying the deposition time. This approach was abandoned because of nonuniform etching of the Al_2O_3. In Fig. 6, scanning electron micrographs are given of the surface of the Al_2O_3 as deposited and after etching a portion of the layer. Microscopic thickness variations (± 500 Å) are evident after etching. Since these variations lead to significant errors in the flatband voltage measurements, the Al_2O_3 thickness was varied only by varying the growth time.

Experimental data for each sample investigated were obtained by means of high-frequency capacitance-voltage (C–V) analysis.[12] Such measurements, obtained using either the mercury probe electrode or thermally evaporated titanium-aluminum thin-film electrodes, enable one to obtain accurate measurements of the insulator thickness and

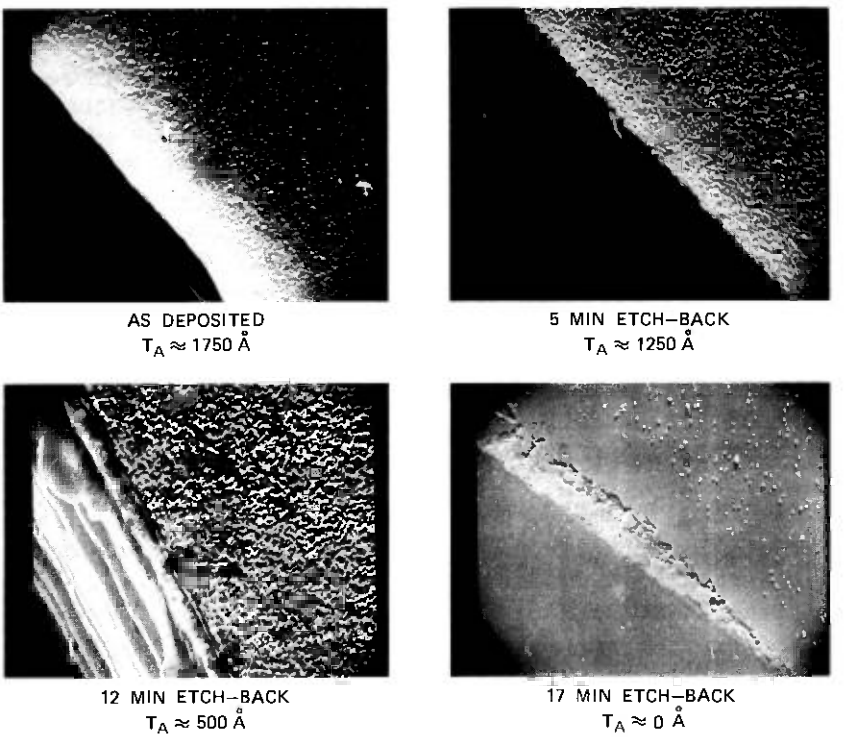

Fig. 6—SEM photographs depicting the increasing roughness of the Al_2O_3 surface as it is etched back using phosphoric acid.

the associated flatband voltage. Each data point reported is the average flatband voltage calculated from at least three measurements performed on each sample. The typical spread in measurements is 0.10 volt. When measurements are performed on the double-insulator (Al_2O_3/SiO_2) structure, a measurement of V_{FB}(D.I.) [eq. (14)] is obtained. The normalized accumulation capacitance C_{acc} (farads/cm^2), which is a measure of the insulator thicknesses, is given by

$$C_{acc}(\text{D.I.}) = \left[\frac{T_{ox}}{\epsilon_{ox}} + \frac{T_A}{\epsilon_A} \right]^{-1}. \tag{16}$$

When the Al_2O_3 is completely etched off and C–V analysis is conducted on the remaining single insulator (SiO_2), then measurements of V_{FB}(S.I.) [see eq. (15)] are obtained. The normalized accumulation capacitance in this case yields a measurement of T_{ox} since

$$C_{acc}(\text{S.I.}) = \frac{\epsilon_{ox}}{T_{ox}}. \tag{17}$$

By combining eqs. (16) and (17), accurate measurements of both T_A and T_{ox} are obtained.

It is possible to obtain independent quantitative values for Q_{ss} for each sample studied after the Al_2O_3 is etched off if the constant term $(\phi_{m,ox} - \phi_s)$ in eq. (14) is known. Measurement of $(\phi_{m,ox} - \phi_s)$ can be accomplished by SiO_2 etch-back experiments in which V_{FB}(S.I.) is measured as the SiO_2 layer is successively thinned by etching in a dilute hydrofluoric acid solution. Typical data obtained with a mercury probe on four different samples are shown in Fig. 7. As expected, there is a linear relationship between V_{FB}(S.I.) and T_{ox} and an extrapolation of this relationship back to $T_{ox} = 0$ indicates that 0.67 volt is the appropriate value of $(\phi_{m,ox} - \phi_s)$ for n-type, 10-ohm-cm Si and a mercury electrode. This value is in excellent agreement with previously determined values.[13] The experiment also provides independent verification of the assumption that there is negligible space charge in the bulk of the SiO_2.

Based upon the value of Q_{ss} for each sample, it is possible to characterize the flatband voltage shift due to the Al_2O_3. Correcting for the Q_{ss} term and, additionally, subtracting the constant term $(\phi_{m,ox} - \phi_s)$ from eq. (14), a corrected differential flatband voltage ΔV_{FB} can be defined as:

$$\begin{aligned} \Delta V_{FB} &= V_{FB}(\text{D.I.}) - (\phi_{m,ox} - \phi_s) + Q_{ss}\left[\frac{T_{ox}}{\epsilon_{ox}} + \frac{T_A}{\epsilon_A} \right] \\ &= (\phi_{m,A} + \phi_{ii} - \phi_{m,ox}) + V_c\left(\frac{\epsilon_{ox}}{\epsilon_A}\right)\left(\frac{T_A}{T_{ox}}\right) + V_m. \end{aligned} \tag{18}$$

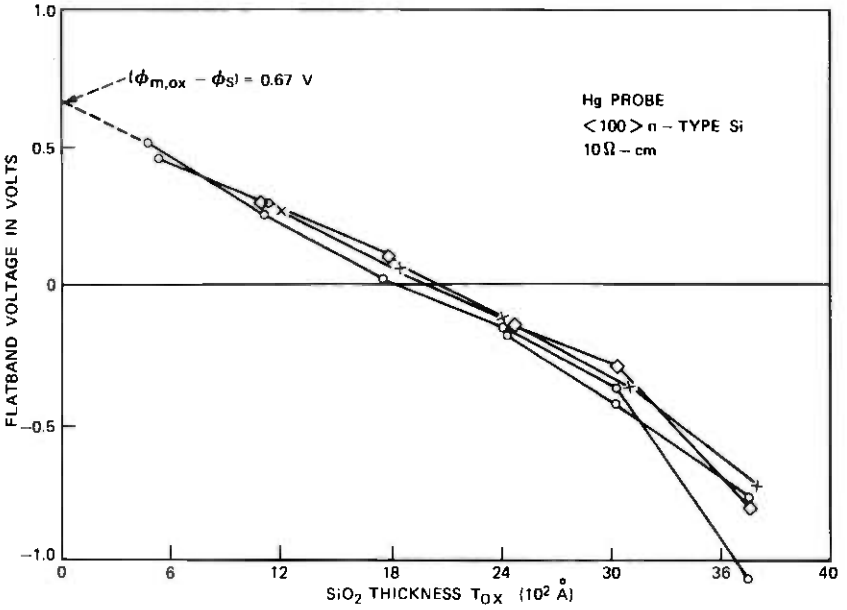

Fig. 7—Plot of the Hg/SiO_2/Si flatband voltage for four wafers as a function of the SiO_2 thickness. Data obtained from SiO_2 etch-back experiments on 10-ohm-cm, n-type, ⟨100⟩, Si substrates.

5.2 Results pertaining to Q_{ii}

Experimental values of V_{FB}(D.I.) for the dual-insulator structure are plotted in Fig. 8 as a function of the Al_2O_3 thickness T_A for a SiO_2 thickness of $\approx$1200 Å on n-type, ⟨100⟩ substrates. These results were obtained with a mercury electrode. Although there is considerable scatter in the data, it is clear that V_{FB}(D.I.) increases monotonically with increasing Al_2O_3 thickness which is in agreement with eq. (14) if the sign of V_c is such that a net negative charge exists at the SiO_2/Al_2O_3 interface. The experimental uncertainties in the V_{FB} measurements are estimated to be $\pm$0.05 volt. Correcting this data for Q_{ss} and subtracting $(\phi_{m,ox} - \phi_s)$, the results are replotted in Fig. 9. This refinement technique leads to a considerable reduction in the scatter in the data and demonstrates that ΔV_{FB} is a linear function of the Al_2O_3 thickness T_A, as predicted by eq. (18). The linear relationship also provides striking evidence that Q_{ii}, the negative charge at the Al_2O_3/SiO_2 interface, is constant from sample to sample for Al_2O_3 thicknesses in the range of 500 Å to 2500 Å if the SiO_2 thickness is held constant. This is in agreement with the postulated model and provides experimental verification of the assumption that the Debye length in Al_2O_3 is small compared to the Al_2O_3 thickness. Given that the Debye length

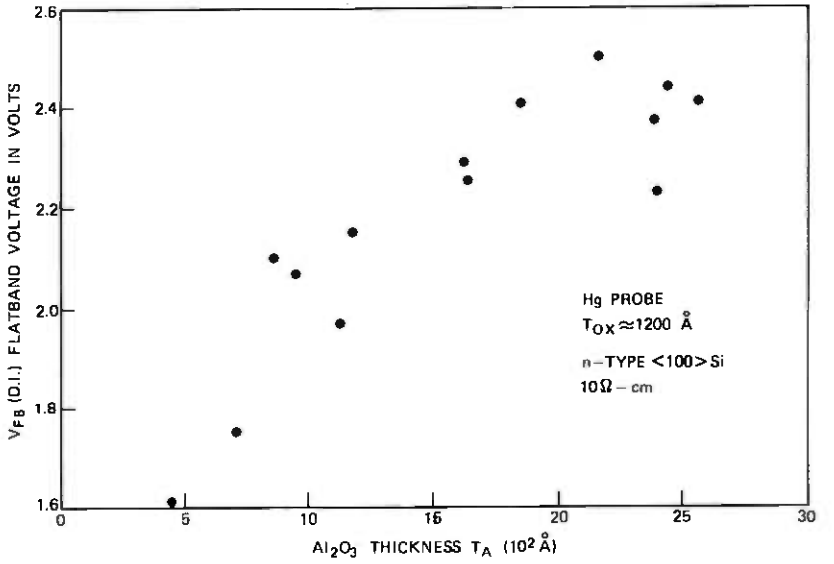

Fig. 8—Plot of the Hg/Al_2O_3/SiO_2/Si flatband voltage as a function of the Al_2O_3 thickness for a constant SiO_2 thickness (T_{ox} = 1200 Å) on n-type, ⟨100⟩, Si substrates.

is small compared to 500 Å at 900°C, estimates of ΔV_A assuming charge densities in excess of 10^{18} cm^{-3} indicate that ΔV_A will be less than 0.1 volt and, thus, negligible as previously assumed.

The data presented so far proves that Q_{ii} is negative and a constant for fixed SiO_2 thickness independent of Al_2O_3 thickness. Another prediction of our model is that Q_{ii} is inversely proportional to the SiO_2

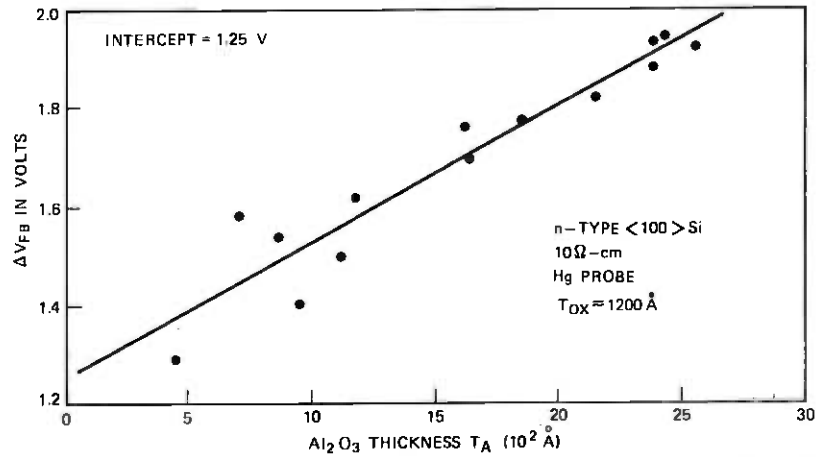

Fig. 9—Plot of the corrected differential flatband voltage as a function of the Al_2O_3 thickness for a constant SiO_2 thickness (T_{ox} = 1200 Å) on n-type, ⟨100⟩, Si substrates.

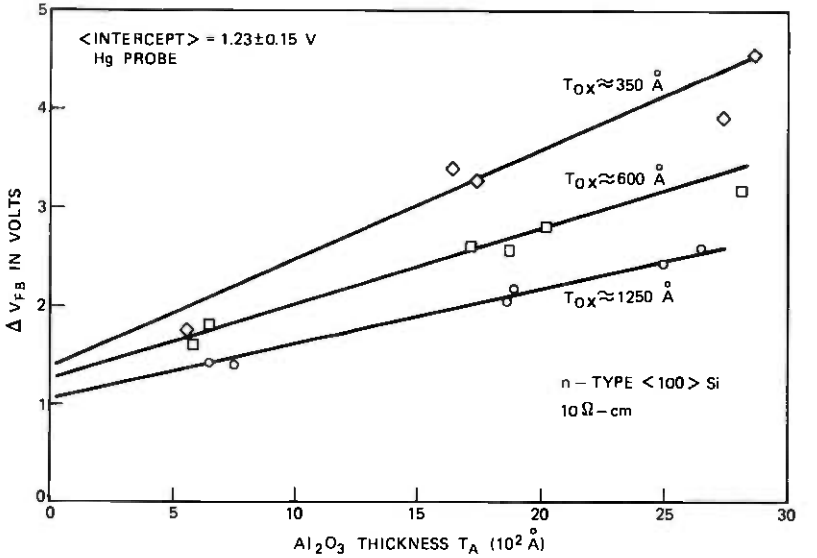

Fig. 10—Plot of the corrected differential flatband voltage as a function of the Al_2O_3 thickness for various values of SiO_2 thickness on n-type, ⟨100⟩, Si substrates.

thickness [eq. (11)]. That this is indeed the case is shown by the data presented in Figs. 10 and 11, which were obtained with a mercury probe and are for ⟨100⟩, n-type, silicon substrates. In Fig. 10, ΔV_{FB} is

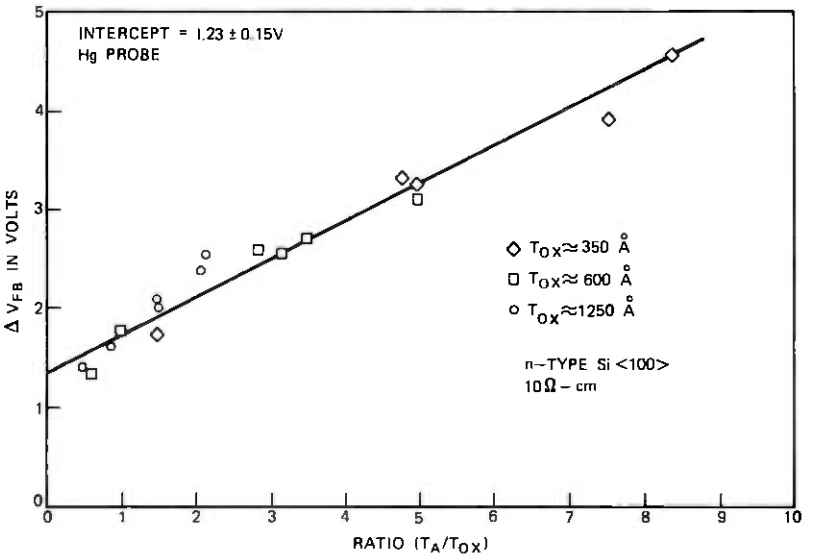

Fig. 11—Plot of the corrected differential flatband voltage as a function of the ratio of Al_2O_3 thickness to SiO_2 thickness on n-type, ⟨100⟩, Si substrates.

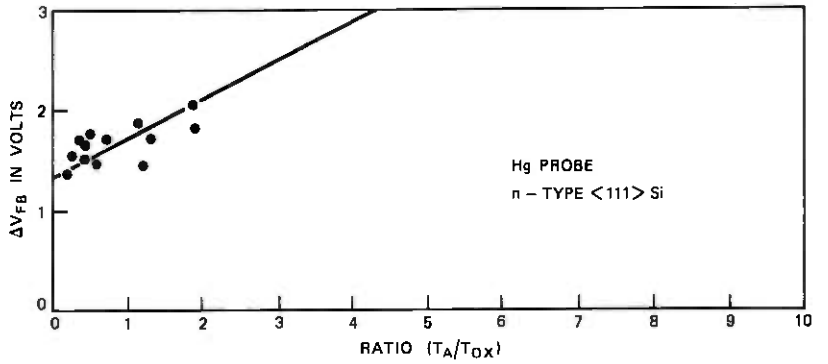

Fig. 12—Plot of the corrected differential flatband voltage as a function of the ratio of Al_2O_3 thickness to SiO_2 thickness on n-type, ⟨111⟩, Si substrates.

plotted versus Al_2O_3 thickness for three different SiO_2 thicknesses. In each case, a linear relationship between ΔV_{FB} and Al_2O_3 thickness is obtained, and the increased slope obtained with smaller SiO_2 thicknesses indicates that Q_{ii} does increase as the SiO_2 thickness is decreased. The data of Fig. 10 are replotted in Fig. 11 as a function of T_A/T_{ox}, the ratio of Al_2O_3 thickness to SiO_2 thickness. As expected from eq. (18), the ΔV_{FB} versus T_A/T_{ox} relationship is accurately represented by a straight line over the T_A/T_{ox} range investigated (0.5 to 8), indicating that Q_{ii} is inversely proportional to the SiO_2 thickness. The slope of the straight line in Fig. 11 corresponds to a V_c value of 0.88 volt.* This value for V_c was obtained in all the measurements on ⟨100⟩ substrates within ±0.1 volt.

Similar measurements were also performed with ⟨111⟩ oriented, n-type substrates. The larger values of Q_{ss} inherent in the ⟨111⟩ orientation meant that the Q_{ss} correction factor was much larger and, hence, the accuracy of the results was somewhat poorer. Results for ⟨111⟩ samples are given in Fig. 12, where ΔV_{FB} is plotted as a function of the T_A/T_{ox} ratio. Again these data were obtained with a mercury probe. The straight line shown in Fig. 12 is a best fit to the data if the slope of the line is restricted to correspond to a V_c value of 0.88 volt. Considering the possible errors due to the Q_{ss} correction, the straight-line fit of the data in Fig. 12 is good enough to conclude that the value of Q_{ii} is independent of the substrate orientation for the two orientations investigated, ⟨100⟩ and ⟨111⟩, and in complete agreement with the predictions of our model.

* For 1000 Å of SiO_2, a V_c value of 0.88 volt corresponds to a Q_{ii} value of 1.9×10^{11} charges/cm^2.

5.3 Results pertaining to Q_{ss}

All of the experimental results presented so far have focused on Q_{ii}, the charge at the Al_2O_3/SiO_2 interface. Some interesting facets of Q_{ss}, the charge at the SiO_2/Si interface, were also discovered during our study and are discussed in the following. As mentioned previously, a value of Q_{ss} was determined for all samples by measuring the flatband voltage V_{FB}(S.I.) of the single-insulator structure after removing the Al_2O_3 and then calculating Q_{ss} using the $(\phi_{m,ox} - \phi_s)$ value for the Hg/SiO_2/Si system as determined in Fig. 7. A plot of V_{FB}(S.I.) versus T_{ox}, the SiO_2 thickness, for n-type, ⟨100⟩ substrates is given in Fig. 13. Previous results published in the literature have shown that for single-insulator (SiO_2/Si) structures, Q_{ss} is independent of the SiO_2 thickness.[14] If this were the case for our structures, we would expect to find a linear relationship between V_{FB}(S.I.) and T_{ox} with an intercept on the V_{FB}(S.I.) axis equal to $(\phi_{m,ox} - \phi_s) = 0.67$ volt. The results given in Fig. 13 indicate that this is not the case. Although the data could be interpreted as being consistent with a linear relationship, they are definitely not consistent with an intercept equal to 0.67 volt. The results are more consistent with the supposition that, to first order, V_{FB}(S.I.) is independent of T_{ox}.

A more detailed study of Q_{ss} was pursued by preparing samples of various SiO_2 thicknesses (n-type, Si, ⟨100⟩) and measuring V_{FB}(S.I.) for each sample. Approximately 500 Å of Al_2O_3 was then deposited

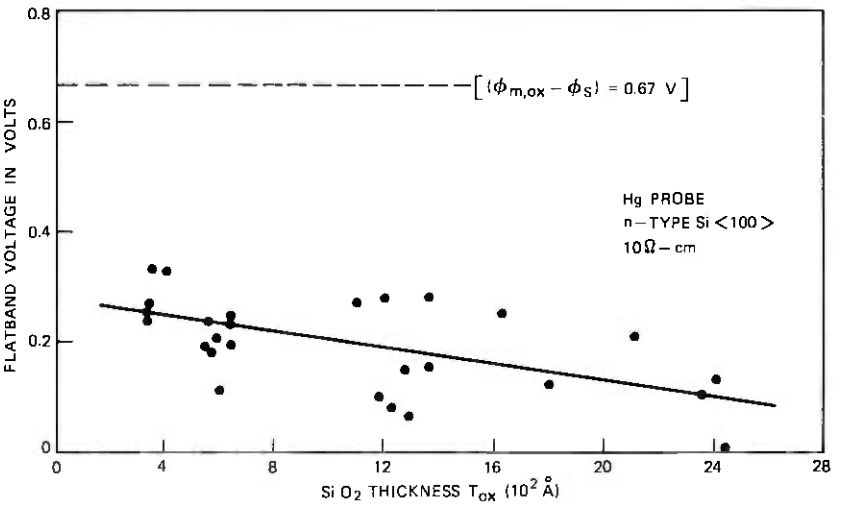

Fig. 13—Plot of the Hg probe single-insulator flatband voltage after Al_2O_3 deposition and etch-off as a function of SiO_2 thickness for n-type, ⟨100⟩, Si substrates.

on all samples, the Al_2O_3 was removed by etching, and V_{FB}(S.I.) was remeasured. Finally, for each of the samples, the SiO_2 was etched back in steps, and V_{FB}(S.I.) was determined as a function of SiO_2 thickness. The results are given in Fig. 14.

Prior to Al_2O_3 deposition, the results are consistent with the samples having a constant value of Q_{ss} independent of T_{ox}, that is, a linear relationship exists between V_{FB}(S.I.) and T_{ox} with an intercept equal to 0.67 volt. However, after Al_2O_3 deposition, V_{FB}(S.I.) is seen to be essentially independent of T_{ox}. Furthermore, if the SiO_2 is now etched back, a linear relationship between V_{FB}(S.I.) and SiO_2 thickness is obtained with an intercept equal to 0.67 volt. In Fig. 14, results of the etch-back study are given for only one representative sample, since the results obtained on the other samples were similar.

The conclusion which follows from the results given in Fig. 14 is that before the deposition of the Al_2O_3, Q_{ss} is independent of T_{ox}, whereas after deposition, the value of Q_{ss} is changed, the amount of change depending upon T_{ox}, the SiO_2 thickness. This effect is further illustrated by the results given in Fig. 15, where Q_{ss} after Al_2O_3 deposition and etch-off is plotted versus $1/T_{ox}$ for both ⟨100⟩ and ⟨111⟩, n-type substrates. For both orientations, the data are seen to fall

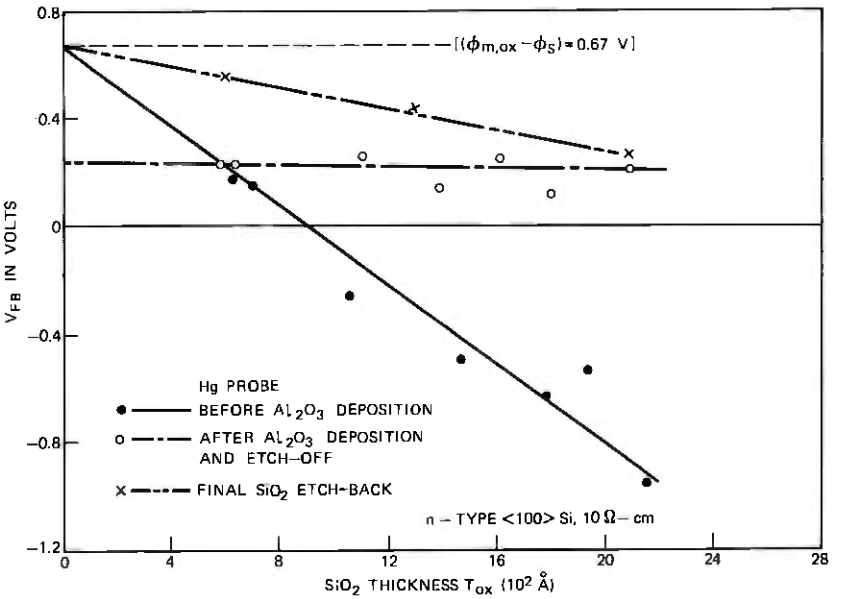

Fig. 14—Plot of the single-insulator flatband voltage before and after Al_2O_3 deposition and after final SiO_2 etch-back, indicating the change in Q_{ss} induced by the Al_2O_3 deposition on n-type, ⟨100⟩, Si substrates.

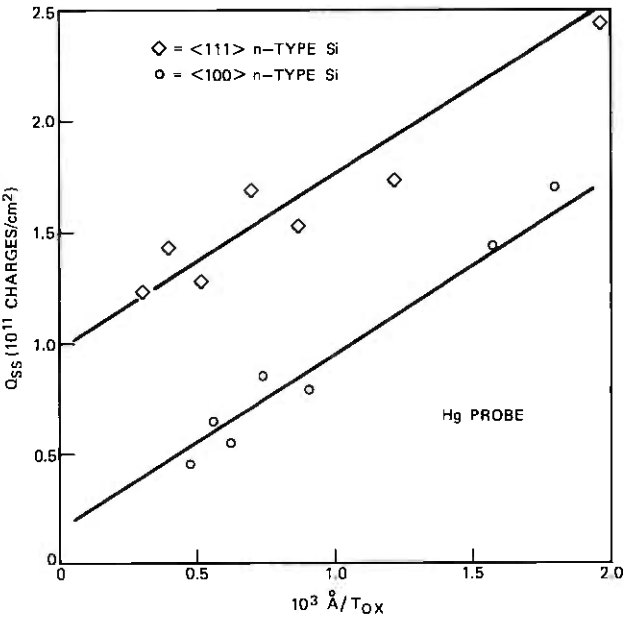

Fig. 15—Plot of Q_{ss} after Al_2O_3 deposition as a function of the SiO_2 thickness for n-type, ⟨100⟩ and ⟨111⟩, Si substrates.

along a straight line consistent with the equation

$$Q_{ss} = (v_o\epsilon_{ox}/T_{ox}) + Q_{sso}. \tag{19}$$

The slopes of the two lines in Fig. 15 were taken to be equal to each other ($v_o = 0.38$ volt), and it is observed that an excellent fit to the two sets of data is obtained with the one v_o value. The background charge densities, Q_{sso}, or equivalently the values of Q_{ss} for large values of T_{ox} are approximately 1.0×10^{11} and 1.0×10^{10} charges/cm^2 for ⟨111⟩ and ⟨100⟩ orientations, respectively. For the ⟨100⟩ orientation, Q_{sso} is negligible.

The results presented so far have established that for the double-insulator structure, both Q_{ii} and Q_{ss} depend on T_{ox} and that these dependencies can be written as:

$$\begin{aligned} Q_{ii} &= \epsilon_{ox}V_c/T_{ox} \\ Q_{ss} - Q_{sso} &= \epsilon_{ox}v_o/T_{ox}, \end{aligned} \tag{20}$$

where $V_c \simeq 0.88$ volt and $v_o \simeq 0.38$ volt. Thus, the single-insulator flatband voltage after Al_2O_3 deposition can be written as

$$V_{FB}(\text{S.I.}) = (\phi_{m,ox} - \phi_s) - v_o - (T_{ox}/\epsilon_{ox})Q_{sso}, \tag{21}$$

which is consistent with the after-deposition results presented in Fig. 14. Considering Q_{sso} as a fundamental property of the SiO_2/Si interface, which is not affected by the Al_2O_3 deposition, it follows that the contribution to the single-insulator flatband voltage, δV_{FB}, induced by the portion of Q_{ss} that is influenced by the Al_2O_3 deposition is

$$\delta V_{FB} = v_o. \tag{22}$$

The voltage drop across the SiO_2 during the deposition of the Al_2O_3 is V_c, and if this is considered as a stress voltage applied to the SiO_2/Si interface, then

$$V_c(\text{stress})/\delta V_{FB} = \alpha = 2.3. \tag{23}$$

It is interesting to note that this result is in good agreement with previously published results relating stress voltage to saturated flatband voltage shift for SiO_2/Si structures. Specifically, in Ref. 15 it was observed that the ratio of stress voltage to saturated flatband voltage shift was given by:

$$\alpha = \begin{cases} 3.33 \text{ at } 350°C \\ 2.38 \text{ at } 450°C. \end{cases} \tag{24}$$

One implication of the relationship given in eq. 20 for Q_{ss} is that for ⟨100⟩ substrates, where Q_{sso} is negligible, the flatband voltage of a double-insulator structure will be insensitive to a SiO_2 thickness variation. Thus, in MOSFET integrated circuits with a ⟨100⟩ substrate, where a thick SiO_2 layer is used to inhibit parasitic inversion, the contribution of the Q_{ss} term to the parasitic inversion voltage will be independent of SiO_2 thickness.

5.4 Results pertaining to p-type substrates

The substrate conductivity type does not appear directly in the model that has been proposed for the magnitude and origin of Q_{ii}, and all of the results presented so far have been for n-type substrates. Since the silicon substrate will be intrinsic at 900°C, the deposition temperature of the Al_2O_3, the voltage drop across the oxide V_c will be the same for both n-type and p-type substrates and, thus, Q_{ii} at 900°C should also be independent of the conductivity type of the substrate. If, during the cool down after Al_2O_3 deposition, Q_{ii} is frozen in at a temperature at which the silicon is still intrinsic, then the value of Q_{ii} measured at room temperature should not depend on whether the substrate is n-type or p-type. Results obtained with ⟨100⟩, p-type substrates are presented below.

A plot for p-type substrates similar to that of Fig. 11 (for n-type substrates) is given in Fig. 16, where ΔV_{FB} is plotted as a function of

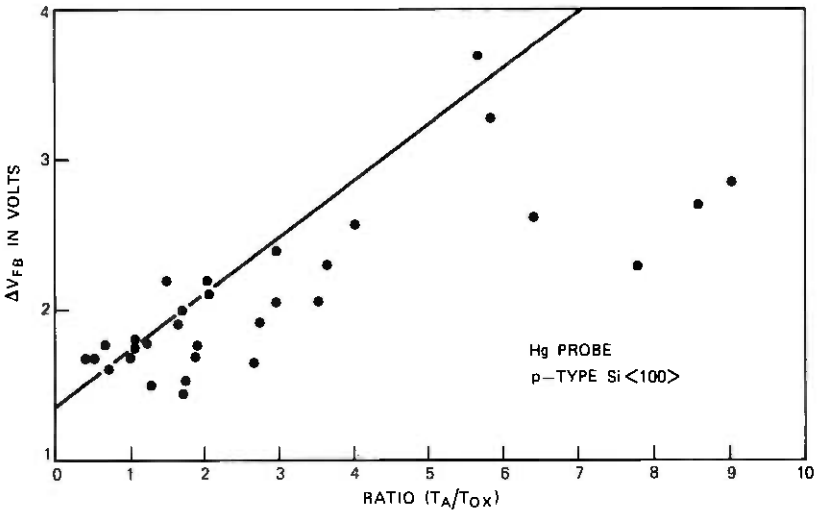

Fig. 16—Plot of the corrected differential flatband voltage as a function of the ratio of Al_2O_3 thickness to SiO_2 thickness on p-type, ⟨100⟩, Si substrates. The solid line is taken from Fig. 11.

T_A/T_{ox}. The solid line corresponds to the linear fit of the data of Fig. 11. Although there is general agreement between the results for the n-type substrate (solid line) and this experimental data, the large amount of scatter in the data must be recognized. Figure 17 is a plot of the single-insulator flatband voltage for the p-type substrates after the Al_2O_3 etch-off. The $(\phi_{m,ox} - \phi_s)$ value was obtained by etch-back experiments, as outlined previously. Here again, a large amount of scatter in the data is evident.

The variations in the above sets of data are not random scatter, but are due to some mechanism unique to the p-type substrates. In Fig. 18, the single-insulator flatband voltage V_{FB}(S.I.) is plotted as a function of T_A for n-type samples after Al_2O_3 etch-off, and it is clear that no dependence on T_A or T_{ox} is evident. In Fig. 19, similar data are plotted for the p-type samples and it is evident that in this case there is a dependence on T_A, but again, no dependence on T_{ox}. For both cases, the conclusions regarding T_{ox} are obtained from the points in Figs. 18 and 19, which explicitly denote points of constant T_{ox}. Although a detailed explanation of this effect cannot be given, it is felt that this effect is due to the fact that boron-doped p-type wafers were used in the experiment. It is known that boron will greatly out-diffuse from a silicon substrate into an SiO_2 layer in the presence of a high-temperature, hydrogen-containing ambient.[16,17] Additionally, the introduction of this impurity into the SiO_2 may enhance its conduc-

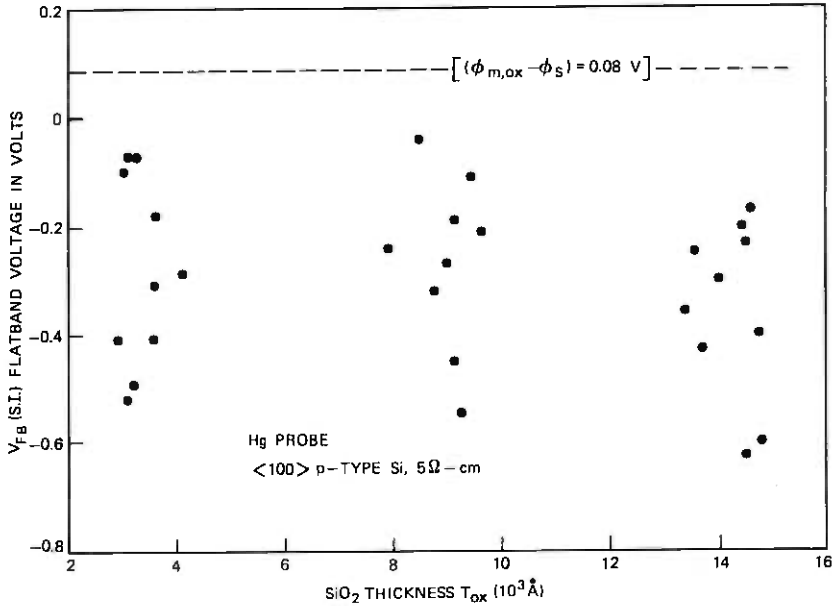

Fig. 17—Plot of the single-insulator flatband voltage after Al_2O_3 deposition and etch-off as a function of the SiO_2 thickness for p-type, ⟨100⟩, Si substrates.

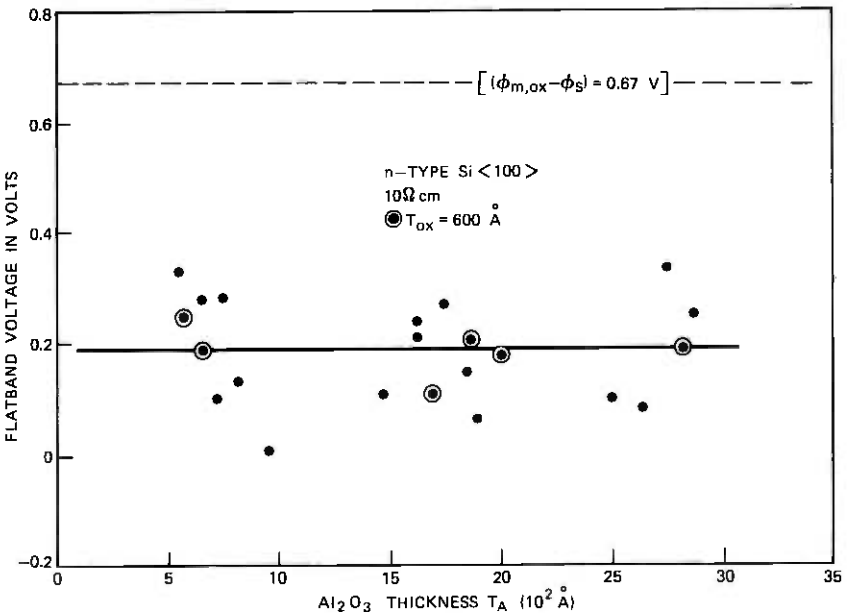

Fig. 18—Plot of the single-insulator flatband voltage after Al_2O_3 deposition and etch-off as a function of the Al_2O_3 thickness for n-type, ⟨100⟩, Si substrates.

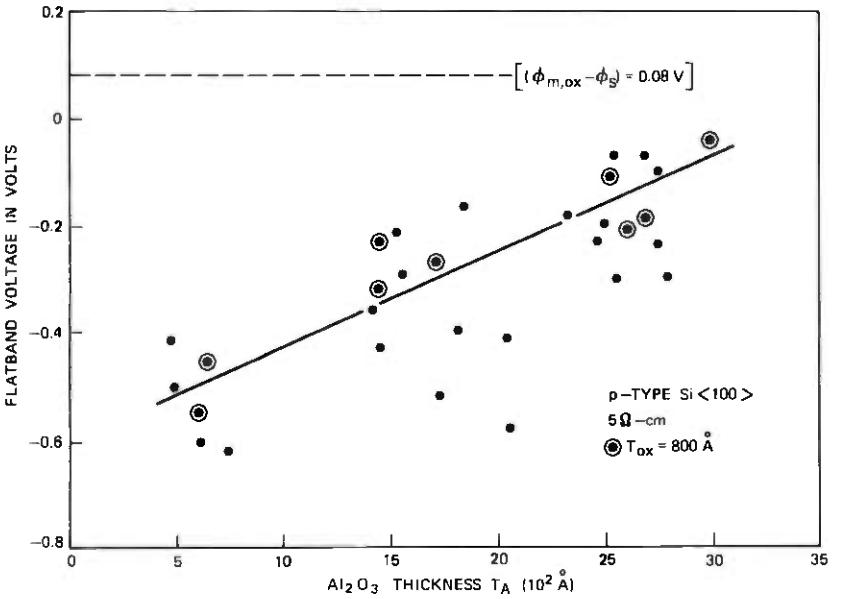

Fig. 19—Plot of the single-insulator flatband voltage after Al_2O_3 deposition and etch-off as a function of the Al_2O_3 thickness for p-type, ⟨100⟩, Si substrates.

tivity at high temperature. The net result will be a drop in the effective stress voltage V_c across the SiO_2 layer as a function of Al_2O_3 deposition time* and a lowering of the Q_{ii} and Q_{ss} term.

This hypothesis is consistent with the following data. In Fig. 20, distributions of the potential drop v_o due to the induced Q_{ss} term (see eqs. 19 and 20) are plotted for both the n-type and p-type ⟨100⟩ samples. It is observed that

(*i*) Lower voltage drops for p-type samples occur than observed for the n-type samples (the lower values are correlated to the thicker Al_2O_3 deposition).
(*ii*) No zero (or negative) voltage drops occur.
(*iii*) The upper range of v_o for the p-type samples (the lowest SiO_2 conductivity region), are bounded by the v_o values observed for the n-type samples.

One final point can also be made to support the hypothesis. It is possible to calculate Q_{ii} values from the experimental data (via eqs. 10 and 18) if an intercept voltage value (i.e., $T_A = 0$) is assumed; and from the experimental data for n-type samples, an intercept value of

* In all experiments, the Al_2O_3 deposition rate was a constant (75 Å/min.).

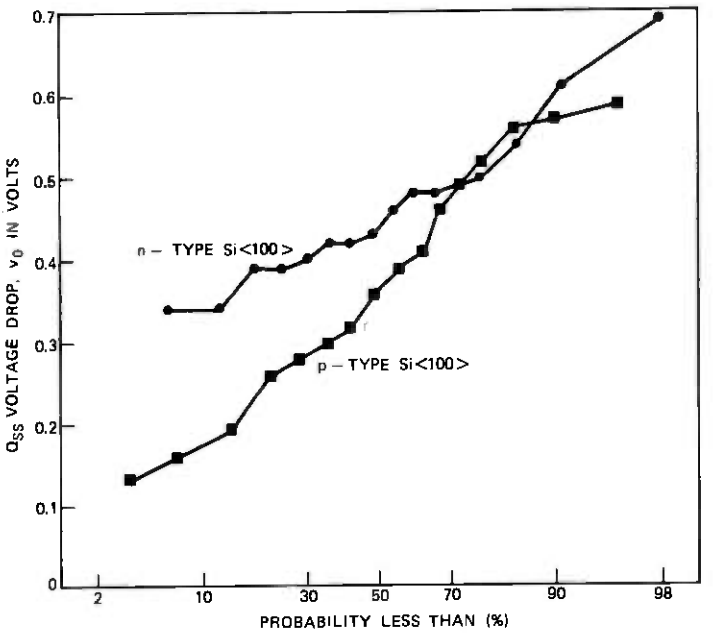

Fig. 20—Distribution plot of the single-insulator flatband voltage observed on n-type and p-type, ⟨100⟩, Si substrates after Al_2O_3 deposition and etch-off.

1.25 volts was obtained. Additionally, Q_{ss} values for each sample may be calculated. If the above hypothesis concerning a lowering of V_c is correct, then the ratio of Q_{ii} and Q_{ss} (given in eqs. 20 and 23) should be independent of the absolute magnitude of V_c for ⟨100⟩ substrates in which Q_{sso} is negligible. That is,

$$Q_{ss} = \epsilon_{ox} v_o / T_{ox} = \epsilon_{ox} \alpha^{-1} V_c / T_{ox}$$

and (25)

$$Q_{ii}/Q_{ss} = \alpha.$$

Figure 21 is a plot of Q_{ii} vs Q_{ss} for both the n-type and p-type samples. It is noted that a linear relationship exists for both sets of data with α the same for both ($\alpha \approx 2.5$). The several data points that deviate from the linear relation are associated with very small Q_{ss} and Q_{ii} values, and the deviation is most likely due to small errors in flatband voltage measurements.

5.5 Results pertaining to observed potential jumps

Extrapolating the linear relationships in Figs. 9, 11, and 12 to $T_A = 0$ gives a value:

$$(\phi_{m,A} + \phi_{ii} - \phi_{m,ox}) + V_m \triangleq \Delta\phi + V_m \simeq 1.25 \text{ volts}. \quad (26)$$

While the experiments cannot determine the relative contributions of V_m and $\Delta\phi$ to the intercept value, it is worth pointing out the consequences of the two limiting possibilities. First, if V_m is zero, then $\Delta\phi$ is non-zero and equal to 1.25 volts. This implies that the work function model for barrier heights must be incorrect (otherwise $\Delta\phi = 0$). The second limiting possibility is that $\Delta\phi = 0$ and V_m is non-zero. In this case, there must be a 1.25-volt band-bending effect at the outer Al_2O_3 interface. Although we have not been able to perform an experiment that unequivocally separates the contributions of $\Delta\phi$ and V_m to the intercept value, it is worthwhile to consider some additional items of relevant experimental information.

First, it was stated previously that all samples are fabricated with a deposited SiO_2 layer on top of the Al_2O_3 layer. An etch-back experiment was conducted on the deposited SiO_2 (using n-type Si, ⟨100⟩, $T_{ox} = 600$ Å, $T_A = 500$ Å) and the flatband voltage was measured as a function of the equivalent SiO_2 thickness T_{eq}:

$$T_{eq} = T_{ox} + (\epsilon_{ox}/\epsilon_A)T_A + T_{SiO_2}, \tag{27}$$

where T_{SiO_2} equals the deposited SiO_2 thickness. The results of this

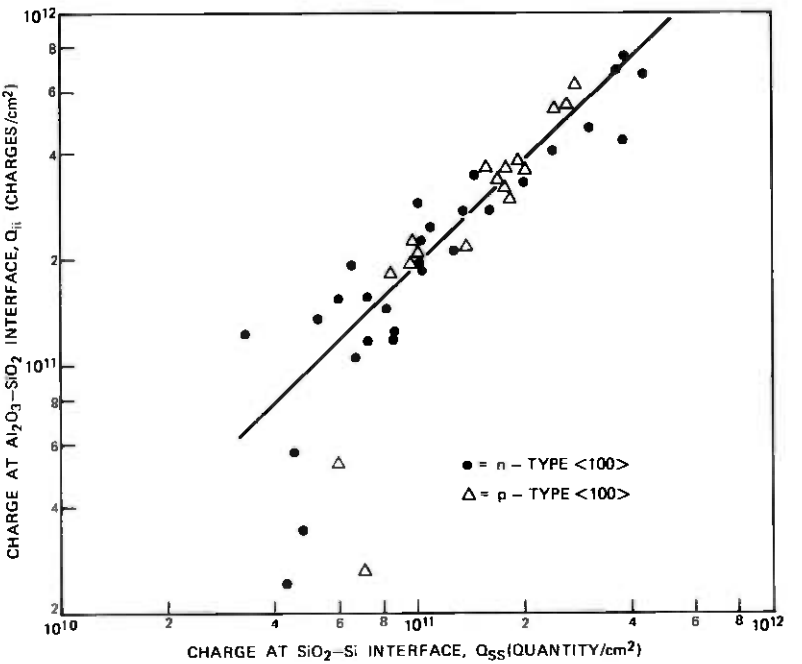

Fig. 21—Plot of Q_{ii}, the charge at the Al_2O_3/SiO_2 interface, versus Q_{ss}, the charge at the SiO_2/Si interface, for both n-type and p-type, ⟨100⟩, Si substrates.

experiment are plotted in Fig. 22. It is evident from the linearity of the flatband voltage that no significant charge density is present in the bulk of this deposited SiO_2. It is interesting to note that a positive potential jump of 1.22 volts in the flatband voltage is associated with the outer Al_2O_3/SiO_2 deposited interface (using the Hg metal electrode). This value is close to the 1.25-volt potential jump associated with the inner Al_2O_3/SiO_2 (thermal) interface.

It is straightforward to show that the potential jump observed in Fig. 22 is given by

$$PJ = (\phi_{m,A} + \phi_{ii} - \phi_{m,ox}) \tag{28}$$

if it is assumed that there is no change in the charge distribution in the Al_2O_3 film when the deposited SiO_2 is completely removed, and that the barrier height of metal-to-deposited-SiO_2 is the same as the barrier height of metal-to-thermal-SiO_2. With these assumptions, the conclusion follows that $V_m \simeq 0$ and $\Delta\phi \simeq 1.25$ volts.

Second, measurements were also made with titanium-aluminum evaporated electrodes. The double-insulator flatband voltage for this metallization system is plotted in Fig. 23 as a function of the Al_2O_3 thickness T_A for a constant SiO_2 thickness $T_{ox} \approx 1200$ Å (n-type Si, ⟨100⟩). It is evident that the scatter in the data is much greater than that found for the Hg metallization. Attempts to refine the data proved

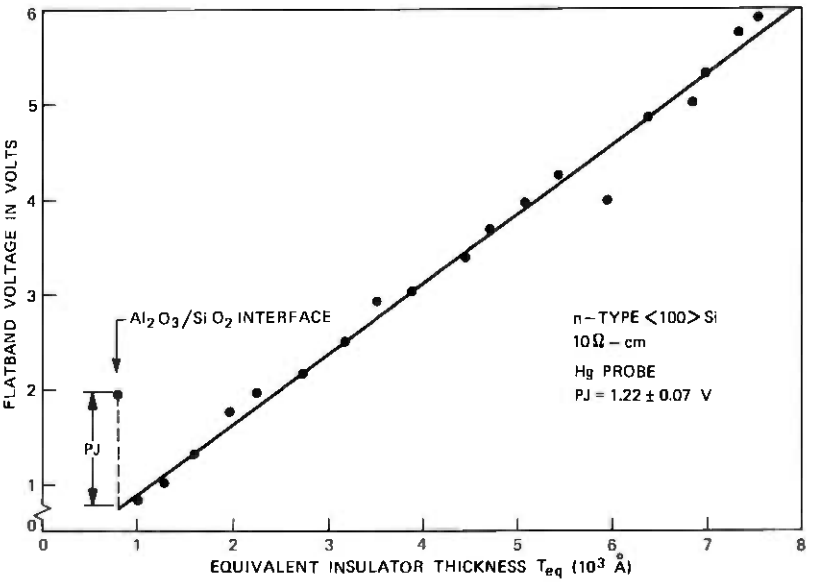

Fig. 22—Plot of the triple-insulator flatband voltage, obtained by means of an etch-back experiment, as a function of the equivalent insulator thickness.

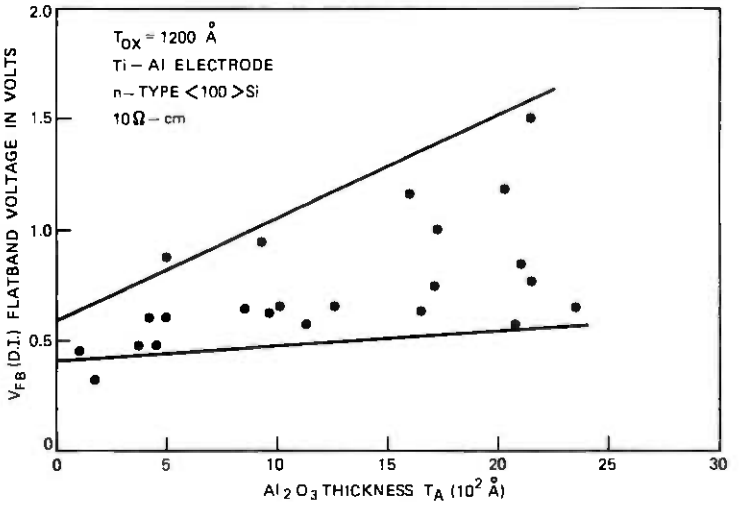

Fig. 23—Plot of the titanium-gate double-insulator flatband voltage as a function of the Al_2O_3 thickness for a constant SiO_2 thickness (T_{ox} = 1200 Å) on n-type, ⟨100⟩, Si substrates.

fruitless due to the additional scatter observed in the single-insulator (SiO_2) flatband voltages (see Fig. 24).

Some general comments can be made, however, concerning these data. The scatter in the double-insulator flatband voltage (Fig. 23) decreases with descreasing Al_2O_3 thickness, indicating that an uncontrolled charging effect must take place in the Al_2O_3 during the

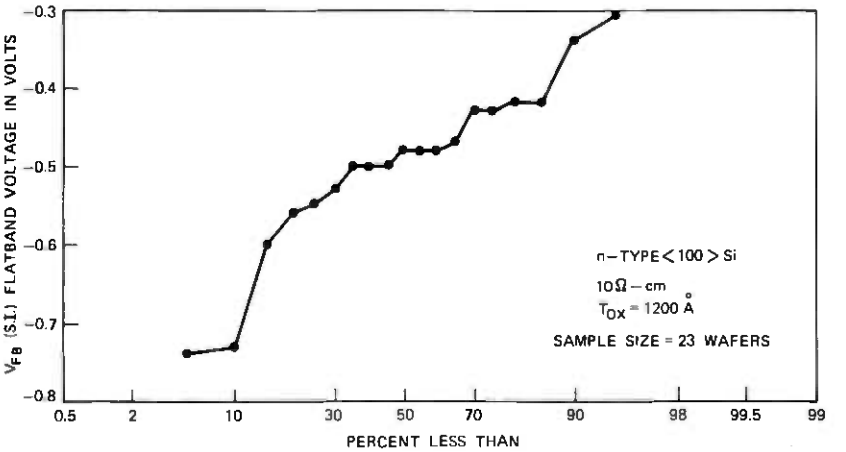

Fig. 24—Distribution plot of the titanium-gate single-insulator flatband voltage after Al_2O_3 etch-off as observed on the samples reported in Fig. 23.

metal deposition. Extrapolating to $T_A = 0$ implies

$$0.40 \text{ V} \leqslant [V_{FB}(\text{D.I.})|_{T_A=0}] \leqslant 0.60 \text{ volt}. \tag{29}$$

The distribution in the single-insulator flatband voltage V_{FB}(S.I.) can be characterized by (see Fig. 24)

$$\begin{aligned} \text{Average } V_{FB}(\text{S.I.}) &= -0.48 \text{ volt} \\ \text{Standard Deviation } V_{FB}(\text{S.I.}) &= 0.10 \text{ volt}. \end{aligned}$$

Thus, a positive potential jump of 0.98 ± 0.14 volt can be associated with the addition of the Al_2O_3 layer. This value is in reasonable agreement with the previous potential shift results found for the Hg metallization system.

Third, in Ref. 4, the authors find on a V_{FB}(D.I.) versus T_A plot for various T_{ox} values an intercept value of -0.80 volt at $T_A = 0$. This result, obtained on 2-ohm-cm, p-type, ⟨100⟩, Si substrates, is interpreted by the authors to be the expected metal-to-silicon work function difference when using aluminum electrodes, a conclusion that may not be valid since Q_{ss} measurements after Al_2O_3 deposition were not reported. We shall now show that the results in Ref. 4 are in excellent agreement with our results by taking the results we have obtained with a Hg electrode and converting them to the results we would have obtained if we had used an Al electrode.

To explicitly denote the use of a Hg electrode, eq. (26) is rewritten as

$$(\phi_{\text{Hg, A}} + \phi_{ii} - \phi_{\text{Hg, }ox}) + V_m = 1.25 \text{ V} \tag{30}$$

and from our measurements on p-type material

$$\phi_{\text{Hg, }ox} - \phi_s = 0.08 \text{ V}. \tag{31}$$

The difference in barrier heights between Hg and Al on the type of Al_2O_3 studied in this paper has been reported by Nigh[18] and is given by

$$\phi_{\text{Hg, A}} - \phi_{\text{Al, A}} = 1.7 \text{ V}. \tag{32}$$

Combining eqs. (30), (31), and (32) yields

$$(\phi_{\text{Al, A}} + \phi_{ii} - \phi_s) + V_m = -0.37 \text{ V}. \tag{33}$$

The intercept value for $T_A = 0$ predicted by (14) is given by

$$\text{Intercept } (T_A = 0) = (\phi_{\text{Al, A}} + \phi_{ii} - \phi_s) + V_m - \frac{Q_{ss}T_{ox}}{\epsilon_{ox}}. \tag{34}$$

Using the values $v_o = 0.38$ volt and $Q_{sso} = 0$ for ⟨100⟩ material in eq. (20), and combining eqs. (20), (33), and (34) yields

$$\text{Intercept } (T_A = 0) = -0.75 \text{ V}.$$

Thus, our corresponding intercept value for Al electrodes is almost identical to that reported in Ref. 4, and it strongly implies that the electrical properties of the different Al_2O_3/SiO_2 films which determine the flatband voltage in MOS structures are identical. More specifically, it indicates that the Q_{ss} dependence on T_{ox} reported in this paper is also true for the structures studied in Ref. 4, and that the value of $\Delta\phi + V_m$ in both cases is the same.

VI. SUMMARY AND DISCUSSION

By varying the thicknesses of both insulators in a silicon/SiO_2/Al_2O_3/mercury MOS structure and accurately measuring changes in flatband voltage, we have established that a net negative space charge exists near the SiO_2/Al_2O_3 interface, which is spatially constrained to a region much less than 500 Å thick. The magnitude of this negative charge varies inversely with SiO_2 thickness and is the same for both ⟨100⟩ and ⟨111⟩ oriented n-type and p-type silicon substrates. These results are consistent with a model for space-charge formation based on work by Simmons on metal-insulator-metal structures.[6] At the elevated temperature (900°C) of Al_2O_3 deposition, the Al_2O_3 is a good enough conductor that thermal equilibrium is established. Since the electrostatic screening or Debye length in Al_2O_3 at this temperature is small compared to the Al_2O_3 thickness of interest, the bulk of the Al_2O_3 is at zero electric field, and a Fermi level can be defined that must align with the Fermi level in the silicon substrate. This requires that a fixed "contact potential", experimentally found to be 0.88 volt, must exist across the SiO_2 at 900°C. The electric field associated with this potential generates a net negative space-charge layer near the SiO_2/Al_2O_3 interface. When the structure is cooled to room temperature, the conductivity of the Al_2O_3 reduces to a negligible value and the space charge is frozen in. The net negative charge can thus be considered to be the charge on the SiO_2 capacitance associated with the constant 0.88 volt contact potential.

When the double-insulator flatband voltage is corrected for the independently measured Si/SiO_2 interface charge Q_{ss} and the barrier heights of the single-insulator system, a corrected differential flatband voltage is generated. Extrapolation of this function to zero Al_2O_3 thickness reveals a potential jump of approximately 1.25 volts when using a mercury electrode. Similarly, a potential jump of approximately 1.0 volt is found with titanium-aluminum electrodes. The interfacial barrier energies that contribute to these jumps are shown not to be derivable from a simple work function argument.

The large amount of scatter observed in the data where titanium-aluminum electrodes are used, compared to the very consistent data

obtained with the mercury electrode, implies that thermal evaporation of a metal onto an Al_2O_3 film introduces significant variation in the flatband voltage. This effect is most probably due to a charging phenomena that occurs during the transient heating of the sample during evaporation.

Measurements of Q_{ss} made on samples before and after Al_2O_3 deposition revealed that during this deposition, the value of Q_{ss} was changed. After Al_2O_3 deposition, it was found that Q_{ss} could be written as the sum of two terms. One term was a constant background charge density that was independent of SiO_2 thickness and that had the values of 1.0×10^{11} and 1.0×10^{10} charges/cm^2 for $\langle 111 \rangle$ and $\langle 100 \rangle$ oriented substrates, respectively. The other term was orientation independent and inversely proportional to the SiO_2 thickness, indicating that it is derivable from a constant contact potential that was experimentally determined to be 0.38 volt. Thus, the electric field that exists in the SiO_2 during the deposition of the Al_2O_3 determines not only Q_{it} but also a portion of Q_{ss}.

The proposed charging model was also found to be correct for p-type substrates except that an additional effect was uncovered in that the effective contact potential decreased with increasing Al_2O_3 thickness. This effect may be due to boron penetration of the thermal SiO_2 layer and an associated increased electrical conductivity at high temperature.

According to the model presented, a contact potential exists across the SiO_2 at the Al_2O_3 deposition temperature, which results in an electric field in a direction to drive mobile positive ions away from the Si/SiO_2 interface. This means that if some mechanism exists for either removing or immobilizing these positive ions when they reach the SiO_2/Al_2O_3 interface, the Al_2O_3 deposition is expected to stabilize the MIS system against ionic drifts. Such a mechanism may indeed be present since HCl, a by-product of the Al_2O_3 formation reaction, is known to be an excellent sodium getter. While the importance of this electric field in accounting for the stability of the SiO_2/Al_2O_3 system is not presently clear, it seems reasonable to assume that net positive charge at the insulator-insulator interface would make it much more difficult to remove or immobilize positive ions in the SiO_2 during second insulator deposition, since the positive ions would then tend to drift to the Si/SiO_2 interface.

Since the presented model for space-charge layer formation at insulator-insulator interfaces is relatively insensitive to the nature of the deposited insulator, provided the assumption of thermal equilibrium at the deposition temperature is correct, considerations similar to those given in this paper for the SiO_2/Al_2O_3 system should apply to other SiO_2/deposited insulator systems.

VII. ACKNOWLEDGMENTS

The authors acknowledge L. P. Adda and H. E. Nigh for many interesting discussions concerning the electrical characteristics of the $Al_2O_3/SiO_2/Si$ structure. In addition, the authors thank E. G. Parks and J. J. Nolen for their help in fabricating and measuring the many samples used in this study.

REFERENCES

1. S. K. Tung and R. E. Caffrey, "The Deposition of Oxide on Silicon by the Reaction of a Metal Halide with a Hydrogen-Carbon Dioxide Mixture," Trans. Met. Soc., AIME, *233* (March 1965), pp. 572–577.
   H. E. Nigh, J. Stach, and R. M. Jacobs, "A Sealed Gate IGFET," IEEE Trans. El. Dev., *14*, No. 9 (September 1967), p. 631.
2. J. T. Clemens and E. F. Labuda, "Semiconductor Silicon 1973," edited by H. R. Huff and R. R. Burgess, Electrochemical Society, Princeton, N. J., pp. 779–790.
3. M. P. Lepselter, "Beam Lead Technology," B.S.T.J., *45*, No. 2 (February 1966), pp. 233–253.
4. J. A. Aboff, D. R. Kerr, and E. Bassous, "Charge in SiO_2/Al_2O_3 Double Layers on Silicon," J. Electrochem. Soc., *120*, No. 8 (August 1973), pp. 1103–1106.
5. E. F. Labuda, J. T. Clemens, and C. N. Berglund, IEEE Dev. Res. Conf., University of Michigan, Ann Arbor, June 28-July 1, 1971.
6. J. G. Simmons, "Electronic Conduction Through Thin Insulating Films," *Handbook of Thin Film Technology*, edited by L. I. Maissel and R. Glang, Chapter 14, New York: McGraw-Hill, 1970.
7. C. N. Berglund and R. J. Powell, "Photoinjection into SiO_2: Electron Scattering in the Image Force Potential Well," J. Appl. Phys., *42*, No. 2 (February 1971), pp. 573–579.
   C. N. Berglund and R. J. Powell, "Photoinjection Studies of Charge Distributions in Oxides of MOS Structures," J. Appl. Phys., *42*, No. 11 (October 1971), pp. 4390–4397.
8. D. A. Mehta, S. R. Butler, and F. J. Feigl, "Electronic Charge Trapping in Chemical Vapor-Deposited Thin Films of Al_2O_3 on Silicon," J. Appl. Phys., *43*, No. 11 (November 1972), pp. 4631–4638.
9. J. J. Curry and H. E. Nigh, "8th Annual Proceedings Reliability Physics," IEEE (April 7–10, 1970), p. 29.
10. R. H. Walden and R. J. Strain, "8th Annual Proceedings Reliability Physics," IEEE (April 7–10, 1970), p. 23.
11. R. Hammer, "A Mercury Contract Probe for MOS Measurements on Oxidized Silicon," Rev. Sci. Instr., *41*, No. 2 (February 1970), pp. 292–293.
12. A. S. Grove, *Physics and Technology of Semiconductor Devices*, Chapter 9, New York: John Wiley & Sons, 1967.
13. H. Kolter, J. J. H. Schatorje, and E. Kooi, "Electric Double Layers in MIS Structures with Multilayered Dielectrics," Philips Res. Repts., *26* (1971), pp. 181–190.
14. B. E. Deal, M. Sklar, A. S. Grove, and E. H. Snow, "Characteristics of the Surface State Charge (Q_{SS}) of Thermally Oxidized Silicon," J. Electrochem. Soc., *114*, No. 3 (March 1967), pp. 266–273.
15. A. Goetzberger and H. E. Nigh, "Surface Charge After Annealing of $Al/SiO_2/Si$ Structures Under Bias," Proc. IEEE, *54*, No. 10 (October 1966), p. 1454.
16. A. S. Grove, O. Leistiko, and C. T. Sah, "Redistribution of Acceptor and Donor Impurities During Thermal Oxidation," J. Appl. Phys., *35*, No. 9 (September 1964), pp. 2695–2701.
17. B. E. Deal, A. S. Grove, E. H. Snow, and C. T. Sah, "Observation of Impurity Redistribution During Thermal Oxidation of Silicon Using the MOS Structure," J. Electrochem. Soc., *112*, No. 3 (March 1965), pp. 308–314.
18. H. E. Nigh, Proc. of Intl. Conf. on Properties and Use of MIS Structures, Grenoble, France (June 17–20, 1969), pp. 77–88.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 4, April 1975
Printed in U.S.A.

A 1-Watt, 6-Gigahertz IMPATT Amplifier for Short-Haul Radio Applications

By J. E. MORRIS and J. W. GEWARTOWSKI

(Manuscript received September 14, 1973)

A 1-watt IMPATT diode amplifier has been developed for short-haul FM radio relay applications in the 6-GHz common-carrier band. The amplifier is used in the new TM-2 system and as part of a retrofit package to upgrade the performance of the existing TM-1 system. Amplification is provided by a single silicon IMPATT diode which is used in an injection-locked mode. A finned heat sink provides IMPATT diode cooling by natural air convection within the radio bay. The diode is expected to have a mean life greater than 10 years, and it can be replaced in the field without the use of special tools or equipment. This microwave-integrated amplifier contains the rf samplers and detectors necessary to monitor both input and output rf power levels. The input power monitor also provides an input to a power-supply squelch circuit that removes dc power from the IMPATT diode if the rf input signal level becomes too low for adequate performance. The influence of the system requirements upon the amplifier design is described, and data on system performance are presented.

I. INTRODUCTION

The IMPATT diode has been developed to the point where several watts of CW power can be generated reliably in the microwave frequency range. This negative-resistance device used in conjunction with a circulator comprises a reflection amplifier suitable as the power amplifier in a microwave communications transmitter. In the present application, the diode operates in the injection-locked oscillator mode. It was demonstrated by Tatsuguchi, Dietrich, and Swan that such an amplifier using a single silicon IMPATT diode could meet the basic performance objectives of a typical short-haul radio-relay system.[1] The amplifier operates with a nominal gain of 20 dB and a noise figure of less than 52 dB. The corresponding system performance is better than 22 dBrnc0 per hop for a 1200-circuit message load. The amplifier's system performance is found to be dominated by thermal noise, with intermodulation distortion negligible. The dc-to-rf efficiency is 4 percent.

To be useful to the system, the amplifier package also contains rf samplers and detectors necessary to monitor the rf input and output power levels. The input power-monitor circuit furnishes the input information for a power-supply squelch circuit. If the input rf level drops low enough so that the locking bandwidth of the amplifier becomes small, the power supply is turned off, preventing the IMPATT oscillator from free-running out of the assigned frequency range. The dc power is automatically restored when the input rf level returns to normal. The amplifier also contains harmonic suppression filters to prevent radiation of spurious tones. The amplifier has standard WR-159 waveguide input and output ports with VSWRs of less than 1.07 across the band.

To be suitable for manufacture, an economical design was evolved based on the microwave integrated-circuit techniques successfully employed in the TH-3 system by Dietrich.[2] The construction consists of a thin-film strip-line pattern on a suspended alumina substrate, which is mounted in a die-cast aluminum housing, connected to a coaxial section containing the IMPATT diode, the tuning mechanism, and a second harmonic filter. In addition, a wide range of tunability had to be incorporated to accommodate a wide range of diode parameters, both for initial manufacture and field replacement of the diode. Both frequency and output power adjustments are provided. All these features have been successfully accomplished in the amplifier designed for manufacture.

II. AMPLIFIER DESIGN

The amplifier design is based upon the use of a single silicon IMPATT diode used in a phase-locked oscillator mode. This mode of operation, described below, is chosen since it permits the relatively high gain of approximately 20 dB to be obtained stably in a single stage.

2.1 Operating point selection

The choice of an operating point for the amplifier follows the method described by Tatsuguchi et al.[1] Figure 1 illustrates typical contours of constant system thermal noise performance, in dBrnc0 per hop, plotted on coordinates of amplifier output power versus amplifier noise figure. The system performance contours shown apply to the highest frequency message slot of one particular short-haul FM system configuration that is operated at a 1200-message circuit loading. The contours assume that a +10-dBm level signal is available to drive the amplifier. This input power level is the minimum value anticipated in one of the systems in which this amplifier will be used. The options open to the amplifier circuit designer are illustrated on the same figure

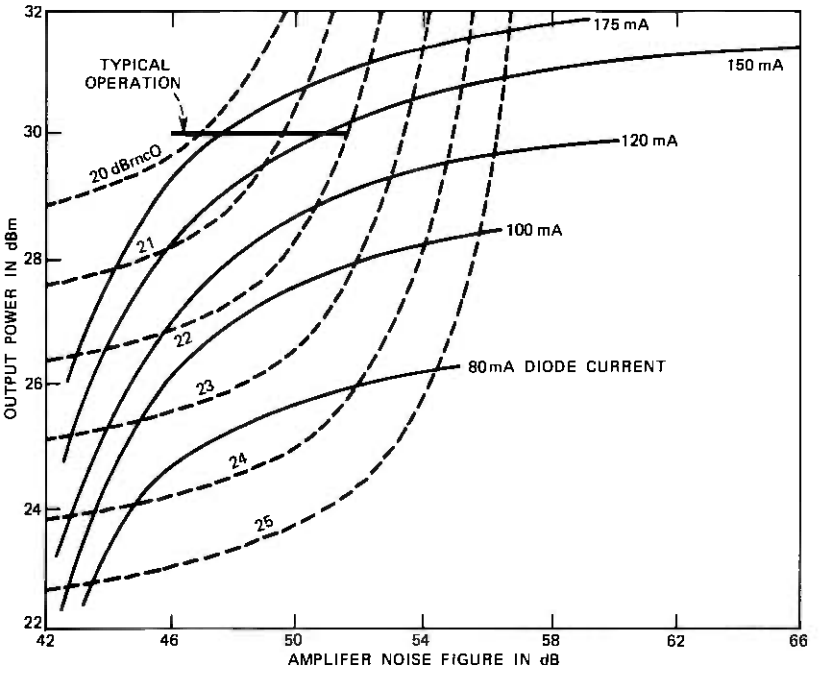

Fig. 1—Contours of constant system thermal-noise performance, in dBrnc0 per hop, plotted on coordinates of amplifier output power versus amplifier noise figure. A particular amplifier's performance is indicated by the solid lines at several IMPATT diode dc currents. Typical performance obtained on a large sample of amplifiers when adjusted for 1-watt output power is shown.

by the superimposed contours of IMPATT amplifier performance at various dc power levels. For a given dc power level, the operating point is a function of the microwave circuit impedance seen by the IMPATT device. The shape of these curves is due to the fact that an IMPATT device becomes noisier as the rf level is increased. From such curves, it becomes apparent that operation at the maximum possible rf power will result in poor system performance. Optimum performance occurs at neither maximum rf power nor minimum noise. It is instructive to note that the optimum performance, i.e., lowest dBrnc0 number, occurs with the largest dc power. The use of high dc powers must be tempered by reliability considerations, which generally dictate the use of lower powers.

For this amplifier application, the trade-off between rf output power, FM noise, and diode reliability formed the basis of the decision to operate at 1-watt rf output with 24 watts of dc supplied to the IMPATT diode. At this operating point, the diode junction temperature is expected to be approximately 200°C in convection-cooled radio bays

operating in room ambient temperature up to 50°C. This operating point is expected to provide a mean diode life greater than 10 years. This reliability is the result of careful device processing combined with low thermal impedances both within the diode package and between the diode case and ambient air.

2.2 Oscillator mode

The IMPATT diode is operated in an injection-locked (phase-locked) oscillator mode, shown schematically in Fig. 2. The IMPATT device and its associated resonating circuitry terminate one port of a circulator in a negative impedance. In the absence of an input signal, a free-running oscillation at frequency f_0 occurs, which is coupled to the output through the circulator. When an appropriate input signal is added, the oscillation frequency locks to the input over a band of frequencies $2\Delta f$, approximately symmetrical about f_0. The free-running frequency is adjusted to the desired operating channel. Figure 2 illustrates the power and phase variations that occur across the locking frequency band. The power levels and oscillator external Q (Q_{ex}) are chosen such that Δf is at least 10 times the highest modulating frequency. In this way, only the center linear portion of the phase variation curve is used, and phase distortion is minimized.

Since the rf output power is fixed from other considerations (system performance and fading margin) and the rf input power available in

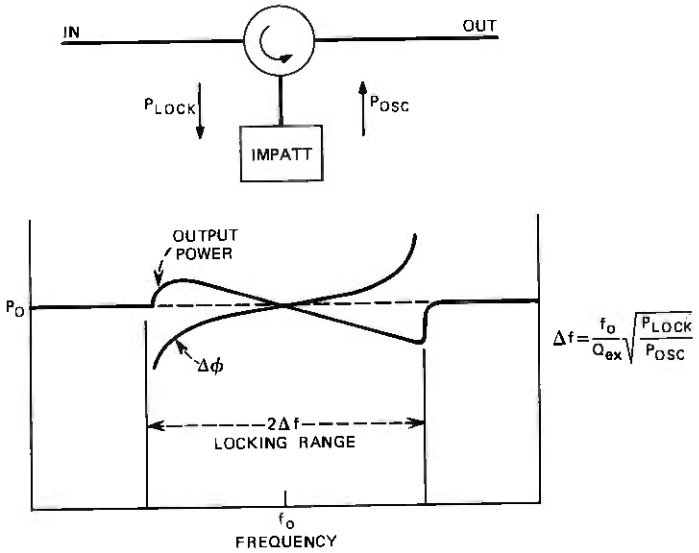

Fig. 2—Simplified representation of an injection-locked oscillator.

the systems in which the amplifier is to be used is limited, the designer is required to provide a circuit of the lowest possible Q. For this amplifier, a low Q circuit is provided by the use of a diode circuit consisting of a coaxial one-quarter-wavelength transformer plus a short section of coaxial line between the transformer and the diode that series-tunes the IMPATT diode's capacitance. With this resonator, the circuit Q is sufficiently low that Q_{ex} is largely determined by the IMPATT device itself.

2.3 IMPATT characteristics

The IMPATT diode used for this amplifier is an n-type silicon diode whose junction side is bonded to a metallized diamond within a copper and ceramic microwave pill package.[3] The large-signal rf characteristics of the diodes are measured near 6 GHz using the method described by Decker et al.[4] The diode wafer admittance is measured on all devices at 24-watts dc, with an rf voltage corresponding to the diode's operating point in the amplifier. Wafer susceptances are specified at 19.0 millimhos. Tuning is provided to accommodate a range of diode susceptances. The wafer Q, defined as the magnitude of the ratio of wafer susceptance to wafer conductance, has values that vary by a factor of 2.5 to 1.

2.4 Circuit description

The requirements of practical radio-relay equipment dictate an amplifier circuit somewhat more complex than the simple circulator, diode, and resonator shown in Fig. 2. A more complete schematic of the amplifier is shown in Fig. 3. The circuit contains three circulators, of which the center circulator corresponds to the one shown in Fig. 2. Additional circulators with one port resistively terminated are used at both the amplifier's input and output to provide isolation from the effects of external reflections and to provide input and output return losses better than 30 dB.

The dc power for the IMPATT diode from the current-regulated power supply is coupled to the oscillator port of the center circulator through a resistor and a band-stop filter tuned to 6 GHz. The resistor is used here to provide the high resistive impedance at low frequencies that has been shown by Brackett to prevent spurious oscillations.[5] The dc power is isolated from the remaining rf circuit by a series capacitor in the main rf circuit adjacent to the band-stop bias filter.

On the output side of the center circulator, a small sample of the amplified output is picked off by a nondirectional coupling probe. This sample of the rf output is detected using a point contact diode

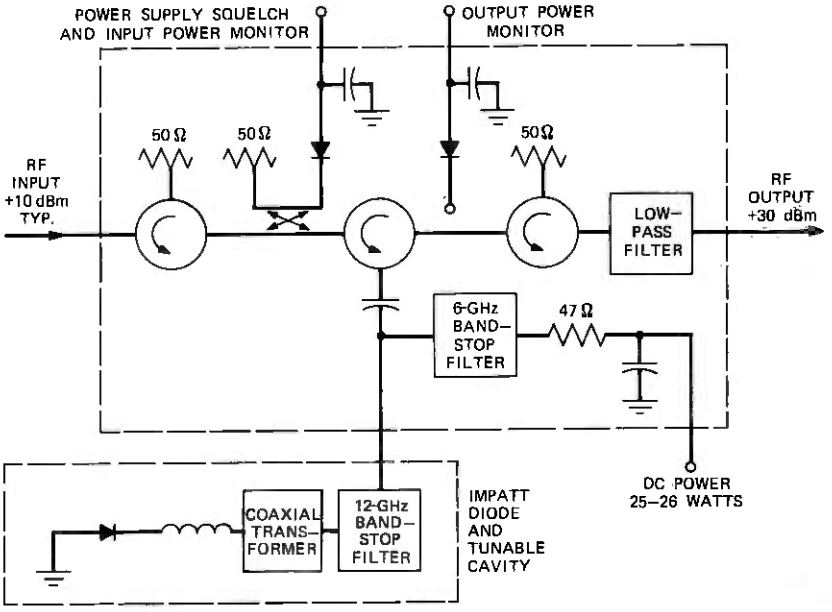

Fig. 3—Schematic diagram of complete IMPATT diode amplifier.

to provide a dc current for the radio bay meter panel and alarm functions.

On the input side of the center circulator, a sample of the input power is taken and detected for bay panel metering and to operate a power-supply squelch circuit. The power-supply squelch operates to remove dc power from the IMPATT diode if the rf input to the amplifier falls below a prescribed level. This effectively prevents free-running oscillations by the IMPATT oscillator, whose free-running frequency is not sufficiently stabilized to prevent interchannel interference. A directional coupler is used for the input coupler to provide a good match on the circulator common-arm and to provide, via its 20-dB directivity, discrimination against leakage of power generated by the IMPATT diode.

A band-stop filter is used on the oscillator port of the center circulator to prevent second-harmonic energy generated by the IMPATT from interfering with the operation of the monitor circuits. A low-pass filter is located at the amplifier's output to ensure that all harmonics are suppressed.

III. CIRCUIT FABRICATION AND TUNING

Most of the circuit is fabricated using the microwave integrated-circuit techniques developed for use in the TH-3 system.[2] Film inte-

grated circuits (FICs) consisting of patterns defined photolithographically on 0.024-inch (0.61-mm) thick unglazed alumina are used as a suspended-substrate strip-line transmission-line medium. The strip-line circuitry, as well as the amplifier's waveguide input and output, are contained in a die-cast aluminum housing, shown in the amplifier photograph, Fig. 4. The IMPATT diode and its resonator are contained in a short section of coaxial line that projects perpendicularly from the housing and is topped by the large, finned heat sink used for IMPATT diode cooling. Adjustments are provided on the coaxial section for field tuning of frequency and power output.

3.1 Strip-line circuits

The layout of the circuitry within the die-cast housing is illustrated in Fig. 5 and shown pictorially in Fig. 6. The ceramic substrates are located within a narrow channel to avoid multimoding problems. The complex substrate shape is fabricated by an automated laser-cutting

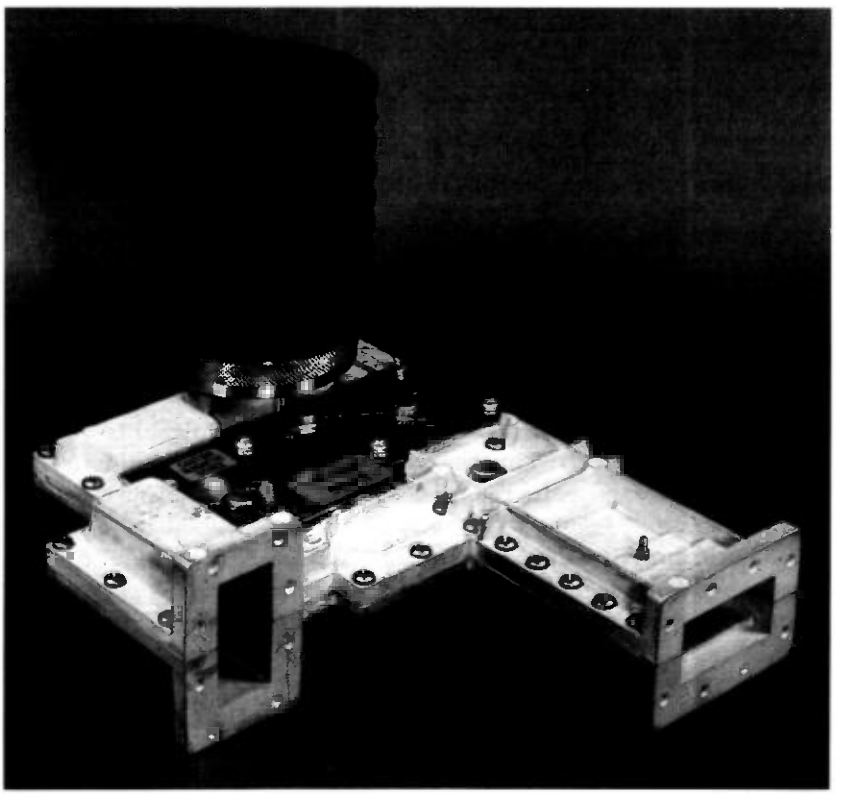

Fig. 4—Complete amplifier.

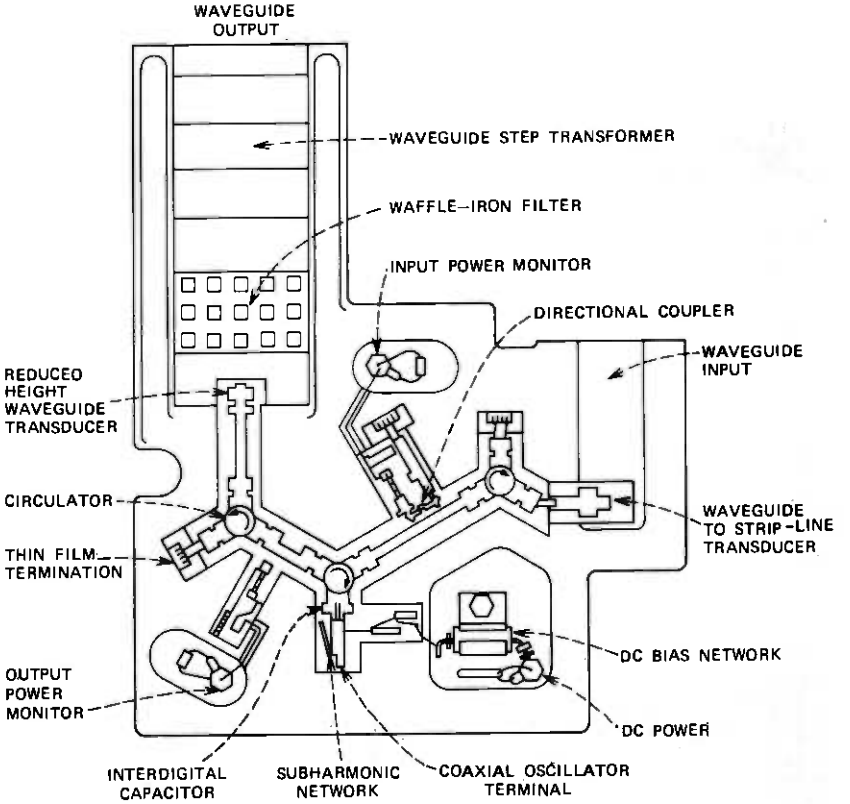

Fig. 5—Arrangement of waveguide and strip-line circuit within die-cast housing.

technique.[6] The amplifier's full-height waveguide input is shown on the right. A thin-film probe transition from the input waveguide to the suspended-substrate strip-line couples the input signal to the first of the three circulators. This circulator, with one port terminated in a thin-film resistor, provides the necessary input isolation. The circulator and termination designs follow closely those described by Dietrich,[2] modified to improve the temperature stability. A directional coupler, located between the input and center circulators, diverts approximately 10 percent of the input rf signal to the input detector diode to generate the dc needed for bay panel metering and power-supply squelch functions.

The remaining input signal is coupled to the oscillator port of the center circulator. The series capacitor, which dc-isolates this port, is realized by a narrow, meandering, interdigital gap in the thin-film conductor. The 6-GHz band-stop bias filter is realized by a high-

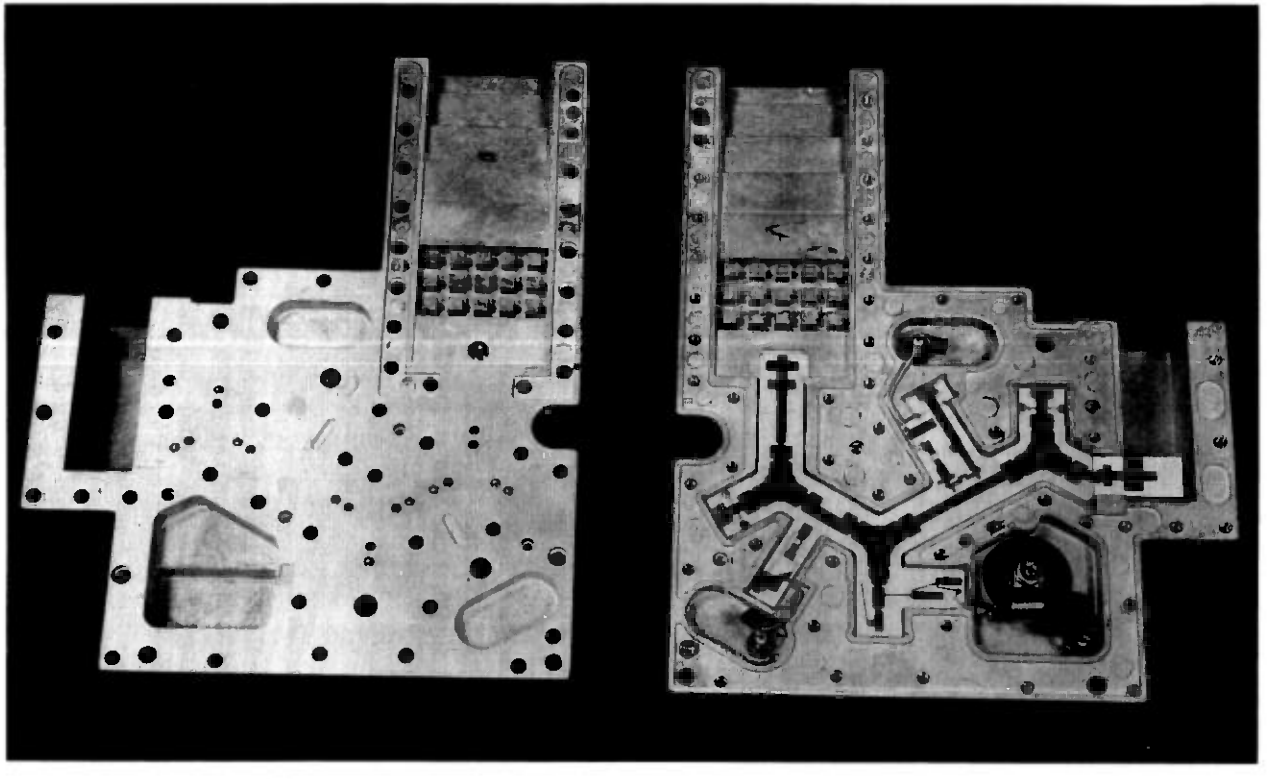

Fig. 6—Pictorial view of open die-casting showing strip-line circuitry.

characteristic-impedance line along which, at quarter-wavelength intervals, are placed two quarter-wavelength-long open-circuited stubs of lower characteristic impedance. The bias circuit is completed by a 47-ohm power resistor that is clamped to the aluminum housing to maximize heat transfer. Several ferrite beads are placed on the leads of this resistor to provide additional stability against bias circuit oscillations.

An open-circuited stub, one-half-wavelength long at 6 GHz, which connects to the oscillator terminal through a thin-film resistor, is used to control the circuit impedance at 3 GHz (the subharmonic of the 6-GHz band) without adding significant loss or mismatch at 6 GHz.[7] This was found to be necessary to eliminate frequency jumps during tuning, which occur when the subharmonic impedance is too high.

At the end of the thin-film pattern (coaxial oscillator terminal), connection is made, using a bellows, to the center conductor of the coaxial line through the top half of the aluminum housing.

The amplified rf signal reflected from the IMPATT diode down the coaxial line is coupled by the center circulator to the output circulator. A small portion of the amplified output is capacitively coupled to the output detector circuit to provide the direct current for bay panel metering and alarm functions. This nondirectional coupling is approximately 28 dB. The amplified signal passes through the output circulator, used as an isolator, and is coupled into a reduced-height waveguide. Within the reduced-height waveguide, a waffle-iron filter[8] having a low-pass characteristic strips the amplified signal of any residual harmonic energy either generated by the IMPATT or contained in the input signal. Following the waffle-iron filter, a four-step transition couples the reduced-height waveguide to standard-height WR-159 waveguide.

3.2 *Coaxial circuit*

A cross section of the coaxial line is shown in Fig. 7. At the bottom, just above the bellows contact to the thin-film circuit, is located a three-resonator, radial-line, band-stop filter[9] that is tuned to the 12-GHz second harmonic of the 6-GHz common-carrier band. The filter prevents the second harmonic energy generated by the IMPATT diode from causing anomalous monitor circuit operation. Appropriate steps in the coaxial center conductor in the filter section provide a good match across the 6-GHz band. The center conductor tip is spring-loaded against the IMPATT diode, which is held centered at the upper end of the coaxial section. A large, finned heat sink contacting

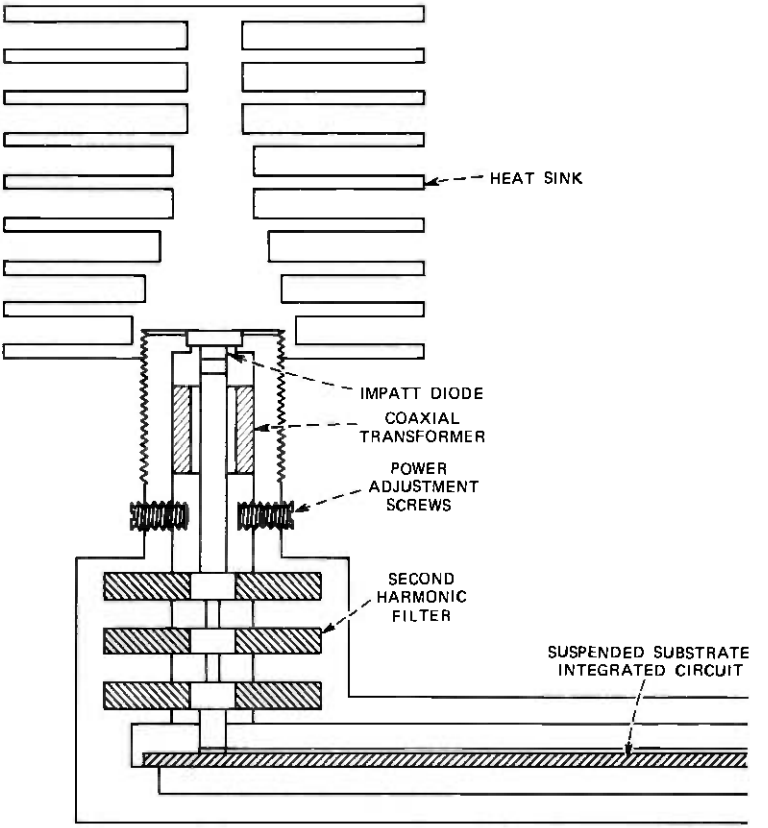

Fig. 7—Cross section of the coaxial line section.

the back of the diode provides diode cooling and eliminates the need for forced convection.

IMPATT tuning is accomplished by a movable quarter-wavelength coaxial transformer and four capacitive power-adjustment screws located radially around the coaxial line at a point nominally an eighth-wavelength from the end of the transformer. The position of the transformer relative to the IMPATT diode primarily determines the frequency of operation. The transformer is moved using a large-diameter knurled ring, shown just below the heat sink in Fig. 4. This ring is mechanically coupled to the transformer through two slots in the coaxial line. These slots are completely covered by the transformer and ring to prevent rf leakage.

The transformer section characteristic impedance is designed to produce 1 watt at the amplifier's output port (nearly 1 dB more at the

diode's location) with the highest Q diode and with the power adjustment screws adjusted flush with the inner diameter of the coaxial outer conductor. For all other diodes that would give greater than 1 watt with this transformer, capacitance is added using the four power-adjustment tuning screws. When transformed through the coaxial circuit to the diode position, the added capacitance appears as an increase in the resistive part of the circuit impedance. Increased circuit resistance reduces the generated power down to the 1-watt level, where optimum system performance occurs. This power adjustment is made on diodes having low values of Q; the increase in circuit Q because of the screw insertion is counterbalanced by the lower diode Q, so that the overall external Q is not increased by this power adjustment when compared with high Q diodes requiring little screw penetration.

3.3 Circuit tuning

The circuitry within the die-cast housing is initially tuned in the factory with a 7-mm precision connector located in place of the IMPATT diode and heat sink. During the initial tuning, the transformer is not installed and the power-adjustment screws are adjusted flush with the coaxial-line inner surface. Tuning of all ports of the three circulators to better than 30-dB return loss is accomplished across the 8-percent common-carrier band. The three ports of the center circulator are tuned over a slightly wider band to include the extremities of the locking bandwidth of amplifiers operated on the end channels of the common-carrier band.

By matching the diode port of the center circulator to achieve this broadband high return loss, the oscillator circuit Q is essentially determined by that of the quarter-wavelength transformer and the short section of 50-ohm line from the transformer to the diode. In practice, Q_{ex} of the oscillator is determined largely by the IMPATT diode wafer. The transformer position and power adjustment screws permit adjustment in the field of any amplifier to 1 watt on any channel assignment with any diode. The IMPATT diode is replaceable in the field by simply removing the heat sink and inserting a new diode in the coaxial line against the spring-loaded center conductor.

The amplifier cost has been kept low by the use of thin-film integration, casting technology, and laser cutting of ceramic substrates.

IV. AMPLIFIER PERFORMANCE

Ten models were constructed in the laboratory, and information was conveyed to the Western Electric Company, who is now producing the unit. Measurements of intermodulation distortion indicate that the distortion products are small and that system performance

can be accurately predicted on the basis of power output and FM thermal noise with no correction for distortion. Performance shown in Fig. 1 can be readily obtained using the silicon IMPATT diodes in manufacture. A detailed evaluation has been completed of a TM-2 system in Ohio that includes eight factory-built IMPATT amplifiers. Satisfactory operation was noted over a 10-month test period.

V. SUMMARY

A 1-watt, 6-GHz silicon IMPATT diode amplifier has been developed and is being manufactured for use as the transmitter power amplifier in short-haul radio systems. The amplifier operates with a nominal gain of 20 dB and a noise figure of less than 52 dB. The noise contribution of the IMPATT amplifier is substantially thermal noise, with intermodulation distortion negligible. The dc-to-rf efficiency is 4 percent. The amplifier includes integrated input and output rf power monitors and harmonic suppression circuitry.

The input monitor circuit furnishes the input information for the power-supply squelch circuit. If the input rf level drops low enough so that the locking bandwidth becomes small, the power supply is turned off, preventing the oscillator from free-running out of the assigned frequency range. The dc power is automatically restored when the input level returns to normal.

The low cost and reliability of this IMPATT amplifier make it an attractive rf output device in short-haul applications.

REFERENCES

1. I. Tatsuguchi, N. R. Dietrich, and C. B. Swan, "Power-Noise Characterization of Phase-Locked IMPATT Oscillators," IEEE Journal of Solid-State Circuits, *SC-7*, No. 1 (February 1972), pp. 2–10.
2. N. R. Dietrich, "TH-3 Microwave Radio System: Microwave Integrated Circuits," B.S.T.J., *50*, No. 7 (September 1971), pp. 2175–2194.
3. R. L. Kuvas and J. A. Rupp, "Design, Characterization and Reliability of 6-GHz Silicon IMPATT Diode," Technical Digest Int. Elec. Devices Meeting, December 3–5, 1973, No. 24.4, pp. 489–492.
4. D. R. Decker, C. N. Dunn, and R. L. Frank, "Large-Signal Silicon and Germanium Avalanche-Diode Characteristics," IEEE Trans. on Microwave Theory and Techniques, *MTT-18*, No. 11 (November 1970), pp. 872–876.
5. C. A. Brackett, "The Elimination of Tuning-Induced Burnout and Bias Circuit Oscillations in IMPATT Oscillators," B.S.T.J., *52*, No. 3 (March 1973), pp. 271–306.
6. J. Longfellow, "Sawing Alumina Substrates With a CO_2 Laser," American Ceramic Society Bulletin, *52*, No. 6 (June 1973), pp. 513–515.
7. C. W. Lee and W. C. Tsai, "High Power GaAs Avalanche Diode Amplifiers," IEEE International Convention Digest, March 1971, pp. 368–369.
8. S. B. Cohn, "Design Relations for the Wide-Band Waveguide Filter," Proc. IRE, *38*, No. 7 (July 1950), pp. 799–803.
9. B. C. DeLoach, Jr., "Radial-Line Coaxial Filters in the Microwave Region," IEEE Trans. on Microwave Theory and Techniques, *MTT-11*, No. 1 (January 1961), pp. 50–55.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 4, April 1975
Printed in U.S.A.

Line-of-Sight Paths Over Random Terrain

By E. N. GILBERT

(Manuscript received September 5, 1974)

Line-of-sight paths are important as VHF radio channels. In a mobile radio system, for example, the landscape determines the communication possibilities in a complicated way. This paper analyzes a simple model of rough terrain to relate statistical terrain properties to line-of-sight paths. The model is constructed from conical hills, all the same height, distributed at random over the surface of a spherical earth.

The parameters of the model are the earth's radius a, the density σ of hills, and the grade g of the hills. Although a simpler planar model is obtained by letting $a \to \infty$, a finite spherical earth is needed for most questions. Assuming that a base station is located at the peak of a hill, the most interesting line-of-sight paths are those from a typical hilltop. A large number of statistics of these paths are then derived, usually as simple functions of a, σ, and g. These include properties of paths to other peaks, to the horizon, and to random points on the ground.

I. INTRODUCTION

Very-high frequency radio propagation is often said to resemble optical propagation. A line-of-sight path provides a good radio channel; a path blocked by the terrain does not. With the aid of a topographic map, one can determine whether a path Q_1Q_2 is a line-of-sight path. Essentially, one must plot the ground elevation profile along the path to see whether the ground intersects the straight line segment Q_1Q_2. This calculation must include the effect of the earth's curvature. Atmospheric refraction is also accounted for by changing the earth's radius to a fictitious value.

Having done the calculation for one path Q_1, Q_2, we learn little about other paths. The region covered by a transmitter at Q_1, i.e., the set of points Q visible from Q_1, would be found by plotting ground elevation profiles along views from Q_1 at every possible azimuth angle. This region might represent the coverage of a TV station or of a base station in a mobile telephone system.

This paper analyzes a statistical model to give insight into the way coverage regions depend on properties of the terrain. The parameters

of the model are the radius a of the earth, a density σ of mountains (or hills) per unit area, and a grade (slope) g of these mountains. Many statistical properties of terrain and paths are then derived as functions of a, σ, g. These properties are means, or in some cases distribution functions, of the random variables that appear in the INDEX. Line-of-sight paths from a typical mountain peak receive special attention because a peak is the most likely site for a base station. Although the exact formulas contain integrals with unwieldy trigonometric integrands, most of these formulas may be replaced by simple expressions, to a very good approximation. The expected area visible from a peak and the expected number of peaks visible from a random point on the ground are more complicated quantities, leading to integrals that are evaluated numerically.

INDEX

Altitude—eqs. (6), (7), (8), Table I, Fig. 6.
Area blocking—eqs. (12), (17) to (20), (23), Table III.
  Visible—eq. (43), Table VII.
  Within horizon—eq. (35).
Number of peaks visible:
  From a peak—eq. (26), Table V.
  From a point on the ground—eq. (37).
Range from a peak:
  To furthest visible peak—eq. (31), Table V.
  To horizon—eqs. (33), (34), Table VI.
  To random visible peak—eqs. (25) to (29), Table IV.
  To random visible point on ground—eqs. (39) to (42), Fig. 15.
Slope—Table II.

The earth's radius a is an important parameter of the model. Although a simpler planar model is obtained by letting $a \to \infty$, the planar model is inadequate for most statistics of interest.

With a and σ fixed, the terrain becomes rougher as g increases. As a rule, the model predicts more long line-of-sight paths and larger expected visible area for rougher terrains. However, in mobile radio these long paths are more important as sources of interference than as useful channels.

II. THE MODEL

The terrain model will use conical mountains distributed at random in a Poisson pattern over the surface of the earth. Begin with a sphere of radius a miles (a may be the true radius of the earth, or something larger if atmospheric refraction effects are to be taken into account).

Place points at random on the surface S of this sphere using a Poisson process with density σ points per square mile. Each Poisson point will represent a mountain peak, and so the sphere of radius a will be called the *peak sphere.*

Each Poisson point P will be associated with a mountain-shaped subset $M(P)$ of the interior of the peak sphere. The subsets $M(P_1)$, $M(P_2)$, $\cdots$ for the various peaks will overlap. Take the union of all the subsets $M(P)$ to represent the earth.

The simplest shape for $M(P)$ is the cone consisting of all rays from P making angle $<\theta$ with the inward-pointing normal to S. This cone has to be truncated to keep it from extending beyond the peak sphere in the direction antipodal to P. The surface of the cone is tangent to an *inner sphere,* concentric with the peak sphere and having radius $a \sin \theta$. Take $M(P)$ to be the inner sphere plus the part of the cone that lies between P and the inner sphere. Figure 1 shows $M(P)$ shaded.

With this construction, the terrain consists of conical mountains, all having the same height and the same grade $g = \cot \theta$. There may also be flat places where the earth's surface coincides with the inner sphere. A flat spot occurs at any point that lies further than $(\frac{1}{2}\pi - \theta)$ radians away from all Poisson points. Flat spots are rare, except when the parameters σ, θ are chosen to produce widely separated mountains having very gentle slope.

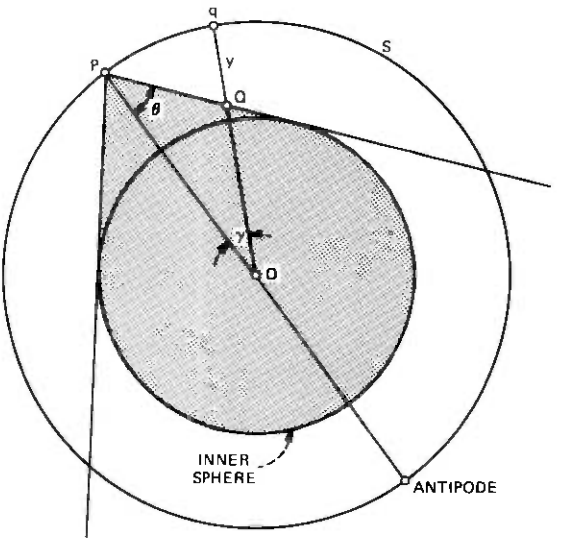

Fig. 1—Construction of $M(P)$.

Figure 2 is an elevation contour map for a typical random terrain. Some unrealistic features of the model are evident. The conically shaped mountains have circular contour lines. The peaks are distributed chaotically instead of being arranged in rows (mountain ranges).

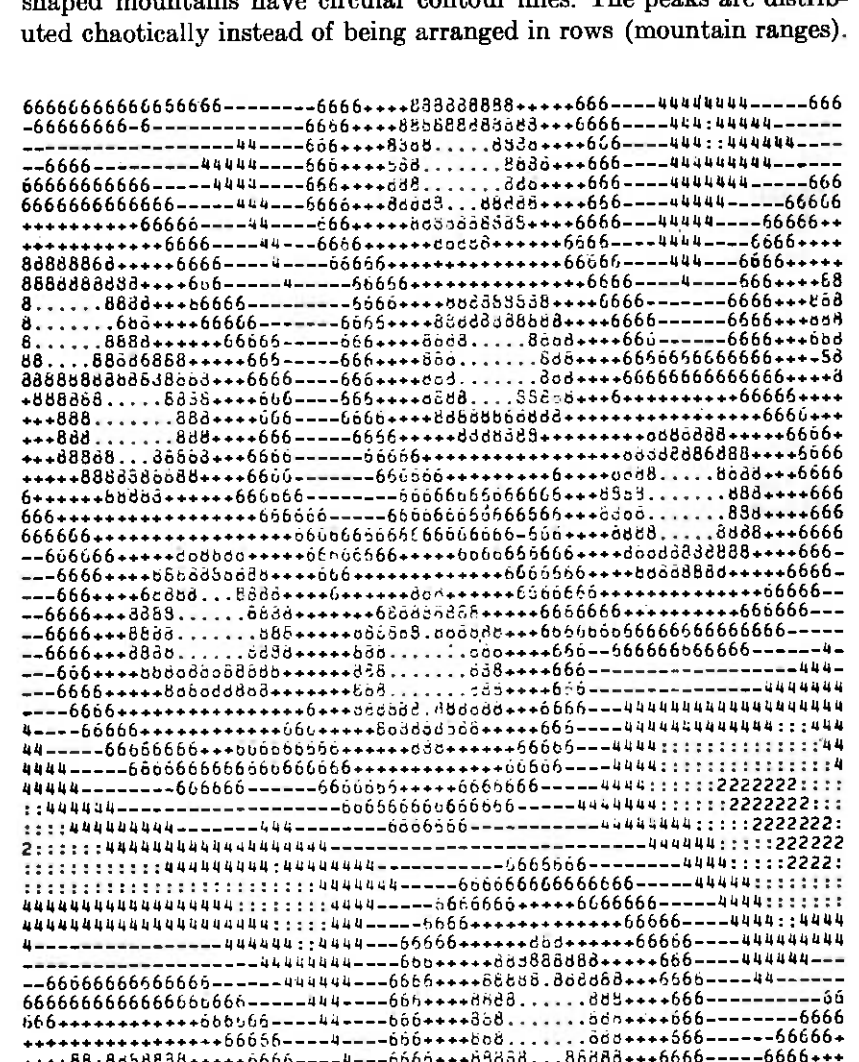

Fig. 2—Contour map. The symbols · 8 + 6 − 4 : 2 denote altitude levels ordered from the peak sphere downward.

Figure 3 is a plane cross section through the earth in the same model. This figure is more interesting for the present problem because the existence of a clear line-of-sight path between two points depends only on the shape of such a profile. Note that the elevation curve in Fig. 3 is composed of convex arcs (hyperbolas) that join in the valleys between mountains. But, at least, the maxima in Fig. 3 have different heights. Figure 4 shows random terrain as seen from one of the peaks looking out toward the horizon. The nearest and furthest peaks shown have ranges of 6 and 150 miles. The parameters were picked to match a particular portion of the Alps for which a panoramic photograph was also available. The deficiencies of the model are less evident in this figure. The curvature of the earth makes it less obvious that all peaks have the same height.

In real terrain, it is sometimes possible to see part of a mountain even though the mountain's peak is obscured from view. That cannot happen in this model, as will now be proved. Suppose that the view of a peak P_1 is blocked when the eye is at E. Then the line segment P_1E contains a blocking point B_2 belonging to another mountain $M(P_2)$. Now consider any other point P of $M(P_1)$. P must lie on some line segment P_1I, where I belongs to the inner sphere. Figure 5 shows the triangle EP_1I. The segments EP and B_2I cross at some point B in the triangle. B belongs to the convex set $M(P_2)$ because B_2 and I belong. Then B is a point of $M(P_2)$ blocking the view of P.

By making $a \to \infty$, one obtains a *planar model* of random topography. The peak sphere S becomes a peak plane. At a point Q, the land surface lies below the peak plane a distance

$$y = g \operatorname{Min}_i \|P_i - Q\|, \tag{1}$$

where the minimization is over all Poisson points P_i. Replacing S by a plane simplifies the analysis considerably but, unfortunately, it produces a much less realistic model. If Fig. 4 has been drawn for a planar model, every peak P_i would have been visible. Even worse, Section VIII shows that the expected area visible from a peak would be infinite. For that reason, the extra complication of a spherical earth is really necessary for some questions about line-of-sight paths.

III. PARAMETER ESTIMATION

The two parameters σ, $g = \cot\theta$ can be chosen to fit the model to terrain measurements. One might estimate the density σ by counting peaks. A difficulty is that one must then decide how big a hill must be to be counted. Surely every bump on the landscape ought not to count as a peak. This decision is avoided by using statistical properties of the point Q lying below a random point q on the peak sphere. The

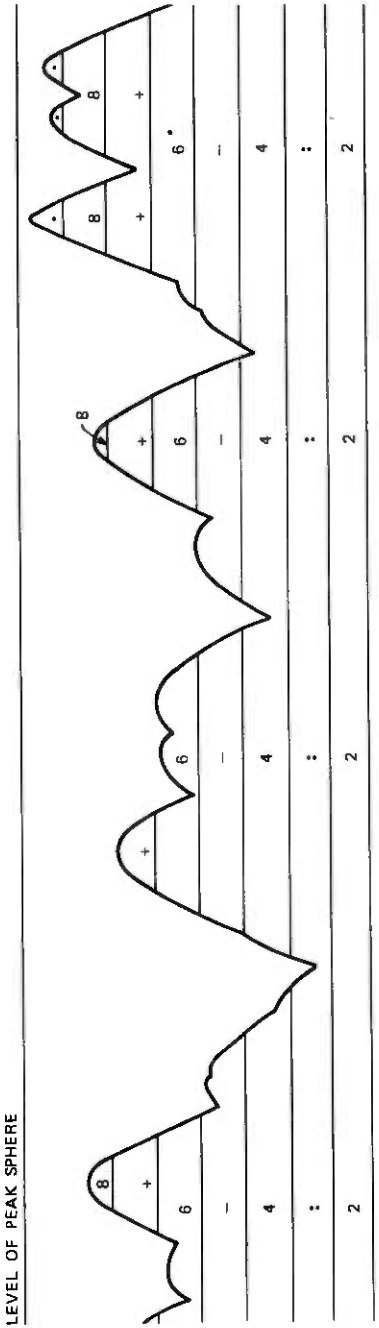

Fig. 3—Cross-section view.

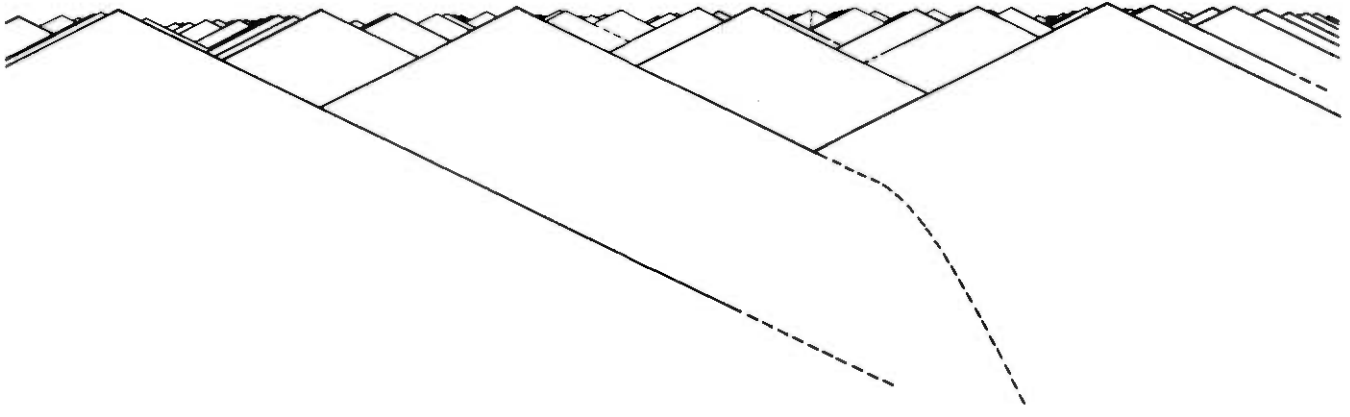

Fig. 4—View of horizon from a peak.

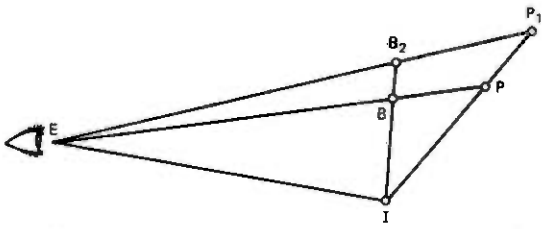

Fig. 5—$M(P_2)$ blocks all of $M(P_1)$ from view if it blocks P_1.

altitude and slope of the terrain at Q are two useful random variables. Both depend on the angle $\gamma = < POQ$ from Q to the nearest peak P (see Fig. 1). Since a circular cap of angle γ_0 on the peak sphere has area $2\pi a^2(1 - \cos \gamma_0)$, the distribution function for $< POQ$ can be written immediately,

$$\text{Prob } (\gamma \leqq \gamma_0) = 1 - \exp [- 2\pi a^2 \sigma (1 - \cos \gamma_0)]. \tag{2}$$

There is no natural sea level in the model, and so it will be convenient to specify the altitude at Q by giving the *depth* y, measured from S down to the land. If $\gamma \geqq \frac{1}{2}\pi - \theta$, then ground level coincides with the inner sphere, i.e.,

$$y = a(1 - \sin \theta), \qquad \gamma \geqq \tfrac{1}{2}\pi - \theta. \tag{3}$$

For smaller angles γ,

$$y = a[1 - \sin \theta / \sin (\gamma + \theta)], \qquad 0 \leqq \gamma < \tfrac{1}{2}\pi - \theta, \tag{4}$$

as is clear from Fig. 1. These formulas, together with the distribution (2) for γ, determine the depth distribution,

$$\begin{aligned} \text{Prob } \{y \leqq a[1 - \sin \theta / \sin (\gamma + \theta)]\} \\ = 1 - \exp [-2\pi a^2 \sigma (1 - \cos \gamma)], \qquad 0 \leqq \gamma < \tfrac{1}{2}\pi - \theta \\ \text{Prob } \{y \leqq a(1 - \sin \theta)\} = 1. \end{aligned} \tag{5}$$

Although one can easily tabulate the distribution function for y by substituting numerical values of γ into (5), the distribution function is easier to visualize in a limiting case. Since a is a large radius, let $a \to \infty$ in (5). As one might expect, the formulas tend toward the depth distribution function in the planar model,

$$\text{Prob } \{y \leqq Y\} = 1 - \exp [-\sigma \pi (Y/g)^2]. \tag{6}$$

In this limit, σ and g enter the distribution only via a single length parameter $\sigma^{-\frac{1}{2}} g$, which is an index of altitude variability. Thus, altitude distribution data alone cannot be expected to supply good estimates of

both g and σ. Some simpler statistics are the median,

$$\begin{aligned}\text{Median } (y) &= (\pi^{-1} \log_e 2g^2/\sigma)^{\frac{1}{2}} \\ &= 0.4697\sigma^{-\frac{1}{2}}g,\end{aligned}$$

and the moments,

$$E(y^k) = \Gamma(1 + \tfrac{1}{2}k)(g^2/\pi\sigma)^{k/2}.$$

Particularly, the mean is

$$\bar{y} = E(y) = \tfrac{1}{2}\sigma^{-\frac{1}{2}}g \tag{7}$$

and the standard deviation is

$$[\text{Var } (y)]^{\frac{1}{2}} = [(\pi^{-1} - 2^{-2})g^2/\sigma]^{\frac{1}{2}} = 0.2683\sigma^{-\frac{1}{2}}g. \tag{8}$$

It is also possible to obtain (6) as an exact result for a spherical model in which the shape of the mountains is only approximately conical. That entails a new choice of the set $M(P)$ in Fig. 1. Define the new shape so that the depth becomes

$$y = 2ga \sin \tfrac{1}{2}\gamma, \tag{9}$$

where again γ is the angle to the peak. At P, $M(P)$ comes to a point approximating a cone of slope g. At the antipode to P, $M(P)$ has depth $2ga$; then this model requires $g < \frac{1}{2}$. Now the depth distribution for all γ is again given by (5) but with the left-hand side replaced by Prob $\{y \leqq 2ga \sin \frac{1}{2}\gamma\}$. But that is (6), exactly.

For many values of g and σ, the planar approximation (6) to the depth distribution (5) is very good. For example, Table I compares the planar approximation with some distributions having $a = 3959$ mi, the earth's radius. In the table, the cones have grades $g = 0.05$, 0.1, and 0.2 and the density σ is adjusted to fix the standard deviation in (8) at 528 ft (0.1 mi). Table I gives percentiles of the distribution as

Table I — Altitude percentiles (in feet)

	Spherical Model			Planar
$g =$ $\sigma =$	0.05 0.0171	0.1 0.0683	0.2 0.2732	$\sigma/g^2 = 6.831$
0.1%	−1899	−1964	−1980	−1986
1%	−1378	−1421	−1432	−1436
10%	−691	−712	−717	−719
25%	−315	−327	−331	−332
50%	70	63	62	61
75%	402	400	399	399
90%	642	641	640	640
99%	896	896	896	896

altitudes measured upward from a common level, corresponding to the depth $\bar{y}$ in (7).

Figure 6 is an altitude distribution for northern New Jersey. It was obtained from a topographic map by reading altitudes at 52 points, 10 km apart in a rectangular grid covering latitudes 40°30′ to 41° and longitudes west of 74°. The altitudes ranged from 0 to 1100 ft. Data for parts of New Jersey further south were not used; the topography of New Jersey is too variable for both north and south to be well represented by a single simple model. The planar model fits the observed points well, except at low altitudes. As an alternative, use the spherical model with $a = 3959$ mi. By taking $g = 0.011$, one obtains a maximum depth (3) near 1100 ft, so that low altitudes can be regarded as occurring on the inner sphere. Then σ remains as a parameter to adjust for a good fit.

The parameter $g = \cot \theta$ is the grade at mountain peaks. At the random point Q, at angle γ away from a peak in Fig. 1, the grade is smaller because the normal to the conical surface makes an angle $\frac{1}{2}\pi - \theta - \gamma$ with the vertical direction OQ. Thus, the grade at Q is

$$g' = \begin{cases} \cot(\theta + \gamma) = (g - \tan\gamma)/(1 + g\tan\gamma), & \text{if } \theta + \gamma \leqq \frac{1}{2}\pi \\ 0 \quad \text{otherwise.} \end{cases}$$

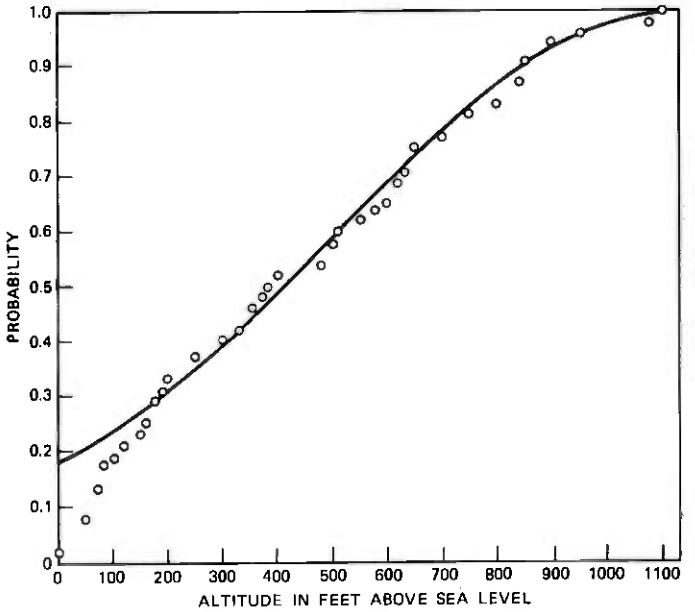

Fig. 6—Altitude distribution for northern New Jersey. Curve is for planar model with peak sphere at 1130-ft altitude and $[\text{Var}\ (y)]^{\frac{1}{2}} = 400$ ft.

Table II — Percentiles of G/g

g =	small	0.1	0.2	0.5
0%	−1.0000	−1.0000	−1.0000	−1.0000
1%	−0.9995	−0.9995	−0.9995	−0.9994
10%	−0.9510	−0.9506	−0.9492	−0.9399
25%	−0.7070	−0.7053	−0.7001	−0.6667
50%	0	0	0	0
75%	0.7070	0.7053	0.7001	0.6667
90%	0.9510	0.9506	0.9492	0.9399
99%	0.9995	0.9995	0.9995	0.9994
100%	1.0000	1.0000	1.0000	1.0000

The grade 0 occurs on the inner sphere. This result, together with (2), determines the distribution of the grade g'. In most cases, the grade g' has high probability of being close to g; one should not expect this distribution to fit observed grade data well.

At q, one might move in a random direction and ask for the slope G along the random path through Q. The slope, which depends on the angle φ between the path direction and the uphill direction, lies in the range $-g' \leqq G \leqq g'$. With some simple geometry, one finds

$$G = g' \cos \varphi/(1 + g'^2 \sin^2 \varphi)^{\frac{1}{2}}.$$

By using the known distribution for g' and assuming a flat distribution for φ, one can obtain a distribution function for G. This would be the distribution of the slopes G seen in cross sections like Fig. 3. A simple distribution is obtained only in the planar model limit, for which $g' = g = \cot \theta$ identically:

$$\text{Prob } \{G \leqq \tan \chi\} = 1 - \pi^{-1} \arccos \{\sin \chi/\cos \theta\}.$$

Table II gives the slope distribution in the planar model for several values of g. In the limit of small g, the distribution function for G/g tends to $1 - \pi^{-1} \arccos G/g$.

IV. BLOCKING REGIONS

Suppose two points Q_1, Q_2 are given, representing the positions of two antennas. In general, Q_1, Q_2 can lie anywhere above the inner sphere. A clear line-of-sight path exists between Q_1 and Q_2 as long as the straight-line segment Q_1Q_2 does not intersect any of the sets $M(P_i)$. The *blocking region* for Q_1, Q_2 is the (open) set of points P on S such that Q_1Q_2 intersects $M(P)$. The area of the blocking region enters into the probability that a line-of-sight path Q_1Q_2 exists. The advantage of the conical mountains $M(P)$ is that blocking regions assume simple shapes.

The simplest blocking region is one for a pair of points Q_1, Q_2 both on S. If the line Q_1Q_2 intersects the inner sphere, all $M(P_i)$ block the path. The blocking region then consists of the entire sphere S. If Q_1Q_2 misses the inner sphere, then blocking occurs at a point Q on the path Q_1Q_2 if a peak P_i lies too close to Q. If the depth of Q is y, then (4) gives the angle γ to peaks P such that Q lies on the surface of $M(P)$. Then blocking occurs at Q if a circular cap of angular radius γ contains a peak. The pole of this cap is the radial projection q of Q onto S. The blocking region for the path Q_1Q_2 is the union of all the blocking caps for points Q on the path. These caps are largest midway between Q_1 and Q_2, shrinking to points at Q_1 and Q_2. Then the blocking region is lens-shaped, as in Fig. 7.

Figure 7 shows two arcs K, K' which form the boundary of the blocking region. The argument that follows shows that K, K' are actually arcs of circles. Figure 8 is another view of the peak sphere projected directly along the line Q_1, Q_2. Two planes, π and π', can be drawn through Q_1, Q_2 and tangent to the inner sphere, say at C and C'. These planes project to lines in Fig. 8. The planes π and π' intersect S in two circles, centered at C and C' and both passing through Q_1 and Q_2. Since $M(P)$ is the convex hull of P and the inner sphere, π is a supporting plane of $M(P)$ as long as P lies below π (i.e., in the half-space containing the inner sphere). Then $M(P)$ does not block the path Q_1Q_2 if P lies below π, or below π'. The part of S lying above both π and π' appears shaded in Fig. 8. Suppose P belongs to the shaded region. Project the triangle $C'PC$, a subset of $M(P)$, onto the plane of Fig. 8. The path Q_1Q_2 projects to a point lying inside this projected triangle. Then Q_1Q_2, a chord of S, must intersect the triangle $C'PC$.

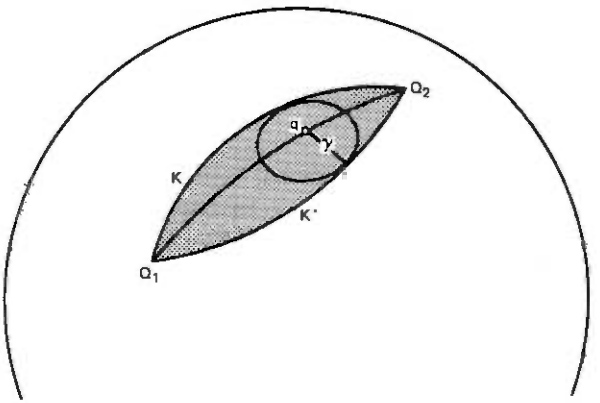

Fig. 7—Blocking region for two points Q_1, Q_2 on the peak sphere.

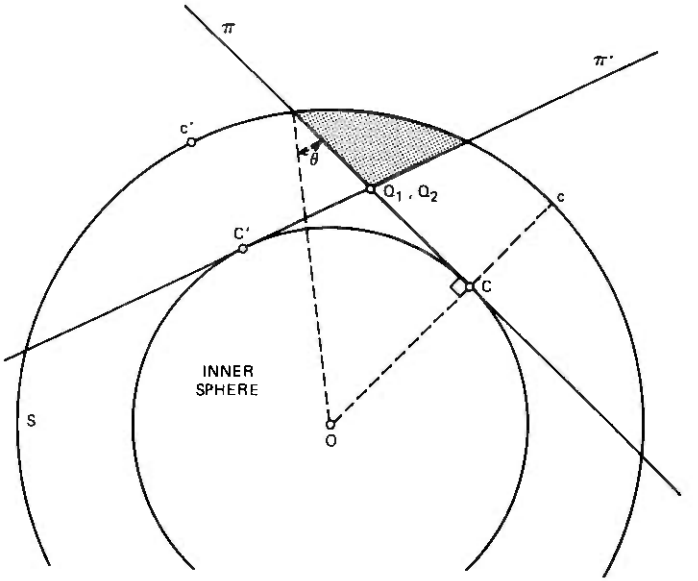

Fig. 8—Another view of the blocking region.

The point of intersection is a point of $M(P)$, which blocks the path. Thus, the shaded region, bounded by arcs of the circles $S \cap \pi$, $S \cap \pi'$, is the blocking region for Q_1Q_2.

The area $A(Q_1, Q_2)$ of the blocking region in Fig. 7 will now be expressed as a function of the angle $2\rho = \angle Q_1OQ_2$. Project the centers C, C' in Fig. 8 radially out to c, c' on S. Figure 9 is another view of S showing c, c' as the poles of two circular caps bounded by $S \cap \pi$ and $S \cap \pi'$. The angular radius of both caps is $\frac{1}{2}\pi - \theta$, as is clear from Fig. 8. The chord Q_1Q_2 subtends some angle $2\alpha = \angle Q_1cQ_2$ at c. Using the spherical sine law in the right triangle Q_1, c, $\frac{1}{2}(Q_1 + Q_2)$, one may determine α from

$$\sin \alpha = \sin \rho / \cos \theta. \tag{10}$$

The cap with pole c has area $2\pi a^2(1 - \sin \theta)$ and the sector included within angle 2α has area

$$A_s = 2\alpha a^2(1 - \sin \theta).$$

Also, the triangle Q_1cQ_2 has area

$$A_T = (2\alpha + 2\beta - \pi)a^2,$$

where $\beta = Q_1Q_2C_1 = Q_2Q_1c$. The sine law may be applied to triangle Q_1cQ_2 to find β

$$\sin \beta = \cos \theta \sin 2\alpha / \sin 2\rho. \tag{11}$$

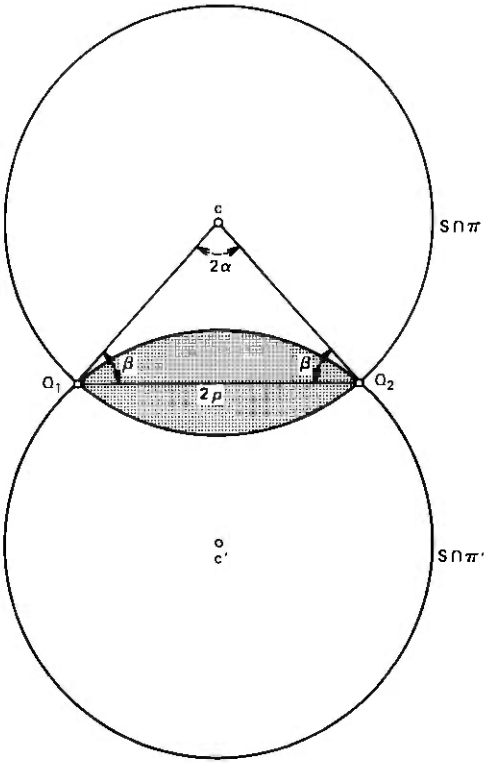

Fig. 9—Angles used in deriving area A of blocking region.

The difference $A_S - A_T$ is $\frac{1}{2}A(Q_1, Q_2)$. Thus,

$$A(Q_1, Q_2) = a^2(2\pi - 4\beta - 4\alpha \sin \theta), \tag{12}$$

where (10), (11) give α, β.

The blocking region is more complicated if Q_1, Q_2 or both are not on S. As in Fig. 7, each point Q on Q_1Q_2 is blocked by peaks lying in a circular cap of radius γ given by (4); the blocking region is the union of these caps. Let Q_1', Q_2' be the points where the extended line Q_1Q_2 meets S. The blocking region for Q_1Q_2 is a subset of the blocking region for $Q_1'Q_2'$. As shown in Fig. 10, the blocking region consists of the caps for blocking at Q_1 and Q_2 plus the part of the blocking region for $Q_1'Q_2'$ that lies between these caps. The centers of the two end caps are the points q_1, q_2 obtained by projecting Q_1, Q_2 radially onto S.

The two end caps have a special role in the blocking. Normally, Q_1, Q_2 are known to lie above ground, and so the two end caps are known to contain no peaks. If the ground levels below q_1, q_2 are known, then peaks P_1, P_2 must exist somewhere at the appropriate angles γ_1,

γ_2 away from q_1, q_2. In Fig. 10, Q_1, Q_2 are assumed to be at ground level; then P_1, P_2 lie on the boundaries of the end caps. The mountains $M(P_1)$, $M(P_2)$ on which Q_1, Q_2 lie can themselves block the path Q_1Q_2. Thus, in Fig. 10, P_2 blocks the path because it lies in the blocking region. To compute the conditional blocking probability for the configuration in Fig. 10, one must know both the area of the part of the blocking region that lies outside the end caps and also angles φ_1, φ_2 that limit where P_1, P_2 can lie to cause blocking. In the applications that follow, it will suffice to let Q_1 lie at a peak $Q_1 = P_1$ and let Q_2 be a point at ground level. That simplifies Fig. 10 to Fig. 11.

Let $z_2 = \angle P_1 0 q_2$, the angular distance along the arc P_1q_2. The depth at Q_2 determines the angle γ_2 of the end cap. The sides of the spherical triangle P_1q_2c are now known, and so its angles $\beta = \angle q_2P_1c$, $\zeta_2 = \angle q_2cP_1$, $\pi - \varphi_2 = \angle P_1q_2c$ are determined. One finds

$$\tan \tfrac{1}{2}\varphi_2 = \left\{ \frac{1 - \cos(z_2 - \gamma_2) + g \sin(z_2 - \gamma_2)}{g \sin(z_2 + \gamma_2) - 1 + \cos(z_2 + \gamma_2)} \right\}^{\frac{1}{2}} \tag{13}$$

$$\sin \beta = \sin \varphi_2 \cos(\theta + \gamma_2)/\cos \theta \tag{14}$$

$$\sin \zeta_2 = \sin \varphi_2 \sin z_2/\cos \theta. \tag{15}$$

The blocking area is twice the area of the half of the blocking region above the line P_1Q_2' in Fig. 11. That half can be obtained as a sum of two parts. One part is a sector of angle $\pi - \varphi_2$ from the end cap; its area is $(\pi - \varphi_2)(1 - \cos \gamma_2)a^2$. The other part is obtained by removing

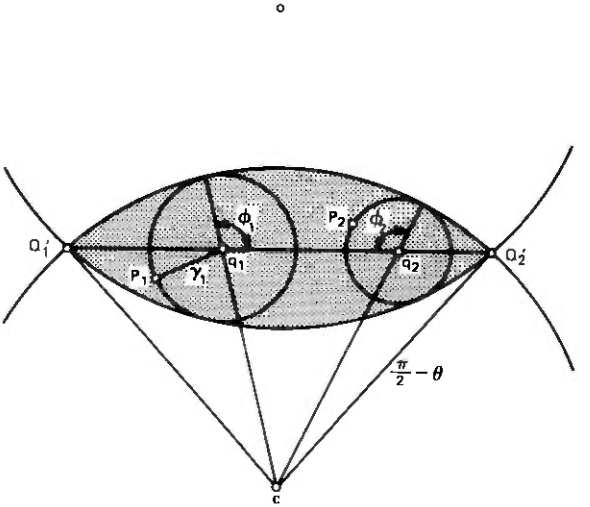

Fig. 10—Blocking region for points Q_1, Q_2 which are not peaks.

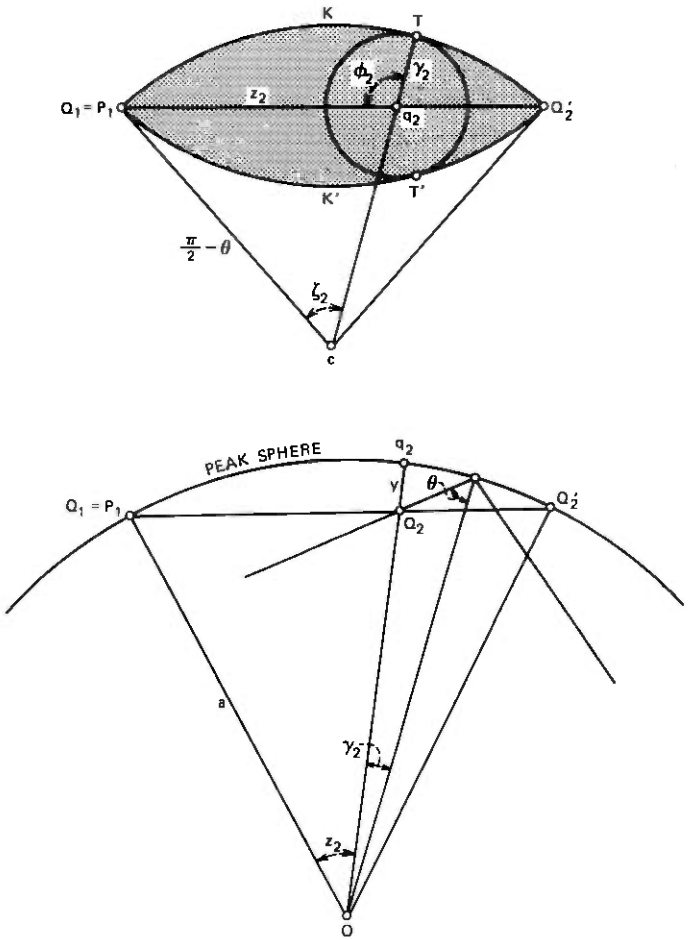

Fig. 11—Blocking region for $Q_1 = P$, a peak, Q_2 not a peak.

the triangle P_1q_2c of area $(\beta + \zeta_2 - \varphi_2)a^2$ from a sector centered at c. The sector has area $\zeta_2(1 - \sin\theta)a^2$. These areas may be combined to express the area of the blocking region in the form

$$A(Q_1, Q_2) = A_0 + A_2,$$

where

$$A_2 = 2\pi(1 - \cos\gamma_2)a^2 \tag{16}$$

and

$$A_0 = (2\varphi_2 \cos\gamma_2 - 2\beta - 2\zeta_2 \sin\theta)a^2. \tag{17}$$

A_2 is the area of the end cap and A_0 is the area of the remainder of the blocking region.

Although the blocking area $A(Q_1, Q_2)$ for the general situation of Fig. 9 will not be needed, it can be obtained in the form

$$A(Q_1, Q_2) = A(Q_1', Q_2) + A(Q_1, Q_2') - A(Q_1', Q_2'). \tag{18}$$

Note that formulas like (16) and (17) give $A(Q_1', Q_2)$, $A(Q_1, Q_2')$ while (10), (11), and (12) give $A(Q_1', Q_2')$. Likewise, with a change of subscripts, (13) gives φ_1 as well as φ_2.

These formulas can now be used to obtain the *path probability* $p(Q_1, Q_2)$, the conditional probability that a clear line-of-sight path exists between given points Q_1, Q_2. When Q_1, Q_2 are on S, as in Fig. 7, $p(Q_1, Q_2)$ is just the probability that the shaded region of area $A(Q_1, Q_2)$ contains no peaks. Then

$$p(Q_1, Q_2) = \exp\left[-\sigma A(Q_1, Q_2)\right], \tag{19}$$

with $A(Q_1, Q_2)$ given by (12).

The situation in Fig. 11 is more complicated. $Q_1 = P_1$, a peak, and Q_2 is supposed to lie on the ground. Then the cap of area A_2 is known to be empty. Two conditions must hold if the entire blocking region is to be empty. One is that the peak P_2 of the mountain on which Q_2 lies causes no blocking. Since P_2 is equally likely to be anywhere on the boundary of the cap around Q_2, there is probability $1 - \varphi_2/\pi$ that P_2 does not block the path. The second condition is that the remainder of area A_0 of the blocking region is empty. Then

$$p(Q_1, Q_2) = (1 - \varphi_2/\pi) \exp(-\sigma A_0), \tag{20}$$

where (13) and (17) give φ_2 and A_0. Formula (20) applies as long as $\gamma_2 < z_2$. It is also possible to have $\gamma_2 = z_2$. In that case, Q_2 lies on the mountain $M(P_1)$; the path in question runs from the peak $Q_1 = P_1$ to Q_2 along the surface of the cone $M(P_1)$. Whether or not such a path is to be considered blocked is a matter of definition. Here $M(P_1)$ is regarded as an open set so that the path is not blocked. As $\gamma_2 \to z_2$, one finds $\varphi_2 \to 0$ and $A_0 \to 0$ so that $p(Q_1, Q_2) \to 1$, i.e., (20) continues to give the correct probability in the limit.

Another limiting situation, $\gamma_2 \to 0$, illustrates an important distinction between Figs. 11 and 7. In the limit $A_2 \to 0$ and A_0 becomes the area of a lens-shaped region, such as shown in Fig. 7. Then the exponential factor in (20) becomes the path probability (19) for the two peaks $Q_1 (= P_1)$, and $\lim Q_2$. However, (20) contains an extra factor $(1 - \varphi_2/\pi)$ which approaches $\frac{1}{2}$, not 1. This disagreement between (19) and (20) is explained as follows. From a point Q_2 near a peak P_2, one can look over a 180-degree view; the mountain $M(P_2)$ blocks the other 180 degrees. Then the factor $\frac{1}{2}$ in (20) is needed to account for possible blocking by $M(P_2)$. But exactly at the peak

$Q_2 = P_2$, the mountain $M(P_2)$ no longer interferes in any direction. Then no factor $\frac{1}{2}$ is needed in (19). The discontinuity in $p(Q_1, Q_2)$ as $Q_2 \to P_2$ could be avoided by assuming that the antenna location Q_2 lies at some known positive height h above ground.

V. PATHS BETWEEN PEAKS

The simplest blocking regions were for paths Q_1Q_2 with both end points on mountain peaks. The path probability $p(Q_1, Q_2)$ in (19) can now be used to derive some interesting properties of peak-to-peak paths. In this section, Q_1 will be a given peak P_1. Q_2 will be another peak selected at random. An element of area $dA(Q_2)$ on the peak sphere S has probability $\sigma dA(Q_2)$ of containing a peak Q_2. Then $\sigma p(Q_1, Q_2)dA(Q_2)$ is the probability that the element contains a peak Q_2 which is visible from Q_1.

Let $d(Q_1, Q_2)$ denote great circle distance between Q_1 and Q_2. Let $\Sigma_k(d)$ denote the random variable which is the sum

$$\Sigma_k(d) = \sum_i d(Q_1, P_i)^k \tag{21}$$

of kth powers of distances from Q_1 to all other visible peaks P_i lying within distance $d[d(Q_1, P_i) \leqq d]$. The element $dA(Q_2)$ contributes a term $d(Q_1, Q_2)^k$ to $\Sigma_k(d)$ with probability $\sigma p(Q_1, Q_2)dA(Q_2)$. Thus, the expected value of $\Sigma_k(d)$ is

$$E[\Sigma_k(d)] = \sigma \int\int d(Q_1, Q_2)^k p(Q_1, Q_2)dA(Q_2), \tag{22}$$

where the integral extends over all points Q_2 in the cap $d(Q_1, Q_2) \leqq d$. Another random variable Σ_k is a sum like (21) extended over all visible peaks, at any distance from Q_1. The mean $E(\Sigma_k)$ is an integral (22) over the entire sphere. Evaluating (22) will give the mean number

Table III — Blocking area $A(Q_1, Q_2)$ in square miles, as given by exact and approximate formulas (12) and (23) with range $d(Q_1, Q_2) = 100$ miles

Grade g	Exact	Approximate
0.01263	7887.7	3333.1
0.02	2430.9	2104.9
0.05	857.6	842.0
0.1	423.2	421.0
0.2	210.2	210.5
0.5	84.1	84.2
1.0	42.0	42.1

of visible peaks $E(\Sigma_0)$ and other information about the distances to visible peaks. That could be done numerically, using (10), (11), (12), and (19). However, the approximation that follows simplifies the evaluation.

The approximation is one which holds when a is so large that angles 2ρ between visible peaks can be considered small. The planar model has $A(Q_1, Q_2) = 0$ and $p(Q_1, Q_2) = 1$, which is too rough to make sense in (22). Instead, the first nonzero term in a series for $A(Q_1, Q_2)$ in powers of ρ will be used. Expansion of the exact formulas (10), (11), and (12) is laborious but straightforward:

$$\begin{aligned} \alpha &= \rho(1 + g^2)^{\frac{1}{2}}/g + \rho^3(1 + g^2)^{\frac{1}{2}}/(6g^3) + 0(\rho^5) \\ \beta &= \tfrac{1}{2}\pi - \rho/g - (2g^2 + 1)\rho^3/(6g^3) + 0(\rho^5) \\ A(Q_1, Q_2) &= a^2(2\rho)^3/(6g) + 0(\rho^5). \\ &= r^3/(6ga) + \cdots, \end{aligned}$$

where $r = 2a\rho$ is the great circle distance $d(Q_1, Q_2)$.

For a simpler, more intuitive, derivation, one may find the size of the circle about a typical point q along the path Q_1, Q_2 in Fig. 7. If z is the great circle distance from Q_1 to q, then the ground level below q lies at depth y satisfying

$$(a - y) \cos (z - \tfrac{1}{2}r)/a = a \cos \tfrac{1}{2}r/a$$

or

$$y = z(r - z)/(2a) + \cdots.$$

The radius $a\gamma$ of the circle about q is approximately y/g, and the blocking area is approximately

$$\begin{aligned} A(Q_1, Q_2) &= \int_0^r a\gamma dz \\ &= \int_0^r z(r - z)dz/(2ag) \\ A(Q_1, Q_2) &= r^3/(6ga) + \cdots, \end{aligned} \tag{23}$$

as before.

From the form of the series used in deriving (23), one may predict that the rate of convergence is determined by the ratio ρ/g. Table III shows that (23) does give a better approximation for large g than for small. In Table III, $a = 3959$ mi; a large range, 100 mi, was used for a severe test of the approximation. At grades g smaller than 0.01263, a 100-mi path between peaks is blocked by the inner sphere. The small ρ/g condition is another way of requiring that the path Q_1Q_2 clears the inner sphere by a safe margin.

Now use the approximation (23) for $A(Q_1, Q_2)$ in (19) to evaluate the integral (22) for the expectation $[\Sigma_k(d)]$:

$$E[\Sigma_k(d)] = 2\pi\sigma \int_0^{d/(2a)} (2a\rho)^{k+1} \exp\{-4a^2\sigma\rho^3/3g\} 2ad\rho$$

$$E[\Sigma_k(d)] = (2\pi\sigma/3)D^{k+2} \int_0^{(d/D)^3} u^{(k-1)/3} \exp(-u)du, \qquad (24)$$

where

$$D = (6ag/\sigma)^{\frac{1}{3}}.$$

The integral in (24) may be expressed in terms of the incomplete gamma function,

$$E[\Sigma_k(d)] = (2\pi\sigma/3)D^{k+2}\{\Gamma[(k+2)/3] - \Gamma[(k+2)/3], (d/D)^3\},$$

or the χ^2 distribution function,

$$E[\Sigma_k(d)] = (2\pi\sigma/3)D^{k+2}\Gamma[(k+2)/3]P(\chi^2|\nu),$$

where

$$\chi^2 = 2(d/D)^3$$

and the number of degrees of freedom is

$$\nu = 2(k+2)/3.$$

Although the approximation (23) becomes poor at long ranges, the integrand is very small there. Thus, (24) can be expected to hold even for long ranges. In particular, the expectation $E(\Sigma_k)$, for visible peaks of all ranges, may be approximated by letting $d \to \infty$ in (24):

$$E(\Sigma_k) = (2\pi\sigma/3)D^{k+2}\Gamma[(k+2)/3]. \qquad (25)$$

For the special value $k = 0$, (25) gives the mean number of peaks visible from Q_1:

$$\begin{aligned} E(\Sigma_0) &= \pi\sigma D^2\Gamma(5/3) \\ &= 9.3645(a^2g^2\sigma)^{\frac{1}{3}} \\ &= 2344(g^2\sigma)^{\frac{1}{3}} \quad \text{if} \quad a = 3959 \text{ mi}. \end{aligned} \qquad (26)$$

Note, as predicted earlier, that the mean number of visible peaks tends to infinity in the limit of large a (planar model). When $k = 1$, (25) simplifies to

$$E(\Sigma_1) = 4\pi ag. \qquad (27)$$

As σ increases, (26) shows that the mean number of visible peaks increases, but (27) shows that the mean sum of distances to visible peaks remains unchanged. This indicates that visible peaks tend to be closer for large σ than for small. One way to define a range for a

"typical" peak is to form the ratio

$$\begin{aligned} D_1 &= E(\Sigma_1)/E(\Sigma_0) = D/\Gamma(2/3) \\ &= 0.738487D \\ &= 1.34190(ag/\sigma)^{\frac{1}{3}}. \end{aligned} \tag{28}$$

VI. RANGES BETWEEN PEAKS

One might ask for a probability distribution for the range d from a peak P_1 to a randomly chosen visible peak $P \neq P_1$. The random process for choosing a peak must be specified with care. Perhaps the most natural process would be this. Construct a random landscape and choose a peak P from the set of Σ_0 visible peaks, all peaks equally likely. Then ask for the probability that P is one of the $\Sigma_0(d)$ peaks within range d of P_1. Given a landscape, the conditional probability that P is within range is $\Sigma_0(d)/\Sigma_0$. Then the unconditioned probability is $E[\Sigma(d)/\Sigma_0]$. Unfortunately, the expectation is hard to obtain [there is also a question of giving an appropriate meaning to $\Sigma_0(d)/\Sigma_0$ when $\Sigma_0(d) = \Sigma_0 = 0$].

By using a different random process, one obtains a simpler distribution. Construct a trial random landscape and pick one of the peaks P at random, this time from the set of all peaks on the entire sphere S. P may not be visible. If not, discard that trial and construct a new landscape. Continue constructing landscapes and choosing peaks until the chosen point P is visible. Then ask for the probability $p(d)$ that P lies within range d.

To determine $p(d)$, note that the total number of peaks on the entire sphere has the Poisson distribution with mean $4\pi a^2\sigma$. The argument to follow assumes that this number is large, so that the number of peaks actually obtained is almost always very close to its mean value. Then the probability that P is visible is $E(\Sigma_0)/(4\pi a^2\sigma)$. If $q[\Sigma_0, \Sigma_0(d)]$ is the joint probability for Σ_0 and $\Sigma_0(d)$ in each trial, then $q[\Sigma_0, \Sigma_0(d)]\Sigma_0/(4\pi a^2\sigma)$ is the probability that a trial has Σ_0 visible peaks, $\Sigma_0(d)$ within range d, and that P is visible. The joint probability for the numbers Σ_0, $\Sigma_0(d)$ of the landscape, selected when P is visible, is $q'[\Sigma_0, \Sigma_0(d)] = q[\Sigma_0, \Sigma_0(d)]\Sigma_0/E(\Sigma_0)$. The probability that P lies within range d is obtained as a sum over $\Sigma_0(d)$ and Σ_0

$$\begin{aligned} p(d) &= \sum q'[\Sigma_0, \Sigma_0(d)]\Sigma_0(d)/\Sigma_0 \\ &= E[\Sigma_0(d)]/E(\Sigma_0) \\ p(d) &= 1 - \Gamma[2/3, (d/D)^3]/\Gamma(2/3), \end{aligned} \tag{29}$$

the last line following from (24).

It is clear from this derivation that the second random process tends to select random landscapes with larger Σ_0 than the landscapes

Table IV — Probability $p(d) = E(\Sigma_0(d))/E(\Sigma_0)$ that a randomly chosen visible peak lies within range d

d/D	Probability
0.25	0.06880
0.30211	0.1
0.48595	0.25
0.5	0.26361
0.72212	0.5
0.75	0.53050
1	0.77518
1.20507	0.9
1.32182	0.95
1.5	0.98440
1.55886	0.99
1.81350	0.999
2	0.99983

of the first process. However, Σ_0 may be expected to have a highly peaked distribution, in which case Σ_0 is nearly always close to $E(\Sigma_0)$. Then $q(\cdot, \cdot)$ and $q'(\cdot, \cdot)$ are nearly the same, and (29) is also a good approximation to the range distribution for the first random process.

Equation (29) provides numerical values for the range distribution in Table IV.

Another random variable of interest is the range to the furthest visible peak. Even the expectation of this maximum range is hard to derive. However, a simpler "typical" maximum range is the range d_m such that the expected number of visible peaks with ranges $d > d_m$ is just $\frac{1}{2}$. Then d_m satisfies

$$E(\Sigma_0) - E[\Sigma_0(d_m)] = \tfrac{1}{2}, \tag{30}$$

and (24) shows that

$$\int_{(d_m/D)^2}^{\infty} u^{-\frac{1}{2}} \exp(-u)du = 3/(4\pi\sigma D^2). \tag{31}$$

Table V — Mean number of visible peaks $E(\Sigma_0)$ and range d_m such that $E(\Sigma_0\text{-}\Sigma_0(d_m)) = \frac{1}{2}$

σD^2	$a^2g^2\sigma$	$E(\Sigma_0)$	d_m/D
3.11	0.84	8.8	1.30
5.69	5.12	16.1	1.40
11.3	40.1	32.1	1.50
24.5	409	69.5	1.60
58.4	5528	165.6	1.70
153.6	100572	435.5	1.80
450.5	2.54×10^6	1278	1.90
1477	8.95×10^7	4189	2.00
5447	4.49×10^9	15448	2.10

Table V gives numerical values of d_m/D as a function of $\sigma D^2 = (36a^2g^2\sigma)^{\frac{1}{4}}$. $E(\Sigma_0)$, which also depends on σD^2 as shown by (26), also appears in the table. Note that d_m is not just a function of a single product of powers of a, g, σ; it has a more complicated form $(ag/\sigma)^{\frac{1}{4}} \times$ function $(a^2g^2\sigma)$.

The integral in (31) is a rapidly decreasing function of d_m/D. Then the numbers in Table V would not change much if d_m were redefined with the term $\frac{1}{2}$ in (30) replaced by any other number of the same order of magnitude. For the same reason, d_m can be expected to be a good approximation to the mean range to the furthest peak.

VII. THE HORIZON

The approximation (23) will now be used to derive properties of the range from a peak P_1 to the horizon at a random azimuth angle. The range to the horizon is a more interesting random variable than the range to a random visible peak. As has been noted, it is not always clear what to count as a peak in a real landscape. But the horizon has no ambiguity.

Look from P_1 with a fixed azimuth angle. One sees only sky at high elevations and ground at low elevations. The *horizon point* is the limiting point at ground level which has the highest elevation angle. The distance z from P_1 to the horizon is the range of interest here.

Figure 11 will now be used to derive the conditions under which Q_2 is the horizon point, as seen from a peak P_1. If Q_2 is the horizon point, the entire straight line path P_1Q_2' in Fig. 11 must intersect the ground only at Q_2. Then the entire lens-shaped blocking region for Q_2' must contain no peaks. But the depth at Q_2 determines the circle on which a peak must lie. This circle appears in Fig. 11 inside the (open) blocking region. The only place that this peak can be now is on the boundary of the blocking region at one of the points of tangency T, T'.

Figure 11 shows the usual situation in which the horizon point is not on the inner sphere. There is small probability that Q_2 is on the inner sphere. In that case, q_2 is at angle $\frac{1}{2}\pi - \theta$ away from P_1, the centers c, c' coincide with q_2, and the blocking region is bounded by the circle through P_1 with center q_2. There is no second peak on the boundary of the blocking region; Q_2 lies on $M(P_1)$.

To find the probability distribution for the horizon range, one may first find the joint distribution for that range and the range r to the intersection point Q_2'. Since r is the largest range for which the corresponding blocking region is empty, the probability distribution function for r is $P(r) = 1 - \exp[-\sigma A(P_1, Q_2')]$. In this derivation, the approximation (23) will be used to write

$$P(r) = 1 - \exp[-(r/D)^3].$$

Figure 12 shows Q_2' lying at a range between r and $r + dr$, an event of probability $dP(r)$. Given this position for Q_2', the band between the boundaries of the blocking regions at r and $r + dr$ contains the peak on which the horizon point lies. The shaded part of this band is the region in which the peak must lie so that the horizon point Q_2 will have range $z_2 \leqq z$. The conditional probability function for the horizon range is just the ratio of the area of the shaded part of the band to the total area of the band. To simplify that calculation, one may replace the dotted line by a great circle that crosses P_1P_2' at right angles. That approximation leads ultimately to the conditional distribution

$$\text{Prob}\{\text{horizon range} \leqq z | r\} = (z/r)^2, \quad 0 \leqq z \leqq r. \tag{32}$$

The details are omitted because the result can almost be guessed immediately from the roughly triangular shapes of the two parts of the shaded region.

Now the unconditional probability distribution for the horizon range is obtainable from (32) by integrating

$$\begin{aligned}\text{Prob}\{\text{horizon range} \leqq z\} &= \int_0^z dP(r) + \int_z^\infty (z/r)^2 dP(r) \\ &= (z/D)^2\Gamma[\tfrac{1}{3}, (z/D)^3] + 1 - \exp[-(z/D)^3]. \end{aligned} \tag{33}$$

Table VI gives numerical values.

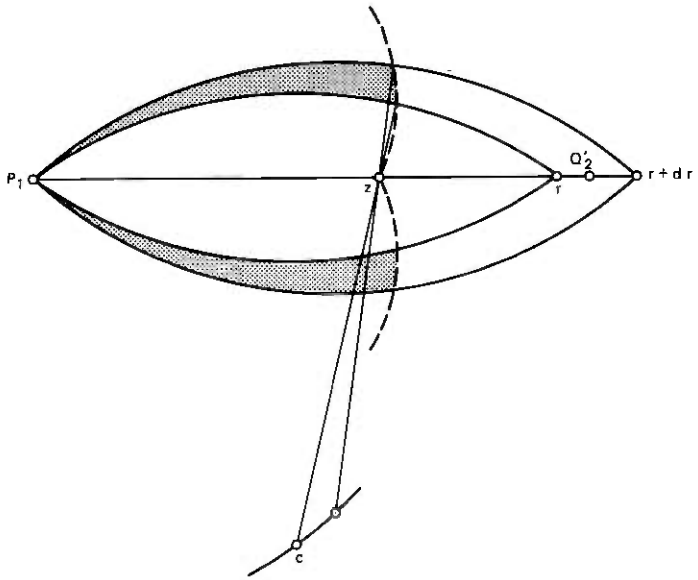

Fig. 12—Horizon at range $\leqq z$.

Table VI — Distribution function for range z to the horizon looking from peak P_1 with a random azimuth angle

z/D	Probability
0.21417	0.1
0.25	0.13625
0.35618	0.25
0.5	0.42355
0.56305	0.5
0.75	0.70432
0.79977	0.75
1.0	0.88853
1.02324	0.9
1.15749	0.95
1.25	0.97109
1.40527	0.99
1.5	0.995247
1.67110	0.999

The moments of the horizon range z are easy to find. From (33), the probability density for z is

$$2z \int_z^\infty r^{-2} dP(r).$$

The kth moment of z is

$$\begin{aligned} E(z^k) &= 2 \int_0^\infty z^{k+1} \int_z^\infty r^{-2} dP(r) dz \\ &= [2/(k+2)] E(r^k). \end{aligned}$$

The last line is obtained by integrating by parts. The expectation on the right is another integral that can be evaluated in the manner of (24) and (25). The final result is

$$E(z^k) = 2D^k \Gamma[1 + (k/3)]/(k+2). \tag{34}$$

Equation (34) with $k = 2$ is particularly interesting. If $z(\varphi)$ is the range to the horizon when looking with azimuth φ, then the mean area within horizon range is

$$\begin{aligned} E(\text{area within horizon}) &= E\left(\frac{1}{2}\int_0^{2\pi} z^2(\varphi) d\varphi\right) \\ &= \pi E(z^2) \\ &= \tfrac{1}{2}\pi \Gamma(5/3) D^2 \\ &= 1.41803 D^2. \end{aligned} \tag{35}$$

This expectation is only an upper bound on the mean area visible. For, as is clear in Fig. 4, there are points within horizon range that are obscured from view.

It is interesting to compare Tables IV and VI. At any given probability level, the range to the horizon is smaller than the range to a randomly chosen visible peak. This may be surprising at first. However, each visible peak is itself a point on the horizon. As seen in Fig. 4, the horizon consists entirely of small line segments extending down the sides of the mountains from the visible peaks. The line segments for distant peaks tend to subtend smaller azimuth angles than the segments for nearby peaks. Picking an azimuth at random, one is more likely to find the horizon point on one of the nearby visible peaks than on one far away.

Another expectation that exhibits the same effect is the mean azimuth angle between the horizon point and the peak of the mountain on which the horizon point lies. The ranges z and r determine this angle. Without belaboring the details, one can approximate this angle by its tangent and make the further approximations by which (33) was derived. The expected angle is found to be

$$\begin{aligned} E(\text{angle}) &= E[(r - z)/(2ag)] \\ &= \Gamma(4/3)D/(6ag). \end{aligned}$$

That result can be stated in a more illuminating way:

$$\begin{aligned} E(\Sigma_0)E(\text{angle}) &= \pi\Gamma(4/3)\Gamma(5/3) \\ &= (2\pi 3^{-\frac{1}{2}})(2\pi) \\ &= 0.40306(2\pi). \end{aligned}$$

By contrast, if $E(\Sigma_0)$ peaks were evenly distributed in azimuth with angular separation $2\pi/E(\Sigma_0)$, one would obtain

$$E(\Sigma_0)E(\text{angle}) = 0.25(2\pi).$$

The larger factor 0.40306 again occurs because of the variability of the angles which visible mountains subtend on the horizon.

VIII. COVERAGE AREA

The *coverage set* for a point P is the set of points Q such that a line-of-sight path PQ exists. In VHF radio applications, the coverage set of P is the set of points Q to which an antenna at P can radiate a strong signal. This section will estimate the *mean coverage area* C, the expected area of the coverage set for $P = P_1$, a peak.

The coverage set can have a very complicated shape. Figure 13 shows one coverage set. In Fig. 13, the peaks are not in a Poisson pattern; to simplify the drawing, the peaks were located at points of a regular lattice. The coverage set contains the entire mountain on which P lies plus parts of adjacent mountains. These points alone

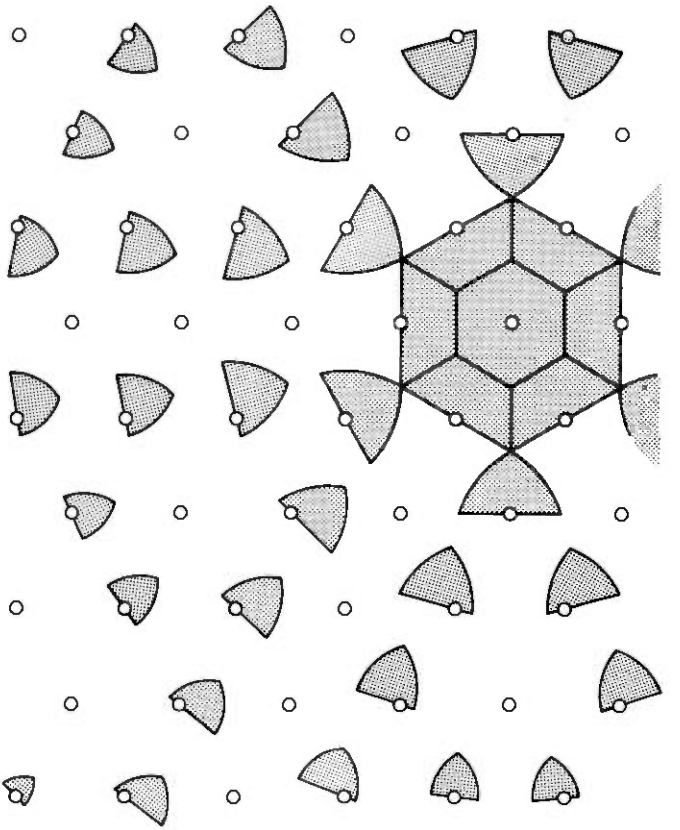

Fig. 13—A coverage set.

constitute a hexagon of area $4/\sigma$. In addition, the coverage set contains many smaller isolated patches on more remote mountains. These small patches can be so numerous that they represent most of the coverage area.

If Fig. 13 represented the coverage set of a base station in a mobile radio telephone system, the station would only serve the hexagon of area $4/\sigma$. The other small patches would lie in the service areas of other stations, and so these patches would represent places where the given station can interfere with other stations.

As in Fig. 11, let P_1 be a given peak and Q_2 another point at ground level. Suppose the distance r from P_1 to Q_2 is known, i.e., the angle $z = r/a$ in Fig. 11 is given. Let $f(r)$ denote the probability that a line-of-sight path P_1Q_2 exists. Since an element of area $dA(Q_2)$ at Q_2 belongs to the coverage set of P_1 with probability $f(r)dA(Q_2)$, the

mean area covered is

$$C = \int\int f(r)dA(Q_2) = 2\pi a^2 \int_0^{\pi} f(r) \sin z_2 dz_2. \tag{36}$$

Before attempting to evaluate $f(r)$ and C, the integral (36) will be given a second interpretation. Now consider Q_2 at a fixed location and count the number of peaks visible from Q_2. The probability that an element of area $dA(P_1)$ contains a peak P_1 visible from Q_2 is $\sigma f(r)dA(P_1)$. Then the mean number of peaks visible from Q_2 is

$$\begin{aligned} E(\text{visible peaks}) &= \sigma \int\int f(r)dA(P_1) \\ &= 2\pi a^2 \sigma \int\int f(r) \sin z_2 dz_2 \\ &= \sigma C. \end{aligned} \tag{37}$$

Equation (37) can be used to derive very simple bounds on C. Clearly, more peaks are visible from a point Q_2 at high altitudes than at low. If Q_2 were itself a peak, the mean number of visible peaks would be $E(\Sigma_0)$, given by (26). But Q_2 has probability zero of being exactly at a peak. If Q_2 is at any slightly lower altitude, Q_2 is on the side of a hill which obscures 180 degrees of the view from Q_2. Thus, $E(\text{visible peaks}) \leqq \frac{1}{2}E(\Sigma_0)$, and (37) shows

$$C \leqq E(\Sigma_0)/(2\sigma). \tag{38}$$

Curiously, the right-hand side of (38) is exactly the mean area within the horizon as given by (35). Then (38) is a bound that was obtained in Section VII.

At the other extreme, Q_2 might be on the inner sphere, where no peak is visible. In most cases, that event will be so unlikely that it will be safe to say that the worst reasonable possibility is that Q_2 is down in a valley near the point where three mountains meet. Here, the three mountain peaks are visible and so one concludes $C \geqq 3/\sigma$.

To obtain $f(r)$, and hence C, recall that (20) is a formula for the path probability $p(Q_1, Q_2)$, depending on the altitude y at Q_2. To get $f(r)$ one may average $p(Q_1, Q_2)$ over y (or γ_2). This average may be expressed as a sum of two terms which account for the possibilities that Q_2 belongs to the same mountain $M(P_1)$ as Q_1 or to a different mountain.

In Fig. 11, if $\gamma_2 = z_2$, then Q_2 lies on the mountain $M(P_1)$. This event has probability $\exp\{-2\pi\sigma a^2(1 - \cos z_2)\}$. The path probability $p(Q_1, Q_2) = 1$ if $r \leqq (\frac{1}{2}\pi - \theta)a$. If $r > (\frac{1}{2}\pi - \theta)a$, then Q_2 lies on the inner sphere, the path P_1Q_2 is blocked, and $p(Q_1, Q_2) = 0$.

If $\gamma_2 < z_2$, then Q_2 lies on a different mountain $M(P_2)$. This possibility contributes a second term to $f(r)$,

$$\int_{\gamma_2=0}^{z_2} p(Q_1, Q_2) d\{1 - \exp(-\sigma A_2)\},$$

where A_2 and $p(Q_1, Q_2)$ are given by (16) and (20).

For $r \leqq (\frac{1}{2}\pi - \theta)a$, the two terms combine into

$$f(r) = \exp\{-2\pi\sigma a^2(1 - \cos z_2)\} + \int_{\gamma_2=0}^{z_2} (1 - \varphi_2/\pi) \exp[-\sigma(A_0 + A_2)]\sigma dA_2, \quad (39)$$

where (16) and (17) provide A_2, A_0. A similar formula applies when r is larger, but $f(r)$ is very small at such large ranges.

One could find $f(r)$ to any desired accuracy by evaluating the integral in (39) numerically. Instead, (39) will be replaced by a simpler approximate formula. Since a is large, the first term of (39) is approximately $\exp(-\sigma\pi r^2)$; also, $A_2 = \pi x^2$ where $x = a\gamma_2$. The approximations to φ and A_0 which follow are not uniformly good but are intended to apply in situations that contribute most of the coverage area C. Except at very short ranges r, a typical blocking region is more elongated than that shown in Fig. 11. Figure 14 is more typical. Then $\varphi_2 = \frac{1}{2}\pi$, approximately. With that approximation, the shaded region in Fig. 14 has area $A_0 + \frac{1}{2}A_2$. It consists of a triangle, of area xr, and two extra lens-shaped pieces. The two extra pieces can fit together into one lens of exactly the shape of the blocking region for two peaks at separation r. Then the two extra pieces combined have area $r^3/(6ga)$, as in (23). Now the exponent $\sigma(A_0 + A_2)$ in (39) is approximately $(r/D)^3 + \frac{1}{2}\pi\sigma x^2$.

To substitute these approximations into (39), write

$$\begin{aligned} Y &= \sigma\pi r^2 \\ X &= r^3/(6ga) = (r/D)^3 = [\Gamma(5/3)Y/E(\Sigma_0)]^{\frac{3}{2}} \end{aligned} \quad (40)$$

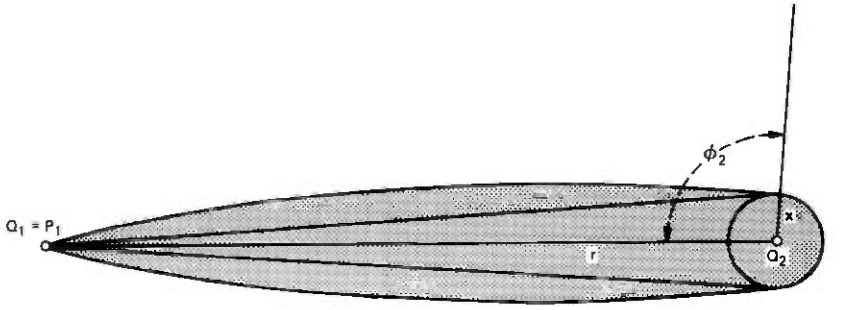

Fig. 14—Approximation of A_0 and φ_2.

and then

$$f(r) = \exp(-Y) + \pi\sigma \exp(-X) \int_0^r \exp\{-\sigma x r - \tfrac{1}{2}\sigma\pi r^2\} r dr$$

$$f(r) = \exp(-Y) + \exp(-X)\{1 - \exp(-Y[\tfrac{1}{2} + \pi^{-1}])\} + \exp(-X + \tfrac{1}{2}Y/\pi^2)(2Y/\pi)^{\frac{1}{2}} \times \{\operatorname{erf}(Y^{\frac{1}{2}}[1 + \pi^{-1}]) - \operatorname{erf}(Y^{\frac{1}{2}}/\pi)\}. \quad (41)$$

Figure 15 shows curves of $f(r)$. The ordinate $Y^{\frac{1}{2}} = (\sigma\pi)^{\frac{1}{2}}r$ was used as a convenient normalized range. It may be interpreted as the square root of the mean number of other peaks within range r of P_1. As (40) shows, the parameter $E(\Sigma_0)$ enters into $f(r)$ through the variable X. The $f(r)$ curves for different values of $E(\Sigma_0)$ lie close together at short ranges. As the range increases, $f(r)$ falls more sharply for small $E(\Sigma_0)$. There is a limiting curve, as $E(\Sigma_0) \to \infty$, which is obtained from (41) by setting $X = 0$. As (40) shows, $X = 0$ also corresponds to the limit $a \to \infty$; i.e., this limit represents the planar model.

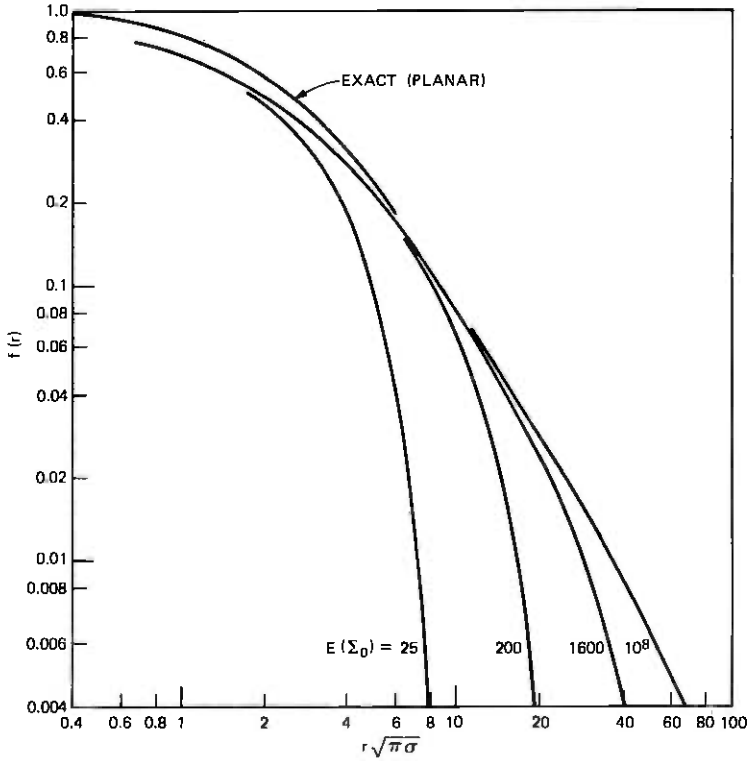

Fig. 15—Probability $f(r)$ that a random point at ground level is visible from a peak r miles away.

Table VII — Mean coverage area C

$E(\Sigma_0)$	σC
25	8.3
50	12.2
100	17.3
200	23.5
400	30.6
800	38.6
1600	47.4

The approximations which led to (41) are poor when r is small. Fortunately, the planar model is good for smaller r. The curve labeled "exact (planar)" in Fig. 15 was obtained by a numerical integration, using an exact equation (39) for the planar model. The planar curve crosses the 0.5 probability level at $Y^{\frac{1}{2}} = 2.3$. Then $r = 1.3\sigma^{-\frac{1}{2}}$ is the range at which the odds of finding a clear path are even.

The behavior of $f(r)$ for large r can be obtained by replacing the error functions in (41) by their asymptotic expansions. The leading terms are

$$f(r) \sim \exp(-Y) + \pi^2 Y^{-1} \exp(-X). \tag{42}$$

In Fig. 14, the $f(r)$ curve starts to depart from the limiting curve at values of Y near $E(\Sigma_0)$. For larger Y, the factor $\exp(-X)$ in (42) becomes small rapidly. In the planar model, $\exp(-X) = 1$ for all Y; (42) then shows that $f(r) \sim \pi^2/Y = \pi/(\sigma r^2)$.

To good approximation, the integral (36) for the mean coverage area can be replaced by

$$C = \int_0^\infty f(r)d(\pi r^2) = \sigma^{-1} \int_0^\infty f(r)dY. \tag{43}$$

The main contribution to C in (36) comes from the range $0 \leqq r \leqq (\frac{1}{2}\pi - \nu)a$, in which (39) holds exactly and (41) approximately. Then (41) will be used for $f(r)$ in (43) and, since $f(r) \to 0$ rapidly for larger r, the range of integration has been extended to $0 \leqq Y < \infty$.

Since (40) and (41) express $f(r)$ in terms of Y and the single parameter $E(\Sigma_0)$, (43) shows that σC is a function of $E(\Sigma_0)$ only. Table VII gives values of this function, obtained from (41) and (43) by numerical integration. These values also represent mean numbers of peaks visible from a random point on the ground, as (37) showed.

In Table VII, σC appears to be a slowly increasing function of $E(\Sigma_0)$. As $E(\Sigma_0) \to \infty$, the model becomes planar and then $f(r) \sim \pi^2/Y$. Then (43) shows $\sigma C \to \infty$ in the planar model limit.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 4, April 1975
Printed in U.S.A.

Two Design Techniques for Digital Phase Networks

By F. J. BROPHY and A. C. SALAZAR

(Manuscript received October 7, 1974)

Two computer-aided algorithms for the design of all-pass digital filters are presented. The first technique is based on a linear programming approach to solving the approximation problem posed by the minimax design of an all-pass digital filter. A new iterative algorithm with stability constraints is offered for direct form design. The second technique implements a gradient search for those quadratic factors of an all-pass transfer function that lead to a locally optimal approximation (in the least-squares sense) of a desired phase function. New initial guess procedures and the parameterization of linear-phase offset enhance the least-squares design procedure. Examples illustrating the result of both procedures are included.

I. INTRODUCTION

The increasing availability of digital signal processors such as those described in Refs. 1 and 2 has generated much interest in the algorithmic design of digital filters. One particular class of recursive digital filters commonly referred to as all-pass digital networks has an important and interesting design problem associated with it. That is, the design objective for this type of filter involves the following transfer function

$$H(z^{-1}) = \frac{\sum_{k=0}^{N} b_{N-k} z^{-k}}{\sum_{k=0}^{N} b_k z^{-k}}.$$

Because of the relationship between numerator and denominator polynomials, the number of degrees of freedom in filter design has been reduced to N from the usual $2N$. Since the magnitude function of $H(z^{-1})$ is precisely 1.0 on the unit circle, the design problem is focused directly on the phase variation of $H(z^{-1})$. The importance of this design problem does not arise from an academic viewpoint.

There are signal processing applications in which an influential factor in signal fidelity is the amount of phase distortion present in

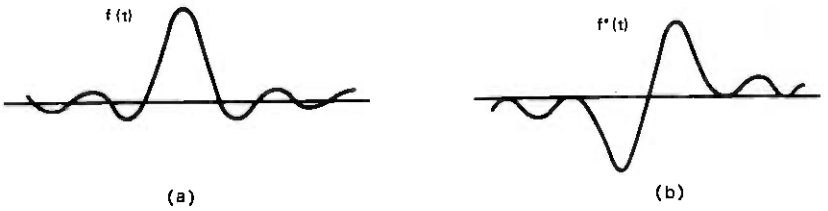

Fig. 1—(a) Original pulse. (b) Phase-distorted pulse.

the medium. The effects of phase distortion in communication systems are illustrated in Refs. 3 and 4. Apart from nonlinear phase equalization applications, all-pass networks can be used to provide a constant phase shift over a specified frequency band or bands. The Hilbert transformer commonly found in bandpass modulation systems is just one example of this application. In constructing phased arrays in radar and seismic research, constant phase shifters are also found to be useful.[5,6] Figure 1 illustrates how a constant phase offset can shape (or distort) the impulse response of a system where $f(t)$ and $f^*(t)$ differ by a constant phase offset of $\pi/2$. A constant phase shift of any amount besides an odd multiple of $\pi/2$ will produce a pulse with a single large lobe. Equalization of this type of distortion is again possible by all-pass networks.

Previous work[7] has addressed the envelope delay design problem. In many cases, this is sufficient but, as seen above, there are applications where the phase function must be treated directly.

Our design techniques are for all-pass structures where the design criteria stem from the phase function directly. The first technique, described in Section II, is a new method for designing all-pass networks using linear programming. This approach allows for fast (at least quadratic), always convergent design of phase networks. For the first time, stability can be treated directly in the design procedure. The second algorithm is based on a gradient search procedure on a least-squares criterion. The basic approach is analogous to those described in Refs. 7 to 9. The all-pass structure reduces the number of variables and simplifies the gradient calculations. In addition to developing the algorithm, we provide initial guess procedures and linear-phase offset parameters that enhance the algorithm. These initial guess procedures are new noniterative filter designs that can serve as excellent all-pass approximations in their own right.

II. A LINEAR PROGRAMMING APPROACH

A need for fast, reliable design of all-pass digital filters has been shown in the previous section. Linear programming techniques have

been found to be useful in rational function approximations[10,11] and have been applied to the magnitude-only design of digital filters.[12,13] Here we show how the all-pass structure in digital filters can be transformed into a problem that can be handled by linear programming techniques also. As we shall see, the rational function differs from the magnitude-only case. Most importantly, this technique allows the question of stability to be handled directly in the design procedure. Other techniques that consider the phase or envelope delay variation of the digital filter (see Refs. 7 and 9 and Section III of this paper) deal with stability with a more heuristic approach.

To develop the linear programming design method, we first recall that the all-pass transfer function is

$$H(z^{-1}) = \frac{P(z^{-1})}{Q(z^{-1})} = \frac{b_N + b_{N-1}z^{-1} + b_{N-2}z^{-2} + \cdots + b_0 z^{-N}}{b_0 + b_1 z^{-1} + \cdots + b_N z^{-N}} \tag{1a}$$

$$\frac{P(z^{-1})}{Q(z^{-1})} = \frac{z^{-N}(b_N z^N + b_{N-1}z^{N-1} + \cdots + b_0)}{(b_N z^{-N} + b_{N-1}z^{-N+1} + \cdots + b_0)}. \tag{1b}$$

Hence, the phase function of (1) on the unit circle is

$$\phi\left(z^N \frac{P(z^{-1})}{Q(z^{-1})}\right)\Big|_{|z|=1} = -2\phi[Q(z^{-1})]|_{|z|=1}. \tag{2}$$

From (2) we note that the phase variation of $H(z^{-1})$ is equivalent (modulo a constant multiplier and an N sample delay term) to the phase of $Q(z^{-1})$. Henceforth, we address the problem of synthesizing $Q(z^{-1})$. The phase variation of $Q(z^{-1})$ on the unit circle is

$$\phi[Q(z^{-1})]|_{|z|=1} = \tan^{-1} \frac{-\sum_{k=1}^{N} b_k \sin 2\pi kf}{\sum_{k=0}^{N} b_k \cos 2\pi kf}$$

$$\tan \phi[Q(z^{-1})]|_{|z|=1} = \text{Imag}\,[Q(e^{-j2\pi f})]/\text{Real}\,[Q(e^{-j2\pi f})]. \tag{3}$$

Further,

$$\tan \phi[Q(e^{-j2\pi f})] = \frac{-\sum_{k=1}^{N} b_k \sin 2\pi kf}{\sum_{k=0}^{N} b_k \cos 2\pi kf} \triangleq \frac{R(f)}{S(f)}. \tag{4}$$

Our design criterion is chosen to be

$$\min_{\{b_k\}} \max_{n} \left| D(f_n) - \frac{R(f_n)}{S(f_n)} \right| \qquad n = 0, 1, 2, \cdots, M,$$

where $D(f)$ is the tangent desired phase function and M is a number

of frequency points* ($\gg N$) chosen to ensure adequate approximation over a subinterval of $|f| \leqq \frac{1}{2}$, namely, $0 < f_0 < f_1 \cdots < f_M < \frac{1}{2}$. We recall that the desired phase function has been scaled down by $-\frac{1}{2}$ because of the factor of -2 appearing in (2) and will have a delay of N samples inherent in its design by the z^{-N} factor of (1b). It is important to note here that the norm is applied to the tangent of the desired phase function instead of the desired phase function itself.†

If we prevent $S(f)$ from assuming the value zero, we seek the minimum value of Δ,

$$|D(f_n)S(f_n) - R(f_n)| \leqq \Delta S(f_n). \tag{5}$$

Using the differential correction idea of Ref. 10, we expand the right-hand side of (5) in an iterative form:

$$\Delta S(f_n) \approx \Delta_k S_k(f_n) + (\Delta - \Delta_k)S_k(f_n) + [S(f_n) - S_k(f_n)]\Delta_k. \tag{6}$$

The intention is to iterate toward those values of $\{b_j\}$ that minimize Δ. The subscript k indicates the kth iteration. We then have, from (5) and (6),

$$|D(f_n)S(f_n) - R(f_n)| - \Delta_k S(f_n) - (\Delta - \Delta_k)S_k(f_n) \leqq 0,$$

which translates into a familiar pair of equations[10]

$$[D(f_n) + \Delta_k]S(f_n) - R(f_n) + (\Delta - \Delta_k)S_k(f_n) \geqq 0 \tag{7}$$

$$[-D(f_n) + \Delta_k]S(f_n) + R(f_n) + (\Delta - \Delta_k)S_k(f_n) \geqq 0. \tag{8}$$

Substituting the series forms for $R(f_n)$ and $S(f_n)$, we have

$$\sum_{j=1}^{N} \{[D(f_n) + \Delta_k] \cos 2\pi j f_n + \sin 2\pi j f_n\} b_j + (\Delta - \Delta_k)S_k(f_n) \geqq -D(f_n) - \Delta_k \tag{9}$$

$$\sum_{j=1}^{N} \{[-D(f_n) + \Delta_k] \cos 2\pi j f_n - \sin 2\pi j f_n\} b_j + (\Delta - \Delta_k)S_k(f_n) \geqq D(f_n) - \Delta_k, \tag{10}$$

where $b_0 \equiv 1$ is the normalization made. We have in (9) and (10) an over-determined set of $2M$ equations in $N + 1$ variables. The objective is to minimize Δ, one of the variables. It would seem that the condition $S(f_n) > 0$ would be necessary to solve (9) and (10). But the phe-

* An extension into a weighted criterion can be handled, but is suppressed in this presentation. M was chosen to be in the range $4N$ (N large) $\leqq M \leqq 10N$ (N small) in our implementation of the algorithm.

† Therefore, the nonlinear nature of the tangent transformation may inhibit designs in the neighborhood of π.

nomenon experienced in Ref. 10 occurs here also. That is, if $S_k(f_n) > 0$, $0 \leqq n \leqq M$, then $S_{k+1}(f_n) > 0$, $0 \leqq n \leqq M$, also.

Standard linear programming techniques can now be used on (9) and (10) to iterate toward a minimum Δ. However, no restriction has been made on the locations of the zeros of $Q(z^{-1})$. Now there exist sufficient conditions for stability that can be written as linear constraints. We have looked at two of these, e.g., a restriction that $b_1, b_2, \cdots, b_N$ of (1) form a monotonic sequence[14] or the restriction that the sum $\sum_{k=0}^{N} b_k \cos 2\pi kf > 0$, $\forall f \in [0, \frac{1}{2}]$.[15] (The formulation of the linear programming problem gives us this condition on the subset of $[0, \frac{1}{2}]$ over which we are approximating.) For an example of a filter designed using this technique and the latter constraint to assure stability, refer to Fig. 2. Curve B is the sixth-order approximation to Curve A (only approximated over $[0.075, 0.425]$).

However, the filter designer may decide that these types of constraints are too restrictive for his particular applications. Nonlinear stability constraints, such as those found in Ref. 14, Chapter 3, can be included via the cutting planes algorithm,[16] but this may require excessive computation times. Another suggestion involves interrupting the standard simplex method for solving the linear programming problem after each iteration. We may then further constrain the b vector used in the next basic feasible solution to a choice (i.e., some

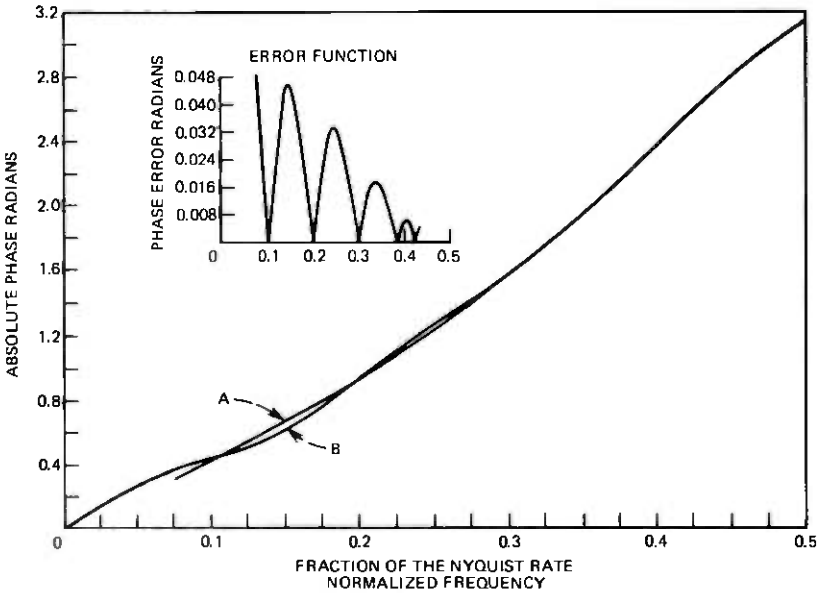

Fig. 2—Sixth-order approximation using linear programming method.

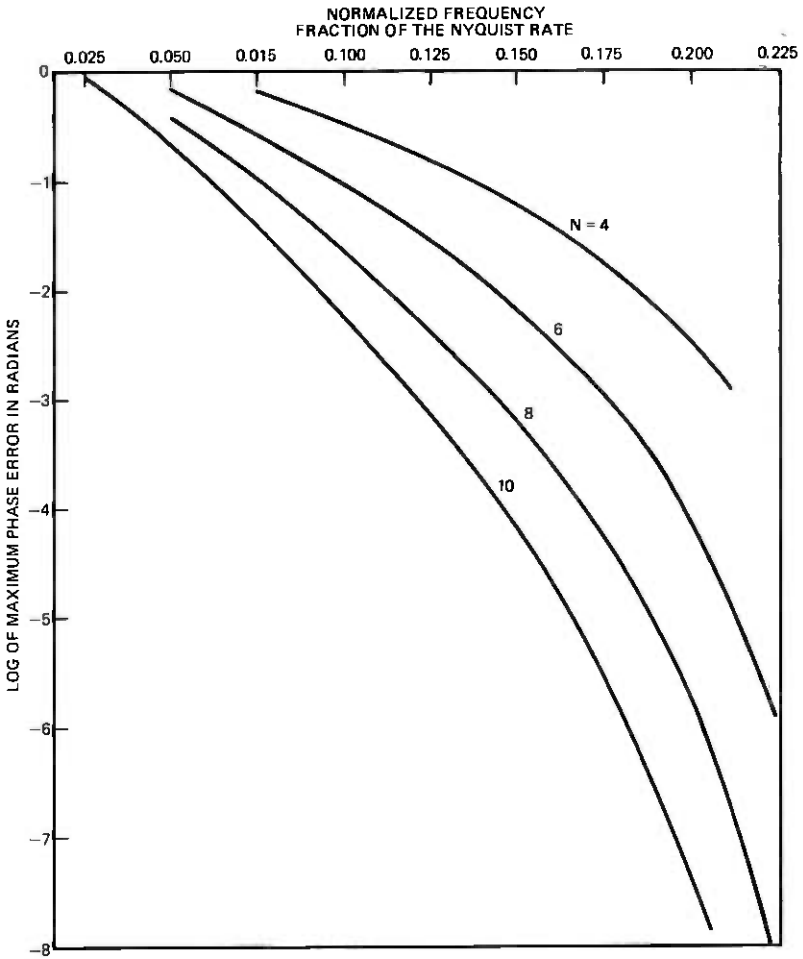

Fig. 3—Phase error vs bandwidth for various orders of Hilbert transformers.

"maximum") from among those vectors that would result in a stable filter in addition to the normal improvement of an object function.

Using the standard formulation of the problem with no additional constraints or techniques necessary to assure stability, we were able to design many Hilbert transformer filters.* Figure 3 shows the relationship between the maximum error (recall that the tangent of the desired function is approximated) and a bandwidth (the filters were

* FIR designs of Hilbert transformers are well documented (see Ref. 17). There, 90-degree phase is guaranteed, and the magnitude of 1.0 is approximated.

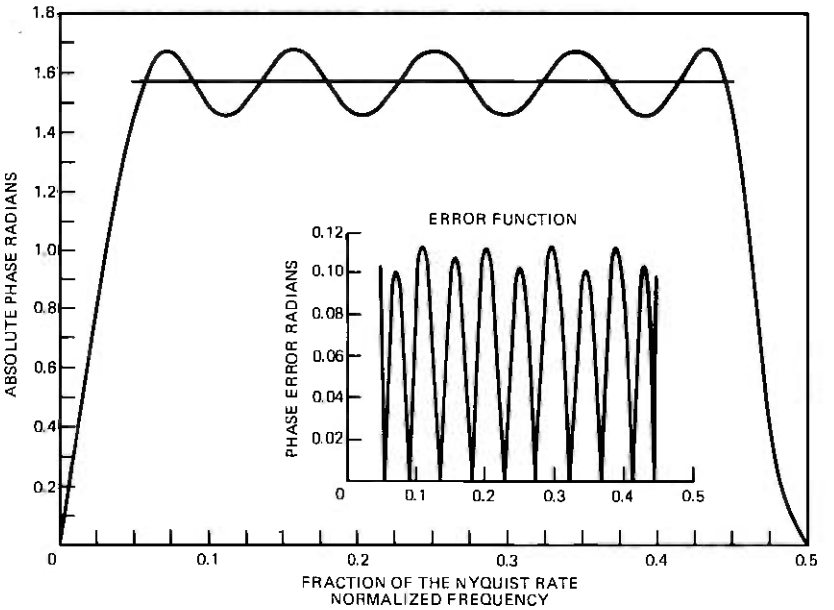

Fig. 4—Tenth-order Hilbert transformer.

designed* over $[f, 0.5 - f]$, $f = 0.075, 0.05, \cdots, 0.225$) for various orders of filters $N = 4, 6, 8, 10$. The log of the maximum error is given in the figure.

The minimax approximation formulated here is performed on the tangent of the desired phase and not on the desired phase itself. For very good approximations, however, no penalties seem apparent. We have briefly looked at methods to design minimax phase approximations based on the algorithm we have presented here. Our conclusions are that a two-stage design algorithm is required to iteratively locate a proper weight function that will "prewarp" the "tangent" design so that the weighted "tangent" design is minimax and the phase approximation is itself equiripple.

Figure 4 illustrates the effect of the tangent transformation in this design procedure. In this figure, we see the phase of the resultant design (and its error function). This is a 10th order approximation to a 90-degree phase shift over $[0.05, 0.45]$. While the design guarantees a minimax solution (equiripple) to the tangent, we can see that the resultant phase approximation is not exactly equiripple.† We have not

* Each design only required a few (e.g., 5) iterations.

† We can see from Fig. 2 that the effect that the tangent transformation has on the error curve also depends on the values of the desired function.

implemented an algorithm to find the minimax solution to the phase, since, for our needs, the improvement in the phase approximation from the method outlined here did not seem to justify the use of a modified algorithm.

III. A GRADIENT SEARCH TECHNIQUE FOR LEAST-SQUARES DESIGN

The next design algorithm we describe involves the computation of the gradient vector relative to the set of coefficients in a product of quadratic factors. The transfer function of an all-pass digital filter, expressed as a product of second-order sections, is:

$$H(z^{-1}) = \prod_{i=1}^{M} \left(\frac{\beta_i + \alpha_i z^{-1} + z^{-2}}{1 + \alpha_i z^{-1} + \beta_i z^{-2}} \right). \tag{11}$$

The least-squares form

$$E = \sum_{k=1}^{L} [D(f_k) - \text{Ang } H(e^{-j2\pi f_k})]^2 w(f_k) \tag{12}$$

will be used as a measure of the approximation error from the desired function $D(f)$ on the set of frequency samples $\{f_k\}_1^L$. Here, $w(f)$ denotes a nonnegative weighting function. A. G. Deczky has also considered gradient techniques applied to the least-square design of all-pass digital filters.[7] In that paper, the emphasis was on envelope delay design. However, as shown in Section I, there are applications where envelope delay approximations are not adequate. Specifically, there are cases where phase distortion (e.g., constant phase offset) must be eliminated with an all-pass structure.

Our design algorithm stems from familiarity with Ref. 8, which considers magnitude-only designs. With the least-squares criterion, the cascade second-order section form can be used. The advantage is that coefficient accuracy problems are minimized. As an alternative to the linear programming approach considered in Section II, this least-squares approach also enables one to more easily control the linear-phase offset permitted in the design. However, a disadvantage of the least-squares approach is that stability of the designed filter cannot be handled directly. Stability is obtained by confining the gradient movement to within the unit circle. This constraint may increase the likelihood of reaching an unsatisfactory local optimum. As we see later, there are initial guess procedures that provide excellent approximations to the desired phase function which, through the design algorithm, increase the likelihood of reaching a satisfactory local optimum.

3.1 Gradient calculations

We find the entries of the gradient vector are

$$\frac{\partial E}{\partial \alpha_i} = -2\sum_{k=1}^{L}[D(f_k) - \text{Ang } H(e^{-j2\pi fk})]w(f_k)\frac{\partial \text{ Ang } H(e^{-j2\pi fk})}{\partial \alpha_i} \quad (13)$$

$$\frac{\partial E}{\partial \beta_i} = -2\sum_{k=1}^{L}[D(f_k) - \text{Ang } H(e^{j2\pi fk})]w(f_k)\frac{\partial \text{ Ang } H(e^{-j2\pi fk})}{\partial \beta_i}. \quad (14)$$

Here we define $\phi(f) = \text{Ang } H(e^{-j2\pi f}) = \tan^{-1}I(f)/R(f)$, where $I(f) = \text{Imag } H(e^{-j2\pi f})$ and $R(f) = \text{Real } H(e^{-j2\pi f})$. We seek

$$\frac{\partial \phi(f)}{\partial \alpha} = R(f)I'_\alpha(f) - I(f)R'_\alpha(f)$$

$$\frac{\partial \phi(f)}{\partial \beta} = R(f)I'_\beta(f) - I(f)R'_\beta(f),$$

where prime (′) denotes the partial derivative relative to the subscript. After some algebra, we find

$$\frac{\partial \phi}{\partial \alpha_i} = 2(1 - \beta_i)F_i(f)\sin 2\pi f \qquad i \leqq i \leqq M \quad (15)$$

$$\frac{\partial \phi}{\partial \beta_i} = 2F_i(f)(\sin 4\pi f + \alpha_i \sin 2\pi f) \qquad 1 \leqq i \leqq M, \quad (16)$$

where $F_i(f) = |1 + \alpha_i e^{-j2\pi f} + \beta_i e^{-j4\pi f}|^{-2}$. Finally, (13) and (14) can be simplified for $1 \leqq i \leqq M$ to

$$\frac{\partial E}{\partial \alpha_i} = -4(1 - \beta_i)\sum_k [D(f_k) - \phi(f_k)]F_i(f_k)w(f_k)\sin 2\pi f_k \quad (17)$$

$$\frac{\partial E}{\partial \beta_i} = -4\sum_k [D(f_k) - \phi(f_k)]F_i(f_k)w(f_k)(\sin 4\pi f_k + \alpha_i \sin 2\pi f_k). \quad (18)$$

The minimization of E in (12) then proceeds with an iterative algorithm that is based on the formula

$$\mathbf{c}^{(n)} = \mathbf{c}^{(n-1)} - \epsilon_{n-1}A_n(\nabla E)_{n-1}, \quad (19)$$

where $\mathbf{c}^{(n)}$ is the coefficient vector $(\alpha_1, \beta_1, \alpha_2, \beta_2, \cdots, \alpha_M, \beta_M)$ at the nth iteration, ϵ_n is the nth step size in the coefficient adjustment, A_n is a positive definite matrix at the nth step ($\equiv I$ in the case of the steepest descent algorithm) and $(\nabla E)_n$ is the gradient vector whose entries are given by (17), (18) at the nth iteration (we use the Fletcher-

Powell algorithm). An initial guess procedure is required to start an iterative algorithm such as that of (19).

3.2 Initial guess procedures for all-pass networks

Convergence to a local minimum at which the approximation to a desired phase function is satisfactory can be made easier if a good initial guess is provided to $\mathbf{c}^{(0)}$ of (19). A desired feature of an initial guess procedure is that it be simple in nature. After all, excessive computation and effort should not be expected in simply starting a complex algorithm. In this section, we consider two procedures in which only a linear set of equations need be solved to obtain initial values for $\{b_k\}_{k=0}^{N}$ of (1). The value of having several initial guess procedures is that the designer may want to exercise his algorithm from multiple starting points to choose the best from a set of local optima. The following initial guess procedures operate on the direct form of $H(z^{-1})$(1) which can be factored to the cascade form (11).

3.2.1 Tangent approximation by Gauss' method

From (4) in Section II we know that a desired phase function can be approximated by considering a monotonic function of the phase, namely the tangent. Hence,

$$\tan \phi(f) = -\frac{\sum_{k=1}^{N} b_k \sin 2\pi kf}{\sum_{k=0}^{N} b_k \cos 2\pi kf} \tag{20}$$

is the approximating function of the tangent of half the desired phase. If we require the estimates of the desired phase tangent $[\tan \phi_d(f)]$ to be "good" at a number of frequencies, we then have the following equations:

$$\begin{aligned}
\tan \phi_d(f_0) \sum_{k=0}^{N} b_k \cos 2\pi f_0 - \sum_{k=1}^{N} b_k \sin 2\pi k f_0 &= r_0 \\
\tan \phi_d(f_1) \sum_{k=0}^{N} b_k \cos 2\pi k f_1 - \sum_{k=1}^{N} b_k \sin 2\pi k f_1 &= r_1 \\
&\vdots \\
\tan \phi_d(f_M) \sum_{k=0}^{N} b_k \cos 2\pi k f_M - \sum_{k=1}^{N} b_k \sin 2\pi k f_M &= r_M.
\end{aligned} \tag{21}$$

If $\{r_n\}_0^M$ were all zero, then the approximation would be exact. The objective then is to minimize $\sum_{n=0}^{M} r_n^2$, where $M > N$. This problem is a least-squares minimization problem for which the solution is

derived from solving a set of normal equations:

$$\begin{bmatrix} (\mathbf{a}_1, \mathbf{a}_1)(\mathbf{a}_1, \mathbf{a}_2)\cdots(\mathbf{a}_1, \mathbf{a}_N) \\ (\mathbf{a}_2, \mathbf{a}_1)(\mathbf{a}_2, \mathbf{a}_2)\cdots(\mathbf{a}_2, \mathbf{a}_N) \\ \vdots \qquad\qquad\qquad \vdots \\ (\mathbf{a}_N, \mathbf{a}_1) \qquad\qquad (\mathbf{a}_N, \mathbf{a}_N) \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \end{bmatrix} = \begin{bmatrix} (\mathbf{a}_1, \mathbf{d}) \\ (\mathbf{a}_2, \mathbf{d}) \\ \vdots \\ (\mathbf{a}_N, \mathbf{d}) \end{bmatrix} \tag{22}$$

or

$$A\mathbf{b} = \mathbf{e},$$

where

$$\begin{aligned} \mathbf{a}_n = (&\tan \phi_d(f_0) \cos 2\pi n f_0 - \sin 2\pi n f_0, \\ &\cdot \tan \phi_d(f_1) \cos 2\pi n f_1 - \sin 2\pi n f_1, \cdots, \\ &\cdot \tan \phi_d(f_M) \cos 2\pi n f_M - \sin 2\pi n f_M) \quad n = 1, 2, \cdots, N \end{aligned}$$

and

$$\mathbf{d} = [-\tan \phi_d(f_0), -\tan \phi_d(f_1), \cdots, -\tan \phi_d(f_M)]$$
$$\mathbf{b} = (b_1, b_2, \cdots, b_N) \text{ and } \mathbf{e} = [(\mathbf{a}_1, \mathbf{d}), \cdots, (\mathbf{a}_N, \mathbf{d})].$$

Let

$$\rho_{\max} = \max_{0 \leq n \leq M} \{|r_n^*|\} \quad \text{and} \quad \bar{\rho} = \sqrt{\frac{(\mathbf{r}^*, \mathbf{r}^*)}{M}},$$

where $r^* = (r_0^*, r_1^*, \cdots, r_M^*)$, the residual values after the least-squares approximation. If $\rho_{\max} - \bar{\rho}$ is large (it is always positive), then a Chebyshev approximation may be desirable.[18]

3.2.2 *Tangent approximation in Chebyshev sense*

It is well known that the minimax solution to (21) requires solving an appropriate subsystem of $N + 1$ equations. Further, the minimax solution of $N + 1$ inconsistent equations can be effected by examining the least-squares solution to the same set of equations and proceeding to solve a set of N linear equations.[19]

For our purposes here, an effective method of obtaining an initial guess for the iterative procedure implied by (19) is that of choosing $M = N$ and obtaining the minimax solution to (21). This can be done by solving (22) for $\mathbf{b} = (b_1, b_2, \cdots, b_N)$ and then evaluating (21) for the residuals $r_0^*, r_1^*, \cdots, r_N^*$. The minimax solution to (21) is then given by the linear set of equations

$$B\mathbf{b} = \boldsymbol{\sigma}, \tag{23}$$

where $B = (b_{jk})$, $N + 1$ by N matrix with $b_{jk} = \tan \phi_d(f_j) \cos 2\pi k f_j - \sin 2\pi k f_j$, $\boldsymbol{\sigma} = \epsilon[\text{sign } (r_0), \text{sign } (r_1), \cdots, \text{sign } (r_N)]$, and

$$\epsilon = \sum_{k=0}^{N} r_k^{*2} / \sum_{k=0}^{N} |r_k^*|.$$

It may be noted that only N of $N + 1$ equations are used in the solution of (23).

3.2.3 Discussion

It should be noted that no constraint has been made on the initial guess procedures of A or B to ensure that the resulting digital filter is stable. In fact, if $\sum_{k=0}^{N} b_k \cos k2\pi f$ should ever change sign in $|f| \leqq \frac{1}{2}$ or at least in the subinterval of approximation $[f_0, f_M]$, then the transition from (20) to (21) is not really valid since a division by zero is implied. Should $\sum_{k=0}^{N} b_k \cos k2\pi f$ be strictly positive over $|f| \leqq \frac{1}{2}$, then stability results.[5] (The interesting point is that stability can result even if the cosine series does change sign in $|f| \leqq \frac{1}{2}$). However, the point to remember is that the resulting initial guess may be unstable. In our experience, we have not encountered any serious problems using these initial guess procedures.

We must further remark that the inherent N sample delay present in these approximations [see (2)] could present a problem when designing filters with $M \neq N$ sample delays. However, we feel, intuitively, that since some delay is unavoidable, a delay of the order of the filter will not, for most applications, be overly restrictive.

The last point to consider is that the initial guess procedure of Sections 3.2.1 and 3.2.2 obtains a direct form estimate of the digital filter coefficients. What is really required for $\mathbf{c}^{(0)}$ of (19) are quadratic factors. We remark that we make the transition from the direct form estimate of (20) to quadratic factors by using a Bairstow quadratic factorization routine.

3.3 Some considerations for least-squares design

Often the engineering systems requirement of a digital filter can tolerate a linear-phase offset. While the systems engineer cannot always adapt to an arbitrary delay, there will usually be a range of delays permissible to him. How then can a designer incorporate these relaxations into the design mechanism? One technique for doing this is to add an acceptable delay to the desired function to create a new desired function and proceed from there. By designing filters for each of the permissible delays, one can choose from among the delays and their associated errors to decide which filter to implement.

In Fig. 5 we can see the error function of a sixth-order filter* (B)

* We have not tested the limit of the order of filters that can be designed by this method, but we have obtained a twentieth-order approximation (20-sample delay) to the desired function in this example. Quality initial guess procedures help us do this without excessive computation times.

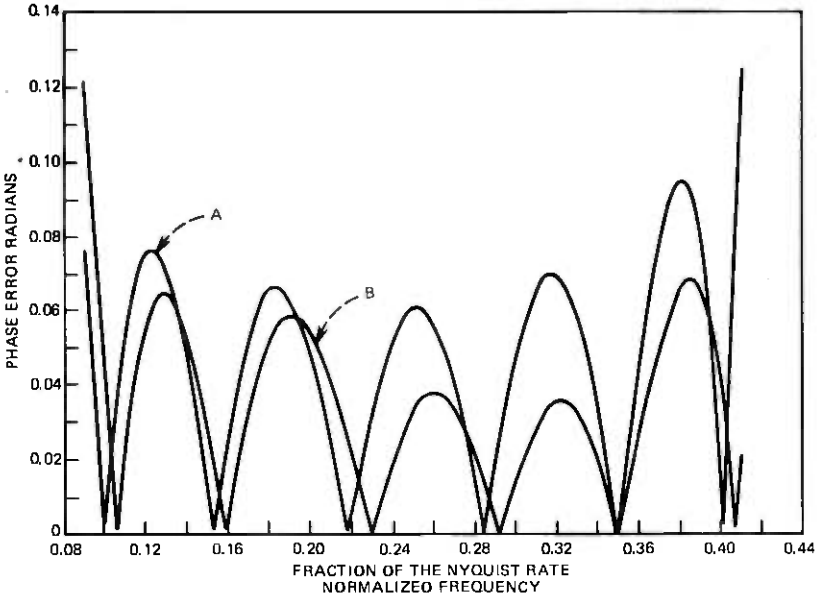

Fig. 5—Error curves for initial and final sixth-order Hilbert transformer designs.

designed for a delay of six samples. The desired function is a 90-degree phase shifting filter with the approximation having weight 1 on [0.08, 0.41] and 0 otherwise. Note the quality of the initial approximation (A) using the first initial guess technique outlined in Section 3.2. Of course, the disadvantage of presetting the delay is obvious; the choice of the optimal delay from those that are acceptable is not automatic but requires a separate design for each delay. However, eq. (12) can be expanded to include delay as parameter A

$$E = \sum_k [D(f_k) - \text{Ang } H(e^{-j2\pi f_k}) - A2\pi f_k]^2 w(f_k).$$

An optimal A can be found analytically at each step in the gradient search and at convergence A will represent the amount of delay which, in conjunction with the filter, produces the best design.* Of course, we cannot expect that this delay will represent an integral number of samples or even a delay that the designer can tolerate. Figure 6 shows a desired function (A) (this curve is only shown where the weight of the approximation is nonzero) and its fourth-order approximation (B),

* It is possible to include a constant phase angle as parameter B similar to the A used here. In such a case, our procedure becomes an envelope delay design technique.

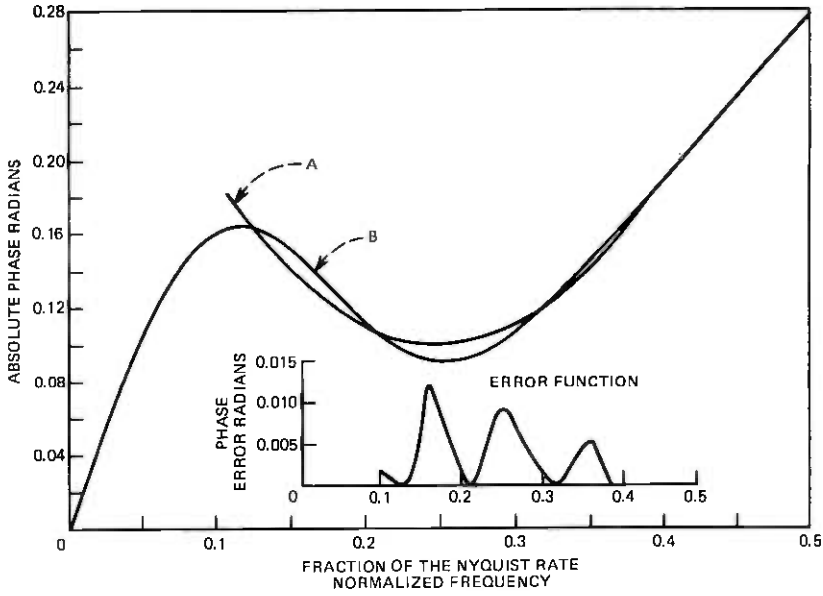

Fig. 6—Fourth-order approximation to desired shape (A).

which is by solving for optimal A. Allowing for arbitrary delay, the algorithm obtained this optimal design with a 4.1-sample delay.

We can offer an heuristic solution to guarantee integer delays in an automatic fashion; namely, at each step (that is, at each calculation of A), the nearest acceptable delay* is used to replace A in the algorithm. This, of course, places a serious strain on optimality, although it does permit an automatic design procedure.

As a footnote to this algorithm, we remark that there is a tendency, when working with procedures for designing filters in the cascade form, to use a previous optimal design of order n as the initial starting point in the design of filters of order $n + 2$. In the case of magnitude-only design, this is easily implemented since the appended second-order section can be initialized with magnitude 1. However, in the all-pass presentation there does not exist any second-order section that can be added which does not distort the overall phase when using a previous optimal design of order n to provide the initial guess for a design of order $n + 2$. And so the user of this algorithm must consider the effect of the appended second-order section if he does not want to obviate the value of a previous design toward providing an initial guess.

* Nearest in the sense of greatest reduction of (12); "acceptable" here means "integer."

REFERENCES

1. G. D. Hornbuckle and E. I. Ancona, "The LX-1 Microprocessor and Its Applications to Real Time Signal Processing," IEEE Transactions on Computers, *C-19*, August 1970.
2. B. Gold, I. L. Lebow, P. G. McHugh, and C. M. Rader, "The FDP, a Fast Programmable Signal Processor," IEEE Transactions on Computers, *C-20*, January 1971.
3. R. W. Lucky, J. Salz, and E. J. Weldon, Jr., *Principles of Data Communication*, New York: McGraw-Hill, 1968, pp. 190–192.
4. F. D. Sunde, *Communication Systems Engineering Theory*, New York: John Wiley, 1969, Chaps. 6–7, notably, pp. 244–246, pp. 229–232, and Figs. 6.4 and 6.6.
5. E. A. Robinson, *Statistical Communication and Detection*, New York: Hafner, 1967, pp. 325–326.
6. L. V. Blake, *Antennas*, New York: John Wiley, 1965, pp. 250–252.
7. A. G. Deczky, "Synthesis of Recursive Digital Filters Using the Minimum p-Error Criterion," IEEE Trans. on Audio and Elect., October 1972.
8. K. Steiglitz, "Computer-Aided Design of Recursive Digital Filters," IEEE Transactions on Audio and Electroacoustics, *AU-18*, June 1970.
9. R. E. King and G. W. Condon, "Frequency Domain Synthesis of a Class of Optimum Recursive Digital Filters," Int. J. on Control, 1973, No. 3.
10. I. Barrodale, M. J. D. Powell, and F. D. K. Roberts, "The Differential Correction Algorithm for Rational L_∞-Approximation," SIAM J. of Num. Math., *9*, September 1972.
11. H. L. Loeb, "Algorithms for Chebyshev Approximations Using the Ratios of Linear Forms," J. Soc. of Ind. Appl. Math., September 1960.
12. F. Brophy and A. Salazar, "Synthesis of Spectrum Shaping Digital Filters of Recursive Design," IEEE Trans. on Circuits and Systems, March 1975.
13. L. R. Rabiner, N. Y. Graham, and H. D. Helms, "Linear Programming Design of IIR Digital Filters with Arbitrary Magnitude Function," IEEE Trans. on Acoustics, Speech, Signal Proc., April 1974.
14. E. I. Jury, *Theory and Applications of the z-transform Method*, New York: John Wiley, 1964, p. 116.
15. Ref. 5, p. 188.
16. B. S. Gottfried and J. Weisman, *Introduction to Optimization Theory*, Englewood Cliffs, N. J.: Prentice-Hall, 1973, p. 255.
17. L. R. Rabiner and R. W. Schafer, "On the Behavior of Minimax FIR Digital Hilbert Transformers," B.S.T.J., *53*, No. 2 (February 1974), pp. 363–390.
18. E. L. Stiefel, *An Introduction to Numerical Mathematics*, New York: Academic Press, 1963, pp. 54–55.
19. E. W. Cheney, *Introduction to Approximation Theory*, New York: McGraw-Hill, 1966, pp. 36–41.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 4, April 1975
Printed in U.S.A.

The Effect of Small Phase Errors Upon Transmission Between Confocal Apertures

By I. ANDERSON

(Manuscript received November 14, 1974)

The effect of small, periodic phase errors upon transmission between two coaxial, circularly symmetric apertures is considered when the aperture phase distributions are confocal and the amplitude distributions are gaussian. The results are applicable to loss calculations in beam waveguide systems with imperfect lenses. When the periods of the phase errors are less than one-half the aperture radii, the total loss is approximately $\frac{1}{2}(\beta_1^2 + \beta_2^2)$, *where* β_1, β_2 *are the peak phase errors (in radians) on the apertures. Phase errors with periods greater than the aperture diameters are found to cause comparatively little transmission loss.*

I. INTRODUCTION

The use of beam waveguide[1] systems for the transmission of information,[2] or for the transmission of power,[3] necessitates the design of lenses (or cylindrical reflectors[4]) as focusing elements. In the design of these elements, it is desirable to estimate the degradation in performance caused by surface profile errors. Such degradation results in transmission loss and, in a communications system, will contribute to interference. Typically, the profile errors are associated with machining operations and, for lenses with circular symmetry, these errors are frequently circularly symmetric. The principal effect of the errors is to impart small, circularly symmetric phase perturbations to the field distribution adjacent to the lenses. The purpose of this paper is to calculate the reduction in transmission, caused by phase errors of this type, in a simple system comprising two coaxial, circular apertures as shown in Fig. 1. The field distributions on the apertures may represent the fields in the aperture planes of two antennas or the fields on adjacent lenses in a beam waveguide system.

In the absence of phase errors the transmission between coaxial apertures has been extensively studied by Kay,[5] Borgiotti,[6] Heurtley,[7] and others, with the principal objective of determining that aperture

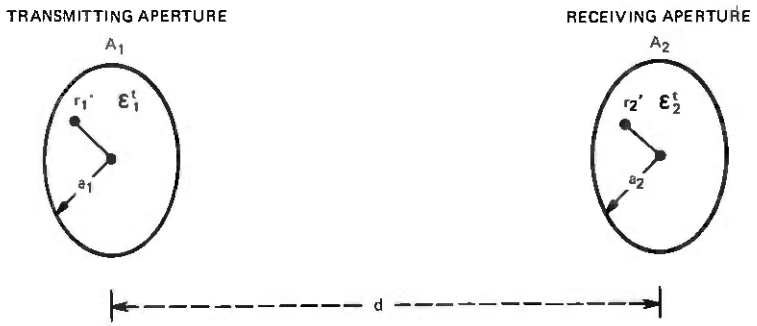

Fig. 1—Coaxial, circular apertures with confocal field distributions $\mathcal{E}_1^t(r_1')$ and $\mathcal{E}_2^t(r_2')$.

field distribution which maximizes the transmission. When the aperture separation is greater than some minimum, it has been found that this field distribution corresponds to that of the lowest order mode in an open, confocal resonator.[8] The phase distribution appropriate to this mode is obtained when the phase fronts on the apertures are confocal, i.e., are spherical with the center of curvature at the center of the other aperture. The appropriate amplitude distribution is well approximated[9] by a gaussian curve. For this (optimum) distribution, the transmission between the apertures can attain surprisingly high values even when the apertures are separated by many aperture diameters (see Refs. 5, 6, and 7). Although the effect of periodic and random phase errors upon antenna gain and side lobe level has been investigated by several authors,[10,11] little information appears to be available regarding transmission between two apertures when each has phase errors. In the case of transmission between two reflector antennas, Chu[12] has obtained an upper bound for the loss resulting from those phase errors produced by displacement of the feeds from the reflector foci. Yoneyama and Nishida[13] have considered transmission through a two-dimensional, confocal beam waveguide consisting of lenses with random phase errors. We compare their results to those of the present study later in this paper.

In the following section the total transmission loss in a confocal system, with small phase errors on apertures with arbitrary amplitude distributions, is expressed in terms of the losses associated with each aperture when the other is free from phase errors. In the next section explicit expressions are derived, for two cases of practical interest, when the phase errors are sinusoidal and when the aperture amplitude distributions are gaussian. These expressions are then discussed and compared with results from the literature.

II. TRANSMISSION BETWEEN CONFOCAL APERTURES WITH PHASE ERRORS

Consider the circular apertures A_1, A_2 of radius a_1, a_2 which are separated by a distance $d > a_1, a_2$, as in Fig. 1. It is assumed that the apertures are focused at each other such that the tangential electric fields in the apertures, when each is transmitting in the absence of the other, have the quadratic phase variation

$$\mathcal{E}_i'(r_i') = E_i(r_i') \exp\left(j \frac{kr_i'^2}{2d}\right), \qquad i = 1, 2, \tag{1}$$

where the r_i' are radial coordinates in the apertures. In the absence of phase errors the $E_i(r_i')$ are real. If interaction is neglected, the transmission between the apertures is readily found[7,12] from the results of Hu[14] and Kay[5]:

$$T = \frac{1}{D}\left|\int_0^1 \int_0^1 F_{12} dr_1 dr_2\right|^2, \tag{2}$$

where

$$F_{12} = E_1(r_1)E_2(r_2)J_0(nr_1r_2)r_1r_2 \tag{3}$$

and

$$D = \frac{1}{n^2}\left[\int_0^1 |E_1(r_1)|^2 r_1 dr_1 \int_0^1 |E_2(r_2)|^2 r_2 dr_2\right]. \tag{4}$$

In these expressions the r_i are normalized so that $r_i = r_i'/a_i$. The Fresnel number $n = ka_1a_2/d$, with k the wave number, and J_0 is the Bessel function of order zero. In the special case when the aperture separation is much greater than the aperture diameters, we see that $n \ll 1$ and that the aperture phases are uniform, i.e., $\mathcal{E}_i'(r_i) = E_i(r_i)$. Substituting the small argument approximation for the Bessel function, $x \ll 1$, $J_0(x) \approx 1$, eq. (2) then reduces to the familiar Friis transmission formula[15]

$$T = \frac{A_1^e A_2^e}{(\lambda d)^2}, \qquad n \ll 1. \tag{5}$$

The effective aperture areas A_i^e are defined by

$$A_i^e = 2\pi a_i^2 \frac{\left|\int_0^1 \mathcal{E}_i'(r) r dr\right|^2}{\int_0^1 |\mathcal{E}_i'(r)|^2 r dr}, \qquad i = 1, 2, \tag{6}$$

e.g., in the case of uniform illumination, $\mathcal{E}_i'(r) = 1$ and $A_i^e = \pi a_i^2$. The far-field transmission, T in (5), is also expressible in terms of the gains (G) of the apertures A_1 and A_2:

$$T = G_1G_2\left(\frac{\lambda}{4\pi d}\right)^2, \qquad n \ll 1, \tag{7}$$

where

$$G_i = \frac{4\pi A_i^2}{\lambda^2}, \qquad i = 1, 2. \tag{8}$$

Returning now to the discussion of phase errors, suppose that the phases in the apertures depart from the ideal (confocal) distributions by amounts $\phi_1(r_1)$ in A_1 and $\phi_2(r_2)$ in A_2. The transmission T_{12} between the apertures in the presence of these phase errors is, from (2),

$$T_{12} = \frac{1}{D}\left| \int_0^1 \int_0^1 F_{12} \exp\left[j(\phi_1 + \phi_2)\right] dr_1 dr_2 \right|^2, \tag{9}$$

where the $\phi_i(r_i)$ are abbreviated to ϕ_i. T_{12} is expressible as

$$T_{12} = T_0 - \Delta T_{12}, \tag{10}$$

where T_0 is the transmission in the absence of phase errors and ΔT_{12} is the loss resulting from the phase errors. Let

$$T_i = T_0 - \Delta T_i, \qquad i = 1, 2 \tag{11}$$

be the transmission between the apertures when there is a phase error on only the aperture A_i. From Appendix A we then find

$$\Delta T_{12} \approx \Delta T_1 + \Delta T_2 + R, \tag{12}$$

where

$$\Delta T_i = \frac{1}{D}\left[\int_0^1 \int_0^1 F_{12} dr_1 dr_2 \int_0^1 \int_0^1 F_{12}\phi_i^2 dr_1 dr_2 - \left\{ \int_0^1 \int_0^1 F_{12}\phi_i dr_1 dr_2 \right\}^2 \right], \quad i = 1, 2, \tag{13}$$

$$R = \frac{2}{D}\left[\int_0^1 \int_0^1 F_{12} dr_1 dr_2 \int_0^1 \int_0^1 F_{12}\phi_1\phi_2 dr_1 dr_2 - \int_0^1 \int_0^1 F_{12}\phi_1 dr_1 dr_2 \int_0^1 \int_0^1 F_{12}\phi_2 dr_1 dr_2 \right]. \tag{14}$$

Equation (12) states that the total loss incurred by (small) phase errors $\phi_1(r_1)$ and $\phi_2(r_2)$ on confocal apertures A_1, A_2 is approximately equal to the sum of the losses associated with each aperture when the other is free of phase errors together with the term R. In the next section we evaluate the expressions for ΔT_i and R when the aperture amplitude distributions are gaussian.

III. GAUSSIAN APERTURES

In the case of gaussian amplitude distributions, $E_i(r_i) = \exp(-\alpha_i r_i^2)$. To simplify the analysis we assume the α_i to be sufficiently large that the upper limits in the integrals may be extended to infinity. The

transmission T_0 in the absence of phase errors may now be obtained from (2), (3), and (4) by noting[16]

$$\int_0^\infty \exp(-\alpha_2 r_2^2) J_0(n r_1 r_2) r_2 dr_2 = \frac{1}{2\alpha_2} \exp\left(-\frac{n^2 r_1^2}{4\alpha_2}\right) \tag{15}$$

to give

$$T_0 = \frac{16 n^2 \alpha_1 \alpha_2}{(n^2 + 4\alpha_1\alpha_2)^2}, \qquad \exp(-\alpha_i) \ll 1, \qquad i = 1, 2. \tag{16}$$

Apart from differences in notation, this expression is identical to the corresponding result obtained by Kogelnik[17] for the coupling of (fundamental) gaussian modes. We find

$$T_0 = 1 \qquad \text{when} \qquad n = 2\sqrt{\alpha_1\alpha_2}, \tag{17}$$

i.e., within the approximation $\exp(-\alpha_i) \ll 1$, there are optimum amplitude distributions that will ensure complete power transfer between the apertures for a given n. A detailed analysis,[9] or numerical integration, indicates this to be a satisfactory approximation when $\alpha_i \gtrsim 2.3$, $i = 1, 2$. For example, when $\alpha_1 = \alpha_2 = 2.36$, the exact[9] results for identical apertures are $T_0 = 0.9931$, $n = 5.00$, and the approximate results are $T_0 = 1.00$, $n = 4.72$. From (6), the effective aperture area, A_i^e, of a gaussian aperture is

$$A_i^e = \frac{2\pi a_i^2}{\alpha_i}, \qquad i = 1, 2. \tag{18}$$

As expected from physical considerations A_i^e decreases as α_i increases, i.e., as the aperture field becomes more concentrated about the aperture center.

In the case of transmission with circularly symmetric, periodic phase errors, we take

$$\phi_i(r_i) = \beta_i \cos(\gamma_i r_i), \qquad i = 1, 2, \tag{19}$$

where

$$\beta_i = k\delta_i, \qquad \gamma_i = \frac{2\pi a_i}{l_i}.$$

β_i is the peak value of the error in radians (with δ_i the peak profile deviation) and l_i is the period of the error. For errors of period much greater than the circumference of the apertures, $\gamma_i \ll 1$ and then, in (9), $\phi_i(r_i) \approx \beta_i$, $i = 1, 2$. It follows, therefore, that $T_{12} = T_0$, i.e., to this order of approximation, small, slowly varying, circularly symmetric phase errors do not affect transmission between the apertures. In the general case of small errors, the transmission loss ΔT_{12} for gaussian apertures is found from (12) with (13) and (14) by

substituting for the $E_i(r_i)$ and $\phi_i(r_i)$. From Appendix B, we have

$$\frac{\Delta T_i}{T_0} = \beta_i^2 \gamma_i' [2\mathfrak{D}(\gamma_i'/2) - \mathfrak{D}(\gamma_i') - \gamma_i' \mathfrak{D}^2(\gamma_i'/2)], \qquad i = 1, 2, \quad (20)$$

where

$$\gamma_i' = \gamma_i \sqrt{\frac{4\alpha_j}{n^2 + 4\alpha_1\alpha_2}} \quad (21)$$

and

$$\mathfrak{D}(x) = \exp(-x^2) \int_0^x \exp(\tau^2) d\tau \quad (22)$$

is the (tabulated) Dawson integral.[18] The index $j = \{{2 \atop 1}\}$ when $i = \{{1 \atop 2}\}$.

The term R in (14) may be evaluated approximately in two cases of practical interest. In the first of these, the apertures are sufficiently far apart that $n \ll 1$ so that $J_0(nr_1r_2) \approx 1$ in (3). F_{12} is then separable in functions of r_1 and r_2 and hence, from (14), $R = 0$. For this case,

$$\Delta T_{12} \approx \Delta T_1 + \Delta T_2, \quad (23)$$

i.e., the total loss is approximately the sum of the losses associated with each aperture when the other aperture is free of phase errors. The total loss is given by (20) and (23) in which γ_i' simplifies to

$$\gamma_i' \approx \frac{\gamma_i}{\sqrt{\alpha_i}}, \qquad n \ll 1. \quad (24)$$

It is noted from (7) and (11) that

$$\frac{\Delta T_i}{T_0} = \frac{\Delta G_i'}{G_i}, \qquad n \ll 1, \quad i = 1, 2, \quad (25)$$

where $G_i' = G_i - \Delta G_i'$ is the gain of aperture A_i with the phase error ϕ_i. Hence, (20) with (24) gives the fractional reduction in gain of the aperture A_i resulting from a (small) periodic phase error.

The second case of practical interest arises when the amplitude distributions on the apertures are optimized in accordance with (17) such that, in the absence of phase errors, the transmission is unity. From Appendix C, the term R in (12) is negligible in this case provided $\gamma_1, \gamma_2 \gg n = 2\sqrt{\alpha_1\alpha_2}$, i.e., the periods ($l_i$) of the phase errors satisfy

$$l_1 \ll \frac{\lambda d}{a_2} \quad \text{and} \quad l_2 \ll \frac{\lambda d}{a_1}. \quad (26)$$

The transmission loss, ΔT_{12}, between the apertures is then the sum of the losses associated with each aperture as given by (23). This result

implies that the transmission through a sequence of confocal lenses, each with small phase errors of period satisfying (26), may be obtained by calculating the transmissions associated with each lens in the absence of phase errors on the other lenses. Furthermore, (23) indicates that it is not possible to compensate for such phase errors on one lens by introducing phase variations on an adjacent lens. When (26) is satisfied, the Dawson integrals in (20) may be replaced by the first terms of the asymptotic expansion (44) to give

$$\Delta T_i \approx \tfrac{1}{2}\beta_i^2, \qquad i = 1, 2. \tag{27}$$

It is of interest to note the physical significance of the condition (26) for the validity of the approximate forms (23) and (27). As expected from the theory of diffraction gratings, a circularly symmetric, periodic phase perturbation on a circular aperture generates[19] two additional side lobes in the aperture radiation pattern. If the period of the phase perturbation is l, then the two side lobes are symmetrically located about the main beam at an angle $\theta = \sin^{-1}(\lambda/l)$. Consider now the two apertures of Fig. 1 with phase errors of period l_1 in A_1 and l_2 in A_2. If the apertures are sufficiently far apart the side lobes, due to the phase error l_1 in A_1, will not intercept A_2 provided $\sin^{-1}(\lambda/l_1) \gg \tan^{-1}(a_2/d)$, i.e., for small angles, $l_1 \ll \lambda d/a_2$. Similarly, the main beam of A_1 will not couple energy to the side lobes of A_2 provided $l_2 \ll \lambda d/a_1$. The condition (26) implies, therefore, that energy is coupled from A_1 to A_2 via the main beams alone.

Figure 2 shows the transmission, as a function of $\gamma = \gamma_1 = \gamma_2$, between two identical apertures as obtained by numerical integration of (9) with (19), and as obtained from the approximate result (23) with (27). The upper curve in the figure applies for $\alpha = 4$, $n = 8$, $\beta = 0.36$ and the lower curve for $\alpha = 2.36$, $n = 5$, $\beta = 0.18$. In the absence of phase errors $T_0 = 1$ for these (optimum) distributions. The dashed lines correspond to the approximation (23) with (27). As anticipated earlier, the transmission is essentially unaffected by phase errors of large period, e.g., when $\gamma \lesssim n/2$, i.e., $l \gtrsim 2\lambda d/a$. The approximate form (23) with (27) is seen to be within about 1 percent of the exact result when $\gamma \gtrsim 2n$, i.e., $l \lesssim \lambda d/2a$. As an illustrative example, consider a beam waveguide system of the type described by Arnaud and Ruscio[2] with $\lambda = 3 \times 10^{-3}m$, $d \approx 80m$, $a \approx 0.5m$. The parameters of this system correspond to those of the lower curve in Fig. 2. Substitution shows that small, circularly symmetric phase errors of period $l \gtrsim 2a$ on the lenses will cause negligible transmission loss, and that the approximation (23) with (27) is applicable for phase errors of period $l \lesssim a/2$.

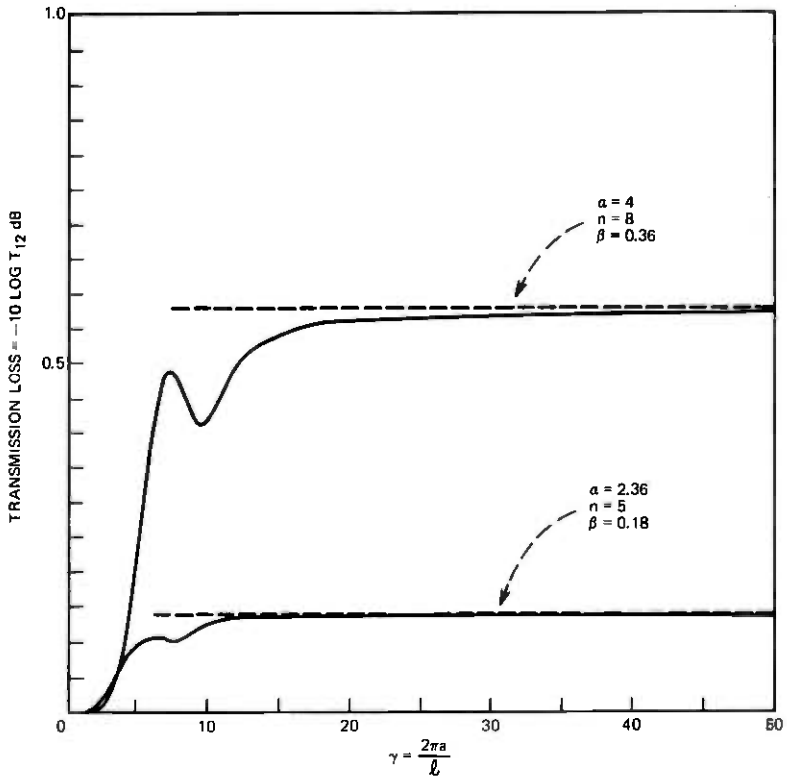

Fig. 2—Dependence on γ of phase error loss.

IV. COMPARISON WITH PREVIOUS RESULTS

To conclude this discussion on the effects of phase errors, we briefly compare the preceding results with the work of others. An expression for the gain (G') of an aperture with small, periodic phase errors was given in (25). Consider the special case in which the period of the phase error is much less than the dimensions of the effective aperture, i.e., $l \ll \sqrt{A^e}$. From (18) and (19), this implies $\gamma \gg \sqrt{\alpha}$ so that the Dawson integrals in (20) with (24) may be replaced by the large argument form (44) to give, with (25),

$$\frac{G'}{G} \approx 1 - \tfrac{1}{2}\beta^2, \tag{28}$$

where G is the gain in the absence of phase errors. Since this resnlt depends only upon the magnitude β of the phase error, it is anticipated that it may apply to random phase errors with correlation lengths that are small compared with the dimensions of the effective aperture. Ruze[11] has examined the reduction in aperture gain caused by such

random errors and we compare (28) with his results. In the particular case of a sinusoidal surface error of rms value ϵ on a parabolic reflector antenna, we have $\beta = k\delta = 2\sqrt{2}k\epsilon$. From (28), the gain with this small phase error is

$$\frac{G'}{G} \approx \exp\left[-\left(\frac{4\pi\epsilon}{\lambda}\right)^2\right]. \tag{29}$$

This expression, derived here for a sinusoidal phase error, is identical to that obtained by Ruze in the case of a random error. As noted in Section I, Yoneyama and Nishida[13] have examined the effect of random phase errors on lenses in a two-dimensional, confocal, beam waveguide system. Their approach is based on the concept of a statistical beam mode and this leads to a description of the field distribution, and transmission loss, in terms of an integral equation. A computer was used to solve the integral equation by numerical iteration from the solution in the absence of phase errors. It is interesting to find that the conclusions of their study, of a two-dimensional system with random errors, are similar to those obtained here for transmission between circular apertures with periodic phase errors. In particular, it was found that the transmission was not appreciably affected by phase errors with large correlation lengths and that the loss for a given error tended to a constant value for increasingly small correlation lengths.

V. CONCLUSIONS

We have examined the effect of small, periodic, radial phase errors upon transmission between two coaxial, circularly symmetric apertures with confocal phase distributions. Two cases of practical interest have been considered when the amplitude distributions on the apertures are gaussian. In the first of these the apertures are widely separated with phase errors of arbitrary period. The total loss is then the sum of the losses associated with each aperture and is given in terms of tabulated Dawson integrals. This result reduces to a known form when the periods of the phase errors are sufficiently small. The second case of interest applies to transmission through a beam waveguide system with imperfect lenses. When the periods (l_i) of the phase errors on the apertures satisfy $l_i \lesssim a_i/2$, ($i = 1, 2$), where a_i is the aperture radius, the total loss resulting from phase errors is approximately $\frac{1}{2}(\beta_1^2 + \beta_2^2)$, where β_1, β_2 are the peak phase errors in radians on the two apertures. A comparison, based on numerical integration, shows this to be within about 1 percent of the exact result in a typical case. Phase errors with periods $l_i \gtrsim 2a_i$ ($i = 1, 2$) have comparatively little effect upon transmission.

VI. ACKNOWLEDGMENTS

The author thanks S. J. Buchsbaum for suggesting this problem. Discussions with J. A. Arnaud and D. Gloge are greatly appreciated. The numerical integrations were kindly computed by D. Vitello.

APPENDIX A

Derivation of (12)

For small phase errors, $(\phi_1 + \phi_2) \ll 1$ and the exponential in (9) may be expanded to second order. Recalling that the E_i are real, i.e., F_{12} is real, we then find

$$T_{12} \approx \frac{1}{D}\left[\left\{\int_0^1 \int_0^1 F_{12}[1 - \tfrac{1}{2}(\phi_1 + \phi_2)^2]dr_1dr_2\right\}^2 + \left\{\int_0^1 \int_0^1 F_{12}(\phi_1 + \phi_2)dr_1dr_2\right\}^2\right]. \quad (30)$$

Expanding the first bracket and noting that

$$\frac{1}{D}\left\{\int_0^1 \int_0^1 F_{12}(\phi_1 + \phi_2)^2 dr_1dr_2\right\}^2 \leqq (\phi_1 + \phi_2)_{\max}^4 T_0, \quad (31)$$

where $(\phi_1 + \phi_2)_{\max}$ is the maximum value of $(\phi_1 + \phi_2)$, we have, to second order in ϕ_1, ϕ_2,

$$T_{12} = T_0 - \Delta T_{12}, \quad (32)$$

where

$$\Delta T_{12} \approx \frac{1}{D}\left[\int_0^1 \int_0^1 F_{12}dr_1dr_2 \int_0^1 \int_0^1 F_{12}(\phi_1 + \phi_2)^2 dr_1dr_2 - \left\{\int_0^1 \int_0^1 F_{12}(\phi_1 + \phi_2)dr_1dr_2\right\}^2\right]. \quad (33)$$

Expanding the brackets gives (12) with (13) and (14).

APPENDIX B

Derivation of (20)

Substituting $E_i(r_i) = \exp(-\alpha_i r_i^2)$, $\phi_i(r_i) = \beta_i \cos(\gamma_i r_i)$ $(i = 1, 2)$ into (13) with (3) and (4) gives, with (15),

$$\frac{\Delta T_i}{T_0} = 2\eta\beta_i^2[I_1(\gamma_i) - 2\eta I_2^2(\gamma_i)] \qquad i = 1, 2, \quad (34)$$

where

$$I_1(\gamma_i) = \int_0^\infty \exp(-\eta r^2)\cos^2(\gamma_i r)rdr, \quad (35)$$

$$I_2(\gamma_i) = \int_0^\infty \exp(-\eta r^2)\cos(\gamma_i r)rdr \quad (36)$$

and

$$\eta = \frac{1}{4\alpha_j}(n^2 + 4\alpha_1\alpha_2), \tag{37}$$

with $j = \{{}^{2}_{1}\}$ when $i = \{{}^{1}_{2}\}$. Expanding $\cos^2(\gamma_i r)$ and integrating:

$$I_1(\gamma_i) = \frac{1}{4\eta} + \tfrac{1}{2}I_2(2\gamma_i). \tag{38}$$

Integrating by parts,

$$I_2(\gamma_i) = \frac{1}{2\eta}\left[1 - \gamma_i \int_0^\infty \exp(-\eta r^2)\sin(\gamma_i r)dr\right], \tag{39}$$

which is expressible[18] in terms of the (tabulated) Dawson integral, i.e.,

$$I_2(\gamma_i) = \frac{1}{2\eta}\left[1 - \frac{\gamma_i}{\sqrt{\eta}}\,\mathfrak{D}\left(\frac{\gamma_i}{2\sqrt{\eta}}\right)\right], \tag{40}$$

where

$$\mathfrak{D}(x) = \exp(-x^2)\int_0^x \exp(\tau^2)d\tau. \tag{41}$$

From (34), (38), and (40), we then obtain (20) and from (19) and (37) we obtain (21).

APPENDIX C

Approximate Evaluation of R in (14)

We derive an approximate expression for R when $\gamma_1, \gamma_2 \gg n$. It is assumed that the amplitude distributions on the apertures are optimized such that $T_0 = 1$ with $n = 2\sqrt{\alpha_1\alpha_2}$, where $\alpha_1, \alpha_2 \geqq 2.3$. Consider the integrals in (14): Extending the integration limits to infinity and substituting $E_i(r_i) = \exp(-\alpha_i r_i^2)$, $i = 1, 2$ gives, with (3), (4), and (15),

$$\int_0^1\int_0^1 F_{12}dr_1dr_2 = \frac{1}{2n^2}; \qquad D = \frac{1}{4n^4}. \tag{42}$$

Similarly, substituting $\phi_i = \beta_i \cos(\gamma_i r_i)$ and using (15) and (40),

$$\int_0^1\int_0^1 F_{12}\phi_1 dr_1dr_2 = \frac{\beta_1}{2n^2}\left[1 - \frac{\gamma_1\sqrt{2\alpha_2}}{n}\,\mathfrak{D}\left(\frac{\gamma_1}{n}\sqrt{\frac{\alpha_2}{2}}\right)\right]. \tag{43}$$

Since $\gamma_1 \gg n$, the Dawson integral may be replaced by the first two terms of the asymptotic expansion,[20]

$$\mathfrak{D}(x) \sim \frac{1}{2x}\left[1 + \sum_{m=1}^{\infty}\frac{1.3\cdots(2m-1)}{(2x^2)^m}\right], \qquad x \gg 1, \tag{44}$$

to give

$$\int_0^1\int_0^1 F_{12}\phi_1 dr_1dr_2 \approx -\frac{\beta_1}{2\alpha_2\gamma_1^2}. \tag{45}$$

Further,

$$\int_0^1 \int_0^1 F_{12}\phi_1\phi_2 dr_1 dr_2 = \beta_1\beta_2 \int_0^1 \exp(-\alpha_1 r_1^2) r_1 \cos(\gamma_1 r_1) g(r_1) dr_1, \quad (46)$$

where

$$g(r_1) = \int_0^\infty \exp(-\alpha_2 r_2^2) J_0(n r_1 r_2) r_2 \cos(\gamma_2 r_2) dr_2. \quad (47)$$

Substituting the integral representation of the Bessel function

$$J_0(x) = \frac{1}{\pi} \int_0^\pi \cos(x \sin\theta) d\theta, \quad (48)$$

interchanging orders of integration and expanding the cosine product,

$$g(r_1) = \frac{1}{2\pi} \int_0^\pi [I(\theta) + I(-\theta)] d\theta, \quad (49)$$

where

$$I(\theta) = \int_0^\infty \exp(-\alpha_2 r_2^2) r_2 \cos[(\gamma_2 + n r_1 \sin\theta) r_2] dr_2. \quad (50)$$

From (40),

$$I(\theta) = \frac{1}{2\alpha_2}\left[1 - \frac{1}{\sqrt{\alpha_2}}(\gamma_2 + n r_1 \sin\theta) \cdot \mathfrak{D}\left\{\frac{1}{2\sqrt{\alpha_2}}(\gamma_2 + n r_1 \sin\theta)\right\}\right]. \quad (51)$$

Since $\gamma_2 \gg n$, both Dawson integrals in (49) may be replaced by the large argument form (44) to give

$$g(r_1) \approx I(\theta) \approx -\gamma_2^{-2}. \quad (52)$$

Evaluating (46) by (40) and using (44),

$$\int_0^1 \int_0^1 F_{12}\phi_1\phi_2 dr_1 dr_2 \approx \frac{\beta_1\beta_2}{\gamma_1^2\gamma_2^2}. \quad (53)$$

Substituting for the integrals in (14) and reducing (20) then gives

$$\frac{R}{\Delta T_1 + \Delta T_2} \approx -\frac{\beta_1\beta_2}{\beta_1^2 + \beta_2^2} \frac{8n^2}{\gamma_1^2\gamma_2^2}. \quad (54)$$

But $|\beta_1\beta_2/(\beta_1^2 + \beta_2^2)| \leqq \frac{1}{2}$, i.e.,

$$\left|\frac{R}{\Delta T_1 + \Delta T_2}\right| \leqq \left(\frac{2n}{\gamma_1\gamma_2}\right)^2. \quad (55)$$

Since $\gamma_1\gamma_2 \gg 2n$, (55) is much less than unity and so, from (12),

$$\Delta T_{12} \approx \Delta T_1 + \Delta T_2. \quad (56)$$

REFERENCES

1. G. Goubau and F. Schwering, "On the Guided Propagation of Electromagnetic Wave Beams," IRE Trans. on Antennas and Propagation, *AP-9*, May 1961, pp. 248–256.
2. J. A. Arnaud and J. T. Ruscio, "Guidance of 100 GHz Beams by Cylindrical Mirrors," to appear in IEEE Trans. on Microwave Theory and Techniques, April 1975.
3. E. C. Okress *et al.*, "Microwave Power Engineering," IEEE Spectrum, October 1964, pp. 76–100.
4. J. A. Arnaud and J. T. Ruscio, "Focusing and Deflection of Optical Beams by Cylindrical Mirrors," Appl. Opt., *9*, October 1970, pp. 2377–2380.
5. A. F. Kay, "Near-Field Gain of Aperture Antennas," IRE Trans. On Antennas and Propagation, *8*, November 1960, pp. 586–593.
6. G. V. Borgiotti, "Maximum Power Transfer Between Two Planar Apertures in the Fresnel Zone," IEEE Trans. on Antennas and Propagation, *AP-14*, No. 2 (March 1966), pp. 158–163.
7. J. C. Heurtley, "Maximum Power Transfer Between Finite Antennas," IEEE Trans. on Antennas and Propagation, *AP-15*, No. 2 (March 1967), pp. 298–300.
8. H. Kogelnik and T. Li, "Laser Beams and Resonators," Appl. Opt., *5*, October 1966, pp. 1550–1567.
9. S. Takeshita, "Power Transfer Efficiency Between Focused Circular Antennas with Gaussian Illumination in Fresnel Region," IEEE Trans. on Antennas and Propagation, *AP-16*, No. 3 (May 1968), pp. 305–309.
10. *Antenna Engineering Handbook*, H. Jasik, editor, New York: McGraw-Hill, 1961, p. 2–39.
11. J. Ruze, "Antenna Tolerance Theory—A Review," Proc. IEEE, *54*, No. 4 (April 1966), pp. 633–640.
12. T. S. Chu, "Maximum Power Transmission Between Two Reflector Antennas in the Fresnel Zone," B.S.T.J., *50*, No. 4 (April 1971), pp. 1407–1420.
13. T. Yoneyama and S. Nishida, "Effects of Random Surface Irregularities of Lenses on Wave Beam Transmission," Rep. Res. Inst. Elec. Comm., Tohoku Univ., *25*, No. 2, 1973, pp. 67–77.
14. M. K. Hu, "Near-Zone Power Transmission Formulas," IRE Nat. Conv. Record, *6*, pt. 8, 1958, pp. 128–138.
15. H. T. Friis, "A Note on a Simple Transmission Formula," Proc. IRE, *34*, May 1946, pp. 254–256.
16. G. N. Watson, *Theory of Bessel Functions*, Cambridge, England: Cambridge University Press, 1948, p. 393.
17. H. Kogelnik, "Coupling and Conversion Coefficients for Optical Modes," Polytechnic Institute of Brooklyn, Microwave Research Institute Symposia Series, *XIV*, 1964, pp. 333–347.
18. *Handbook of Mathematical Functions*, M. Abramowitz and I. A. Stegun, editors, New York: Dover, 1965, pp. 302, 319.
19. C. Dragone and D. C. Hogg, "Wide-Angle Radiation Due to Rough Phase Fronts," B.S.T.J., *42*, No. 5 (September 1963), pp. 2285–2296.
20. Reference 18, pp. 297, 298 by substituting equation (7.1.23) into (7.1.3).

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 55, No. 4, April 1975
Printed in U.S.A.

Toward a Group-Theoretic Proof of the Rearrangeability Theorem for Clos' Network

By V. E. BENEŠ

(Manuscript received October 8, 1974)

Methods from group theory and combinatorics are used to prove the (Slepian-Duguid) rearrangeability theorem for Clos' three-stage network. The nr-permutations realizable in such a network can be represented as a product $G\varphi^{-1}H\varphi G$, where G, H are subgroups realized by stages and φ is the special cross-connect field used in making frames. Thus, rearrangeability can be cast as $G\varphi^{-1}H\varphi G = S_{nr}$ = symmetric group of degree nr. Since it is an elementary theorem that a permutation group containing all transpositions is symmetric, it is enough to show that the product $G\varphi^{-1}H\varphi G$ is closed under multiplication and contains all transpositions. We prove that closure of the product is equivalent to a property of suitable partitions: existence of systems of common representatives. This property, formulated by J. B. Kruskal, is a consequence of Hall's theorem on distinct representatives. It is easily seen that $G\varphi^{-1}H\varphi G$ contains all transpositions, so the Slepian-Duguid theorem follows.

I. INTRODUCTION

In this paper we continue the exploration begun in previous work[1–3] of the relationships between permutation groups and connecting networks that are made of stages, frames, and cross-connect fields. Our results concern a well-known theoretical result of this area, the Slepian-Duguid theorem, which states that Clos' three-stage network with square switches is rearrangeable, i.e., realizes any permutation. Since the permutations realizable by a stage form a special kind of subgroup, the theorem has been viewed in terms of group theory as a factorization of the symmetric group S_{nr} of degree nr into a product of three subgroups or, alternatively, into a product of two mutually inverse double cosets.[3]

We further illuminate this basic rearrangeability theorem by giving it as nearly group-theoretic a proof as we have been able to find. This proof starts from the known characterization[1] of the nr-permutations

realizable by a Clos' three-stage network as a product $G\varphi^{-1}H\varphi G$, where G, H are subgroups realized by stages and φ is a "canonical" cross-connect field. It then shows that this product is closed under multiplication, and that it contains all nr-transpositions, whence immediately, by an elementary theorem, that it contains any nr-permutation, i.e., that $S_{nr} = G\varphi^{-1}H\varphi G$.

In the course of this proof we show that the basic combinatorial backbone of the rearrangeability theorem is really the existence of systems of common representatives (SCRs) for pairs of partitions. Since, in apparent contrast, Duguid's original proof[4] used Hall's theorem on systems of distinct representatives (SDRs) of subsets, we have also sought to clarify just how the rearrangeability result depends on Hall's theorem. The contrast above is apparent only because there are standard ways of proving SCR results from SDR results. In the present context, the two approaches are equivalent and lead to the same results. However, the SCR formulation is closer to the group-theoretic aspects than is Duguid's original SDR proof: it provides an SCR property that is a consequence of Hall's theorem and is necessary and sufficient for the product $G\varphi^{-1}H\varphi G$ to be closed. The property was first formulated by J. B. Kruskal in unpublished notes about rearrangeable networks dating from 1964.

II. SETTING AND FORMULATION

We now sketch the group-theoretic interpretation of the Slepian-Duguid theorem in some detail, as has been done in earlier work.[3] Figure 1 shows Clos' three-stage network, composed of three sym-

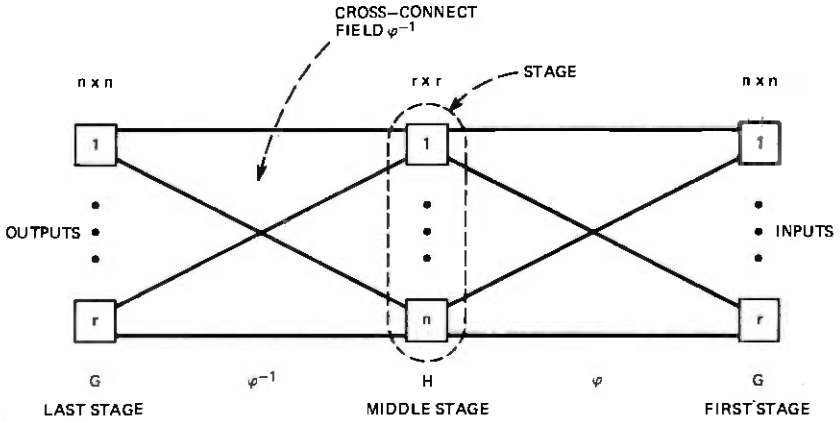

Fig. 1—$G\varphi^{-1}H\varphi G$ describes the permutations realizable by Clos' three-stage network.

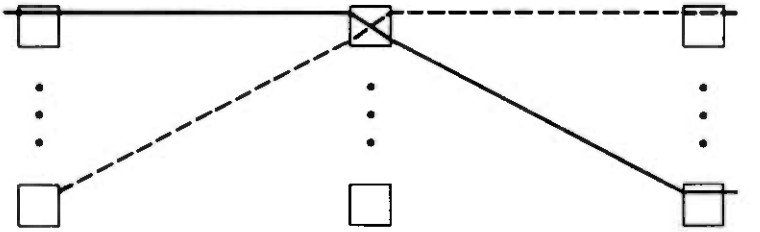

Fig. 2—Getting transpositions in $G\varphi^{-1}H\varphi G$: terminals on different outer switches.

metrically placed stages interconnected by the "canonical" cross-connect field φ and its inverse. Each stage can realize precisely those permutations from a certain subgroup of S_{nr}, depending on the size and number of switches in the stage. The r $n \times n$ switches of each outer stage realize a subgroup G isomorphic to $(S_n)^r$, viz., all those that permute the sets $\{kn + 1, kn + 2, \cdots, (k + 1)n\}$, $k = 0, \cdots, r - 1$, within themselves. A similar statement holds for the center stage, but with n and r interchanged, to define a subgroup H isomorphic to $(S_r)^n$.

Thus, if we think of the network in Fig. 1 as acting from right to left, and if we interpret composition of permutations as left-multiplication of the inner permutation by the outer, then the permutations realizable by Clos' three-stage network with square switches are precisely those in the complex

$$G\varphi^{-1}H\varphi G.$$

The Slepian-Duguid theorem says that this complex is exactly the symmetric group S_{nr} of degree nr. We note for future reference that all transpositions are realizable; this can be seen from Figs. 2 and 3, in which the remaining terminals (not shown) are connected through to "themselves," as is possible and indeed necessary to realize a transposition.

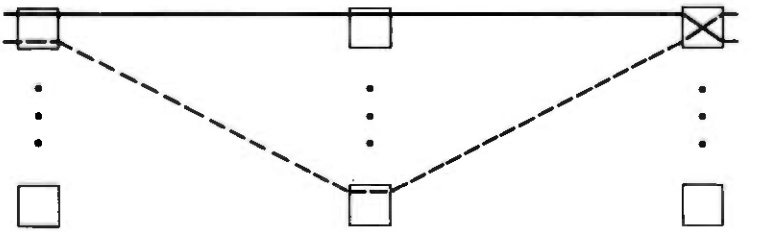

Fig. 3—Getting transpositions in $G\varphi^{-1}H\varphi G$: terminals on same outer switch.

III. SYSTEMS OF DISTINCT REPRESENTATIVES

Let X be a set, and $X_1, \cdots, X_m$ finite subsets of X. We make the following definition.

Definition 1: Elements $x_1, \cdots, x_m$ from X form a system of distinct representatives (SDR) *of $X_1, \cdots, X_m$ iff $x_i \in X_i$ and $x_i \neq x_j$ if $i \neq j$, for $i, j = 1, \cdots, m$.*

Hall's theorem[5] gives a necessary and sufficient condition for the X_i to have an SDR, thus:

Theorem 1 (Hall): $X_1, \cdots, X_m$ have an SDR *iff for $k = 1, \cdots, m$, the union of any k X_i has at least k elements.*

This result was used by Duguid in his proof of the rearrangeability of Clos' network with square switches. It enabled him to decompose any permutation into a union of submaps each of which, in switching terminology, carried exactly one terminal on each input switch onto images that were spread over all the output switches. These submaps could then be accommodated, one each on a middle switch.

IV. SYSTEMS OF COMMON REPRESENTATIVES

Let $P = \{P_i\}$ and $Q = \{Q_i\}$ be partitions of a set X with $|P| = |Q|$.

Definition 2: A subset $E \subset X$ is called a system of common representatives (SCR) *for P and Q iff*

$$\begin{aligned} |E \cap P_i| &= 1, \qquad P_i \in P \\ |E \cap Q_j| &= 1, \qquad Q_j \in Q. \end{aligned}$$

Ryser[6] gives an SDR argument to prove a necessary and sufficient condition for two partitions as above to have an SCR. In the cases of interest to us here, a sufficient condition can be given in a particularly simple way. We make

Definition 3: Q is an (r, n)-partition iff $|Q| = r$, and $|Q_i| = n$ for $Q_i \in Q$. An (r, n)-partition of X is one into r sets each having n elements.

We use substantially Ryser's argument[6] to prove the following special case (Theorem 2.2, p. 51, of Ref. 5) of his result:

Theorem 2: Let P, Q be (r, n)-partitions of X. Then P and Q have an SCR.

Proof: For $j = 1, \cdots, r$, let $A_j = \{i : P_i \text{ meets } Q_j\}$. Take any union of k of these sets, $A_{j_1} \cup \cdots \cup A_{j_k}$, and observe that $Q_{j_1} \cup \cdots \cup Q_{j_k}$ has precisely nk elements in it. Hence, at most $r - k$ integers in the range $1, \cdots, r$ fail to be in some $A_{j_1}, \cdots, A_{j_k}$. Thus,

$$|A_{j_1} \cup \cdots \cup A_{j_k}| \geqq k,$$

so, by Hall's theorem, $\{A_j\}$ has an SDR $\{i_j\}$, and $P_{i_j} \cap Q_j \neq \phi$. Hence, P and Q have an SCR.

V. ORTHOGONAL PARTITIONS

We now prove a property of partitions that will later turn out to be equivalent to the closure of the permutations realizable by Clos' network.

Definition 3: Partitions P, Q are orthogonal, written $P \perp Q$, iff $P_i \in P$ and $Q_j \in Q$ imply $|P_i \cap Q_j| = 1$.

Remark: If $P \perp Q$, and π is a permutation, then $\pi P \perp \pi Q$.

The next result was first given by J. B. Kruskal.

Theorem 3: If P, R are both (r, n)-partitions, then there is an (n, r)-partition Q orthogonal to each of P and R.

Proof: By Theorem 2, P and R have an SCR Q_1. Remove all elements of Q_1 from the P_i and the Q_j to give new $(r, n-1)$-partitions P' and Q'. Repeat to find $Q_2, Q_3, \cdots, Q_n$, and then take $Q = \{Q_i\}$.

It is convenient to have notations for three special partitions which arise naturally from the switching applications we are making. Clearly, the inlets (or outlets) of the network in Fig. 1 can be partitioned according to what last (or first) stage switch they are on. Similarly, the "wires" of the cross-connect fields between the stages can be partitioned according to what middle switch they impinge on. Accordingly, we define the (r, n)-partition S (by "outer" switches) as

$$S = \{S_j, j = 1, \cdots, r\}, \qquad S_j = \{k : (j-1)n < k \leqq jn\},$$

and the (n, r)-partition M (by "middle" switches) as

$$M = \{M_j, j = 1, \cdots, n\}, \qquad M_j = \{k : (j-1)r < k \leqq jr\}.$$

It is also convenient to partition by terminal position on outer switches, so we define the (n, r)-partition T by $T = \{T_j, j = 1, \cdots, n\}$ with

$$T_j = \{k : k = ln + j \quad \text{for some} \quad 0 \leqq l \leqq r-1\}.$$

The canonical cross-connect field is defined by

$$\varphi : j \to 1 + \left[\frac{j-1}{n}\right] + r[(j-1) \bmod n] \qquad j = 1, 2, \cdots, nr.$$

The following properties can be verified: $\varphi T = M$, $S \perp T$. Intuitively, φ takes the jth terminal on the ith switch into the ith terminal in the jth switch.

VI. CHARACTERIZATION OF REALIZABLE PERMUTATIONS

The next theorem will give a necessary and sufficient condition on a permutation π to be realizable in Clos' network, i.e., to belong to $G\varphi^{-1}H\varphi G$. We start with a lemma.

Lemma: *Let P be any (r, n)-partition. If there is an (n, r)-partition R such that*

$$P \perp R \perp S,$$

then there exists an element $g \in G$ such that

$$\varphi g P \perp M.$$

The practical import of this result is as follows: Consider a frame of r $n \times n$ switches followed by n $r \times r$ switches, with the canonical cross-connect field φ in between (Fig. 4); then, under the hypothesis there is a setting of the right-hand switches (i.e., the r $n \times n$), which has the effect of connecting each set of P to some terminal on every switch of the left-hand stage of n $r \times r$, i.e., it images each P_i so as to reach every left switch (exactly once).

Proof of lemma: Let $R = \{R_i\}$. Each R_i is simultaneously an SDR of P and one for S. Thus, if we connect the terminals of R_1 to the first left-hand stage switch, we will have used up one terminal from each P-set and also one from each switch on the right. This procedure can be repeated with $R_2, R_3, \cdots, R_n$ to give the result. Evidently, this set of connections defines an element $g \in G$ such that each set of $\varphi g P$ is spread over the left-hand stage switches, i.e., such that $\varphi g P \perp M$.

Theorem 4: *$\pi \in G\varphi^{-1}H\varphi G$ iff there is an (n, r)-partition R such that*

$$S \perp R \perp \pi^{-1}S.$$

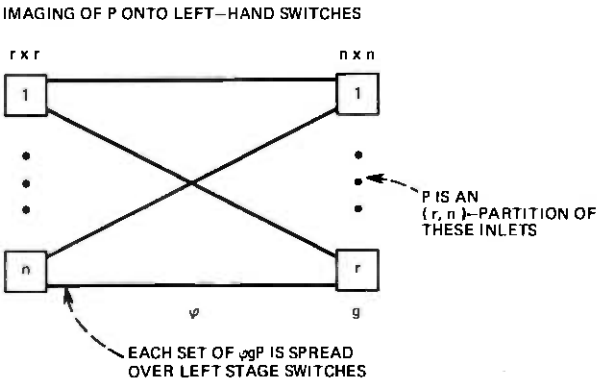

Fig. 4—Import of the lemma.

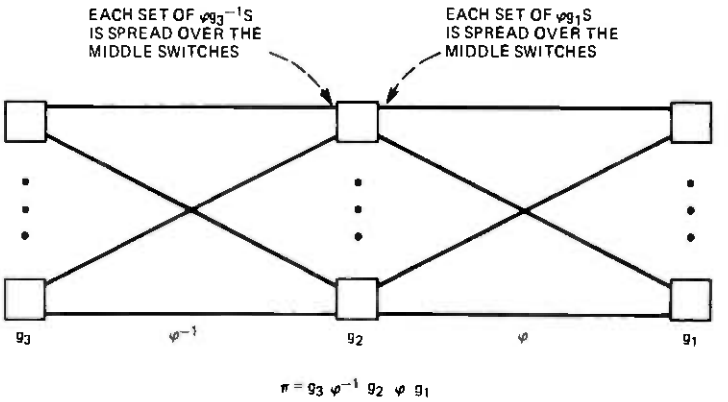

Fig. 5—$\varphi g_1 S \perp M \perp \varphi g_3^{-1} S$.

Proof: Let M be the partition of nr by middle switches, i.e., the (n, r)-partition consisting of the n sets

$$\{jr + 1, jr + 2, \cdots, (j + 1)r\} \qquad J = 0, 1, \cdots, n - 1,$$

and note that $hM = M$ for $h \in H$. Suppose now that $\pi \in G\varphi^{-1}H\varphi G$ with $\pi = g_3\varphi^{-1}g_2\varphi g_1$ and $g_1, g_3 \in G$, and $g_2 \in H$. It can be seen from Fig. 5 that each set of $\varphi g_3^{-1}S$ is spread over all the middle switches. Similarly, each set of $\varphi g_1 S$ is spread over the middle switches. Combinatorially, and without the help of pictures, these facts follow from $\varphi T = M$, from $gS = S$ for $g \in G$, and from $S \perp T$, and they can be rendered as

$$\varphi g_1 S \perp M$$
$$\varphi g_3^{-1} S \perp M.$$

It follows from the observation above that $g_2 M = M$, and thus, by the remark after Definition 3,

$$g_2\varphi g_1 S \perp M \perp \varphi g_3^{-1} S,$$

whence

$$\pi S \perp g_3\varphi^{-1} M \perp S$$

or

$$S \perp g_1^{-1}\varphi^{-1}g_2^{-1} M \perp \pi^{-1} S.$$

For R, we take $g_1^{-1}\varphi^{-1}g_2^{-1}M$, and the necessity is proved.

For the sufficiency, we use the lemma, according to which the hypothesis implies that there is an element $g_1 \in G$ such that

$$\varphi g_1 \pi^{-1} S \perp M.$$

Thus, in Fig. 5, by setting up g_1 in the right-hand stage, we can connect,

for each $j = 1, \cdots, n$, the terminals of $\pi^{-1}S_j$, one each to a middle switch. It remains to define g_2 for the middle stage by collecting those destined for $S_1, S_2, \cdots$, and g_3 for the left-hand stage by distributing within each of the sets $S_1, S_2, \cdots$ in the left-hand stage. This is done precisely as follows: Define g_2 by switching a terminal l to third stage switch j iff

$$l \in \varphi g_1 \pi^{-1} S_j.$$

It follows that $\varphi^{-1} g_2 \varphi g_1 \pi^{-1} S_j = S_j$. Then define g_3 by switching, within each final switch, $\varphi^{-1} g_2 \varphi g_1 \pi^{-1} i$ to i. Then $\pi = g_3 \varphi^{-1} g_2 \varphi g_1 \in G \varphi^{-1} H \varphi G$, as was to be proved.

VII. CLOSURE AND FACTORIZATION

Theorem 5: *$G \varphi^{-1} H \varphi G$ is closed under multiplication iff, for any two (r, n)-partitions P, Q, there is an (n, r)-partition R such that $P \perp R \perp Q$.*

Proof: Let P, Q be given (r, n)-partitions. If $G \varphi^{-1} H \varphi G$ is closed, then it is a group that contains all transpositions, and so equals S_{nr}. Hence, there exist permutations π_1 and π_2 such that

$$\pi_1 S = P, \qquad \pi_2^{-1} S = Q.$$

Since $G \varphi^{-1} H \varphi G$ is closed, it is clear that $\pi_2 \pi_1$ belongs to it. By Theorem 3, or by inspection of Fig. 5, with $\pi = \pi_2 \pi_1$, we see there is a partition N such that

$$S \perp N \perp (\pi_2 \pi_1)^{-1} S;$$

that is,

$$\pi_1 S \perp \pi_1 N \perp \pi_2^{-1} S.$$

For the requisite partition R, take $\pi_1 N$, and the necessity is proved.

For the sufficiency, let $\pi_1, \pi_2 \in G \varphi^{-1} H \varphi G$, and let $P = \pi_1 S$, $Q = \pi_2^{-1} S$. Then, by the hypothesis, there is an (n, r)-partition R such that

$$P \perp R \perp Q;$$

that is,

$$\pi_1 S \perp R \perp \pi_2^{-1} S$$
$$S \perp \pi_1^{-1} R \perp (\pi_2 \pi_1)^{-1} S.$$

Hence, by Theorem 4, $\pi_2 \pi_1 \in G \varphi^{-1} H \varphi G$, and we have proved that $G \varphi^{-1} H \varphi G$ is closed.

Theorem 6 (Slepian-Duguid):

$$S_{nr} = G \varphi^{-1} H \varphi G.$$

Proof: Immediate from Theorems 3 and 5, since the right-hand side contains all transpositions and is closed.

VIII. FURTHER PROBLEMS AND COMMENTS

Since H is a group, it follows that $\varphi^{-1}H\varphi$ is also a group, one conjugate to H, and that the Slepian-Duguid theorem can be cast as a decomposition

$$S_{nr} = \bigcup_{\pi \in \varphi^{-1}H\varphi} G\pi G$$

into disjoint double cosets, similar to the classical Frobenius[7] decomposition. It is tempting to expect some sort of connection with Frobenius' theorem here. One can speculate, in particular, that there is a proof of the Slepian-Duguid theorem from Frobenius', obtained by specializing the requisite cosets to those of the form $G\pi G$ with π in the conjugate $\varphi^{-1}H\varphi$, and showing that only these need be considered.

In conversation, Richard Stanley has indicated that, in another problem, also concerned with showing that a certain set of generated permutations was all of S_{nr}, he had used the known result that a primitive group containing a transposition is a symmetric group. His remark stimulated our original approach to a "group-theoretic" proof of the rearrangeability theorem: one easily shows that, if $G\varphi^{-1}H\varphi G$ is a group, then it is a primitive group containing a transposition; the problem then became to show that it was closed, a property that turned out to be equivalent to Kruskal's orthogonal partitions result (Theorem 3). Since closure was by comparison difficult to prove, and since it became clear that $G\varphi^{-1}H\varphi G$ contains all transpositions, the simpler proof presented here could be used, making the original side trip via primitive groups gratuitous. Stanley's idea, however, is still a possible proof method for other networks that lead to less transparent groups of realizable permutations.

REFERENCES

1. V. E. Beneš, "Permutation Groups, Complexes, and Rearrangeable Connecting Networks," B.S.T.J., *43*, No. 4, Part 2 (July 1964), pp. 1619–1640.
2. V. E. Beneš, "Proving the Rearrangeability of Connecting Networks by Group Calculations," B.S.T.J., *54*, No. 2 (February 1975), pp. 423–434.
3. V. E. Beneš, "Applications of Group Theory to Connecting Networks," B.S.T.J., *54*, No. 2 (February 1975), pp. 409–422.
4. V. E. Beneš, *Mathematical Theory of Connecting Networks and Telephone Traffic*, New York: Academic Press, 1965, pp. 86 ff.
5. H. J. Ryser, *Combinatorial Mathematics*, Carus Monograph 14, Math. Assoc. of America, New York: John Wiley, 1963, p. 48.
6. Ref. 5, p. 50.
7. W. Ledermann, *Introduction to the Theory of Finite Groups*, Edinburgh and London: Oliver and Boyd, 1957, p. 60.

Contributors to This Issue

Iain Anderson, A.H.-W.C., 1964, Heriot-Watt College, Edinburgh, Scotland; M.Sc., 1966, University College, London, England; Ph.D., D.I.C., 1969, Imperial College, London, England; Bell Laboratories, 1970—. Mr. Anderson is in the Radio Research Laboratory and has studied topics in diffraction theory, antenna analysis, and radome design. Associate, IEE.

Václav E. Beneš, A.B., 1950, Harvard College; M.A. and Ph.D., 1953, Princeton University; Bell Laboratories, 1953—. Mr. Beneš has pursued mathematical research on traffic theory, stochastic processes, frequency modulation, combinatorics, servomechanisms, and stochastic control. In 1959–60, he was visiting lecturer in mathematics at Dartmouth College. In 1971, he taught stochastic processes at SUNY Buffalo, and from 1971–72, he was Visiting MacKay Lecturer in electrical engineering at the University of California in Berkeley. He is the author of two books in his field. Member, American Mathematical Society, Association for Symbolic Logic, Institute of Mathematical Statistics, SIAM, Mathematical Association of America, Mind Association.

C. N. Berglund, B.Sc. (E.E.), 1960, Queen's University, Kingston, Ontario; M.S.E.E., 1961, Massachusetts Institute of Technology; Ph.D. (E.E.), 1964, Stanford University. Research Assistant, M.I.T., 1960–61; Research Associate, Department of Electrical Engineering, Queen's University, Kingston, 1961–62; Research Assistant, Stanford Electronics Laboratories, 1962–64. Bell Laboratories, 1964–72. At Bell Laboratories, Mr. Berglund was a supervisor in the Semiconductor Device Laboratory. Member, APS.

Francis J. Brophy, B.S. (Mathematics), 1968, St. Joseph's College; M.S. (Mathematics), 1971, Stevens Institute of Technology; Bell Laboratories, 1968—. Mr. Brophy has been involved with software simulation of various data transmission techniques and most recently with the design of digital filters.

James T. Clemens, B.S. (Physics), 1965, Ph.D. (Physics), 1970, Polytechnic Institute of New York; Bell Laboratories, 1969—. Mr. Clemens has worked on MOS integrated-circuit development and in the characterization and development of the (Al_2O_3/SiO_2) MOS technology. He is currently supervisor of a Si-Gate MOS technology development group responsible for MOS integrated-circuit memory technology.

James W. Gewartowski, B.S., 1952, Illinois Institute of Technology; S.M., 1953, Massachusetts Institute of Technology; Ph.D., 1958, Stanford University; Bell Laboratories, 1957—. Mr. Gewartowski was initially concerned with the development of high-power microwave tubes and electron guns. From 1962 to 1971, he supervised a group studying varactor harmonic generators and upconverters and circuit properties of IMPATT diodes. Since 1971, he has supervised the group developing IMPATT amplifiers for radio relay systems. Member, IEEE, Tau Beta Pi, Eta Kappa Nu, Sigma Xi. Recipient, 1960 IEEE Browder J. Thompson Memorial Prize.

Edgar N. Gilbert, B.S. (Mathematics), 1943, Queens College; Ph.D. (Mathematics), 1948, Massachusetts Institute of Technology; M.I.T. Radiation Laboratory, 1944–46; Bell Laboratories, 1948—. Mr. Gilbert is a member of the Mathematics and Statistics Research Center at Bell Laboratories, specializing in communication theory and other applications of probability and combinatorial mathematics.

Edward F. Labuda, B.S. (Physics), 1959, Case Western Reserve University; M.S.E.E., 1961, New York University; Ph.D. (Electrophysics), 1967, Polytechnic Institute of New York; Bell Laboratories, 1959—. Mr. Labuda has been engaged in the development of low-noise microwave tubes, gas lasers, and silicon diode array camera tubes for *Picturephone*® video telephone systems. He now supervises a group concerned with integrated-circuit technology development.

James McKenna, B.Sc. (Mathematics), 1951, Massachusetts Institute of Technology; Ph.D. (Mathematics), 1961, Princeton University; Bell Laboratories, 1960—. Mr. McKenna has done research in quantum mechanics, electromagnetic theory, and statistical mechanics. He has recently been engaged in the study of nonlinear partial differential equations that arise in solid-state device work and in the theory of stochastic differential equations.

James E. Morris, Bell Laboratories, 1959—. Mr. Morris' early work was in the design and development of microwave strip-line circuits for military applications. Since 1968, he has been engaged in studies of IMPATT devices and their application as amplifiers in radio-relay equipment. Member, IEEE.

Andres C. Salazar, B.A. (Math), B.S.E.E., 1964, M.S., 1965, University of New Mexico; and Ph.D., 1967, Michigan State University; Bell Laboratories, 1967—. Mr. Salazar has been engaged in the statistical evaluation of data set performance on the switched telephone network. His current interests are in the areas of digital filter design and equalization techniques for voiceband data transmission systems. Member, IEEE, Phi Kappa Phi.

N. L. Schryer, B.S., 1965, M.S., 1966, and Ph.D., 1969, University of Michigan; Bell Laboratories, 1969—. Mr. Schryer has worked on the numerical solution of parabolic and elliptic partial differential equations. He is currently studying problems of this type that arise in semiconductor device theory.