THE BELL SYSTEM TECHNICAL JOURNAL

JANUARY 1981
VOL. 60, NO. 1

ISSN0005-8580

THE BELL SYSTEM TECHNICAL JOURNAL

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

Comments on the technical content of any article or brief are welcome. These and other editorial inquiries should be addressed to the Editor, The Bell System Technical Journal, Bell Laboratories, 600 Mountain Avenue, Murray Hill, N.J., 07974. Comments and inquiries, whether or not published, shall not be regarded as confidential or otherwise restricted in use and will become the property of the American Telephone and Telegraph Company. Comments selected for publication may be edited for brevity, subject to author approval.

THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

Volume 60 January 1981 Number 1

An Overflow System in Which Queuing Takes Precedence

By J. A. MORRISON

(Manuscript received July 1, 1980)

When calls offered to a primary group of trunks find all of them busy, provisions are often made for these calls to overflow to other groups of trunks. Such traffic overflow systems have been of interest for a long time, but recently overflow systems that allow for some calls to be queued have been of importance. In this paper we analyze a traffic overflow system with queuing, which consists of a primary and a secondary group. The system which we consider here differs from the two systems we investigated earlier, in that no overflow from the primary to the secondary is permitted if there is a waiting space available in the primary queue. As with the earlier investigations, we adopt an analytical approach which considerably reduces the dimensions of the problem, and simplifies the calculation of various steady-state quantities of interest. Our results include expressions for the loss probabilities, the average waiting times in the queues, and the average number of demands in service in each group.

I. INTRODUCTION

In this paper a particular overflow system with queuing is analyzed. The system consists of two groups, a primary and a secondary, with n_k servers and q_k waiting spaces, which receive demands from independent Poisson sources S_k with arrival rates $\lambda_k > 0$, $k = 1$ and 2, respectively, as depicted in Fig. 1. The service times of the demands are independent, and exponentially distributed with mean service rate $\mu > 0$. If all n_2 servers in the secondary are busy when a demand from S_2 arrives, the demand is queued if one of the q_2 waiting spaces is

In Memoriam: Joanne B. Fry, Associate Editor of The Bell System Technical Journal since 1978, died in an automobile accident January 2, 1981.

available, otherwise it is lost (blocked and cleared from the system). Demands waiting in the secondary queue enter service (in some prescribed order) as servers in the secondary become free.

If all n_1 servers in the primary are busy when a demand from S_1 arrives, the demand is queued in the primary, if one of the q_1 waiting spaces is available. No overflow is permitted from the primary queue, so that a demand in the primary queue must wait for a server in the primary to become free. If all n_1 servers in the primary are busy and all q_1 waiting spaces are occupied, when a demand from S_1 arrives, the demand is served in the secondary, if there is a free server and there are no demands waiting in the secondary queue, otherwise it is lost.

The overflow system described above differs from the two systems which we investigated earlier,[1,2] in that no overflow from the primary to the secondary is permitted if there is a waiting space available in the primary queue. This restriction was one invoked by Anderson.[3] In the two systems investigated earlier, arriving calls can overflow when the primary queue is not full. The system considered in this paper is a particular case of the one considered by Rath,[4] which was composed of two queues, one of which is allowed to overflow to the other under specified conditions involving the queue lengths. He obtained some numerical solutions using a Gauss–Seidel iteration technique, but none of these correspond to the particular system that we are considering. He also developed an approximate procedure for analyzing his system, based on the use of the Interrupted Poisson Process.

Here we analyze the overflow system using a technique analogous to the one introduced in the earlier paper.[1] Let p_{ij} denote the steady-state probability that there are i demands in the primary and j demands in the secondary, either in service or waiting. These probabilities satisfy a set of generalized birth-and-death equations, which take the form of partial difference equations connecting neighboring states. We carry out an analysis that reduces the dimensions of the problem, which may be considerable in cases of interest. The basic technique is to separate variables in the region away from a certain boundary of state space. This leads to an eigenvalue problem for the separation constant. The eigenvalues are roots of a polynomial equation. The probabilities p_{ij} are then represented in terms of the corresponding eigenfunctions. The constant coefficients in these representations are determined from the boundary conditions and the normalization condition that the sum of the probabilities is unity.

Various steady-state quantities are of interest, which may be expressed in terms of the probabilities p_{ij}. The quantities include the loss (or blocking) probabilities, the average waiting times in the queues, and the average number of demands in service in each group. These quantities may be expressed directly in terms of the constant coefficients which occur in the representations for the probabilities p_{ij}. Thus

the steady-state quantities of interest may be calculated directly, without having to calculate the probabilities p_{ij}. Here again the reduction in the dimensions of the problem is valuable.

Only the theoretical results are presented in this paper. Numerical results will be presented in a forthcoming paper by Kaufman, Seery, and Morrison.[5] Results will be given there for the two overflow systems considered previously, based on the earlier analysis,[1] as well as for the system considered in this paper.

Section II discusses the representation of the probabilities p_{ij} in terms of the eigenfunctions, and the boundary and normalization conditions. Various steady-state quantities of interest are calculated in Section III. The appendix gives properties of the eigenfunctions.

We assume throughout the analysis that $q_1 \geq 1$, since the system considered in this paper, and the two systems analyzed earlier, are identical if $q_1 = 0$, i.e., if there is no primary queue. However, the results of this paper reduce to those obtained earlier[1] if $q_1 = 0$. If q_2 is large, or even infinite, an alternate analysis, analogous to that presented for the other two systems,[2] may be carried out for the present system, but we do not pursue that here.

II. REPRESENTATION AND BOUNDARY CONDITIONS

We let p_{ij} denote the steady-state probability that there are i demands in the primary and j demands in the secondary, either in service or waiting. These probabilities satisfy a set of generalized birth-and-death equations,[6] which may be derived in a straightforward manner. We define the traffic intensities

$$a_1 = \lambda_1/\mu, \qquad a_2 = \lambda_2/\mu, \tag{1}$$

and let the total number of servers and waiting spaces in each group be

$$k_1 = n_1 + q_1, \qquad k_2 = n_2 + q_2. \tag{2}$$

It is convenient to introduce the function

$$\chi_l = \begin{cases} 1, & l \geq 0, \\ 0, & l < 0, \end{cases} \tag{3}$$

as well as the Kronecker delta

$$\delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases} \tag{4}$$

Then the birth-and-death equations are

$$[a_1(1 - \delta_{ik_1}\chi_{j-n_2}) + a_2(1 - \delta_{jk_2}) + \min(i, n_1) + \min(j, n_2)]p_{ij}$$
$$= a_1(1 - \delta_{i0})p_{i-1,j} + (1 - \delta_{j0})(a_1\delta_{ik_1}\chi_{n_2-j} + a_2)p_{i,j-1}$$

$$+ (1 - \delta_{ik_1})\min(i + 1, n_1)p_{i+1,j} + (1 - \delta_{jk_2})\min(j + 1, n_2)p_{i,j+1}, \quad (5)$$

for $0 \le i \le k_1$, $0 \le j \le k_2$. The normalization condition is

$$\sum_{i=0}^{k_1} \sum_{j=0}^{k_2} p_{ij} = 1. \quad (6)$$

For $i \ne k_1$, the variables in (5) may be separated, and there are solutions of the form $\alpha_i\beta_j$, where

$$[a_1 + \min(i, n_1) + \rho]\alpha_i = a_1(1 - \delta_{i0})\alpha_{i-1} + \min(i + 1, n_1)\alpha_{i+1}, \quad (7)$$

for $0 \le i \le k_1 - 1$, and

$$[a_2(1 - \delta_{jk_2}) + \min(j, n_2) - \rho]\beta_j = a_2(1 - \delta_{j0})\beta_{j-1} + (1 - \delta_{jk_2})\min(j + 1, n_2)\beta_{j+1}, \quad (8)$$

for $0 \le j \le k_2$, and ρ is a separation constant. Hence, from (7),

$$(a_1 + i + \rho)\alpha_i = a_1(1 - \delta_{i0})\alpha_{i-1} + (i + 1)\alpha_{i+1}, \quad (9)$$

for $0 \le i \le n_1 - 1$. The solution of (9) may be expressed in terms of Poisson-Charlier[7,8] polynomials. We here denote the solution of (9) for which $\alpha_0 = 1$ by $s_i(\rho, a_1)$. The properties of $s_i(\rho, a)$ which we will need are given in the appendix.

We assume that $q_1 \ge 1$. Then, from (7),

$$(a_1 + n_1 + \rho)\alpha_i = a_1\alpha_{i-1} + n_1\alpha_{i+1}, \quad (10)$$

for $n_1 \le i \le k_1 - 1$. The solution of (10) may be expressed in terms of Chebyshev polynomials of the second kind,[9] $U_l(x)$. It is convenient to define

$$\Omega_l(\rho) = \left(\frac{n_1}{a_1}\right)^{l/2} U_l\left(\frac{a_1 + n_1 + \rho}{2\sqrt{a_1 n_1}}\right). \quad (11)$$

The appendix gives the properties of these functions that we need. We note here, however, that $U_0(x) \equiv 1$ and $U_{-1}(x) \equiv 0$. From (9), (10), (53), and (64), with a suitable normalization, it follows that

$$\alpha_i(\rho) = \begin{cases} \left(\dfrac{n_1}{a_1}\right)^{q_1} s_i(\rho, a_1), & 0 \le i \le n_1, \\ \left(\dfrac{n_1}{a_1}\right)^{k_1 - i} [s_{n_1}(\rho, a_1)\Omega_{i-n_1}(\rho) - s_{n_1-1}(\rho, a_1)\Omega_{i-n_1-1}(\rho)], & n_1 \le i \le k_1. \end{cases} \quad (12)$$

Next, from (8),

$$(a_2 + j - \rho)\beta_j = a_2(1 - \delta_{j0})\beta_{j-1} + (j + 1)\beta_{j+1}, \quad (13)$$

for $0 \le j \le n_2 - 1$. It follows from (53) that β_j is proportional to $s_j(-\rho, a_2)$ for $0 \le j \le n_2$. Also,

$$[a_2(1-\delta_{jk_2})+n_2-\rho]\beta_j = a_2\beta_{j-1}+n_2(1-\delta_{jk_2})\beta_{j+1}, \tag{14}$$

for $n_2 \le j \le k_2$. Corresponding to (11), we define

$$\Psi_l(\rho) = \left(\frac{n_2}{a_2}\right)^{l/2} U_l\left(\frac{a_2+n_2-\rho}{2\sqrt{a_2 n_2}}\right). \tag{15}$$

We also define

$$\phi_j(\rho) = \Psi_{k_2-j}(\rho) - \Psi_{k_2-j-1}(\rho). \tag{16}$$

The appendix gives the properties of these functions that we need. It follows from (14) and (62) that β_j is proportional to $\phi_j(\rho)$ for $n_2 - 1 \le j \le k_2$.

Consequently, we take

$$\beta_j(\rho) = \begin{cases} s_j(-\rho, a_2)\phi_{n_2}(\rho), & 0 \le j \le n_2, \\ s_{n_2}(-\rho, a_2)\phi_j(\rho), & n_2 - 1 \le j \le k_2, \end{cases} \tag{17}$$

where

$$s_{n_2-1}(-\rho, a_2)\phi_{n_2}(\rho) = s_{n_2}(-\rho, a_2)\phi_{n_2-1}(\rho). \tag{18}$$

This equation may be written in the form

$$\rho[s_{n_2}(1-\rho, a_2)\Psi_{q_2}(\rho) - s_{n_2-1}(1-\rho, a_2)\Psi_{q_2-1}(\rho)] = 0. \tag{19}$$

The expression in the square brackets in (19) is a polynomial in ρ of degree $k_2 = n_2 + q_2$. It was shown[1] that its zeros are positive and distinct, and we denote them by ρ_m, $m = 1, \cdots, k_2$. We also let $\rho_0 = 0$. It follows that we may represent the probabilities p_{ij} in the form

$$p_{ij} = \begin{cases} \sum_{m=0}^{k_2} c_m \alpha_i(\rho_m) s_j(-\rho_m, a_2)\phi_{n_2}(\rho_m), & 0 \le j \le n_2, \\ \sum_{m=0}^{k_2} c_m \alpha_i(\rho_m) s_{n_2}(-\rho_m, a_2)\phi_j(\rho_m), & n_2 \le j \le k_2, \end{cases} \tag{20}$$

for $0 \le i \le k_1$, where $\alpha_i(\rho)$ is defined in (12), and the constants c_m are to be determined.

It remains to satisfy the boundary conditions corresponding to $i = k_1$ in (5), as well as the normalization condition (6). If we set $i = k_1$ in (5), we obtain

$$(a_1+a_2+n_1+j)p_{k_1,j} = a_1 p_{k_1-1,j} + (a_1+a_2)(1-\delta_{j0})p_{k_1,j-1} + (j+1)p_{k_1,j+1}, \tag{21}$$

for $0 \le j \le n_2 - 1$,

$$[a_2(1-\delta_{q_2,0}) + n_1 + n_2]p_{k_1,n_2} = a_1 p_{k_1-1,n_2} + (a_1+a_2)p_{k_1,n_2-1} + n_2(1-\delta_{q_2,0})p_{k_1,n_2+1}, \tag{22}$$

and, if $q_2 \geq 1$,

$$[a_2(1 - \delta_{jk_2}) + n_1 + n_2]p_{k_1,j} = a_1 p_{k_1-1,j} + a_2 p_{k_1,j-1} + n_2(1 - \delta_{jk_2})p_{k_1,j+1}, \quad (23)$$

for $n_2 + 1 \leq j \leq k_2$.

If we substitute (20) into (21), after reduction with the help of the recurrence relations in the appendix, we find that

$$\sum_{m=0}^{k_2} c_m\{\rho_m[s_{n_1}(1 + \rho_m, a_1)\Omega_{q_1}(\rho_m) - s_{n_1-1}(1 + \rho_m, a_1)\Omega_{q_1-1}(\rho_m)]s_j(-\rho_m, a_2) + a_1[s_{n_1}(\rho_m, a_1)\Omega_{q_1}(\rho_m) - s_{n_1-1}(\rho_m, a_1)\Omega_{q_1-1}(\rho_m)]s_j(-1 - \rho_m, a_2)\} \cdot \phi_{n_2}(\rho_m) = 0, \quad (24)$$

for $0 \leq j \leq n_2 - 1$. Also, from (23), it is found that

$$\sum_{m=0}^{k_2} c_m \rho_m s_{n_2}(-\rho_m, a_2)[s_{n_1}(1 + \rho_m, a_1)\Omega_{q_1}(\rho_m) - s_{n_1-1}(1 + \rho_m, a_1)\Omega_{q_1-1}(\rho_m)]\phi_j(\rho_m) = 0, \quad (25)$$

for $n_2 + 1 \leq j \leq k_2$. It may be shown that the boundary condition (22) is redundant, as is to be expected. Thus the constants c_m are determined by (24) and (25) only to within a multiplicative constant, which is determined from the normalization condition (6).

From (20), with the help of (16), (19), (57), and (58), it is found that

$$\sum_{j=0}^{k_2} p_{ij} = c_0 \alpha_i(0)[s_{n_2}(1, a_2)\Psi_{q_2}(0) - s_{n_2-1}(1, a_2)\Psi_{q_2-1}(0)], \quad (26)$$

for $0 \leq i \leq k_1$. But, from (12) and (66),

$$\alpha_i(0) = \begin{cases} \left(\dfrac{n_1}{a_1}\right)^{q_1} s_i(0, a_1), & 0 \leq i \leq n_1, \\ \left(\dfrac{n_1}{a_1}\right)^{k_1 - i} s_{n_1}(0, a_1), & n_1 \leq i \leq k_1. \end{cases} \quad (27)$$

Hence, from (26) and (27), with the help of (57), (58), and (65), the normalization condition (6) implies that

$$c_0[s_{n_1}(1, a_1)\Omega_{q_1}(0) - s_{n_1-1}(1, a_1)\Omega_{q_1-1}(0)] \cdot [s_{n_2}(1, a_2)\Psi_{q_2}(0) - s_{n_2-1}(1, a_2)\Psi_{q_2-1}(0)] = 1. \quad (28)$$

Once the constants c_m have been determined, the steady-state probabilities p_{ij} may be calculated from (20). We remark that the number

of constants to be determined is only $k_2 + 1$, whereas the number of probabilities p_{ij} is $(k_1 + 1)(k_2 + 1)$.

III. SOME STEADY-STATE QUANTITIES

We proceed now to the calculation of various steady-state quantities of interest. These quantities are depicted in the diagram of Fig. 1, which indicates the mean flow rates. The loss probabilities L_1 and L_2 are given by

$$L_1 = \sum_{j=n_2}^{k_2} p_{k_1,j}, \qquad L_2 = \sum_{i=0}^{k_1} p_{i,k_2}, \tag{29}$$

and the probabilities that a demand from the primary, or secondary, source is queued on arrival are

$$Q_1 = \sum_{i=n_1}^{k_1-1} \sum_{j=0}^{k_2} p_{ij}, \qquad Q_2 = (1 - \delta_{q_2,0}) \sum_{i=0}^{k_1} \sum_{j=n_2}^{k_2-1} p_{ij}. \tag{30}$$

The probability that a demand arriving from the primary source

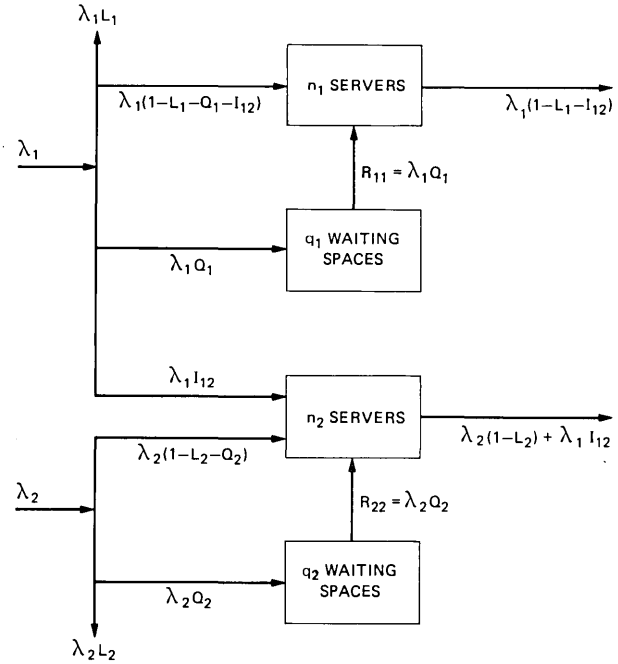


Fig. 1—Mean flow rates for an overflow system with queuing; Poisson arrival rates λ_1 and λ_2, loss probabilities L_1 and L_2, queuing probabilities Q_1 and Q_2, and overflow probability I_{12}.

overflows (immediately) is

$$I_{12} = \sum_{j=0}^{n_2-1} p_{k_1,j}. \tag{31}$$

Since the mean service rate is μ, the mean departure rate from the primary queue to the primary servers is

$$R_{11} = n_1\mu \sum_{i=n_1+1}^{k_1} \sum_{j=0}^{k_2} p_{ij}, \tag{32}$$

while the mean departure rate from the secondary queue to the secondary servers is

$$R_{22} = n_2\mu(1 - \delta_{q_2,0}) \sum_{i=0}^{k_1} \sum_{j=n_2+1}^{k_2} p_{ij}. \tag{33}$$

The average number of demands in the primary and secondary queues are

$$V_1 = \sum_{i=n_1}^{k_1} \sum_{j=0}^{k_2} (i - n_1)p_{ij}, \qquad V_2 = \sum_{i=0}^{k_1} \sum_{j=n_2}^{k_2} (j - n_2)p_{ij}. \tag{34}$$

Also, the average number of demands in service in the two groups are

$$X_1 = \sum_{i=0}^{k_1} \sum_{j=0}^{k_2} \min(i, n_1)p_{ij}, \qquad X_2 = \sum_{i=0}^{k_1} \sum_{j=0}^{k_2} \min(j, n_2)p_{ij}. \tag{35}$$

Now, according to Little's theorem,[6] the average number of demands in a queuing system is equal to the average rate of arrival of demands to that system times the average time spent in that system. If we apply this result to the primary and secondary queues, we find that the average waiting times of the demands which are queued in the primary or in the secondary are given by

$$W_1 = \frac{V_1}{\lambda_1 Q_1}, \qquad W_2 = \frac{V_2}{\lambda_2 Q_2} \qquad (q_2 \geq 1), \tag{36}$$

respectively, independently of the order of service within each queue. Also, if we apply Little's theorem to the primary and secondary groups of servers, we obtain

$$\lambda_1(1 - L_1 - I_{12}) = \mu X_1, \qquad \lambda_2(1 - L_2) + \lambda_1 I_{12} = \mu X_2. \tag{37}$$

The steady-state quantities of interest may be expressed in terms of the constants c_m with the help of the representations in (20). From (29) it is found, with the help of (12) and (16), that

$$L_1 = \sum_{m=0}^{k_2} c_m[s_{n_1}(\rho_m, \alpha_1)\Omega_{q_1}(\rho_m) - s_{n_1-1}(\rho_m, \alpha_1)\Omega_{q_1-1}(\rho_m)]$$
$$\cdot s_{n_2}(-\rho_m, \alpha_2)\Psi_{q_2}(\rho_m). \tag{38}$$

We define

$$d_0 = c_0[s_{n_2}(1, a_2)\Psi_{q_2}(0) - s_{n_2-1}(1, a_2)\Psi_{q_2-1}(0)]$$
$$= [s_{n_1}(1, a_1)\Omega_{q_1}(0) - s_{n_1-1}(1, a_1)\Omega_{q_1-1}(0)]^{-1}, \tag{39}$$

from (28). Then, from (30) it is found, with the help of (26), (27), and (65), that

$$Q_1 = d_0 s_{n_1}(0, a_1)[\Omega_{q_1}(0) - 1]. \tag{40}$$

Moreover, from (29) and (31), it follows that

$$L_1 + I_{12} = d_0 s_{n_1}(0, a_1), \tag{41}$$

and from (32) it follows that $R_{11} = \lambda_1 Q_1$, as is to be expected, since in the steady state the mean departure rate from the queue is equal to the mean arrival rate to it.

We define

$$\Delta_q(\xi) = \sum_{l=1}^{q} l\xi^{q-l} = \begin{cases} [q - (q+1)\xi + \xi^{q+1}]/(1-\xi)^2, & \xi \neq 1, \\ \tfrac{1}{2} q(q+1), & \xi = 1. \end{cases} \tag{42}$$

Then, from (34), with the help of (26), (27), and (39), it is found that

$$V_1 = d_0 s_{n_1}(0, a_1)\Delta_{q_1}\left(\frac{n_1}{a_1}\right). \tag{43}$$

Also, from (35), with the help of (56), (58), (59), and (65), it follows that

$$X_1 = d_0 a_1 \left(\frac{n_1}{a_1}\right)^{q_1} s_{n_1-1}(1, a_1) + s_{n_1}(0, a_1)[\Omega_{q_1}(0) - 1]\Bigg\}. \tag{44}$$

It may be verified, with the help of (39), (57), and (65), that (41) and (44) are consistent with (37). The explicitness of the expressions for the quantities in (40), (41), (43), and (44) is due to the fact that these quantities are not affected by the secondary. This, of course, is not the case for the loss probability L_1, which is given by (38).

Next, from (29), since $\phi_{k_2}(\rho) \equiv 1$, it is found, with the help of (20) and (68), that

$$L_2 = \sum_{m=0}^{k_2} c_m s_{n_2}(-\rho_m, a_2)[s_{n_1}(1 + \rho_m, a_1)\Omega_{q_1}(\rho_m)$$
$$- s_{n_1-1}(1 + \rho_m, a_1)\Omega_{q_1-1}(\rho_m)]. \tag{45}$$

Also, from (35), with the help of (6) and (59), it follows that

$$X_2 = n_2 - \sum_{m=0}^{k_2} c_m s_{n_2-1}(2 - \rho_m, a_2)\phi_{n_2}(\rho_m)$$
$$\cdot[s_{n_1}(1 + \rho_m, a_1)\Omega_{q_1}(\rho_m) - s_{n-1}(1 + \rho_m, a_1)\Omega_{q_1-1}(\rho_m)]. \tag{46}$$

In view of (38), (41), (45), and (46), the second relationship in (37) provides a useful numerical check.

We now define

$$r_j = \sum_{i=0}^{k_1} p_{ij}, \qquad n_2 \le j \le k_2. \tag{47}$$

If we sum on i in (5), we obtain

$$[a_2(1 - \delta_{jk_2}) + n_2]r_j = a_2 r_{j-1} + n_2(1 - \delta_{jk_2})r_{j+1}, \tag{48}$$

for $n_2 + 1 \le j \le k_2$. It follows that

$$n_2 r_j = a_2 r_{j-1}, \qquad n_2 + 1 \le j \le k_2. \tag{49}$$

Hence, since $L_2 = r_{k_2}$, from (29) and (47),

$$r_j = \left(\frac{n_2}{a_2}\right)^{k_2 - j} L_2, \qquad n_2 \le j \le k_2. \tag{50}$$

Then, from (30) and (33), with the help of (63), we obtain

$$Q_2 = [\Psi_{q_2}(0) - 1]L_2, \tag{51}$$

and $R_{22} = \lambda_2 Q_2$, as is to be expected. Also, from (34) and (42), it follows that

$$V_2 = \Delta_{q_2}\left(\frac{n_2}{a_2}\right)L_2. \tag{52}$$

This completes the calculation of expressions for the steady-state quantities of interest.

IV. ACKNOWLEDGMENT

The author is grateful to G. M. Anderson for bringing this problem to his attention.

APPENDIX

We define $s_i(\lambda, a)$ by the recurrence relation

$$(a + i + \lambda)s_i(\lambda, a) = a(1 - \delta_{i0})s_{i-1}(\lambda, a) + (i + 1)s_{i+1}(\lambda, a);$$
$$s_0(\lambda, a) = 1, \tag{53}$$

for $i = 0, 1, \cdots$. Thus $s_n(\lambda, a)$ is a polynomial of degree n in both λ and a, and it may be related to a Poisson–Charlier polynomial.[7,8] However, we give here the properties of $s_n(\lambda, a)$ which we will need. An explicit formula is[1]

$$s_i(\lambda, a) = \sum_{r=0}^{i} \frac{(\lambda)_r a^{i-r}}{r!(i - r)!}, \tag{54}$$

where

$$(\lambda)_0 = 1, \qquad (\lambda)_r = \lambda(\lambda + 1) \cdots (\lambda + r - 1), \qquad r = 1, 2, \cdots. \tag{55}$$

It was also shown[1] that

$$(i + 1)s_{i+1}(\lambda, a) = as_i(\lambda, a) + \lambda s_i(\lambda + 1, a) \tag{56}$$

and

$$s_i(\lambda, a) = s_i(\lambda + 1, a) - (1 - \delta_{i0})s_{i-1}(\lambda + 1, a). \tag{57}$$

From (57) it follows that

$$\sum_{i=0}^{n} s_i(\lambda, a) = s_n(\lambda + 1, a), \tag{58}$$

and, from (56) and (58), we deduce that

$$\sum_{i=0}^{n} (n - i)s_i(\lambda, a) = (1 - \delta_{n0})s_{n-1}(\lambda + 2, a). \tag{59}$$

We now turn our attention to the Chebyshev polynomials of the second kind,[9] $U_l(x)$. They may be defined by the recurrence relation

$$2xU_l(x) = U_{l+1}(x) + U_{l-1}(x); \qquad U_{-1}(x) \equiv 0, \quad U_0(x) \equiv 1, \tag{60}$$

for $l = 0, 1, \cdots$. From (15) and (60) it follows that

$$(a_2 + n_2 - \rho)\Psi_l(\rho) = a_2\Psi_{l+1}(\rho) + n_2\Psi_{l-1}(\rho),$$

$$\Psi_{-1}(\rho) \equiv 0, \qquad \Psi_0(\rho) \equiv 1. \tag{61}$$

From (16) and (61) we deduce that

$$[a_2(1 - \delta_{jk_2}) + n_2 - \rho]\phi_j(\rho) = a_2\phi_{j-1}(\rho) + n_2(1 - \delta_{jk_2})\phi_{j+1}(\rho), \tag{62}$$

for $j \leq k_2$. Also, from (61), it may be shown by induction that

$$\Psi_l(0) = \sum_{r=0}^{l} \left(\frac{n_2}{a_2}\right)^r. \tag{63}$$

Next, from (11) and (60) it follows that

$$(a_1 + n_1 + \rho)\Omega_l(\rho) = a_1\Omega_{l+1}(\rho) + n_1\Omega_{l-1}(\rho),$$

$$\Omega_{-1}(\rho) \equiv 0, \qquad \Omega_0(\rho) \equiv 1. \tag{64}$$

In particular, it is found by induction that

$$\Omega_l(0) = \sum_{r=0}^{l} \left(\frac{n_1}{a_1}\right)^r. \tag{65}$$

Since $s_i(0, a) = a^i/i!$, from (54), it follows that

$$s_{n_1}(0, a_1)\Omega_l(0) - s_{n_1-1}(0, a_1)\Omega_{l-1}(0) = s_{n_1}(0, a_1), \qquad l = 0, 1, \cdots. \tag{66}$$

Next, from (9) and (10), we deduce that

$$\rho \sum_{i=0}^{k_1} \alpha_i(\rho) = (n_1 + \rho)\alpha_{k_1}(\rho) - a_1\alpha_{k_1-1}(\rho). \tag{67}$$

Then, with the help of (12), (56), (57), and (64), it may be shown that

$$\sum_{i=0}^{k_1} \alpha_i(\rho) = s_{n_1}(1 + \rho, a_1)\Omega_{q_1}(\rho) - s_{n_1-1}(1 + \rho, a_1)\Omega_{q_1-1}(\rho), \tag{68}$$

for $\rho \neq 0$. Moreover, this result holds for $\rho = 0$ also, from continuity.

REFERENCES

1. J. A. Morrison, "Analysis of Some Overflow Problems with Queuing," B.S.T.J., *59*, No. 8 (October 1980), pp. 1427–1462.
2. J. A. Morrison, "Some Traffic Overflow Problems with a Large Secondary Queue," B.S.T.J., *59*, No. 8 (October 1980), pp. 1463–1482.
3. G. M. Anderson, "Facilities Design for Automatic Route Selection with Queuing," unpublished work.
4. J. H. Rath, "An Approximation for a Queueing System with Two Queues and Overflows," unpublished work.
5. L. Kaufman, J. B. Seery, and J. A. Morrison, "Numerical Results for Some Overflow Problems with Queueing," unpublished work.
6. L. Kleinrock, *Queueing Systems, Volume I: Theory*, New York: Wiley, 1975.
7. A. Erdélyi et al., *Higher Transcendental Functions, Volume II*, New York: McGraw-Hill, 1953, p. 226.
8. J. Riordan, *Stochastic Service Systems,* New York: Wiley, 1962.
9. W. Magnus, F. Oberhettinger, and R. P. Soni, *Formulas and Theorems for the Special Functions of Mathematical Physics,* New York: Springer-Verlag, 1966, p. 256.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 1, January 1981
Printed in U.S.A.

Absenteeism of Operators: A Statistical Study with Managerial Applications

By Y. VARDI

(Manuscript received July 3, 1980)

The need to assess attendance behavior often arises, at the line-management level, when an employee is considered for a transfer or a promotion. A sound assessment should, of course, take into account the statistical behavior and distributional properties of absenteeism. The first part of this paper is a detailed statistical analysis of attendance records of a sample of 112 telephone operators. We use exploratory and confirmatory statistical techniques to suggest possible theoretical models that can parsimoniously describe the behavior of the variables of interest. Methodological difficulties that often arise in cross-sectional studies and are caused by biased sampling are pointed out and treated. We explore the relation between age and attendance; in particular it is evident that (for this data set) the frequency of "incidental" absences tends to decrease with age, and that the duration of "disability" absences tends to increase with age. In the second part of the paper we suggest an attendance evaluation method based on the statistical analysis of the first part. The method is designed to reflect the current-year attendance as well as a longer-run attendance behavior, interpreted as a personal characteristic, and its properties are demonstrated via examples.

I. INTRODUCTION AND SUMMARY

Management policy regarding absenteeism has two major aspects: a global one spelled out in the various company rules and applied evenly to all employees and a local one, generally less formal, in which line management is concerned about individual's attendance. A question like how many "paid days off" per year an employee should be allowed for unexpected and unavoidable absences is often a subject for union negotiations and is a good example of what we mean by management's global policy. On the other hand, the need to decide whether

a given operator has exhibited satisfactory attendance arises when that operator is considered for a transfer or promotion and is a good example of management's local policy. Whether local or global, a sound policy should consider the statistical characteristics and the distributional properties of absenteeism.

Section II gives a detailed statistical analysis of absenteeism (on the basis of a sample of 112 telephone operators). Such an analysis can enhance our understanding of absenteeism, and can be used as a basis for answering questions of the type described above. For example, the distribution of the duration of incidental absences (Section 2.6) and the frequency of incidental absences per year (Fig. 6, or more generally Section 2.6), can be used to answer how many paid days off per year an employee should be allowed. An answer based on such a statistical analysis is more likely to satisfy the true needs of the average employee than any decision which makes no reference to the distributional properties of absenteeism. (Note that the Bell System's allowance for personal time started after our data were taken.)

In Section III we suggest a method for assessing absenteeism, based on our statistical findings of Section II, and discuss its properties. The analysis of Section 2.4 indicates that one year is too short a period to decide whether an operator is intrinsically "good," "bad," etc., regarding attendance. Thus, if management is interested in assessing attendance as a personal characteristic, the follow-up period needs to be longer than one year. The conflict between the viewpoint that past years' attendance should not affect the present evaluation (for any type of performance rating), and the statistical observation that one year is too short a period to assess attendance, are resolved by basing our evaluation method (Section III) on two indices. One index rates the current year attendance, while the other index rates attendance behavior as a personal characteristic, and it depends on the attendance during the three most recent years.

Various aspects of absenteeism have been studied in recent years (particularly in the fields of labor relations, applied and industrial psychology, and management science). The major contributions of our paper to this area of research, and the relation to other studies, as we see them, are summarized below:

(*i*) We suggest an intuitively appealing method for assessing absenteeism, which reflects the current year attendance, as well as attendance behavior, as personal characteristics. With suitable modifications, the method is adaptable to other occupations.

(*ii*) Often in cross-sectional studies a certain sampling bias is introduced because the sampling is done along the time axis. The detailed analysis of Section 2.4 shows how to identify this bias (and in some instances how to estimate the underlying model in the presence of this

bias). This technique can be of use to other researchers analyzing cross-sectional data. (A more detailed paper devoted entirely to statistical questions that arise in the analysis of this type of data is forthcoming.)

(*iii*) In the course of our analysis in Section II, we use some graphical techniques that are common tools in exploratory data analysis, but are not yet familiar to most social scientists. These tools are useful in the tedious chore of identifying patterns and models in large data sets, and we hope that exposing them to researchers in the social sciences will help make them popular.

(*iv*) Throughout the paper we distinguish between two types of absences, disability and incidental (definitions in Section II). This classification enables us to shed some light on the relation between absenteeism and age. Several authors have tried to relate absenteeism to age and conflicting findings are often reported. Indeed, in a recent study based on a survey of blue-collar production workers, Nicholson et al. (Ref. 1, pp. 319–320) report on a marked inverse relation (especially for male employees) between absence frequency and age which, as they point out, contrasts the conclusions of Porter and Steers[2] (a review of the literature on the subject of absenteeism and turnover) and Cooper and Payne,[3] that absence frequency increases with age. Our data suggest that for telephone operators (all of whom in our sample are females) the truth lies somewhere in the middle. That is, the frequency of incidental absences is higher for younger operators, while the frequency and duration of disability absences is higher for older operators.

For readers who are interested in aspects of absenteeism that are not directly related to this work (such as economic, psychological, etc.), we include a supplementary reference list (which is by no means complete).

II. DATA ANALYSIS AND STATISTICAL MODELING

2.1 Introduction

We distinguish between two types of absences: incidental absences (IA), which are usually short, more frequent, and (to a certain extent) controllable, and disability absences (DA), which are usually long, less frequent, and uncontrollable. Formally, a DA is any absence that lasts six or more days and is due to an illness (an exception is an on-the-job accident in which case the DA period can be shorter than six days); any other absence is defined as an IA. Periods of attendance at work will be referred to as showing up (SU) periods.

Our data are made up of attendance records of 112 New England Telephone operators, for variable periods $t_1, \cdots, t_{112}$. Out of the 112 records, 6 cover approximately 1 year (between 0.8 and 1.4 years), 63

cover approximately 2 years (between 1.6 and 2.4 years), and 43 cover approximately 3 years (between 2.5 and 3.1 years). Here we take a year to be 240 working days. The attendance records in our sample do not usually start, or end, at a beginning of a DA, IA, or SU period and thus two censored (i.e., incomplete) periods typically exist (these are usually SU periods) for each of the 112 records, one at each end of the record. This situation is demonstrated in Fig. 1, which gives a schematic example of an attendance record in our data. Note that holidays, weekends, vacations, etc., have been deleted from the time axis. The large proportion of censored SU periods, among the total number of SU periods, requires special attention and leads to an interesting analysis.

Frequency of absences, duration of absences, duration of SU periods, relations between absence and age, etc., are all parts of the complete picture of "attendance behavior" of operators. We analyze these variables below. In cases where our analysis suggests possible theoretical models that can adequately describe the behavior of the variables in question, we point out these models.

Our analysis suggests that operators older than 35 are different from operators younger than 35 with regard to certain aspects of absence behavior; for the sake of brevity, we refer to the first group as older operators and to the second group as younger operators.

2.2 Duration of IA's

A histogram of the duration of the 560 observed IA's is given in Fig. 2a. A simple theoretical model that fits these data to a remarkable degree of accuracy is

$$P[\text{duration of } IA = j] \equiv P_j = \begin{cases} p & \text{if } j = 1, \\ (1-p)2^{-j+1}, & \text{if } j = 2, 3, \cdots, \end{cases} \quad (1)$$

with $p = \hat{p} = 346/560 = 0.6179$ (note that if $x_1, \cdots, x_n$ is a random sample with a probability density function (pdf) (1), then the maxi-

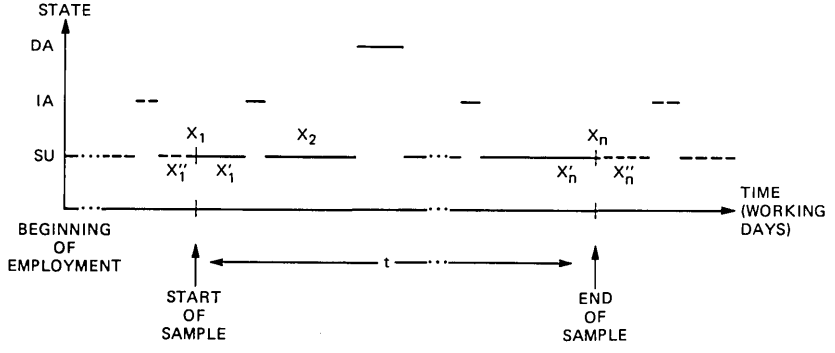
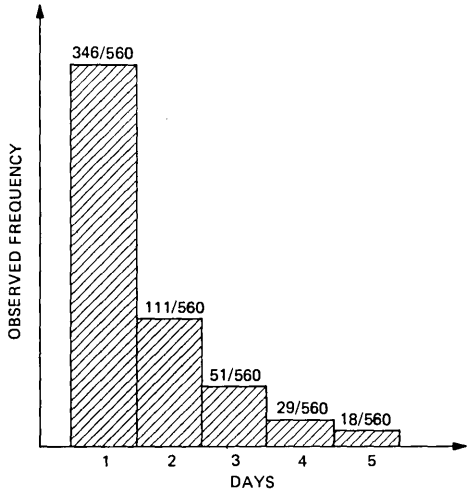


Fig. 1—A schematic description of an attendance record. Note that the first and last SU periods (X_1 and X_n) are censored (only X_1' and X_n' are recorded in the sample).



(a)

DURATION OF IA (DAYS)	1	2	3	4	5	6 OR MORE
OBSERVED FREQUENCY	0.6179	0.1982	0.0911	0.0518	0.0321	0.0089
P_j ACCORDING TO (1)	0.6179	0.1911	0.0955	0.0478	0.0239	0.0239
DIFFERENCE	0	0.0071	−0.0044	0.0040	0.0082	−0.0150

DURATION OF IA (DAYS)	1	2	3	4	5 OR MORE
OBSERVED COUNT	346	111	51	29	23
EXPECTED COUNT, 560 x P_j	346	107	53	28	27
DIFFERENCE	0	4	−2	1	−4

(b)

Fig. 2—(a) A histogram of the durations of IA's (avg = 1.72, stdv = 1.16). (b) Comparison between the pdf of (1) and the observed durations of IA's.

mum likelihood and the minimum variance unbiased estimator of p is $\hat{p} = \sum_{i=1}^{n} I[X_i = 1]/n$, where $I[\ ]$ denotes the indicator function; in our case this gives $\hat{p} = 346/560$). Figure 2b compares the model of (1) with the observed data, and the adequacy of the model is transparent. Nevertheless, it is interesting to note that a chi-square goodness of fit test with size α does not reject the hypothesis that the durations of IA's have the pdf (1), even when α is as high as 0.90!

A random variable, say X, with the pdf (1) has the following interesting property:

$$P[X = k + j \mid X > k] = 2^{-j}, \qquad j = 1, 2, \cdots, \quad k = 1, 2, \cdots. \tag{2}$$

The interpretation of this property, when X stands for the duration of an IA, is the following: On the second day of an incidental absence, the employee tosses a coin; if the result is heads the employee returns to

work on the next day, otherwise he remains absent. The experiment is repeated daily until the first time the result is heads, in which case the employee returns to work on the following day. Should one try to interpret this interesting property, exhibited by the data, in terms of human behavior in regard to short absences?

We remark that the distribution of the duration of IA's for younger operators is approximately the same as for older operators and neither deviate much from (1).

2.3 Duration of DA's

The following are summary statistics for the 78 DA occasions in our data:

lower quartile = 9.0, median = 13.5 upper quartile = 41.0,

mean (with six most extreme observations removed) = 26.1,

standard deviation (with six most extreme observations removed) = 24.2.

Out of the 78 observations, 18 were incurred by operators younger than 35. Figure 3 compares, by means of box plots,[4] the distributions of the duration of DA's in the three different cases; younger operators (18 DA occasions), older operators (60 DA occasions), and the combined

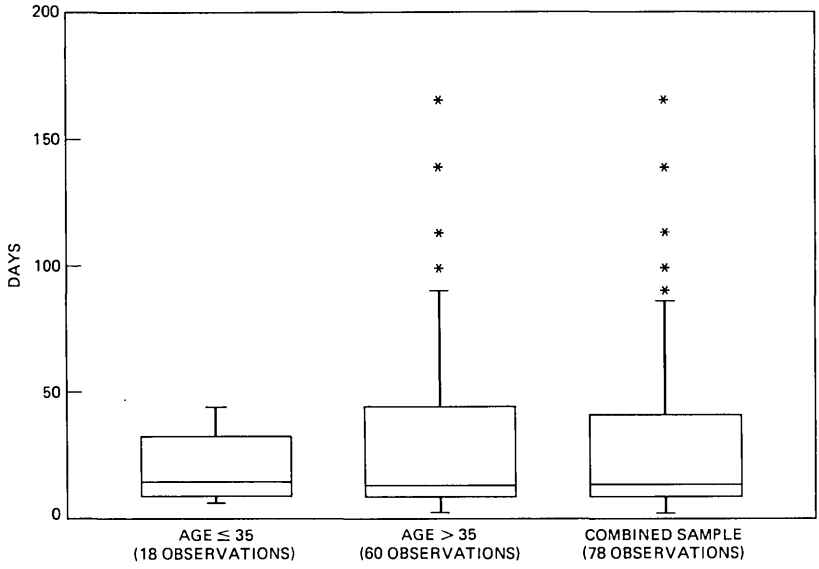


Fig. 3—Box plots for the duration of DA's.

sample (78 DA occasions). The lower and upper sides of each box are the lower and upper quartiles, respectively, and the segment inside the box is the median. If d denotes the distance between the quartiles, then the box whiskers are drawn to the nearest data value with $1.5d$ from the nearest quartile. Points lying outside this range are plotted individually. The figure suggests that long DA's are more frequent among the older operators. For instance, the upper quartile for the older operators is 44.0, with the most extreme observation being 165, while the corresponding figures for the younger operators are 32.0 and 44.0. This difference cannot be accounted for by differences in the total sampling durations, because the distribution of the t_i's is approximately the same for the two age groups.

2.4 Duration of SU periods

We use Fig. 1 as a vehicle to explain some basic concepts regarding the censoring of SU periods. Suppose the length of each individual SU period (X_i in Fig. 1) is distributed according to the cumulative distribution function (cdf) $F(u)$. Then, since the probability of any individual period that covers the point t_0 is directly proportional to its length u, the distribution function of the length of the interval that covers t_0 (X_1, in Fig. 1) is $H(x) = \int_0^x u dF(u)/m$ (m is a normalizing constant that is equal to the mean of F). Given X_1, however, the distribution function of X_1', which is the observable part of X_1, is uniform on the interval $[0, X_1]$ so that (using Bayes' theorem) the unconditional distribution of X_1' is

$$G(y) \equiv P[X_1' \leq y] = m^{-1} \int_0^y [1 - F(u)]\, du. \tag{3}$$

Since (3) is usually derived in the context of renewal processes, in which case an assumption about the independence of different X_i's (SU periods in our application) is built in, it is important to note that this assumption is not used in the derivation of (3) (cf. Ref. 5, p. 66), and therefore it is not assumed in our discussion. In the analysis that follows, however, we assume (unless otherwise stated) that SU periods of different operators have the same cdf F, as long as they are in the same age group.

It is clear that the argument leading to (3) applies also to X_n', so that (3) is the distribution of the censored SU periods. An important property of (3) is

$$G = F \quad \text{if, and only if,} \quad F \text{ is an exponential distribution}$$
$$[\text{i.e.,} \quad F(x) = 1 - e^{-x/\mu}, \quad x > 0, \quad \mu > 0]. \tag{4}$$

Or, in words,

the censored SU's and the completed SU's have the same distribution if, and only if, the distribution of the SU's is exponential. (4′)

2.4.1 Analysis for operators younger than 35

For the younger operators, Fig. 4a compares censored SU's with completed SU's, by means of a Q–Q plot (see Ref. 6, chapter 6). The deviation of the plot from a 45-degree line through the origin is not very large and for practical purposes one can assume that the censored SU's and the completed SU's follow the same distribution. Being more formal, if we test the hypothesis that the two samples have the same distribution, using a Wald-Wolfowitz runs test, we observe 92 runs while the mean and standard deviation under the null hypothesis are 99.0 and 6.0, respectively, so that the hypothesis is not rejected at significance levels of 0.12 or less. Thus, from (4′), we are led to the conclusion that the distribution of the SU's is exponential (or, at least, that this is an adequate description of the data). Figure 4b is a comparison of the combined SU sample (censored and completed) versus quantiles from exponential distribution. The striking closeness to linearity of this plot strongly supports the conclusion that the SU's are exponentially distributed. The estimated mean of the combined sample is 60.2 days, and the standard deviation is 61.2 (which is very close to the mean, as is to be expected from a sample from exponential distribution). In summary, for the purpose of fitting a parsimonious

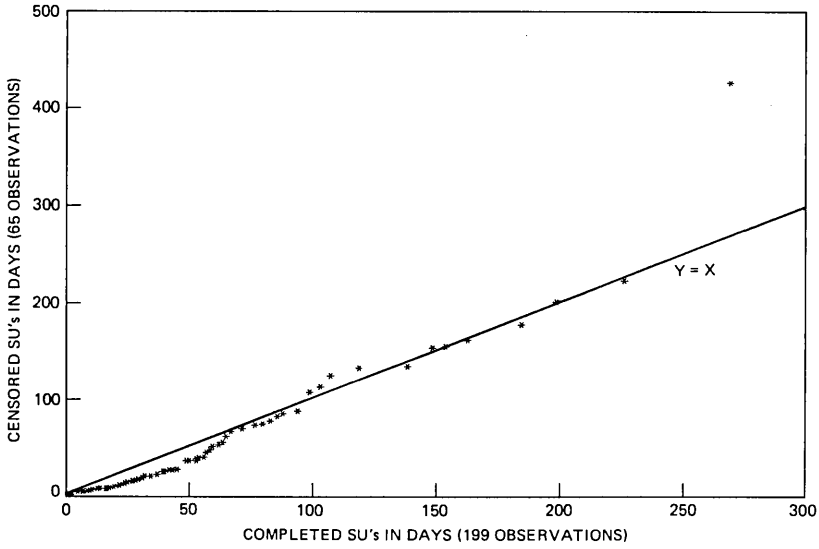


Fig. 4a—Q–Q plot of censored SU's (Y axis, 65 observations) versus completed SU's (X axis, 199 observations), for operators younger than 35.

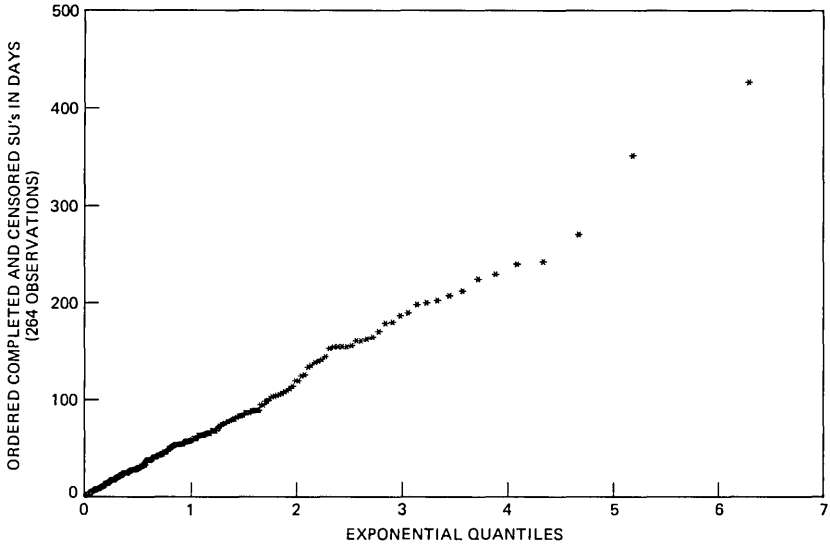


Fig. 4b—Q–Q plot of completed and censored SU periods (Y axis, 264 observations) versus $-\log(1-u)$ (quartiles of standard exponential distribution), for operators younger than 35.

model, one can assume that the periods between consecutive absences, for operators of age 35 or less, follow an exponential distribution with mean = 60 days.

2.4.2 Analysis for operators older than 35

Applying a similar analysis as in the previous case, we point out an interesting "data paradox" exhibited by the two SU samples (censored and completed), and we give possible explanations for this paradoxical behavior of the data.

Figure 5a compares the censored SU's with the completed SU's. The deviation of the Q–Q plot from the 45-degree line through the origin is marked, and it is evident, therefore, that the completed and the censored SU's have different distributions. Specifically, the censored SU's appear to be stochastically bigger than the completed SU's (the Q–Q plot is on or above the 45-degree line through the origin) and for comparison we look also at their summary statistics:

(lower quartile, median, upper quartile, mean, stdv) =

(20.0, 46.0, 88.0, 66.3, 71.2) for the completed SU's, and

(21.0, 47.0, 206.0, 113.3, 130.4) for the censored SU's.

In view of this situation and the assumption that SU's of different operators have the same cdf, (4′) suggests that the distribution of the

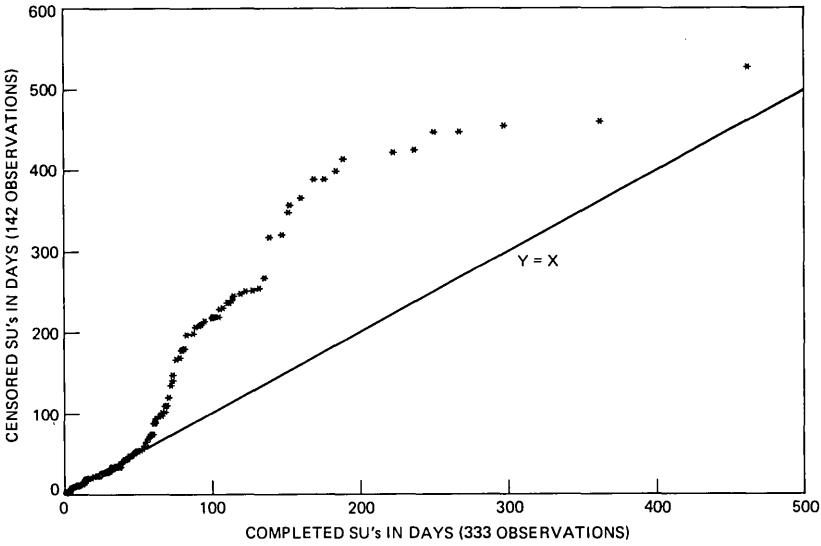


Fig. 5a—Q-Q plot of censored SU's (Y axis, 142 observations) versus completed SU's (X axis, 333 observations), for operators older than 35.

completed SU's cannot be exponential. Figure 5b, however, in which we compare the completed SU's with exponential quantiles, points in the opposite direction. The closeness to linearity of the Q-Q plot (with the exception of the upper 17 points) suggests that the completed SU's do follow an exponential distribution (or perhaps an exponential with a 5-percent contamination).

We give two possible explanations to this data paradox. The first is that intrinsic differences in absence behavior might exist among the 79 operators of age greater than 35, so that any attempt to fit a single cdf to the SU periods of these operators is meaningless [in mathematical language this means that, in eq. (3), different operators are associated with different distribution functions F, while we try to fit a single F], and a more complicated model is needed. One possible model is that operators can be naturally classified into classes according to their attendance behavior (good, bad, etc.). Nevertheless, the sampling periods (t_i's) in our data are not long enough to enable us to decide whether a given operator is intrinsically good, bad, etc., and thus we have not pursued this model.

In the second possible explanation, we show that a certain sampling-bias effect could have been the source of our data paradox. Suppose the cdf of SU periods, which is F of eq. (3), is a mixture of two cdf's, an exponential cdf with mean A, which is small relative to the sampling periods t_i, and a degenerate cdf which assigns a unit mass to a point B

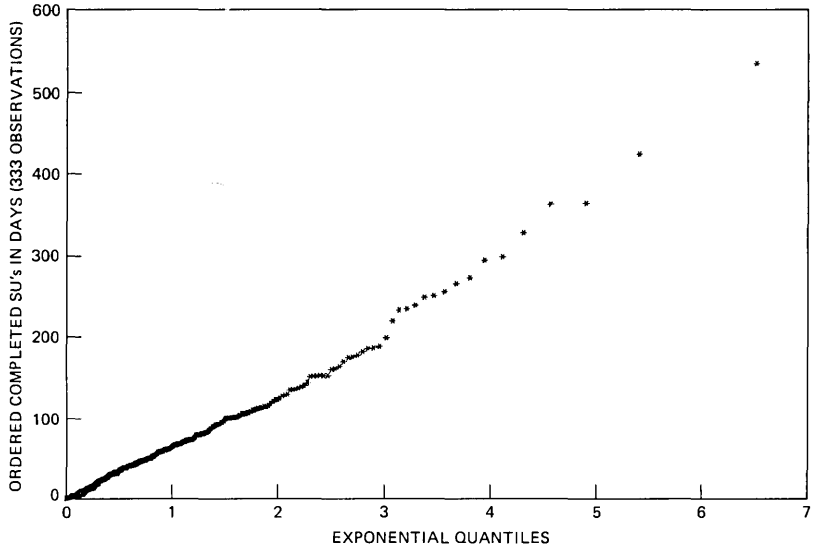


Fig. 5b—Q-Q plot of completed SU's (Y axis, 333 observations) versus $-\log(1-u)$ (quartiles of standard exponential distribution), for operators older than 35.

which is big relative to the t_i's. This means that over a long period of time, a certain proportion, say α, of the SU's have an exponential distribution, while the other SU's last a fixed length B. Now any sampling period of length t satisfying $A \ll t < B$ cannot possibly contain a completed SU period of length B, so that all the completed SU's must be from the exponential population and only censored SU's could possibly be from the B population. In addition to being a model that accommodates the "paradoxical" behavior of our data, this model provides a useful framework for estimation. Under the model's assumptions

$$P[\text{SU} > x] \equiv 1 - F(x) = \alpha e^{-x/A} + (1-\alpha)I[x < B], \tag{5}$$

where $I[\ ]$ denotes the indicator function, so that, using (5) and (3), the moments of the censored SU's are

$$E[\text{censored SU}]^n = \int_0^\infty X^n \, dG(x)$$

$$= \frac{\alpha(n+1)!A^{n+1} + (1-\alpha)B^{n+1}}{(\alpha A + (1-\alpha)B)(n+1)}, \qquad n = 0, 1, \cdots. \tag{6}$$

Since our model assumes $A \ll t$ we have, to a good approximation,

$$E[\text{completed SU}]^n \approx \int_0^t x^n A^{-1} e^{-x/A} \, dx \sim n!A^n, \qquad n = 0, 1, \cdots, \tag{7}$$

and therefore the moments method, applied to (6) and (7), gives

$$\begin{aligned} \hat{E}[\text{completed SU}] = \quad 66.3 \quad &= \hat{A}, \\ \hat{E}[\text{censored SU}] = \quad 113.3 \quad &= \frac{2\alpha\hat{A}^2 + (1-\alpha)\hat{B}^2}{2(\alpha\hat{A} + (1-\alpha)\hat{B})}, \\ \hat{E}[\text{censored SU}]^2 = 29841.05 &= \frac{6\alpha\hat{A}^3 + (1-\alpha)\hat{B}^3}{3(\alpha\hat{A} + (1-\alpha)\hat{B})}, \end{aligned} \tag{8}$$

which yield the estimates $\hat{A} = 66.3$, $\hat{B} = 500.0$, $\hat{\alpha} = 0.96$.

Though the above model accommodates the type of behavior demonstrated by our data, so do other models based on a contaminated exponential distribution and the question of finding a model that fits our data well has not been answered yet. Toward this end we derived a nonparametric estimate of F, denoted $\hat{F}$, by tailoring the Kaplan–Meier estimator[7] to our application, in which each completed observation has to be counted with multiplicity two. (The exact details of

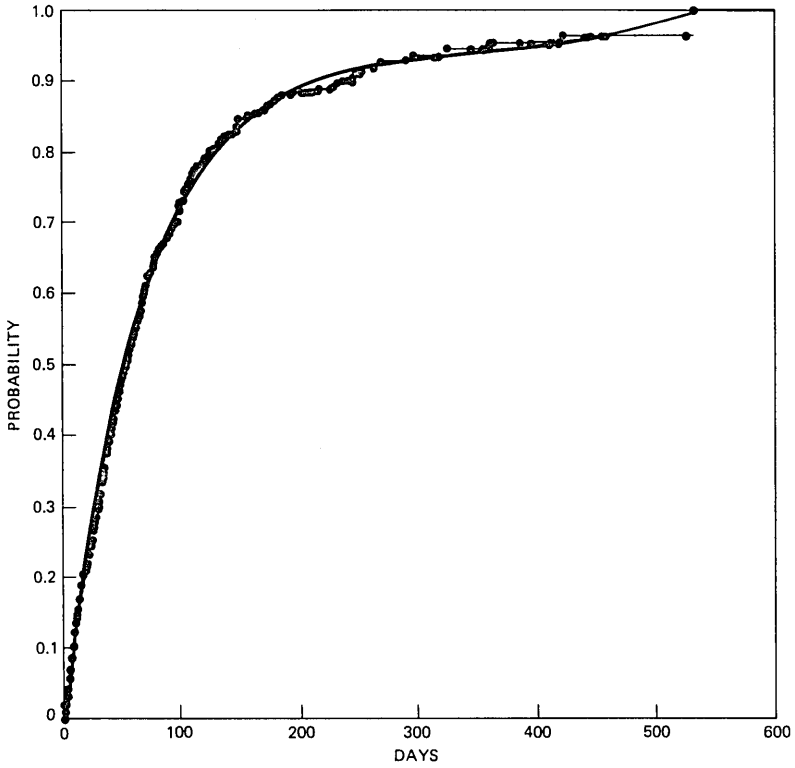


Fig. 5c—The dotted line is $\hat{F}(x) = \hat{P}_{\text{older}}[\text{SU} \leq x]$ (using a modification of the Kaplan–Meier estimator, Section 2.4). The solid line is the contaminated exponential distribution of eq. (1).

this estimator will be discussed in a separate paper.) The result is given in Fig. 5c. The contaminated exponential model

$$F(x) = 0.94(1 - e^{-x/70}) + 0.06\left(\frac{x}{540}\right)^6 I[0 \leq x \leq 540], \qquad x \geq 0, \quad (9)$$

is superimposed on this figure, and it seems to fit the data rather well. Figure 5d compares (9) with $\hat{F}(x)$ by means of a Q-Q plot, and the closeness of the plot to the 45-degree line reassures us about the adequacy of the model.

The behavioral interpretation of (9) is that usually (i.e., 94 percent of the time) the duration of SU periods follows an exponential distribution with mean 70 days, while occasionally (i.e., 6 percent of the time) an SU period can be much longer (perhaps 500 to 540 days). We note that the average SU period, according to (9), is approximately 94 days, which is substantially bigger than the corresponding number for the younger operators (60 days).

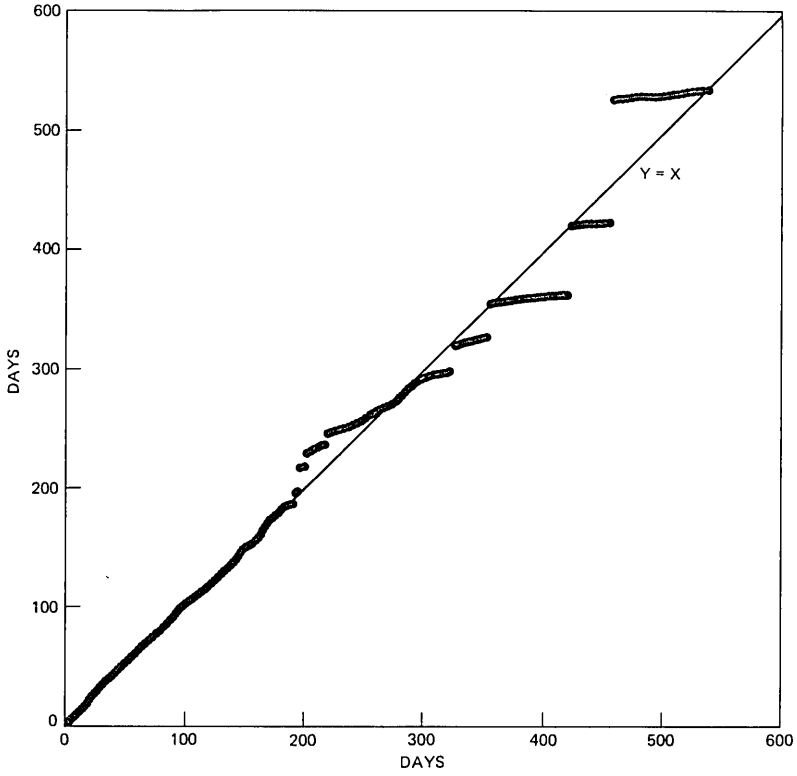


Fig. 5d—A Q-Q plot of $\hat{F}(x) = \hat{P}_{\text{older}}[\text{SU} \leq x]$ (using a modification of the Kaplan-Meir estimator, Section 2.4) versus the contaminated exponential distribution of eq. (10).

An important aspect of our data-paradox, and the contaminated exponential model, is that it indicates that a follow-up period of one or two years is not sufficiently long for evaluating the attendance behavior of operators. This observation has implications to our discussion of evaluation procedures.

2.5 Frequency of absences and total time lost (TTL) due to absences

Figure 6 gives box plots of the frequency of IA's (occasions per year) for the two age groups and for the combined sample. (The nonoverlapping of the notches in the first and second boxes indicates a difference at the rough 5-percent significance level between the two medians.[4]) One can immediately see that younger operators tend to have substantially more IA's. Note again that this difference cannot be accounted for by differences in the total sampling durations, because the distribution of the t_i's is approximately the same for the two age groups.

The situation regarding DA's is somewhat reversed, as one can see from Tables I and II. For example, while the proportion of the younger operators in the sample is 29 percent, the proportion of the DA occasions incurred by them is only 23 percent. We also see in Table II that the ratio "TTL due to DA" to "TTL due to IA" is 0.022/0.021 for younger operators, while it is 0.045/0.014 for older operators. This, plus the fact that the probability distribution of the DA duration for older operators has a substantially longer tail than the corresponding quantity for younger operators (Fig. 3), explains the fact that the TTL

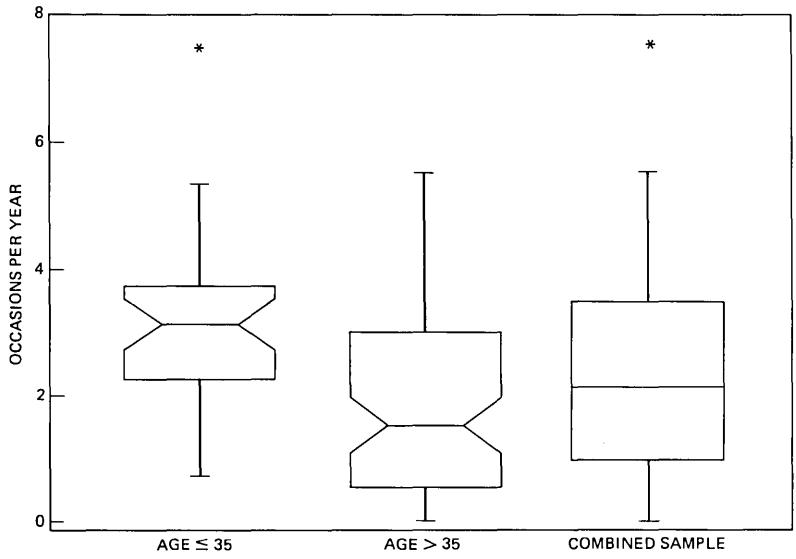


Fig. 6—Box plots for the frequency of IA (occasions per year).

Table I—DA occasions and age

	No. of Individuals with Occasions of DA	No. of Occasions of DA	No. of Operators in the Entire Sample
Age ≤ 35	13 (25%)	18 (23%)	33 (29%)
Age > 35	40 (75%)	60 (77%)	79 (71%)
	53 (100%)	78 (100%)	112 (100%)

caused by absences is somewhat higher for older operators (5.9 percent) than for younger operators (4.2 percent).

The first line of Table II shows that a period that starts at the end of the IA and ends at the end of the following IA (including the possible DA's) lasts, on the average, 72 days for younger operators, and 107 days for older operators.

2.6 The number of IA occasions (IAO) over a fixed period of time

For later applications we want to derive an estimate for the probability distribution of the number of IAO over a fixed period of time, for an arbitrary operator in our sample. To keep the analysis and the presentation simple we ignore, for the time being, the differences between younger and older operators. Later we will comment on the corresponding analysis when the difference in attendance between the two age groups is taken into account. It is well known (e.g., Ref. 5, p. 104) that under fairly weak assumptions about the statistical behavior of the periods between consecutive IA's, the quantity $\sqrt{t}[N(t)b/t]$ has a limiting distribution as $t \to \infty$. Here $N(t)$ denotes the number of IAO over a time period of length t and b is the average number of IAO per unit time. We take this theoretical model as a framework for

Table II—Age comparison of certain absence characteristics (DA = disability absence, IA = incidental absence, TTL = total time lost)

	Age ≤ 35	Age > 35	Combined Sample
$\frac{\text{total sampling periods*}}{\text{total no. of absence occasions}}$	$\frac{16601}{232} = 71.56$	$\frac{43439}{406} = 106.99$	$\frac{60040}{638} = 94.11$
$\frac{\text{no. of DA occasions}}{\text{no. of IA + DA occasions}}$	$\frac{18}{232} = 0.078$	$\frac{60}{406} = 0.148$	$\frac{78}{638} = 0.122$
$\frac{\text{TTL due to IA}}{\text{total sampling periods}}$	$\frac{342}{16601} = 0.021$	$\frac{620}{43439} = 0.014$	$\frac{962}{60040} = 0.016$
$\frac{\text{TTL due to DA}}{\text{total sampling periods}}$	$\frac{363}{16601} = 0.022$	$\frac{1941}{43439} = 0.045$	$\frac{2304}{60040} = 0.038$
$\frac{\text{TTL due to absences}}{\text{total sampling periods}}$	$\frac{705}{16601} = 0.042$	$\frac{2561}{43439} = 0.059$	$\frac{3266}{60040} = 0.054$

* Total sampling periods = the sum of all the observation periods across operators (i.e., sum of the t's of Fig. 1).

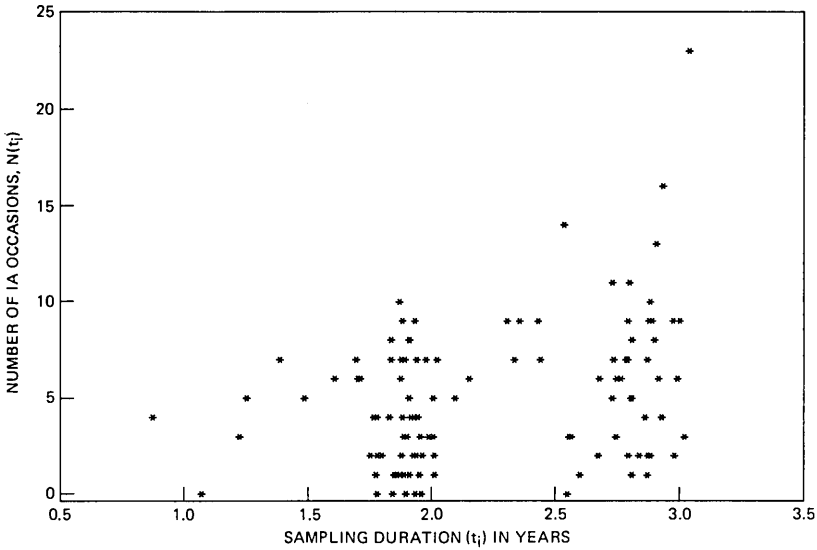


Fig. 7—A scatter plot of the number of IA occasions, $N(t_i)$ (Y axis) versus the sampling duration, t_i years (X axis), for the 112 operators.

producing estimates of the distribution of $N(t)$ for a given t. A scatter plot of $(t_i, N(t_i))$, $i = 1, \cdots, 112$, is given in Fig. 7, and one can see immediately that $VAR(N(t))$ increases with t (this is usually referred to as heteroscedasticity), as to be expected from our model. Note that

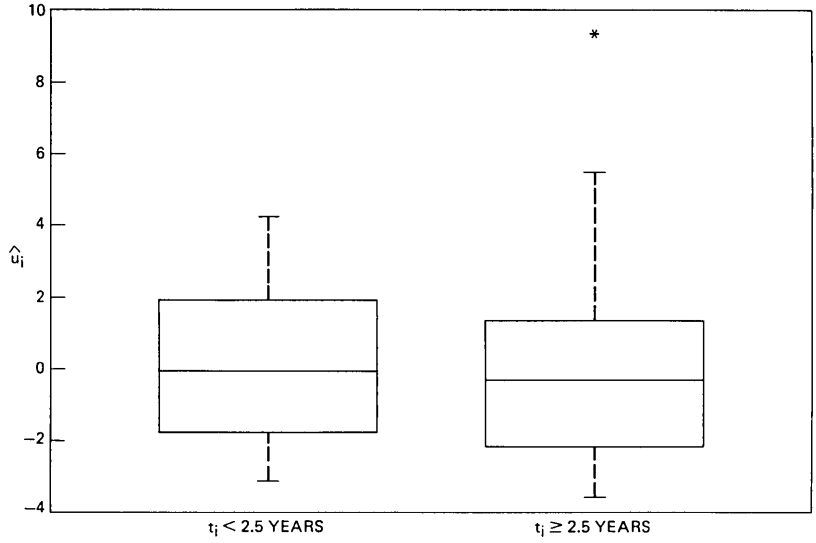


Fig. 8—Comparison of $\hat{u}_i = \sqrt{t_i}[N(t_i)/t_i - \hat{b}]$ for small and large values of t_i.

our model implies that the mean and the variance of $N(t)$ are approximately linear in t, for large values of t. The regression estimate of b in the model

$$N(t_i)/\sqrt{t_i} = b\sqrt{t_i} + U_i \tag{10}$$

is $\hat{b} = 2.24$ (with a t value of 15.36). Figure 8 compares (by means of box plots) the residuals, $\hat{U}_i$'s, of the regression (10) for periods of length $t_i < 2.5$ years with the $\hat{U}_i$'s for periods of length $t_i \geq 2.5$ years. The choice of $t = 2.5$ as a cutoff point seems natural from the distribution of t's (see second paragraph of Section 2.1); other t's in the neighborhood of 2.5 gives similar results. The comparison shows that the heteroscedasticity of the data (Fig. 7) is eliminated and the $\hat{U}_i$'s can be considered as random variables satisfying $VAR\ \hat{U}_i = \sigma_u^2 > 0$, independent of t, so that the empirical distribution of the $\hat{U}_i$'s can be used to estimate the distribution of $N(t)$.

Theoretically, if (i) our assumptions (e.g., all operators behave according to the same probability law) were completely realistic and (ii) t were very large, then the distribution of U would be close to a normal distribution and we could use this fact to estimate the distribution of $N(t)$. Since, however, neither (i) or (ii) is entirely correct, we do not rely on the asymptotic normality of U. Instead we use the empirical distribution of the $\hat{U}_i$'s as an estimate of the distribution of U, and hence obtain an estimate for the distribution of $bt + \sqrt{t}U$. In practice, however, since $N(t)$ is restricted to the nonnegative integers,

Table III—$\hat{P}[N(t) = j]$, estimated probability that the number of IA occasions, over a period of length t, equals j

j \ t	1 year	2 years	3 years
0	0.29	0.09	0.01
1	0.13	0.16	0.08
2	0.14	0.11	0.08
3	0.12	0.07	0.09
4	0.17	0.11	0.11
5	0.08	0.07	0.07
6	0.03	0.10	0.07
7	0.02	0.12	0.10
8	0.02	0.07	0.05
9	0.0	0.04	0.06
10	0.0	0.02	0.13
11	0.0	0.02	0.05
12	0.0	0.02	0.04
13	0.0	0.0	0.02
14	0.0	0.0	0.02
15	0.0	0.0	0.02
Total	1.00	1.00	1.00

we look at

$$m_i(t) = \max\{0, [\hat{b}t + \sqrt{t}\hat{U}_i + \tfrac{1}{2}]\}, \tag{11}$$

where $[x]$ denotes the integer part of x, and we estimate the distribution of $N(t)$ by

$$\hat{P}_t(j) \equiv \hat{P}[N(t) = j] = (\text{number of } m_i(t) = j)/112. \tag{12}$$

Table III gives $\hat{P}_t(j)$ for $t = 1, 2, 3$, years.

Comment: In view of the difference between younger and older operators, it would have been more appropriate to estimate $P[N(t) = j]$ separately for younger and for older operators, and then to use their relative weights in the entire sample to obtain a final estimate. That is,

$$\hat{P}[N(t) = j] = \frac{33}{122}\hat{P}_{\text{younger}}[N(t) = j] + \frac{79}{112}\hat{P}_{\text{older}}[N(t) = j].$$

The actual values of the estimates using this method are close to the values of the estimates we obtained (Table III) without partitioning the sample, and therefore we do not give the details of this calculation. Note that the estimates of (12) are motivated by a model that imposes

Table IV—$\hat{P}[L(t) \leq j]$, estimated probability that the TTL from IA's, over a period of length t, is at most j days

t	1 year	2 years	3 years
0	0.29	0.09	0.01
1	0.37	0.19	0.06
2	0.45	0.26	0.11
3	0.52	0.32	0.16
4	0.60	0.38	0.21
5	0.68	0.43	0.26
6	0.75	0.48	0.31
7	0.81	0.53	0.36
8	0.86	0.58	0.41
9	0.90	0.63	0.46
10	0.93	0.68	0.50
11	0.95	0.73	0.54
12	0.97	0.77	0.58
13	0.98	0.81	0.62
14	0.99	0.84	0.66
15	1.00	0.87	0.70
16		0.89	0.74
17		0.91	0.78
18		0.93	0.81
19		0.95	0.84
20		0.96	0.87
21		0.97	0.90
11		0.98	0.92
12		0.99	0.94
24		1.00	0.96

very few assumptions on the data. Other possible frameworks for estimation, which impose more conditions on the data (e.g., Poisson arrivals of the IA's) result in estimates for which we feel that the assumptions, rather than the data, determine the actual values of the estimates. However, with more absenteeism data (in particular, longer t_i's) it is possible to identify a useful parametric model for estimating $P[N(t) = j]$.

Let $L(t)$ denote the TTL from the $N(t)$ occasions of IA. Clearly

$$L(t) = X_1 + X_2 + \cdots + X_{N(t)}, \tag{13}$$

where X_i denotes the duration of the ith IA. Assuming that the X_i's are independent of $N(t)$, we have

$$P[N(t) = n, L(t) = l] = P\left[\sum_1^n X_i = l\right] P[N(t) = n]. \tag{14}$$

Combining the estimates (12) and (14) with (15), we obtain the joint probabilities of $N(t)$ and $L(t)$ for $t = 1, 2, 3$ years. The marginal distributions of $N(t)$ and $L(t)$ (Tables III and IV, respectively) are then used to construct Tables Va, b, and c, as described below.

2.7 Constructing Tables Va, b, and c

Tables Va, b, and c are the building blocks of our proposed evaluation procedure (Section III) and understanding their construction enables the user to interpret the ratings R_a, R_b, and R_c which make up the attendance evaluation scheme.

To each possible value of $N(1)$, the number of IA's in a single year, and to each possible value of $L(1)$, the total number of days lost in these IA's, we attach a grade and a score. Values of $N(1)$ which lie in the lower 5 percent of the distribution of $N(1)$, which is given in Table

Table Va—Scoring table on the basis of one-year attendance

Number of days	Number of Occasions								
	t \ n	0	1	2	3	4	5	6	
	0	100							E 100
	1		74						G 74
	2		68	55					G 62
	3		60	49	43				F 49
	4		56	46	40	32			F 43
	5		52	43	37	30	23		F 37
	6		48	39	34	27	21	0	F 31
	7		42	34	30	24	18	0	P 24
	8		38	31	27	22	17	0	P 20
	9		34	28	24	20	15	0	P 16
	10		30	24	21	17	13	0	P 12
	11		24	20	17	14	11	0	P 8
	12		0	0	0	0	0	0	U 0
		E	G	F	F	P	P	U	
		100	74	49	37	24	14	0	

Table Vb—Scoring table on the basis of two-years attendance

Number of days t \ Number of Occasions n	0	1	2	3	4	5	6	7	8	9	10		
0	100											E	100
1		94										V	94
2		83	74									G	74
3		81	71	65								G	69
4		78	69	63	56							G	64
5		74	66	60	54	49						G	59
6		71	63	58	51	47	42					G	54
7		68	60	55	49	45	40	34				F	49
8		64	57	52	46	42	38	32	28			F	44
9		61	54	49	44	40	36	31	26	22		F	39
10		57	50	46	41	37	33	29	25	20	0	F	34
11		52	46	42	38	34	31	26	23	19	0	F	29
12		47	42	39	34	31	28	24	21	17	0	P	24
13		45	40	37	33	30	27	23	20	16	0	P	22
14		43	38	35	31	29	26	22	19	15	0	P	20
15		41	36	33	30	27	24	21	18	15	0	P	18
16		39	34	31	28	26	23	20	17	14	0	P	16
17		36	32	29	26	24	21	18	16	13	0	P	14
18		34	30	27	24	22	20	17	15	12	0	P	12
19		31	27	25	22	20	18	15	13	11	0	P	10
20		0	0	0	0	0	0	0	0	0	0	U	0
	E	V	G	G	F	F	F	P	P	P	U		
	100	94	74	62	49	41	33	24	18	12	0		

III, are given the grade Excellent and their scores vary between 100 and 95; values of $N(1)$ which lie between the 6th and the 25th percentile of the distribution of $N(1)$ are given the grade Very Good and their scores vary between 94 and 75, etc. [The particular score depends on how many values of $N(1)$ fall in this range. For example, if only one value of $N(1)$ lies between the 6th and 25th percentile, its score is 94 (e.g., the rightmost column of Table Vb); if there are two values, their scores are 94 and $84 \doteq 94 - (\frac{1}{2})(94 - 75)$ (e.g., the lower-most row of Table Vc); if there are three values, they get the scores 94, $88 \doteq 94 - (\frac{1}{3})(94 - 75)$ and $81 \doteq 94 - (\frac{2}{3})(94 - 75)$, and so on.] We treat $L(1)$ similarly, using the estimated distribution in Table VI. Table VI gives the details of the grading and scoring method, and it is used for $N(t)$ and $L(t)$, $t = 1, 2, 3$.

An exception to Table VI is made when $N(t) = 0$ [and hence $L(t) = 0$], in which case the grade is Excellent and the score is 100 regardless of whether 0 is in the lower 5 percent of the distribution of $N(t)$ [note that $\hat{P}(N(1) = 0) = 0.29$ and $\hat{P}(N(2) = 0.09)$; see Table III]. The scores (and grades) associated with each value of $N(1)$ and $L(1)$ are written on the margins of Table Va and each entry in the body of the table is the geometric mean of the marginal scores; for example, $R_a(N(1) = 3, L(1) = 4) = \sqrt{37 \times 43} \doteq 40$. The reason for choosing the geometric mean to combine the marginal scores is to achieve the desirable shape of the equicontours of the resulting table. More specifically, we observe

the following attractive properties: (*i*) The ratings decrease along the west-east and north-south directions. (*ii*) Each entry in the table is slightly greater than (or equal to) its north-east neighboring entry. (This implies that a reduction in the number of occasions of IA's is desirable even at the expense of a slight increase in the TTL.) (*iii*) An operator is rated Unsatisfactory whenever at least one margin is rated as such.

Tables Vb and Vc are constructed similarly with the obvious substitutions of $(N(2), L(2))$ and $(N(3), L(3))$ for $(N(1), L(1))$.

III. EVALUATING ATTENDANCE AT WORK

3.1 Introduction

The attendance behavior of an operator is one of the most important components in the operator's overall performance, so it is evaluated regularly. In particular, it is weighed very carefully when the operator is considered for a transfer or promotion. So far, however, attendance has been assessed in local terms (compared to other operators in the office) and naturally this is done in a subjective and informal way. Though the informality is an advantage both for management and employees, this is not so for the subjectivity of the evaluation. A

Table Vc—Scoring table on the basis of three-years attendance

		Number of Occasions														
Number of days	t \ n	0	1	2	3	4	5	6	7	8	9	10	11	12	13	
	0	100														E 100
	1		94													V 94
	2		91	86												V 89
	3		89	84	79											V 84
	4		86	81	76	72										V 79
	5		83	79	74	70	66									G 74
	6		81	77	72	68	64	60								G 70
	7		79	74	70	66	62	57	53							G 66
	8		76	72	68	64	60	55	52	48						G 62
	9		74	70	66	62	58	53	50	46	42					G 58
	10		71	67	63	60	56	51	48	45	41	36				G 54
	11		68	64	60	57	53	49	46	43	39	34	30			F 49
	12		65	61	58	54	51	47	44	41	37	33	28	23		F 45
	13		62	59	55	52	49	45	42	39	36	31	27	22	0	F 41
	14		59	56	52	49	46	43	40	37	34	30	26	21	0	F 37
	15		56	53	49	47	44	40	38	35	32	28	24	20	0	F 33
	16		52	49	46	44	41	38	35	33	30	26	23	19	0	F 29
	17		47	45	42	40	37	34	32	30	27	24	21	17	0	P 24
	18		44	42	39	37	32	35	30	28	26	22	19	16	0	P 21
	19		41	39	36	34	32	30	28	26	24	21	18	15	0	P 18
	20		38	35	33	31	29	27	25	24	22	19	16	13	0	P 15
	21		34	32	30	28	26	24	23	21	19	17	15	12	0	P 12
	22		29	27	26	24	23	21	20	18	17	15	13	10	0	P 9
	23		24	22	21	20	19	17	16	15	14	12	10	8	0	P 6
	24		0	0	0	0	0	0	0	0	0	0	0	0	0	U 0
		E	V	V	G	G	G	F	F	F	F	P	P	P	U	
		100	94	84	74	66	58	49	43	37	31	24	18	12	0	

Table VI—Grades and scores for $N(t)$ and $L(t)$

Percentile range	Grade	Score Range
0–5	Excellent	100–95
6–25	Very-good	94–75
26–50	Good	74–50
51–75	Fair	49–25
76–95	Poor	24–5
96–100	Unsatisfactory	0

scheme that allows an objective and consistent evaluation of attendance would be of potential use to line management.

In Section II we studied in detail the statistical aspects of absenteeism. Our analysis, in particular Section 2.4, suggested that if one is interested in attendance behavior as a personal characteristic, then one year is too short for evaluating it. The far past, on the other hand, bears little relevance to recent attendance behavior and thus should not be included in the attendance evaluation. In this section we suggest an evaluation method based on the present and near past (three most recent years) that reflects the current year attendance as well as attendance behavior in a more general sense. We recall, however, from the analysis of Section II that disability absences (DA's) are intrinsically different from incidental absences (IA's). The high variation in the distribution of the duration of DA's and their low frequency of occurrences make it hard to give meaningful statistical guidelines as to what can be considered good, bad, etc., behavior regarding DA's. Furthermore, management can do practically nothing to control DA's. We therefore base our attendance evaluation on IA's only.

In a sensitive issue such as absenteeism from work, the numerical values of the attendance rating do not always tell the whole story. Any method for evaluation might occasionally misjudge good employees, if it is used in a formal and rigid manner. Thus, the best way to avoid these effects is to use it as an informal tool. One has to keep in mind that for every absence there is a reason, and these reasons are not reflected in the formal attendance ratings.

3.2 The evaluation procedure

The proposed scheme is best explained with an example. Consider an operator who started to work on January 1970 and whose IA occurrences and total time lost (TTL) are given in Table VII.

Table VII—Record of IA's

Year	1970	1971	1972	1973	1974	1975	1976	1977	1978
No. of IA occasions	1	4	2	0	0	4	2	0	1
TTL due to IA's	1	6	3	0	0	5	4	0	1

Table VIII—Example of the evaluation scheme (Lines 4, 5, 6 are read off Tables Va, b, c)

	1970	1971	1972	1973	1974	1975	1976	1977	1978
1. $(N(1), L(1))$	(1, 1)	(4, 6)	(2, 3)	(0, 0)	(0, 0)	(4, 5)	(2, 4)	(0, 0)	(1, 1)
2. $(N(2), L(2))$	–	(5, 7)	(6, 9)	(2, 3)	(0, 0)	(4, 5)	(6, 9)	(2, 4)	(1, 1)
3. $(N(3), L(3))$	–	–	(7, 10)	(6, 9)	(2, 3)	(4, 5)	(6, 9)	(6, 9)	(3, 5)
4. R_a	74	27	49	100	100	30	46	100	74
5. R_b	–	45	36	71	100	54	36	69	94
6. R_c	–	–	48	53	84	70	53	53	74
7. R	74	36	44	75	95	51	51	74	81

To evaluate the operator's attendance we propose the following simple procedure:

(*i*) Determine the first three lines of Table VIII as follows: For each year use Table VII to calculate $N(t)$ and $L(t)$ ($t = 1, 2, 3$), the number of IA occasions, and the TTL due to IA's during the t most recent years, respectively.

(*ii*) For each year, read the ratings associated with $(N(1), L(1))$, $(N(2), L(2))$, and $(N(3), L(3))$ from Tables Va, b, and c. These ratings are written in lines 4, 5, and 6 of Table VIII, respectively, and their interpretation [in terms of percentiles of the marginal distributions of $N(t)$ and $L(t)$, $t = 1, 2, 3$] is described in Section 2.7.

(*iii*) Determine line 7 of Table VIII, the attendance index of the jth year, according to the following formula:

$$R_j = \begin{cases} \text{avg}(R_{a,j}, R_{b,j}, R_{c,j}), & \text{if } R_{a,j-1} \geq R_{a,j} \\ \max\{R_{j-1}, \text{avg}(R_{a,j}, R_{b,j}, R_{c,j})\}, & \text{if } R_{a,j-1} < R_{a,j} \end{cases}.$$

For example, in calculating R_{1975} we first compare $R_{a,1975}$ with $R_{a,1974}$. Since $R_{a,1974} = 100 \geq 30 = R_{a,1975}$, we take $R_{1975} = \text{avg}(R_{a,1975}, R_{b,1975}, R_{c,1975}) = (30 + 54 + 70)/3 = 51$. On the other hand, in calculating R_{1976}, comparing $R_{a,1975}$ with $R_{a,1976}$ shows that $R_{a,1975} = 30 < 46 = R_{a,1976}$, so that $R_{1976} = \max\{51, (46 + 36 + 53)/3\} = \max\{51, 45\} = 51$.

(*iv*) The formal evaluation consists of two indices, R_a (line 4) which is the current year rating and R (line 7) which can be considered as an index for attendance behavior (here we view attendance behavior as a personal characteristic of the operator), or in short attendance index.

3.3 Properties of the proposed procedure

(*i*) While the current year rating, R_a, reflects the attendance in the most recent year, the attendance index, R, takes the near past into account, enabling the operator to build up credit. For instance, while 1975 itself was a Fair year ($R_a = 30$), in the example of Table VIII the attendance index, R, for 1975 was Good ($R = 51$). This is due to the perfect attendance during the previous two years. And indeed, if in 1975 this operator was considered for promotion, then the score 51 is a better indicator of her attendance behavior (considered as a personal characteristic) than her current year rating of 30. Similarly, the effect of bad attendance cannot be entirely erased in a single year of perfect attendance, as can easily be seen in the years 1972 and 1973.

By the nature of its definition, R is much smoother than R_a and is a better indicator of attendance. To reemphasize this point, consider an operator whose attendance record fluctuates from $(N(1) = 0, L(1) = 0)$ to $(N(1) = 6, L(1) = 12)$ to $(N(1) = 0, L(1) = 0)$, $\cdots$, etc. The current-year rating then fluctuates from $R_a = 100$ (Excellent) to $R_a = 0$ (Unsatisfactory) while the attendance index fluctuates from $R = 58$ (Good) to $R = 9$ (Poor), which seems more appropriate overall.

(*ii*) If $R_{a,j-1} < R_{a,j}$, then $R_{j-1} \leq R_j$, or, in other words, if the current year rating has improved, then the attendance index will not decrease. This follows from the definition of R and is done to avoid negative reinforcement. The situation is exemplified in moving from 1975 to 1976 in Table VIII. There we have $R_{a,1976} = 46 > 30 = R_{a,1975}$ (an improvement in the current-year rating), so we take $R_{1976} = 51 = R_{1975}$ despite the fact that $\text{avg}(R_{a,1976}, R_{b,1976}, R_{c,1976}) = 45 < 51$.

Since the procedure allows the operator to build up credit, the reverse situation does not hold and one can have $R_{a,j-1} > R_{a,j}$ (deterioration in the current-year rating) with $R_{j-1} < R_j$ (improvement in the attendance index). This is exemplified in the ratings of 1977 and 1978 in Table VIII. And indeed, even though the attendance in 1978 was worse than the attendance in 1977, the period 1976–1978 as a whole reflects better attendance than the period 1975–1977.

IV. REMARKS

(*i*) Though the technical details (such as the length of the periods to be used for rating, and the specific values in Tables Va, b, and c) are tuned to telephone operators (more specifically to our sample), the method itself can be adapted to other occupations. In occupations with substantially higher absence rate, such as auto workers [see, for example, the data collected from 60 blue-collar employees of an automobile-parts foundry, reported in Morgan and Herman (Ref. 8, pp. 739)], periods of 1, 2, and 3 years are too far in the past to affect the current attendance index and should be replaced with shorter periods (e.g., 6, 12, and 18 months).

(*ii*) As pointed out by a Bell Laboratories referee, the choice of the scoring bands in Table VI is somewhat arbitrary, and these bands differ from the HOLU (high, objective, low, unsatisfactory) bands that were recommended by the AT&T Measurements Task Force. Since, however, our main contribution here is the general approach for evaluating attendance (i.e., weighing the recent past in the attendance index) rather than the particular details, we prefer to leave the exposition as is.

(*iii*) In view of the difference between younger and older operators in regard to IA's, note that the proposed scheme is tuned to a population of approximately 30-percent younger operators and 70-percent older operators (as in our sample). This proportion emphasizes the better behavior of the older operators without setting unattainable standards for the younger operators.

V. ACKNOWLEDGMENT

This study was initiated through the encouragement of J. Suzansky of AT&T, who supplied unpublished analyses of previous data which suggested that operator absence behavior might be susceptible to

statistical modeling. I am grateful to AT&T for interesting me in the problem and for their interest in the study. Paul Tukey helped me with the computer, particularly in transcribing the raw data and creating the data base. His help is greatly appreciated. Thanks are also due to Yoav Vardi of Cleveland State University, for bringing some of the references to my attention.

REFERENCES

1. N. Nicholson, C. A. Brown, and J. K. Chadwick-Jones, "Absence from Work and Personal Characteristics," J. Appl. Psychol., *62,* No. 3 (1977), pp. 319–27.
2. L. M. Porter and R. M. Steers, "Organizational, Work, and Personal Factors in Employee Turnover and Absenteeism," Psychol. Bull., *80* (1973), pp. 151–176.
3. R. Cooper and R. L. Payne, "Age and Absence: A Longitudinal Study in Three Firms," Occup. Psychol., *39* (1961), pp. 31–35.
4. R. McGill, J. W. Tukey, and W. A. Larsen, "Variations of Box Plots," The American Statistician, *32,* No. 1 (1978), pp. 12–16.
5. D. R. Cox, *Renewal Theory,* London: Methuen and Co. Ltd., 1962.
6. R. Gnanadesikan, *Methods For Statistical Data Analysis of Multivariate Observations,* New York: Wiley, 1977.
7. E. L. Kaplan and P. Meier, "Nonparametric Estimation from Incomplete Observations," J. Am. Stat. Assoc., *53* (1958), pp. 457–481.
8. L. G. Morgan and J. B. Herman, "Perceived Consequences of Absenteeism," J. Appl. Psychol., *61,* No. 6 (1976), pp. 738–742.

SUPPLEMENTARY REFERENCES

R. W. Beaty and J. R. Beaty, "Longitudinal Study of Absenteeism of Hard Core Unemployed," Psychol. Rep., *36,* No. 2 (1975), pp. 395–406.
H. Behrend and S. Pocock, "Absence and the Individual: A Six-Year Study in One Organization," Int. Labor Rev., *114,* No. 3 (1976), pp. 311–327.
F. Dansereau, Jr., J. A. Alutto, and S. Markham, "An Initial Investigation into the Suitability of Absenteeism Rates as Measures of Performance," Acad. Manage. Proc. (1977), pp. 230–234.
J. N. Hedges, "Absence From Work—Measuring the Hours Lost" (a special labor force report), Mon. Labor Rev. (October 1977), pp. 16–23.
T. A. Jeswald, "The Cost of Absenteeism and Turnover in a Large Organization," in W. C. Hamner and F. L. Schmidt, eds., *Contemporary Problems of Personnel,* Chicago: St. Claire Press, 1974.
G. Johns, "Attitudinal and Nonattitudinal Predictors of Two Forms of Absence from Work," Acad. Manage. Proc. (1978), pp. 69–73.
G. Latham and E. Pursell, "Measuring Absenteeism from the Opposite Side of the Coin," J. Appl. Psychol., *60,* No. 3 (1975), pp. 369–371.
M. G. Miner, "Job Absence and Turnover: A New Source of Data," Mon. Labor Rev. (October 1977), pp. 24–31.
J. Newman, "Predicting Absenteeism and Turnover. A Field Comparison of Fishbein's Model and Traditional Job Attitude Measures," J. Appl. Psychol., *59,* No. 5 (1974), pp. 610–615.
N. Nicholson, "Management Sanctions and Absence Control," Hum. Relat., *29,* No. 2 (1976), pp. 131–151.
E. Pedalino and V. Gamboa, "Behavioral Modification and Absenteeism: Intervention in One Industrial Setting," J. Appl. Psychol. *59,* No. 6 (1974), pp. 694–498.
A. Stecker, "Two Factor Behavior Theory of Industrial Absenteeism," Diss. Abstr. Int. *34* (3-A) (Sept. 1973), p. 951.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 1, January 1981
Printed in U.S.A.

Resource Sharing for Efficiency in Traffic Systems

By D. R. SMITH and W. WHITT

(Manuscript received June 27, 1980)

Experience has shown that efficiency usually increases when separate traffic systems are combined into a single system. For example, if Group A contains 10 trunks and Group B 8 trunks, there should be fewer blocked calls if A and B are combined into a single group of 18 trunks. It is intuitively clear that the separate systems are less efficient because a call can be blocked in one when trunks are idle in the other. Teletraffic engineers and queuing theorists widely accept such efficiency principles and often assume that their mathematical proofs are either trivial or already in the literature. This is not the case for two fundamental problems that concern combining blocking systems (as in the example above) and combining delay systems. For the simplest models, each problem reduces to the proof of an inequality involving the corresponding classical Erlang function. Here the two inequalities are proved in two different ways by exploiting general stochastic comparison concepts: first, by monotone likelihood-ratio methods and, second, by sample-path or "coupling" methods. These methods not only yield the desired inequalities and stronger comparisons for the simplest models, but also apply to general arrival processes and general service-time distributions. However, it is assumed that the service-time distributions are the same in the systems being combined. This common-distribution condition is crucial since it may be disadvantageous to combine systems with different service-time distributions. For instance, the adverse effect of infrequent long calls in one system on frequent short calls in the other system can outweigh the benefits of making the two groups of servers mutually accessible.

I. INTRODUCTION AND SUMMARY

From extensive experience in teletraffic engineering, it is well known that congestion can often be reduced by sharing resources. The block-

ing probability in a loss system and the average waiting time in a delay system are usually much less when separate facilities serving separate streams of traffic are combined to serve all the streams together. Alternatively, for a given level of congestion, fewer facilities are usually required to serve the streams together. Sometimes such results are trivial: Whenever the combined system may be managed as if it were in fact separate systems, the optimal performance of the combined system is at least as good as that of the separate systems. However, such management is not allowed in the models treated here. In any case, the efficiency of shared resources is certainly a fundamental principle of teletraffic engineering.

The purpose of this paper is to establish versions of this efficiency principle mathematically. Our first two results verify conjectures by Arthurs and Stuck.[1] To state our first result, let $L(s, \lambda\ \mu)$ denote the stationary loss or overflow rate in an $M/M/s$ loss system (no waiting room) with s servers, arrival rate λ, and individual service rate μ. (See Kleinrock[2] for background on the queuing models.) It is well known that $L(s, \lambda, \mu) = \lambda B(s, a)$, where $a = \lambda/\mu$ and $B(s, a)$ is the familiar Erlang blocking formula:

$$B(s, a) = (a^s/s!)/\sum_{k=0}^{s} (a^k/k!); \tag{1}$$

see Jagerman[3] and references there. The first efficiency principle we establish says that $L(s, \lambda, \mu)$ is a subadditive function of (s, λ) for each fixed μ:

Theorem 1: For all positive integers s_1 and s_2 and all positive real numbers λ_1, λ_2 and μ,

$$L(s_1 + s_2, \lambda_1 + \lambda_2, \mu) \leq L(s_1, \lambda_1, \mu) + L(s_2, \lambda_2, \mu). \tag{2}$$

This yields immediately that

$$G\left(\sum_{i=1}^{n} s_i, \sum_{i=1}^{n} a_i\right) \leq \sum_{i=1}^{n} G(s_i, a_i)$$

for each integer n, where $G(s, a) = aB(s, a)$, which is the version of Theorem 1 actually conjectured by Arthurs and Stuck.[1]

Of course, Theorem 1 should not surprise teletraffic engineers, since it can be inferred from common tables and graphs, but it has apparently not been proved before. It appears that all previous mathematical results can be described as "one-parameter" results. The relation (2) has been deduced for special cases in which quantities such as the blocking probability or the load per server are held constant. For example, it is known that if one combines separate groups of j and k trunks, each operating at a blocking probability 0.01, then the new blocking probability will be less than 0.01 (or, alternatively, the combined system can handle an increased total load and retain the same

0.01 blocking probability); see p. 68 of Cooper.[4] Such results are often presented without rigorous mathematical support.

From Paul Burke we learned about another special case that has been known for a long time. It is not difficult to show that $B(ts, ta)$ is strictly decreasing in t (see the appendix), from which (2) easily follows in the case $\lambda_1/s_1 = \lambda_2/s_2$. Herbert Shulman has also shown that Theorem 1 follows easily from the monotonicity of $B(ts, ta)$ in t and the convexity of $B(s, a)$ in s for $s \geq 1$, but such convexity has not yet been established (see the appendix). For further discussion of other related work, see Section 5.1 of Kleinrock.[5]

To state our second result, let $D(s, \lambda, \mu)$ denote the mean steady-state delay in an $M/M/s$ queue with infinite waiting room, FCFS (first-come, first-served) queue discipline, s servers, arrival rate λ, and individual service rate μ. It is well known that $D(s, \lambda, \mu) = C(s, \lambda/\mu)/(s\mu - \lambda)$, where $C(s, a)$ is the Erlang delay function:

$$C(s, a) = \frac{a^s/(s-1)!\,(s-a)}{\sum_{k=0}^{s-1} (a^k/k!) + a^s/(s-1)!\,(s-a)}. \tag{3}$$

The following result establishes subadditivity of $D(s, \lambda, \mu)$ as a function of (s, λ) for each fixed μ. Note that

$$[\lambda_1/(\lambda_1 + \lambda_2)]D(s_1, \lambda_1, \mu) + [\lambda_2/(\lambda_1 + \lambda_2)]D(s_2, \lambda_2, \mu)$$

is the overall average delay experienced in the separate systems because $\lambda_1/(\lambda_1 + \lambda_2)$ is the long-run proportion of customers to enter the first system.

Theorem 2: For all positive integers s_1 and s_2 and all positive real numbers λ_1, λ_2, and μ,

$$D(s_1 + s_2, \lambda_1 + \lambda_2, \mu) \leq [\lambda_1/(\lambda_1 + \lambda_2)]D(s_1, \lambda_1, \mu) + [\lambda_2/(\lambda_1 + \lambda_2)]D(s_2, \lambda_2, \mu). \tag{4}$$

This yields immediately that

$$H\left(\sum_{i=1}^{n} s_i, \sum_{i=1}^{n} a_i\right) \leq \sum_{i=1}^{n} H(s_i, a_i)$$

for each integer n, where $H(s, a) = aC(s, a)/(s - a)$, which is the version of Theorem 2 actually conjectured by Arthurs and Stuck.[1]

In order to prove Theorems 1 and 2, we found it convenient to prove stronger results. It is helpful to see how the loss rate $L(s, \lambda, \mu)$ and the mean delay $D(s, \lambda, \mu)$ are related to the steady-state number of customers in the system, say Q. In the $M/M/s$ loss system

$$L(s, \lambda, \mu) = \lambda - \mu EQ \tag{5}$$

because λ is the arrival rate and μEQ is the service-completion rate;

the loss rate is that fraction of the arrivals that are not served. In the $M/M/s$ delay system

$$\lambda(D(s, \lambda, \mu) + \mu^{-1}) = EQ \tag{6}$$

by virtue of the fundamental relation $L = \lambda W$; see Stidham.[6]

Let Q_i be the steady-state number of customers in the ith system ($i = 1, 2$) and let Q be the steady-state number of customers in the combined system. Then Theorem 1 is equivalent to

$$EQ \geq EQ_1 + EQ_2 \tag{7}$$

for the loss systems, and Theorem 2 is equivalent to

$$EQ \leq EQ_1 + EQ_2 \tag{8}$$

for the delay systems.

Instead of comparing the means in (7) and (8), we prove Theorems 1 and 2 by making more general stochastic comparisons. We do this in two different ways. Our first method of proof is to compare the distribution of Q with the distribution of $Q_1 + Q_2$. It turns out to be very easy to establish an appropriate ordering for the entire distributions, which in turn implies the desired inequality for the means. The appropriate order is the monotone likelihood-ratio ordering. We define this ordering and prove the more general theorems implying Theorems 1 and 2 in Section II.

Our second method of proof is to compare entire stochastic processes rather than just stationary distributions. As corollaries we obtain stochastic-order relations for the stationary distributions which in turn also imply the desired inequalities (7) and (8) for the means. This approach has the advantage that the arrival processes can be arbitrary rather than Poisson and the service-time distributions can be general instead of exponential. The argument is also remarkably simple. The idea in this approach is to construct artificially the two stochastic processes being compared on the same probability space. The construction is carried out so that each stochastic process individually has the correct distribution (family of finite-dimensional distributions) as originally specified. We choose a special joint distribution so that each sample path of one process always lies below the corresponding sample path of the other process. Because the construction is artificial, the joint distribution of the two processes is not directly meaningful, but it implies a strong stochastic ordering for the processes. Such special constructions have been used previously to compare queuing processes; see Sonderman,[7] Whitt,[8] Wolff,[9] and references there. In fact, the generalization of Theorem 2 is a direct consequence of Wolff's theorem and the other proofs involve similar reasoning. We present our results using this approach in Section III.

In Theorems 1 and 2 we assume equal service rates in the two systems. It is natural to ask whether extensions of (2) and (4) hold when the service rates are unequal. In Section IV we show that, with unequal service rates, combining resources need not be more efficient; in fact it can substantially degrade performance. Infrequent "bad" customers from one system can adversely affect a large number of "good" customers from the other system.

II. MONOTONE LIKELIHOOD-RATIO COMPARISONS

Let X and Y be random variables assuming values in the nonnegative integers. We say X is less than or equal to Y in the monotone likelihood-ratio ordering and write $X \leq_r Y$ if

$$\frac{P(X = k+1)}{P(X = k)} \leq \frac{P(Y = k+1)}{P(Y = k)} \tag{9}$$

for all integers k; see page 208 of Ferguson.[10] We say X is stochastically less than or equal to Y and write $X \leq_{st} Y$ if $Ef(X) \leq Ef(Y)$ for all nondecreasing real-valued functions f for which the expectations are well defined. Obviously, $EX \leq EY$ whenever $X \leq_{st} Y$. What is important for Theorems 1 and 2 is that $X \leq_r Y$ implies $X \leq_{st} Y$. This is well known and not difficult to show. In fact, the monotone likelihood-ratio ordering is equivalent to stochastic order for all conditional distributions obtained by conditioning on subsets, i.e., $E(f(X) \mid X \in A) \leq E(f(Y) \mid Y \in A)$ for all subsets A and all nondecreasing real-valued functions f; this property is discussed in Whitt[11,12]; see Keilson and Sumita[13] for additional material.

Returning to the notation of (7) and (8), we obtain the following results which imply Theorems 1 and 2.

Theorem 3: For the M/M/s loss systems,

$$Q_1 + Q_2 <_r Q.$$

Theorem 4: For the M/M/s delay systems,

$$Q <_r Q_1 + Q_2.$$

Theorems 3 and 4 can each be proved by simple calculations since the stationary distributions are known and easy to work with. To illustrate, we do one proof.

Direct Proof of Theorem 3: Let $a_i = \lambda_i/\mu_i$ for $i = 1, 2$. Then, using convolution, we obtain for some constant C

$$P(Q_1 + Q_2 = k+1) = C \sum_{\substack{0 \leq i_1 \leq s_1 \\ 0 \leq i_2 \leq s_2 \\ i_1 + i_2 = k+1}} a_1^{i_1} a_2^{i_2} / i_1! i_2!$$

$$= \frac{Ca_1}{k+1} \sum_{\substack{1 \le i_1 \le s_1 \\ 0 \le i_2 \le s_2 \\ i_1 + i_2 = k+1}} a_1^{i_1 - 1} a_2^{i_2} / (i_1 - 1)! i_2!$$

$$+ \frac{Ca_2}{k+1} \sum_{\substack{0 \le i_1 \le s_1 \\ 1 \le i_2 \le s_2 \\ i_1 + i_2 = k+1}} a_1^{i_1} a_2^{i_2 - 1} / i_1! (i_2 - 1)!$$

$$< \left(\frac{a_1 + a_2}{k+1} \right) C \sum_{\substack{0 \le i_1 \le s_1 \\ 0 \le i_2 \le s_2 \\ i_1 + i_2 = k}} a_1^{i_1} a_2^{i_2} / i_1! i_2!$$

$$= \frac{P(Q = k+1)}{P(Q = k)} P(Q_1 + Q_2 = k). \qquad \square$$

It is also significant that both Theorems 3 and 4 can be viewed as trivial corollaries of a more general theorem. This more general theorem is especially useful for comparisons when the limiting distributions are not known. To state our general result, consider two stochastic processes on the integers, $Y_1(t)$ and $Y_2(t)$, that move only by jumps up or down in unit steps to one of the neighboring states. Let all the transitions be governed by birth-and-death rates, but in contrast to those in birth-and-death processes, these rates may depend on information other than the current state such as the history of the process or other relevant variables. Let $\lambda_i(k, I_t)$ and $\mu_i(k, I_t)$ be the birth-and-death rates, respectively, for the ith process ($i = 1, 2$) in state k with additional information I_t at time t. By having transitions governed by birth-and-death rates, we mean that

$$P(Y_i(t+h) = k+1 \mid Y_i(t) = k, I_t) = h\lambda_i(t, I_t) + o(h),$$

$$P(Y_i(t+h) = k-1 \mid Y_i(t) = k, I_t) = h\mu_i(t, I_t) + o(h),$$

and

$$P(Y_i(t+h) = k \mid Y_i(t) = k, I_t) = 1 - h[\lambda_i(t, I_t) + \mu_i(t, I_t)] + o(h),$$

where $o(h)$ means a quantity that converges to zero after division by h as $h \to 0$. Let X_1 and X_2 be random variables with the limiting distributions of these two stochastic processes, which we assume exist as proper distributions. Here is our general monotone likelihood-ratio comparison result.

Theorem 5: Consider the processes $Y_1(t)$ and $Y_2(t)$ defined above. Suppose there exist sequences of constants $\{\alpha_i(k))\}$ and $\{\beta_i(k))\}$ such that

$$\lambda_1(k, I_t) \le \alpha_1(k), \qquad \lambda_2(k, I_t) \ge \alpha_2(k),$$
$$\mu_1(k, I_t) \ge \beta_1(k), \qquad \mu_2(k, I_t) \le \beta_2(k),$$

for all k and I_t. If $\alpha_1(k)/\beta_1(k+1) \leq \alpha_2(k)/\beta_2(k+1)$ for all k, then $X_1 \leq_r X_2$.

Corollary: If

$$\lambda_1(k, I_t) \leq \lambda_2(k, I_t')$$

and

$$\mu_1(k, I_t) \geq \mu_2(k, I_t')$$

for all k, I_t, and I_t', then $X_1 \leq_r X_2$.

Proof of Theorem 5: Look at the stationary flow between states k and $k+1$. The flow from k to $k+1$ is less than or equal to $P(X_1 = k)\alpha_1(k)$ for process 1 and greater than or equal to $P(X_2 = k)\alpha_2(k)$ for process 2. Similarly, the stationary flow from $k+1$ to k is greater than or equal to $P(X_1 = k+1)\beta_1(k)$ in process 1 and less than or equal to $P(X_2 = k+1)\beta_2(k)$ in process 2. Since the stationary flow from k to $k+1$ must equal the stationary flow in the opposite direction,

$$P(X_1 = k)\alpha_1(k) \geq P(X_1 = k+1)\beta_1(k+1)$$

and

$$P(X_2 = k)\alpha_2(k) \leq P(X_2 = k+1)\beta_2(k+1).$$

Consequently,

$$\frac{P(X_1 = k+1)}{P(X_1 = k)} \leq \frac{\alpha_1(k)}{\beta_1(k+1)} \leq \frac{\alpha_2(k)}{\beta_2(k+1)} \leq \frac{P(X_2 = k+1)}{P(X_2 = k)}. \qquad \square$$

We can now apply the corollary to Theorem 5 to prove Theorems 3 and 4.

Second Proof of Theorem 3: Note that the processes depicting the number of customers being served satisfy the hypotheses of Theorem 5. In the case of two separate facilities, the sum is not a birth-and-death process because the rates depend not only on the total number but how many are in the individual facilities. When k customers are present, the death (service) rates are identical, but the birth (arrival) rates can be higher in the combined system because if one of the separate facilities is full, then it cannot accept any more arrivals. Hence, the hypotheses of the corollary to Theorem 5 are satisfied. $\square$

Proof of Theorem 4: Again we apply the corollary to Theorem 5. The reasoning is similar except here when k customers are present, the birth (arrival) rates are always identical, but the death (service) rates can be less with the separate facilities because there can be idle servers in one facility while there are customers waiting in the other facility. $\square$

III. SAMPLE PATH COMPARISONS

Let $\{X(t), t \geq 0\}$ and $\{Y(t), t \geq 0\}$ be real-valued stochastic processes. We call a real-valued function f defined on the space of all sample paths of $X(t)$ and $Y(t)$ nondecreasing if $f(\{x(t), t \geq 0\}) \leq f(\{y(t), t \geq 0\})$ for all sample paths $\{x(t), t \geq 0\}$ and $\{y(t), t \geq 0\}$ such that $x(t) \leq y(t)$ for all $t \geq 0$. We say the stochastic process $\{X(t), t \geq 0\}$ is stochastically less than or equal to the stochastic process $\{Y(t), t \geq 0\}$ and write $\{X(t), t \geq 0\} \leq_{st} \{Y(t), t \geq 0\}$ if $f(\{X(t), t \geq 0\}) \leq_{st} f(\{Y(t), t \geq 0\})$ for all nondecreasing real-valued functions f defined on the sample paths of $X(t)$ and $Y(t)$. Clearly, stochastic order of the processes implies $X(t) \leq_{st} Y(t)$ for each t [just use the projection: $f(\{x(u), u \geq 0\}) = x(t)$], but it is much stronger, applying to many other nondecreasing functionals. In fact, since the queuing processes have sample paths with left and right limits everywhere, stochastic order of the processes is equivalent to stochastic order for all finite-dimensional (joint) distributions; see Section 4 of Kamae, Krengel, and O'Brien.[14] Moreover, stochastic order of the processes here is equivalent to the possibility of a strong sample-path comparison. In particular,

$$\{X(t), t \geq 0\} \leq_{st} \{Y(t), t \geq 0\}$$

holds if and only if it is possible to construct stochastic processes $\{\tilde{X}(t), t \geq 0\}$ and $\{\tilde{Y}(t), t \geq 0\}$ on a common probability space such that $\{\tilde{X}(t), t \geq 0\}$ has the same distribution as $\{X(t), t \geq 0\}$, $\{\tilde{Y}(t), t \geq 0\}$ has the same distribution as $\{Y(t), t \geq 0\}$, and every sample path of $\{\tilde{X}(t), t \geq 0\}$ lies below the corresponding sample path of $\{\tilde{Y}(t), t \geq 0\}$; see Theorem 1 of Kamae, Krengel, and O'Brien.[14] What we do is apply the easy half of this equivalence—the fact that the sample-path construction implies stochastic order—to make stochastic comparisons between the queuing processes. The proofs here are done by actually constructing processes with the sample-path ordering. Previous uses of such constructions appear in Sonderman,[7] Whitt,[8] Wolff,[9] and references therein. The approach is also closely related to the so-called "coupling" techniques; see Lindvall[15] and references therein.

We begin with the generalization of Theorem 2 for delay systems because it follows directly from Wolff.[9] As before, we assume the FCFS discipline, but now we allow the arrival streams in the two separate systems to be arbitrary. We assume the service times are independent of the arrival processes and mutually independent and identically distributed, but they need not be exponentially distributed. Since the arrival process is assumed to be independent of the service-time sequence, the evolution of the arrival process cannot depend on the state of the system. This excludes finite-source models, for which counterexamples to the efficiency of sharing are easy to construct; for

example, see page 1377 of Beneš.[16] Let $Q_i(t)$ be the number of customers in the ith system and let $Q(t)$ be the number of customers in the combined system at time t.

Theorem 6: (Wolff) If $Q_1(0) = Q_2(0) = Q(0) = 0$, *then* $\{Q(t), t \geq 0\} \leq_{st} \{Q_1(t) + Q_2(t), t \geq 0\}$.

Remarks: (*i*) Wolff[9] was actually interested in comparing the FCFS discipline with the cyclic assignment discipline in a single delay-system. He showed that the queue length process with the FCFS discipline is stochastically less than the queue length process in the same system with any other discipline. This result applies here because the two separate facilities can be interpreted as a single system with a special queue discipline: Just label the arrivals in the special system according to the stream from which they came and then assign them according to the FCFS discipline to one of the servers in the corresponding subgroup of servers.

(*ii*) We can obtain corresponding results if the systems are not empty initially. For more general initial conditions, we can assume appropriate stochastic order for the residual service times at $t = 0$.

(*iii*) Wolff[9] also obtained similar comparison results for other processes, all of which hold here too: the departure epochs, the number of customers in queue, the total work (in service time) in the system, and the total work in queue. By the sample-path construction, the stochastic order jointly holds for all these processes. See Theorem 8 here.

(*iv*) As a consequence of Theorem 6, $Q(t) \leq_{st} Q_1(t) + Q_2(t)$ for each t. With the general assumptions here, steady-state distributions need not exist, but if $Q_i(t)$ and $Q(t)$ converge in law to Q_i and Q, respectively, as $t \to \infty$, then $Q \leq_{st} Q_1 + Q_2$; see Proposition 3 of Kamae, Krengel, and O'Brien.[14] The convergence of course holds in the setting of Theorem 2, so Theorem 6 implies (8) and thus Theorem 2.

(*v*) Since Theorem 2 concerns the mean-waiting time, it is natural to ask if the steady-state waiting-time distribution is also stochastically less in the combined system. Unfortunately, in general it is not. The counterexample in Whitt[17] applies here too; the cyclic discipline there can be interpreted as arrivals to separate facilities.

(*vi*) When the arrival streams are not Poisson, which we now permit, a new phenomenon occurs. Then the customers in the different streams experience different congestion when the systems are combined, even if the service times are independent and identically distributed. This phenomenon can be an important consideration in combining systems, but we do not consider it here; it has been studied by Kuczura.[18,19]

We now turn to our generalization of Theorem 1 for loss systems. In addition to allowing arbitrary arrival streams and general service-time distributions, we allow a finite waiting room. The number of waiting spaces in the combined system is the sum of the numbers of waiting

spaces in the separate systems. Let $N_i(t)$ $[N(t)]$ be the number of customers lost in the interval $(0, t)$ in the ith separate system (in the combined system); let $S_i(t)$ $[S(t)]$ be the number of service completions in the interval $(0, t)$ in the ith separate system (in the combined system); and let $C_i(t)$ $[C(t)]$ be the amount of work performed—service given—in the interval $(0, t)$ in the ith separate system (in the combined system).

Theorem 7: If $Q_1(0) = Q_2(0) = Q(0) = 0$ *in these systems with finite waiting rooms, then*

$$\{N(t), t \geq 0\} \leq_{st} \{N_1(t) + N_2(t), t \geq 0\},$$

$$\{S(t), t \geq 0\} \geq_{st} \{S_1(t) + S_2(t), t \geq 0\},$$

and

$$\{C(t), t \geq 0\} \geq_{st} \{C_1(t) + C_2(t), t \geq 0\}.$$

Now assume that $N_i(t)/t$ and $N(t)/t$ converge (either in probability or with probability one) as $t \to \infty$. Let the limits be denoted $L(s_i, k_i, A_i(t), F)$ and $L(s_1 + s_2, k_1 + k_2, A_1(t) + A_2(t), F)$, respectively, with k_i denoting the number of waiting spaces, $A_i(t)$ the arbitrary arrival process and $F(x)$ the general service-time c.d.f. From Theorem 7 we immediately obtain the following generalization of Theorem 1.

Corollary: For all positive integers s_1, s_2, k_1 *and* k_2; *all arrival processes* $A_1(t)$ *and* $A_2(t)$; *and all service time c.d.f.'s* $F(x)$ *such that the loss-rate limits exist,*

$$L(s_1 + s_2, k_1 + k_2, A_1(t) + A_2(t), F) \leq L(s_1, k_1, A_1(t), F) + L(s_2, k_2, A_2(t), F).$$

To prove Theorem 7, we establish a finite-waiting-room generalization of Wolff's[9] comparison theorem. Following Wolff, we shall state the result in terms of the sample-path comparison. Since the joint distribution of the two systems being compared is artificially obtained, the appropriate conclusion is the general stochastic order as in Theorems 6 and 7.

We carry out the artificial construction by letting the systems being compared have identical arrival processes and service times. Note that we are now focusing on a single (arbitrary) sample path. We let the nth service time v_n be associated with the nth customer to enter service in each system rather than the nth arrival. Let a_n be the arrival epoch of the nth arrival, $0 \leq a_1 \leq a_2 \leq \cdots$. We assume there are s servers operating in parallel and k extra waiting spaces in both systems. We also assume the systems are initially empty.

One system, called the original system, will be the conventional system where the servers are fed by a single queue using a FCFS discipline. Moreover, there are k extra waiting spaces and arriving

customers enter the system if the number of customers in the system is less than $s + k$, and are lost otherwise.

The other system, called the modified system, is any alternative to the original system which assigns customers to servers in some manner, independent of the sequence of service times $\{v_n\}$, and which loses arrivals whenever the system is full and in some manner otherwise.

Let a_n be the arrival epoch of the nth customer. For the original system, let t_n be the time that the nth customer to enter the system enters; obviously $t_n = a_k$ for some k, $k \geq n$. Also, for the original system, let b_n be the time the nth customer to begin service begins and let d_n be the nth ordered departure epoch from the system. Let a_n, t'_n, b'_n, and d'_n be the corresponding quantities for the modified system.

Theorem 8: For all integers n, $t_n \leq t'_n$, $b_n \leq b'_n$, and $d_n \leq d'_n$.

Proof: The sets of unordered departure epochs in the two systems are clearly $\{(b_n + v_n)\}$ and $\{(b'_n + v_n)\}$, respectively. For the original system,

$$d_1 = \min_{1 \leq i \leq s} \{b_i + v_i\} = \min_{i \leq i \leq s} \{t_i + v_i\},$$

$$d_n = n\text{th-order statistic from } \{(b_i + v_i)\colon i < n + s\} \tag{10}$$

and

$$b_n = \max\{t_n, d_{n-s}\}, \qquad n \geq 1, \tag{11}$$

where $d_j = 0$ if $j = 0$. For the modified system,

$$d'_n = n\text{th-order statistic from } \{(b'_i + v_i)\colon i < n + s\} \tag{12}$$

and

$$b'_n \geq \max\{t_n, d'_{n-s}\}, \qquad n \geq 1, \tag{13}$$

because in the modified system it is possible to have a positive queue and an idle server.

Since the nth-order statistic is a monotonic function, to prove Theorem 8 it suffices to show that $t_n \leq t'_n$ and $b_n \leq b'_n$ for all n. We show this induction. Obviously $b_i = t_i = a_i \leq t'_i \leq b'_i$, $1 \leq i \leq s$. Suppose $t_i \leq t'_i$ and $b_i \leq b'_i$ for all i, $i \leq n - 1$. We first show that $t_n \leq t'_n$. Suppose not; then

$$t_n > t'_n \geq t'_{n-1} \geq t_{n-1},$$

and thus $n - 1$ customers have entered both systems before the arrival associated with t'_n. (Note that customers could arrive in batches, i.e., $a_k = a_{k+1}$ is a possibility, but this presents no serious difficulty.) However, by the induction hypothesis $b_i + v_i \leq b'_i + v_i$, $i \leq n - 1$, so the original system has at most the same number of customers as the modified system before the arrival associated with t'_n. Thus, $t_n > t'_n$ cannot occur. Hence $t_n \leq t'_n$ as claimed.

To continue the induction proof for b_n, note that (10) and (12) imply that $d_i \leq d'_i$ for $i \leq n - s$. Then, from (11) and (13), we have

$$b_n = \max\{t_n, d_{n-s}\} \leq \max\{t'_n, d'_{n-s}\} \leq b'_n,$$

which completes the proof. □

Remark: Our proof of Theorem 8 is closely related not only to Wolff's proof,[9] but also to Sonderman's comparison proofs.[20,21] Sonderman was concerned with the effect of different service-time distributions instead of different queue disciplines.

We close this section with another result about pure-loss systems. With waiting rooms or with general service times it is easy to show that the stochastic processes representing the number of customers in the system need not be stochastically ordered, but we do get stochastic order with exponential distributions and no waiting rooms.

Theorem 9: In the setting of Theorem 7, if there are no waiting rooms, if the service-time distribution is exponential and if $Q(0) =_{st} Q_1(0) + Q_2(0)$, *then*

$$\{Q(t), t \geq 0\} \geq_{st} \{Q_1(t) + Q_2(t), t \geq 0\}$$

and

$$\{N(t), t \geq 0\} \leq_{st} \{N_1(t) + N_2(t), t \geq 0\}.$$

Proof: Here the argument follows Sonderman[7,20,21] and Whitt.[8] As the first step in constructing the two systems on the same probability space, we let the two systems being compared have identical arrival processes; i.e., we let the arrival process to the combined group of $s_1 + s_2$ servers be the sum of the two arrival processes to the separate groups of s_i servers. This not only means that the arrival processes have the same joint distributions, but that they have the same sample paths. Similarly, we let both systems start off with the same number of customers in the system; i.e., given the pair $[Q_1(0), Q_2(0)]$, we let $Q(0) = Q_1(0) + Q_2(0)$. We now show how to construct the departures so that

$$N(t) \leq N_1(t) + N_2(t) \tag{14}$$

and

$$Q(t) \geq Q_1(t) + Q_2(t) \tag{15}$$

for all $t \geq 0$. We generate departures from both systems using a single Poisson process with rate $(s_1 + s_2)\mu$. Each point in this Poisson process corresponds to a potential departure. Suppose the point occurs at time t. With probability $Q_1(t)/(s_1 + s_2)$, the point corresponds to a departure from both the single group of s_1 servers and the combined group of $s_1 + s_2$ servers; with probability $Q_2(t)/(s_1 + s_2)$, the point corresponds to a departure from both the single group of s_2 servers and the combined

group of $s_1 + s_2$ servers; with probability $[Q(t) - Q_1(t) - Q_2(t)]/(s_1 + s_2)$ the point corresponds to a departure from only the combined group of $s_1 + s_2$ servers; and finally, with probability $[s_1 + s_2 - Q(t)]/(s_1 + s_2)$, the point corresponds to no departure at all. This can be shown to yield the proper distributions for each system; see Sonderman[7] for more detail. This also guarantees that there is a departure in the combined group of $s_1 + s_2$ servers whenever there is a departure from one of the groups with s_1 and s_2 servers. There also cannot be a departure from the combined group alone when $Q(t-) = Q_1(t-) + Q_2(t-)$, so inequality (15) is maintained. This means that all departures and losses from the combined group of $s_1 + s_2$ servers that are not matched by corresponding departures or losses from one of the groups of s_1 and s_2 servers can be matched with earlier losses from one of the groups of s_1 and s_2 servers. Mathematical induction on the arrival index establishes (14) and (15) and formally completes the proof. □

Remark: For the special case of $M/M/s$ systems, the stochastic order in Theorems 6 and 9 can also be established under the conditions in the corollary to Theorem 5 using existing comparison theorems for continuous-time Markov chains; see Sonderman.[7] However, we know of no direct connections between the monotone likelihood-ratio orderings and the sample-path orderings.

IV. DIFFERENT SERVICE RATES

In this section we let the service rates in the two separate systems be different. One way to extend (2) and (4) occurs when the service times are associated with the arrivals. If two independent Poisson streams with rates λ_1 and λ_2 and associated service-time c.d.f.'s $F_1(x)$ and $F_2(x)$ are combined, then the resultant stream is a Poisson stream with rate $\lambda_1 + \lambda_2$ and associated service-time c.d.f.:

$$F(x) = [\lambda_1 F_1(x) + \lambda_2 F_2(x)]/(\lambda_1 + \lambda_2).$$

Of course, when $F_i(x)$ is exponential with mean μ_i^{-1} for each i, $F(x)$ is not exponential unless $\mu_1 = \mu_2$. However, the blocking probability for an $M/G/s$ loss system depends only on the mean service time. Thus the loss rate for the combined system is $L(s_1 + s_2, \lambda_1 + \lambda_2, (\lambda_1 + \lambda_2)/(a_1 + a_2))$, and a natural extension of (2) to conjecture is

$$L(s_1 + s_2, \lambda_1 + \lambda_2, (\lambda_1 + \lambda_2)/(a_1 + a_2)) \leq L(s_1, \lambda_1, \mu_1) + L(s_2, \lambda_2, \mu_2).$$

Unfortunately, this conjectured inequality is not valid. To see this, let $\lambda_1 = 1$, $\mu_1 = \epsilon^{-1}$, $\lambda_2 = \epsilon$, and $\mu_2 = \epsilon^2$; then $a_1 = \epsilon$ and $a_2 = \epsilon^{-1}$. Obviously,

$$\begin{aligned} L(s_1 + s_2, \lambda_1 + \lambda_2, (\lambda_1 + \lambda_2)/(a_1 + a_2)) &= (\lambda_1 + \lambda_2)B(s_1 + s_2, a_1 + a_2) \\ &= (1 + \epsilon)B(s_1 + s_2, \epsilon + \epsilon^{-1}) \\ &\to 1 \quad \text{as} \quad \epsilon \to 0, \end{aligned}$$

whereas

$$\begin{aligned} L(s_1, \lambda_1, \mu_1) + L(s_2, \lambda_2, \mu_2) &= \lambda_1 B(s_1, a_1) + \lambda_2 B(s_2, a_2) \\ &= B(s_1, \epsilon) + \epsilon B(s_2, \epsilon^{-1}) \\ &\to 0 \quad \text{as} \quad \epsilon \to 0. \end{aligned}$$

Consequently, in this case

$$L(s_1 + s_2, \lambda_1 + \lambda_2, (\lambda_1 + \lambda_2)/(a_1 + a_2)) \geq L(s_1, \lambda_1, \mu_1) + L(s_2, \lambda_2, \mu_2)$$

for sufficiently small ϵ. The previous measure, rate of customer loss, is not the only reasonable way to evaluate system performance in this case. For example, one might be interested in the rate of loss of service time. (Note that there is no real difference between these measures when the mean service times of the systems are identical.) With this new measure, the natural extension of (2) to conjecture is:

$$\frac{a_1 + a_2}{\lambda_1 + \lambda_2} L(s_1 + s_2, \lambda_1 + \lambda_2, (\lambda_1 + \lambda_2)/(a_1 + a_2)) \leq \frac{1}{\mu_1} L(s_1, \lambda_1, \mu_1) + \frac{1}{\mu_2} L(s_2, \lambda_2, \mu_2).$$

This inequality is in fact always true, since substitution of $L(s, \lambda, \mu) = \lambda B(s, \lambda/\mu)$ quickly reduces it to the second version of the inequality of Theorem 1. Thus the server occupancy is always increased for the combined system.

Turning to delay systems, we again find examples where sharing can be counterproductive. To see this, consider two $M/M/1$ delay systems with $\lambda_1 = 1$, $\mu_1 = 2$, $\lambda_2 = \epsilon$, and $\mu_2 = 2\epsilon$. Then $EQ_1(\infty) = EQ_2(\infty) = \rho/(1 - \rho) = 1$, but $EQ(\infty)$ can be shown to be of order ϵ^{-1} as $\epsilon \to 0$: Consider the interval following a low-intensity arrival. With probability $\lambda_2/(\lambda_2 + \mu_2) = \frac{1}{3}$, a second low-intensity arrival occurs before the first departs. Then there follows an exponentially distributed interval of mean length $1/4\epsilon$ during which the combined system fills up with high-intensity customers. In computing the average number of customers in the system, we get a term of order ϵ^{-2} (the total area in the plot of the number of customers in the system versus time, starting from the moment the second low-intensity customer arrives and ending when one of the two low-intensity customers departs), divided by a term of order ϵ^{-1}. In other words, with the mean steady-state delays held fixed in the two separate systems, the mean steady-state delay in the combined system can be arbitrarily large.

Note that the combined system can be modeled as an $M/G/s_1+s_2$ delay system where the service-time distribution is the mixture of two exponential distributions, but in contrast to the case of loss systems the mean delay does not depend only on the mean of the service-time

distribution. Hence, the appropriate generalization of (4) involves a system which is not $M/M/s$.

Another possible extension for $\mu_1 \neq \mu_2$ occurs when the service-time distributions are associated with the servers. Here the combined system is not $M/M/s$ because there are heterogeneous servers, so there are no equations similar to (2) and (4). In this case, it can be shown that with exponential service-time distributions and no waiting rooms, resource sharing is always better if customers always are sent to the fastest available server. In particular, as in Theorems 6 to 8, it can be shown for any single system that assigning customers to the fastest available server produces fewer losses than any other rule, where by "fewer losses" we mean in the sample-path ordering of Section III. One other rule, corresponding to the two separate systems, is to assign the customer only to servers associated with their original separate arrival streams.

When we focus on delay systems with heterogeneous servers, it is easy to give counterexamples showing that resource sharing can again be counterproductive. Related literature on the assignment of customers to heterogeneous servers appears in Winston,[22-24] Smith,[25] and references therein.

This section shows that, with unequal service-time distributions, resource sharing can be counterproductive. However, with unequal service-time distributions, much depends on the criterion of system performance. Also, it should be noted that such counterexamples have been observed before; others have discovered that infrequent "bad" customers can affect adversely a large number of "good" customers.

V. ACKNOWLEDGMENT

We are grateful to Ed Arthurs and Bart Stuck for suggesting this problem.

APPENDIX

Here we give two results that are due to others. First, we present Paul Burke's proof that $B(ts, ta)$ is strictly decreasing in t for $t \geq 0$. This result implies Theorem 1 when $\lambda_1/s_1 = \lambda_2/s_2$, because then $B(s_1 + s_2, a_1 + a_2) = B(ts_i, ta_i)$ for some $t \geq 1$, so $B(s_i, a_i) \geq B(s_1 + s_2, a_1 + a_2)$ for each i and

$$\frac{\lambda_1 B(s_1, a_1)}{\lambda_1 + \lambda_2} + \frac{\lambda_2 B(s_2, a_2)}{\lambda_1 + \lambda_2} \geq B(s_1 + s_2, a_1 + a_2),$$

which is equivalent to (2).

To see that $B(ts, ta)$ is strictly decreasing in t, first recall the following equation relating two different expressions for the tail of the

gamma distribution:

$$e^{-a}\sum_{k=0}^{k=n} a^k/k! = \int_a^\infty \frac{x^n e^{-x}}{n!}\,dx.$$

Then note that

$$\begin{aligned}\frac{1}{B(ts, ta)} &= \frac{\int_{ta}^{\infty} e^{-x}x^{ts}\,dx}{e^{-ta}(ta)^{ts}} \\ &= \int_{ta}^{\infty} e^{-(x-ta)}(x/ta)^{ts}\,dx \\ &= \int_0^{\infty} e^{-x}(1 + [x/ta])^{ts}\,dx;\end{aligned}$$

also see Theorem 3 of Jagerman.[3] Finally, $(1 + [x/ta])^{ts}$ is strictly increasing in t.

Second, Herbert Shulman has shown that Paul Burke's result and the convexity of $B(s, a)$ in s for $s \geq 1$ imply a version of Theorem 1. Such convexity has frequently been conjectured but has been proved only for lattices of points with unit spacing, see Messerli[26] and references therein. These versions of convexity are not strong enough to make the following valid even when s_1 and s_2 are integers; however, general convexity would establish the proof for all real numbers s_1 and $s_2 \geq 1$, a more general mathematical result. We reproduce Shulman's argument here:

$$\begin{aligned}B(s_1 + s_2, a_1 + a_2) &\leq \frac{a_1}{a_1 + a_2} B\left(\frac{a_1 + a_2}{a_1} s_1, a_1 + a_2\right) \\ &\quad + \frac{a_2}{a_1 + a_2} B\left(\frac{a_1 + a_2}{a_2} s_2, a_1 + a_2\right) \\ &\leq \frac{a_1}{a_1 + a_2} B(s_1, a_1) + \frac{a_2}{a_1 + a_2} B(s_2, a_2).\end{aligned}$$

REFERENCES

1. E. Arthurs and B. W. Stuck, "Subadditivity of Teletraffic Special Functions," SIAM Rev., to be published.
2. L. Kleinrock, *Queueing Systems, Volume 1: Theory*, New York: John Wiley, 1975.
3. D. L. Jagerman, "Some Properties of the Erlang Loss Function," B.S.T.J., *53*, No. 3 (March 1974), pp. 525–51.
4. R. B. Cooper, *Introduction to Queueing Theory*, New York: Macmillan, 1972.
5. L. Kleinrock, *Queueing Systems, Volume 2: Computer Applications*, New York: John Wiley, 1976.
6. S. Stidham, Jr., "A Last Word on $L = \lambda W$," Oper. Res., *22*, No. 2 (March–April 1974), pp. 417–22.

7. D. Sonderman, "Comparing Semi-Markov Processes," Math. Oper. Res. *5*, No. 1 (February 1980), pp. 110–20.
8. W. Whitt, "Comparing Counting Processes and Queues," Adv. Appl. Probab., *13*, No. 1 (March 1981), to be published.
9. R. W. Wolff, "An Upper Bound for Multi-Channel Queues," J. Appl. Probab., *14*, No. 4 (December 1977), pp. 884–8.
10. T. S. Ferguson, *Mathematical Statistics: A Decision Theoretic Approach*, New York: Academic Press, 1967.
11. W. Whitt, "A Note on the Influence of the Sample on the Posterior Distribution," J. Am. Statist. Assoc., *74*, No. 366 (June 1979), pp. 424–6.
12. W. Whitt, "Uniform Conditional Stochastic Order," J. Appl. Probab., *17*, No. 1 (March 1980), pp. 112–23.
13. J. Keilson and U. Sumita, "Uniform Stochastic Ordering," Working Paper, The Graduate School of Management, University of Rochester, 1980.
14. T. Kamae, U. Krengel, and G. L. O'Brien, "Stochastic Inequalities on Partially Ordered Spaces," Ann. Probab., *5*, No. 6 (December 1977), pp. 899–912.
15. T. Lindvall, "A Note on Coupling of Birth and Death Processes," J. Appl. Probab., *16*, No. 3 (September 1979), pp. 505–12.
16. V. E. Beneš, "Programming and Control Problems Arising from Optimal Routing in Telephone Networks," B.S.T.J., *45*, No. 9 (November 1966), p. 1373.
17. W. Whitt, "On Stochastic Bounds for the Delay Distribution in the *GI/G/s* Queue," Oper. Res., to be published.
18. A. Kuczura, "Queues with Mixed Renewal and Poisson Inputs," B.S.T.J., *51*, No. 6 (July–August 1972), pp. 1305–26.
19. A. Kuczura, "Loss Systems with Mixed Renewal and Poisson Inputs," Oper. Res., *21*, No. 3 (May–June 1973), pp. 787–95.
20. D. Sonderman, "Comparing Multi-Server Queues with Finite Waiting Rooms, I: Same Number of Servers," Adv. Appl. Probab., *11*, No. 2 (June 1979), pp. 439–47.
21. D. Sonderman, "Comparing Multi-Server Queues with Finite Waiting Rooms, II: Different Numbers of Servers," Adv. Appl. Probab., *11*, No. 2 (June 1979), pp. 448–55.
22. W. L. Winston, "Assignment of Customers to Servers in a Heterogeneous Queueing System with Switching," Oper. Res., *25*, No. 3 (May–June 1977), pp. 468–83.
23. W. L. Winston, "Optimal Dynamic Rules for Assigning Customers to Servers in a Heterogeneous Queueing System," Nav. Res. Logis. Q., *24*, No. 2 (June 1977), pp. 293–300.
24. W. L. Winston, "Optimal Assignment of Customers in a Two-Server Congestion System with No Waiting Room," Manage. Sci., *24*, No. 6 (February 1978), pp. 702–5.
25. D. R. Smith, "Optimal Repair of a Series System," Oper. Res., *26*, No. 4 (July–August 1978), pp. 653–62.
26. E. J. Messerli, "Proof of a Convexity Property of the Erlang B Formula," B.S.T.J., *51*, No. 4 (April 1972), pp. 951–3.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 1, January 1981
Printed in U.S.A.

Transient Behavior of the Kendall Birth-Death Process—Applications to Capacity Expansion for Special Services

By R. N. NUCHO

(Manuscript received October 16, 1979)

In this paper we derive explicit expressions for the transient state probabilities of the Kendall birth-death process, with and without immigration, for any initial condition. We then propose this process as a model for special services point-to-point demand, in which the births represent circuit "connects" and the deaths represent "disconnects." This choice of model is based on intuitive arguments and on the fact that the model can represent the growth and turnover characteristics of special services demand. Thus, the model provides a means by which special services demand, with its inherent uncertainty, may be approximately represented in various facility network studies, to obtain, at the very least, useful qualitative results. In particular, we evaluate the probability of a held order (i.e., the probability that a service request is held for lack of spare facilities) with Blocked Customers Held (BCH) as the queue discipline. We also apply the model to capacity expansion problems, introduce the concept of margin, the extra capacity needed to meet the demand within a given held-order probability, and examine its sensitivity with respect to growth, turnover (or churn), and system size. We find that aggregating small demands into a single larger demand produces significant reduction of the margin, because of improved statistical properties.

I. INTRODUCTION

In this article, the transient behavior of the Kendall birth-death process* with immigration is examined, and some applications of the

* The Kendall birth-death process is one in which the transition rates are proportional to the state.

process to capacity expansion problems are discussed. The choice of such a process was motivated by the search for a model for special services point-to-point circuit demand, a model which would be used as a tool for determining facility network circuit routing strategies. Special services demand generally consists of demand for full-time dedicated circuits (e.g., foreign exchange lines, WATS lines, data lines), as opposed to the message-traffic offered load which consists of demand for the use of common facilities for a relatively short period of time. Thus, the system examined is characterized by the stochastic process $\mathcal{N}(t)$ with realizations (states) $n = 0, 1, \cdots, \infty$, where n might refer to the number of working circuits or some other facility, rather than to the number of busy trunks, as in the message-traffic case. By definition, the birth-death process[1] allows transitions from some state n to $n + 1$ via a birth (circuit connect), or to $n - 1$ via a death (circuit disconnnect). The transition rates are λ_n for the births and μ_n for the deaths, both of which are chosen proportional to n for the following reasons.

It is clear, for special services, that the rate of disconnects, μ_n, is state dependent. There are, in fact, indications[2] that μ_n is a monotonically increasing function of n. The simplest such function is $n\mu$, which implies that the probability of disconnects is proportional to the size of the system. With this choice for the death rate, a number of possible choices exist for the birth rate. Choosing it to be a constant causes the mean number of circuits to saturate in time, while choosing it to be proportional to n causes the mean to grow or decline exponentially. Since special services are presently characterized by significant net growth, it would seem that a plausible model for special services demand is a birth and death process in which *both* the birth and death rates are proportional to the state.

One consequence, however, of assuming $\lambda_n = n\lambda$ is that if the process reaches the state $n = 0$ at any time, by a succession of disconnects, it will stay there forever, since the birth rate is zero. To eliminate this characteristic, the concept of immigration may be introduced by taking $\lambda_n = n\lambda + \beta$, where β is the immigration factor. The cases with and without immigration will be discussed below.

It must be emphasized that it is not the intent of this paper to validate the model based on an examination of actual special services data. Such statistical data analysis is important for a final assessment of the accuracy of the model and is currently being undertaken. For the purposes of this paper, it shall be assumed that a study of the proposed model is justified, based on the intuitive arguments given above and on the fact that the model captures the growth and turnover characteristics of special services (see Section 5.1). The model provides a means by which special services demand, with its inherent uncer-

tainty, may be approximately represented in various special services facility network studies, to obtain, at the very least, useful qualitative results.

Since the situation of interest is that of net growth, it is clear that statistical equilibrium does not exist, and that it is the problem of the transient solutions of the Kolmogorov birth-death equations that is of prime importance. Much literature exists on the subject of transient solutions for birth-and-death processes[1,3–11] and the case in which the transition rates are state independent is completely solved.[7,8] The case in which λ_n and μ_n are proportional to the population is solved when immigration is not included: The results for a specific initial condition, namely starting from the state $n = 1$, are derived in Refs. 4 and 10 and the expressions for the general initial condition are quoted in Ref. 10. For the nonzero immigration case, the form of the generating function for the state probabilities is known,[10] but it seems that explicit expressions for the state probabilities have not previously appeared in the literature. In this paper, these expressions are derived for any nonnegative value of β.

In special services, if an order for service is delayed because of lack of spare facilities, the order is said to be held. Thus, in order to study capacity expansion problems, the probability of a held order is introduced, as well as the concept of margin, the extra capacity needed to meet the demand within a given held-order probability. This held-order probability is similar but not identical to the transient time congestion of the process (see Appendix B). The queue discipline followed here is Blocked Customers Held (BCH), in which an arriving customer spends a total time T (random variable) in the system, after which he departs regardless of whether he is waiting to be served (i.e., his service order has been delayed) or is actually being served (i.e., he has been assigned a circuit).

A fundamental difference between this analysis and teletraffic must be emphasized. This difference arises because of the respective time scales in the two cases. Whereas the mean lifetime of a call in traffic ($\tau = 1/\mu$) is of the order of a few minutes, the mean lifetime of a circuit in the process described here is of the order of a few years. It is this fact, coupled with the relatively fast growth of special services demand, that makes it impossible to even approximately treat the process in a statistical equilibrium mode (no growth) with a slowly varying envelope representing the growth. Thus, the transient aspect of the problem is to be contrasted to the more conventional assumption, in teletraffic theory, that statistical equilibrium prevails (it must be mentioned, however, that some work has been done concerning nonstationary telephone traffic with time-varying Poisson-offered load, e.g., Refs. 12 to 14). It must be further noted that, although the model is being

proposed for special services demand, nevertheless, it may be applied, with an appropriate choice of parameters λ, β, and μ, to any process that behaves in a similar manner.

The held-order probability having been defined and the concept of margin introduced, questions concerning capacity expansion problems are addressed. Capacity expansion is a problem that has been studied by many. In this paper, optimal capacity-expansion policies are not sought; only very specialized aspects of the problem are considered. For instance, the effects of aggregating demands into a larger single demand are examined, and the minimum capacity increment which would meet the demand within a specified interval of time and within a given held-order probability is determined. In addition, the relationship between spare capacity and lead time is discussed (see summary of results in Section II). Some relevant work has been done by Freidenfelds[15,16] in which the author computes first-passage times to various levels of demand using a general birth-death process, and discusses briefly fill-at-relief problems. Work by Luss and Whitt[17] studies the impact of both deterministic and stochastic models on utilization. The authors use Brownian motion to model the stochastic demand and follow a scheme similar to ours for determining the margin needed at a future time.

The organization of this paper is as follows. Section II sets up the problem and gives a summary of results. The explicit solutions for the general case are derived in Section III, and their properties are examined in Section IV. In Section V, growth, turnover, and churn are defined, the concept of margin is introduced, and some of its applications to capacity expansion problems are discussed. Finally, Section VI contains the conclusions.

II. BACKGROUND AND SUMMARY OF RESULTS

2.1 General birth-death equations

Consider a system described by a set of states $n = 0, 1, \cdots, \infty$, and a birth and death process defined by a set of transition rates $\{\lambda_n, \mu_n\}$. The quantity $\lambda_n\delta(\mu_n\delta) + o(\delta)$* is the probability of a birth (death) in the small interval $[t, t + \delta]$, given that the system is in state n at time t.[1] The probability of more than one birth or death in $[t, t + \delta]$ is $o(\delta)$. The probabilities $P_n(t)$ of finding the system in state n at time t must satisfy the well-known infinite set of difference-differential equations (p. 454 of Ref. 1)

$$\frac{d}{dt}P_n(t) = -(\lambda_n + \mu_n)P_n(t) + \lambda_{n-1}P_{n-1}(t) + \mu_{n+1}P_{n+1}(t) \qquad [\text{for} \quad n \geq 0, \quad P_{-1}(t) = 0, \quad \mu_0 = 0]. \quad (1)$$

* $o(\cdot)$: $R^1 \to R^1$ is such that $\lim_{\delta \to 0} \frac{o(\delta)}{\delta} = 0$.

If the initial number of circuits is n_0, the initial condition may be written

$$P_n(0) = \delta_{nn_0}, \tag{2}$$

where δ_{nn_0} is the Kronecker delta.

The particular birth-death processes considered in this paper are the cases in which the transition rates are proportional to the population, n, with or without immigration.[5,6,10] The corresponding transition rates, defined for all nonnegative integers, n, are

$$\lambda_n = n\lambda + \beta, \qquad \mu_n = n\mu, \tag{3}$$

where λ, μ, and β are nonnegative constants. In the following sections, results for the case with no immigration may be easily obtained by setting $\beta = 0$.

2.2 Mean and variance

It has been shown[1,4,10] that the mean, $m(t)$, and the variance, $v(t)$, of processes such as those described by eqs. (1) and (3) may be obtained without solving explicitly for the $P_n(t)$. The resulting expressions, satisfying initial condition (2), may be easily found to be

(*i*) Case $\lambda \neq \mu$

$$m(t) = \left(n_0 + \frac{\beta}{\lambda - \mu}\right)e^{(\lambda-\mu)t} - \frac{\beta}{\lambda - \mu}, \tag{4}$$

$$v(t) = n_0 \frac{\lambda + \mu}{\lambda - \mu} e^{2(\lambda-\mu)t}[1 - e^{-(\lambda-\mu)t}] + \frac{\beta}{(\lambda - \mu)^2}[\lambda e^{2(\lambda-\mu)t} - (\lambda + \mu)e^{(\lambda-\mu)t} + \mu]. \tag{5}$$

(*ii*) Case $\lambda = \mu$

$$m(t) = \beta t + n_0, \tag{6}$$

$$v(t) = \lambda\beta t^2 + (2\lambda n_0 + \beta)t. \tag{7}$$

2.3 Solutions for $P_n(t)$

To simplify the notation, define the following quantities:

$$\Delta = \lambda - \mu,$$

$$A = A(t) = \frac{\Delta^2 e^{\Delta t}}{\mu\lambda(e^{\Delta t} - 1)^2},$$

$$B = B(t) = \frac{e^{\Delta t} - 1}{\lambda e^{\Delta t} - \mu},$$

$$C = C(t) = \frac{\Delta}{\lambda e^{\Delta t} - \mu}, \tag{8}$$

and

$$\binom{r}{m} = \frac{r(r-1)\cdots(r-m+1)}{m!}. \tag{9}$$

This definition of the binomial coefficient is valid for any real number r and any positive integer m (see p. 50 of Ref. 1). For $m = 0$, one defines $\binom{r}{0} = 1$, and for negative integers m, one defines $\binom{r}{m} = 0$. The symbol $\binom{r}{m}$ is not used if m is not an integer. Denoting $\nu = \beta/\lambda$, where ν is any nonnegative real number, the solutions derived in this paper are

(*i*) Case $\lambda \neq \mu$

$$P_n(t) = \begin{cases} C^{\nu}(\mu B)^{n_0}(\lambda B)^n \sum_{i=0}^{\min\{n_0,n\}} \binom{n_0}{i} \binom{n_0+n+\nu-i-1}{n-i} \\ \quad \cdot (A-1)^i, \text{ if } n_0+\nu>0, & (10) \\ \delta_{n0}, \quad \text{if } n_0+\nu=0. & (11) \end{cases}$$

(*ii*) Case $\lambda = \mu$

$$P_n(t) = \begin{cases} \left(\frac{1}{1+\lambda t}\right)^{\nu}\left(\frac{\lambda t}{1+\lambda t}\right)^{n+n_0} \sum_{i=0}^{\min\{n_0,n\}} \binom{n_0}{i} \\ \quad \cdot \binom{n_0+n+\nu-i-1}{n-i}\left(\frac{1}{\lambda^2 t^2}-1\right)^i, \text{ if } n_0+\nu>0, & (12) \\ \delta_{n0}, \quad \text{if } n_0+\nu=0. & (13) \end{cases}$$

Equations (10) and (11) with no immigration ($\nu = 0$) are identical to the results quoted by Bailey [Eqs. (8.47) of Ref. 10].

2.4 *Application to capacity expansion*

In Section V, margin is defined as the capacity which must be built in excess of the mean to meet certain service requirements, and the percent margin is defined as the ratio of the margin to the mean in percent. The following is a summary of the main results:

(*i*) By aggregating demands, less percent margin is needed than in the nonaggregated case. This effect is especially significant for small demands.

(*ii*) Given a minimum desired time, T, between successive expansions, a procedure is established for determining the minimum capacity increment which would meet the given service requirements.

(*iii*) Given a lead time, τ, between the moment facilities are ordered and the time they are available for use, a procedure is established for

determining the threshold value of the remaining spare corresponding to the time at which new facilities should be ordered.

(*iv*) By introducing immigration, the absorbing zero state is eliminated and the percent margin needed to meet the service requirements is reduced for moderately large to large times (of the order of two years or more for the particular values examined).

III. DERIVATION OF THE STATE PROBABILITIES

The approach followed to solve the set of equations in (1) is the generating function technique.[5,10] In Ref. 10, a differential equation for the generating function, $F(s, t)$, defined below, is established and its solution is derived. The results are quoted in Section 3.1. Three well-known identities are given in Section 3.2 and are then used in Section 3.3 to derive explicit expressions for the state probabilities. The procedure followed in Section 3.3 is to identify $F(s, t)$ as the generating function for a convolution of two functions.

3.1 The generating function

The generating function, $F(s, t)$, is related to the state probabilities through the following expression:

$$F(s, t) = \sum_{n=0}^{\infty} s^n P_n(t), \qquad 0 \leq s \leq 1. \tag{14}$$

The differential equation for $F(s, t)$, given in eq. (8.63) of Ref. 10 with $e^{\theta} = s$, is

$$\frac{\partial F(s, t)}{\partial t} + H(s) \frac{\partial F(s, t)}{\partial s} = \beta(s - 1)F(s, t), \tag{15}$$

where

$$H(s) = -(s - 1)(\lambda s - \mu).$$

The solutions to this equation are

$$F(s, t) = \begin{cases} \left(\dfrac{\Delta}{d + cs}\right)^{\nu} \left(\dfrac{b + as}{d + cs}\right)^{n_0}, & \lambda \neq \mu, \qquad (16) \\ \left(\dfrac{1}{\tilde{d} + \tilde{c}s}\right)^{\nu} \left(\dfrac{\tilde{b} + \tilde{a}s}{\tilde{d} + \tilde{c}s}\right)^{n_0}, & \lambda = \mu,^{*} \qquad (17) \end{cases}$$

* An alternative approach for obtaining this result is to substitute $\lambda - \Delta$ for μ in the $\lambda \neq \mu$ expression and to take the limit $\Delta \to 0$.

where

$$0 \le s \le 1,$$

$$a = \lambda - \mu e^{\Delta t},$$

$$b = -\mu(1 - e^{\Delta t}),$$

$$c = \lambda(1 - e^{\Delta t}),$$

$$d = \lambda e^{\Delta t} - \mu, \tag{18}$$

$$\tilde{a} = 1 - \lambda t,$$

$$\tilde{b} = \lambda t,$$

$$\tilde{c} = -\lambda t,$$

$$\tilde{d} = 1 + \lambda t. \tag{19}$$

The above solutions may be verified by direct substitution. Equation (16) agrees with eq. (8.71) of Ref. 10 and eq. (17) with $\nu = 0$ agrees with eq. (8.52) of Ref. 10.

3.2 Useful identities

In Section 3.3, use will be made of the three following well-known identities.

3.2.1 Binomial identity

For any α and β and for any nonnegative integer n, the following identity holds (p. 51, Ref. 1):

$$(\alpha + \beta)^n = \sum_{m=0}^{n} \binom{n}{m} \alpha^{n-m} \beta^m. \tag{20}$$

3.2.2 Negative binomial identity

For any α and β such that $|\beta/\alpha| < 1$ and for any real number r, the following identity holds (see pp. 51 and 269 of Ref. 1):

$$(\alpha - \beta)^{-r} = \sum_{m=0}^{\infty} (-1)^m \binom{-r}{m} \beta^m \alpha^{-(m+r)}.$$

If r is strictly positive, identity (12.4) on p. 63 of Ref. 1 may be used to write

$$(\alpha - \beta)^{-r} = \sum_{m=0}^{\infty} \binom{r+m-1}{m} \beta^m \alpha^{-(m+r)}. \tag{21}$$

3.2.3 Generating function for a convolution

Let $F_1(s)$ and $F_2(s)$ be the generating functions for the sequences

$\{P_n^{(1)}\}_{n=0,1,\cdots}$ and $\{P_n^{(2)}\}_{n=0,1,\cdots}$, respectively,

$$F_1(s) = \sum_{n=0}^{\infty} s^n P_n^{(1)}, \qquad F_2(s) = \sum_{n=0}^{\infty} s^n P_n^{(2)}. \tag{22}$$

The function $F(s) = F_1(s)F_2(s)$ is then the generating function for $\{P_n\}_{n=0,1,\cdots}$, the convolution of $P_n^{(1)}$ and $P_n^{(2)}$, and may be written as

$$F(s) = \sum_{n=0}^{\infty} s^n P_n, \tag{23}$$

where

$$P_n = \sum_{i=0}^{n} P_i^{(1)} P_{n-i}^{(2)} = \sum_{i=0}^{n} P_i^{(2)} P_{n-i}^{(1)}.$$

The proof of this theorem is elementary (e.g., see Chapter 11 of Ref. 1).

Note: This theorem applies to arbitrary sequences $\{P_n^{(1)}\}$ and $\{P_n^{(2)}\}$ (not necessarily probability distributions) as long as their respective generating functions exist. Thus, the series in eqs. (22) must converge. For the purpose of this theorem, however, it is assumed that $F_1(s)$ and $F_2(s)$ do exist.

3.3 Derivation of explicit expressions

3.3.1 Case $\lambda \neq \mu$

The generating function in eq. (16) may be rewritten as follows:

$$F(s, t) = F_1(s, t)F_2(s, t), \tag{24}$$

where

$$F_1(s, t) = \Delta^{\nu}(d + cs)^{-(n_0+\nu)},$$

$$F_2(s, t) = (b + as)^{n_0}.$$

(a) $n_0 + \nu > 0$

Applying identity (20), it may be seen that $F_2(s, t)$ is the generating function for a binomial type function,

$$F_2(s, t) = \sum_{m=0}^{n_0} \binom{n_0}{m} b^{n_0-m}(as)^m$$

$$= \sum_{m=0}^{\infty} s^m P_m^{(2)}(t), \tag{25}$$

where

$$P_m^{(2)}(t) = \begin{cases} \binom{n_0}{m} b^{n_0-m} a^m, & \text{if } m \le n_0, \\ 0 & \text{if } m > n_0. \end{cases} \tag{26}$$

In a similar manner, it may be seen from identity (21) that $F_1(s, t)$ is the generating function for a negative binomial type function,

$$\begin{aligned} F_1(s, t) &= \Delta^\nu \sum_{m=0}^{\infty} \binom{n_0 + \nu + m - 1}{m} (-cs)^m d^{-(m+n_0+\nu)} \\ &= \sum_{m=0}^{\infty} s^m P_m^{(1)}(t), \end{aligned} \tag{27}$$

where

$$P_m^{(1)}(t) = \Delta^\nu \binom{n_0 + \nu + m - 1}{m} (-c)^m d^{-(m+n_0+\nu)}. \tag{28}$$

It may be shown that $|cs/d| < 1$ for all values of $t \ge 0$, $0 \le s \le 1$, and $\lambda, \mu \ge 0$. Thus, identity (21) applies in all the relevant cases.

It now follows from eq. (23) that $F(s, t)$ is the generating function of the convolution,

$$\begin{aligned} P_n(t) &= \sum_{i=0}^{n} P_i^{(2)}(t) P_{n-i}^{(1)}(t) \\ &= \sum_{i=0}^{\min\{n_0,n\}} \left[\binom{n_0}{i} b^{n_0-i} a^i \right] \\ &\quad \cdot \left[\Delta^\nu \binom{n_0 + \nu + n - i - 1}{n - i} (-c)^{n-i} d^{-(n_0+\nu+n-i)} \right], \end{aligned} \tag{29}$$

where the upper limit on the sum arises from the condition $i \le n_0$ for $P_i^{(2)}(t)$ established by eq. (26).

Rearranging, one obtains

$$\begin{aligned} P_n(t) &= \left(\frac{\Delta}{d}\right)^\nu \sum_{i=0}^{\min\{n_0,n\}} \\ &\quad \cdot \binom{n_0}{i} \binom{n_0 + n + \nu - i - 1}{n - i} \left(\frac{a}{d}\right)^i \left(\frac{b}{d}\right)^{n_0-i} \left(-\frac{c}{d}\right)^{n-i}. \end{aligned} \tag{30}$$

From the definitions of a, b, c, and d in eqs. (18) and from eqs. (8), the above expression for $P_n(t)$ reduces immediately to eq. (10).

(b) $n_0 + \nu = 0$

For this case, $F(s, t) = 1$. From the definition in eq. (14), it is then

apparent that all $P_n(t)$ for $n \neq 0$ must be zero and that $P_0(t) = 1$, which is the result shown in eq. (11).

3.3.2 Case $\lambda = \mu$

The solutions may be obtained by using the same procedure followed in Section 3.3.1. The only difference is that the starting equation should be eq. (17) rather than (16). Since eq. (17) can be simply obtained from eq. (16) by letting $\Delta \to 1$, $a \to \tilde{a}$, $b \to \tilde{b}$, $c \to \tilde{c}$, and $d \to \tilde{d}$, it follows that the final results for the case $\lambda = \mu$ can be obtained from the results of the case $\lambda \neq \mu$ [i.e., eq. (30)] by making the above substitutions. Alternatively, eqs. (12) and (13) may be obtained by taking the limits of eqs. (10) and (11) as $\mu \to \lambda$. The procedure is the following: Consider λ to be a constant, then replace μ by $\lambda - \Delta$ wherever it appears, and finally take the limits $\Delta \to 0$ using l'Hospital's rule whenever necessary. The results of this limiting procedure are found to be

$$\begin{aligned} \lim_{\mu \to \lambda} A(t) &= \left(\frac{1}{\lambda t}\right)^2, \\ \lim_{\mu \to \lambda} B(t) &= \frac{t}{1 + \lambda t}, \\ \lim_{\mu \to \lambda} C(t) &= \frac{1}{1 + \lambda t}. \end{aligned} \tag{31}$$

IV. PROPERTIES

In this section, the zero-state probability and the cumulative probability distributions, for several choices of the parameters λ, μ, and β, are examined as a function of time. In addition, some cases in which the state probabilities are especially simple are indicated.

4.1 Probability of ultimate extinction

The birth-death process without immigration is characterized by an absorbing state at $n = 0$. If the system reaches that state at some time t_0, it will stay there for all $t > t_0$ since the birth rate is zero. The probability of hitting that state at time t is given by $P_0(t)$. In the limit $t \to \infty$, this probability tends to

$$\lim_{t \to \infty} P_0(t) = \begin{cases} \left(\dfrac{\mu}{\lambda}\right)^{n_0}, & \lambda > \mu, \\ 1, & \lambda \leq \mu. \end{cases} \tag{32}$$

Thus, for any $\mu \neq 0$, there is a nonzero probability of ultimate extinction

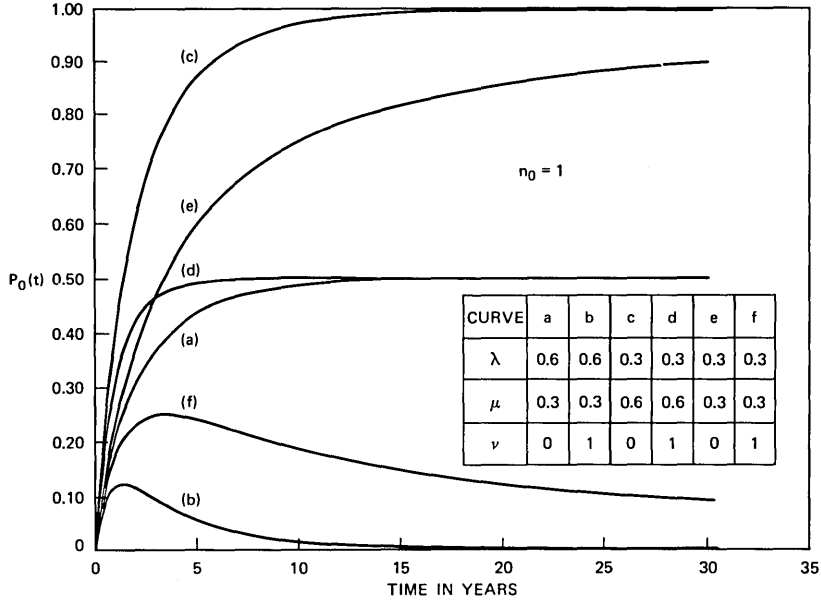


Fig. 1—Zero-state probability.

of the process, while for $\mu \geq \lambda$, ultimate extinction is a certainty. This feature may be removed, if desired, by introducing immigration. In this case, the limiting value of $P_0(t)$ tends to a nonzero value for $\lambda < \mu$ and to zero for $\lambda \geq \mu$,

$$\lim_{t\to\infty} P_0(t) = \begin{cases} \left(\frac{\Delta}{\lambda}\right)^{\nu}\left(\frac{\mu}{\lambda}\right)^{n_0} \lim_{t\to\infty} e^{-\nu\Delta t} = 0, & \lambda > \mu, \\ \left(1 - \frac{\lambda}{\mu}\right)^{\nu}, & \lambda \leq \mu. \end{cases} \tag{33}$$

The effect of immigration on $P_0(t)$ is shown graphically in Fig. 1, where $P_0(t)$ is plotted for various values of λ, μ, and ν, and for $n_0 = 1$. The case $n_0 = 1$ was chosen for clarity of the figure, since the effect of immigration is larger for smaller values of n_0.

4.2 Cumulative probability distribution

The cumulative probability distribution is defined as

$$F_n(t) = \sum_{i=0}^{n} P_i(t).$$

In order that $\lim_{n\to\infty} F_n(t) = 1$ for all t, it is necessary and sufficient

that $\sum_{n=0}^{\infty} \lambda_n^{-1}$ diverges, which is the case for the process discussed in this paper (see Theorem on p. 452 of Ref. 1). The function, $F_n(t)$, for various choices of λ, μ, and β, and for $n_0 = 5$ is shown in Figs. 2 through 5. The lowest curve plotted in each figure is $P_0(t)$, and is consequently a measure of the extinction probability.

The values of λ (0.6) and μ (0.3) chosen in Figs. 2 and 3 correspond to net positive growth (see discussion of growth in Section 5.1). As may be verified from eqs. (32) and (33), the $\lim_{t\to\infty} P_0(t)$ is nonzero in Fig. 2 and zero in Fig. 3. In addition, for each $n > 0$, the $\lim_{t\to\infty} P_n(t) = 0$, although the $\lim_{t\to\infty} \sum_{n=1}^{\infty} P_n(t)$ is nonzero. Thus, for any $n > 0$, the $\lim_{t\to\infty} F_n(t)$ is nonzero for $\beta = 0$ and zero for $\beta \neq 0$, reflecting the fact that the extinction probability is nonzero in the first case and zero in the second. (The $n = 0$ curve in Fig. 3 is essentially flat and is hard to distinguish on the graph.)

The case in which the death rate is larger than the birth rate is shown in Figs. 4 and 5. This situation corresponds to negative growth. As may be verified from eqs. (32) and (33), the $\lim_{t\to\infty} P_0(t)$ is 1.0 in Fig. 4 and 0.5 in Fig. 5. It may be shown that, for the $\beta = 0$ case, $\lim_{t\to\infty} P_n(t) = 0$ for all $n > 0$, while for the $\beta \neq 0$ case, $\lim_{t\to\infty} P_n(t) \neq 0$ for all $n \geq 0$. In both cases, however, the $\lim_{t\to\infty} F_n(t)$ is nonzero, reflecting the fact that the extinction probability is nonzero.

Finally, the case $\lambda = \mu$, $\beta = 0$, is similar to Fig. 4 (with extinction probability equal to unity), and the case $\lambda = \mu$, $\beta \neq 0$, is similar to Fig. 3 (with extinction probability zero). These cases are not shown.

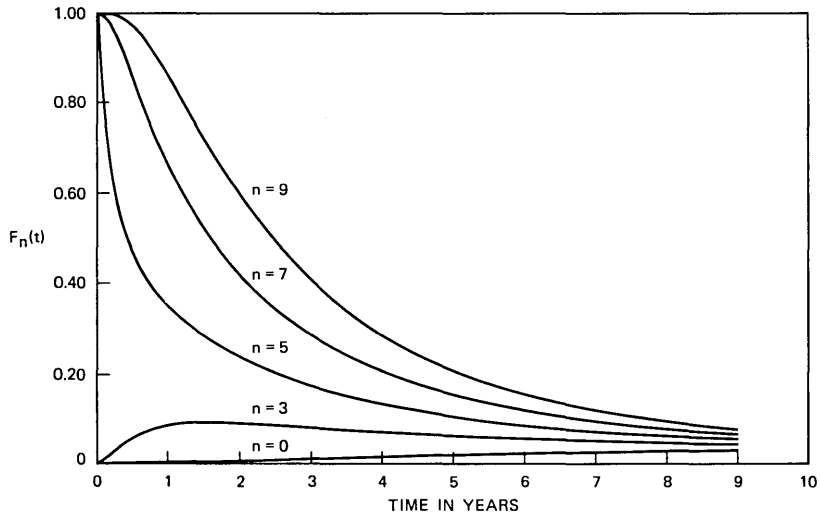


Fig. 2—Cumulative distribution without immigration ($\lambda > \mu$).

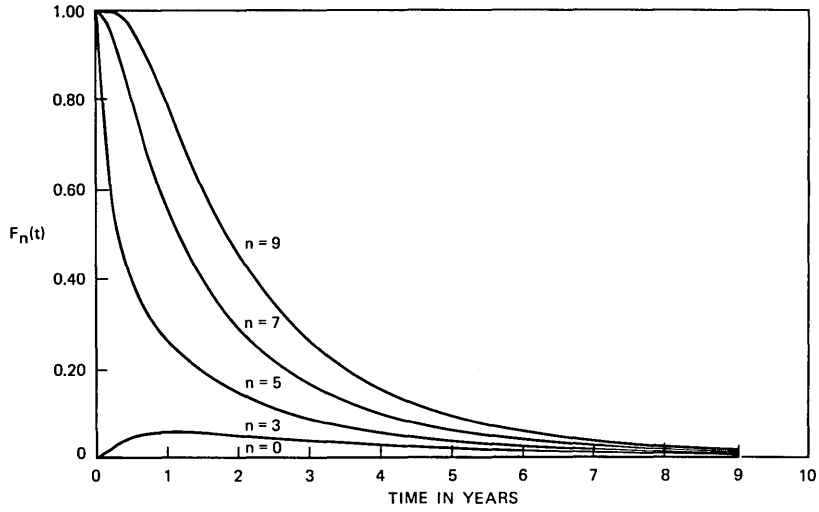


Fig. 3—Cumulative distribution with immigration ($\lambda > \mu$).

4.3 Special case solutions

For certain initial conditions, the general solutions reduce to simple analytical forms. For $n_0 = 0$, the interesting process is the one including immigration ($\nu \geq 1$). The state probabilities of eq. (10) may then be

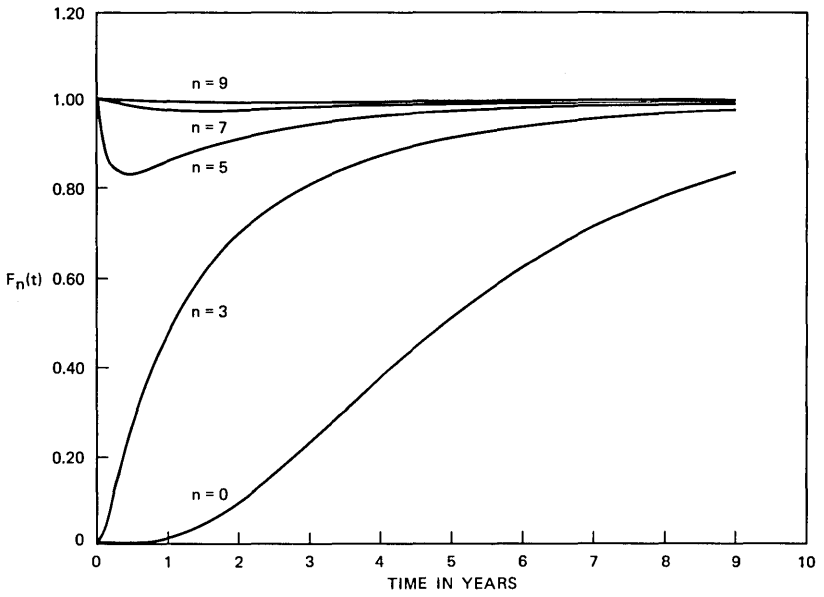


Fig. 4—Cumulative distribution without immigration ($\lambda < \mu$).

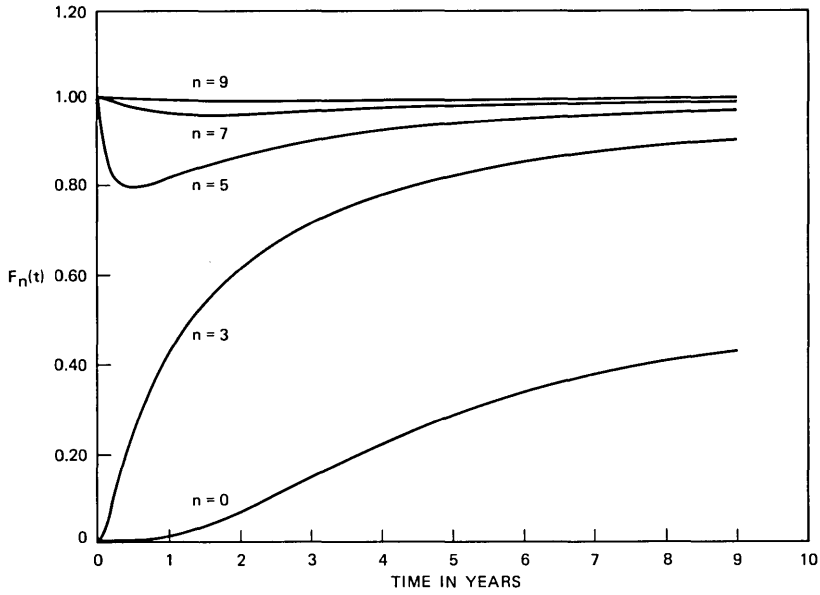


Fig. 5—Cumulative distribution with immigration ($\lambda < \mu$).

written

$$P_n(t) = C^{\nu}(\lambda B)^n \binom{n+\nu-1}{\nu-1}. \tag{34}$$

For $n_0 = 1$, the process without immigration becomes interesting. Equation (10) with $\nu = 0$ reduces to the well-known form[4,10]

$$P_0(t) = \mu B,$$

$$P_n(t) = (\mu B)(\lambda B)^n A = (1 - \lambda B)(1 - \mu B)(\lambda B)^{n-1} \qquad (n \neq 0). \tag{35}$$

V. APPLICATIONS TO CAPACITY EXPANSION

In this section, the λ, μ, and β parameters of the model are related to more physically intuitive quantities such as growth and turnover (or churn). The concept of margin, the extra capacity needed to meet the demand within a given held-order probability, is introduced. The effects of randomness and immigration on the margin are then examined, and finally, several capacity expansion problems are addressed.

5.1 Growth, churn, and turnover

The model described in the preceding sections is completely specified once the parameters λ, μ, and β are known. In this section, quantities that are more physically intuitive than the birth and death rates, namely growth and turnover (or churn), are introduced and

related to λ, μ, and β. First define the following quantities:

$$b(t) = EB(t) = \text{mean number of births in } [0, t], \tag{36}$$

$$d(t) = ED(t) = \text{mean number of deaths in } [0, t]. \tag{37}$$

Then,

$$\begin{aligned} m(t) - n_0 &= b(t) - d(t) \\ &= \text{mean net population increase in } [0, t], \end{aligned} \tag{38}$$

where E refers to the expected value and where $m(t)$ is the mean value of the population, as defined in eqs. (4) and (6). Differential equations for $b(t)$ and $d(t)$ are derived in Appendix A and exact analytical solutions for these equations are found.

The annual rate of growth, g, is defined as the change in the mean number of circuits in one year divided by its value at the beginning of the year. Thus,

$$g(t) = \frac{m(t+1) - m(t)}{m(t)}. \tag{39}$$

From eqs. (4) and (6) it may be seen that

$$g(t) = \begin{cases} \dfrac{[n_0 + \beta/(\lambda - \mu)]e^{(\lambda-\mu)t}(e^{\lambda-\mu} - 1)}{[n_0 + \beta/(\lambda - \mu)]e^{(\lambda-\mu)t} - \beta/(\lambda - \mu)} & \text{for } \lambda \neq \mu, \\ \dfrac{\beta}{n_0 + \beta t} & \text{for } \lambda = \mu. \end{cases} \tag{40}$$

By observation, it may be noted that if $\beta = 0$, the growth is time independent, whereas if $\beta > 0$, the growth depends on time. A time average value of the growth may be defined to be

$$\bar{g} = \lim_{T \to \infty} \frac{1}{T} \int_0^T g(t)\, dt. \tag{41}$$

It may be easily found that

$$\bar{g} = \begin{cases} 0, & \text{for } \lambda < \mu, \quad \beta \neq 0, \\ 0, & \text{for } \lambda = \mu, \\ e^{\lambda-\mu} - 1 & \text{for all other cases.} \end{cases} \tag{42}$$

When $\lambda = \mu$, it must be borne in mind that $\bar{g} = 0$ does not necessarily mean no growth. In fact, for the $\beta > 0$ case there is linear growth at the rate β [recall eq. (6)]. Thus, it is the immigration factor that represents the growth in this case.

In summary, throughout this paper, the last expression in (42) will be used as the definition of growth with the provision that it is the variable β that actually describes the growth in the $\lambda = \mu$ case. The case $\lambda < \mu$ and $\beta > 0$ will be disregarded.

Churn may be defined in various ways. Its purpose is to quantify the "activity" of the process, i.e., to compare the number of "connects" with the number of "disconnects." For instance, churn may be defined as the ratio of the mean number of *total* connects in one year to the mean number of *net* connects in the same period. This ratio has also been called in-to-net, or in-to-gain ratio.[15] Thus,

$$c(t) = \frac{b(t+1) - b(t)}{m(t+1) - m(t)} = \frac{\text{IN}}{\text{NET}}. \tag{43}$$

This quantity has the easy interpretation of being the expected number of connects in one year for a net increase of one circuit. For instance, a churn of four (which seems to be a typical number[18]) means that four connects are expected for every net increase of one circuit. Of course, it follows that three disconnects are also expected for consistency. The problem with this definition of churn is that for low-growth cases (i.e., when the net increase is almost zero) the ratio of total connects to net may become a very large number. Furthermore, in the case of negative growth, this ratio becomes negative. Thus, the range of values which the churn may take is very large, which makes it a difficult number to work with in data analysis.

For this reason, an alternative definition of churn is introduced and is called turnover. Turnover is the expected number of connects (disconnects) needed to replace the number of disconnects (connects) that occurred in one year, divided by the expected number of circuits in place at the beginning of the year. The words outside the parentheses refer to the positive-growth case in which there are more expected connects than disconnects, and the words in the parentheses refer to the negative-growth case when the reverse is true. The turnover may be written as

$$a(t) = \frac{1}{2}\frac{\text{IN} + \text{OUT} - |\text{IN} - \text{OUT}|}{\text{MEAN}} \tag{44}$$

$$= \frac{\min(\text{IN}, \text{OUT})}{\text{MEAN}}, \tag{45}$$

where

$$\begin{aligned} \text{IN} &= b(t+1) - b(t), \\ \text{OUT} &= d(t+1) - d(t), \\ \text{MEAN} &= m(t). \end{aligned} \tag{46}$$

Thus, a turnover of 0.3 with positive growth indicates that over the next year the expected number of disconnects will be equal to 30 percent of the mean at the beginning of the year. The expected number of connects depends on the growth and will be greater than or equal to the disconnects. (Note that it is not necessarily the *same* circuits that

are connected and disconnected.) From the expressions in Appendix A, eq. (44) may be written as

$$a(t) = \begin{cases} \frac{1}{2}(n_{\text{eff}}e^{\Delta t} - \beta/\Delta)^{-1}[n_{\text{eff}}e^{\Delta t}(e^{\Delta} - 1)(\lambda + \mu)/\Delta \\ \qquad - 2\beta\mu/\Delta - n_{\text{eff}}e^{\Delta t}|(e^{\Delta} - 1)|] & (\lambda \neq \mu), \\ \frac{1}{2}\dfrac{2\lambda\beta t + \lambda\beta + 2\lambda n_0}{n_0 + \beta t}, & (\lambda = \mu), \end{cases} \tag{47}$$

where

$$n_{\text{eff}} = n_0 + \frac{\beta}{\lambda - \mu}.$$

Again, by observation, it may be noted that if $\beta = 0$, the turnover is time independent, whereas if $\beta > 0$, it is time dependent. As in the case of the growth, a time-average value of the turnover may be defined to be

$$\bar{a} = \lim_{T \to \infty} \frac{1}{T} \int_0^T a(t)\, dt. \tag{48}$$

It may be easily found that

$$\bar{a} = \begin{cases} \dfrac{\mu}{\Delta}(e^{\Delta} - 1) & \text{if } \lambda > \mu, \\ \dfrac{\lambda}{\Delta}(e^{\Delta} - 1) & \text{if } \lambda < \mu, \quad \beta = 0, \\ \mu & \text{for all other cases.} \end{cases} \tag{49}$$

Throughout this paper, the above expressions for the turnover will be used. As mentioned before, the case $\lambda < \mu$ and $\beta > 0$ will be disregarded.

From preliminary data analysis, it has been found that typical values of growth and turnover for special services fall into the range -0.15 to $+0.15$ for the growth and 0 to 1 for the turnover. Nominal values of $\bar{g} = 0.1$ and $\bar{a} = 0.3$ were chosen in this paper.

To do a study using stochastic special services demand, numerical values of the parameters of the model are needed. Accurate values, if such values exist, may only be found by careful data analysis of historical demands. Approximate values, however, may be found as follows. First equate the mean [eqs. (4) or (6)] to the special services forecast to obtain values for Δ and β. Then equate the variance [eqs. (5) or (7)] to some measure of the forecast uncertainty to determine λ and μ. For some cases, given below, one may conveniently use eqs. (42) and (49) to write the mean as a function of growth alone, and the variance as a function of both growth and turnover. The results are as

follows.

(*i*) Exponential growth with no immigration ($\lambda \neq \mu$, $\beta = 0$)

$$m(t) = n_0(1 + \bar{g})^t, \qquad (50)$$
$$v(t) = n_0 \left(\frac{2\bar{a}}{\bar{g}} \pm 1 \right) (1 + \bar{g})^t[(1 + \bar{g})^t - 1],$$

where the positive (negative) sign refers to $\lambda > \mu$ ($\lambda < \mu$),

(*ii*) Linear growth ($\lambda = \mu$, $\beta \geq 0$)

$$m(t) = \beta t + n_0, \qquad (51)$$
$$v(t) = \bar{a}\beta t^2 + (2\bar{a}n_0 + \beta)t.$$

5.2 Margin and minimum capacity increments

Define the quantity $h(t)$ as the probability of a held order, i.e., the probability that at least one service order is delayed due to lack of spare facilities. Then

$$h(t) = \sum_{n=d+1}^{\infty} P_n(t) = 1 - F_d(t),$$

where $d = d(t)$ is the total number of servers (facilities) at time t. The quantity $h(t)$ is similar but not identical to the transient time congestion function (see Appendix B). Computationally, $d(t)$ may be determined from the state probabilities by requiring $h(t)$ to be less than or equal to some predetermined number, **h**. Thus

$$d(t) = \min\{d = 0, 1, \cdots | \sum_{n=d+1}^{\infty} P_n(t) \leq \mathbf{h}\}. \qquad (52)$$

Of course, the actual sum involved in the computation is not infinite, since the condition in eq. (52) is equivalent to $\sum_{n=0}^{d} P_n(t) \geq 1 - \mathbf{h}$. The level $d(t)$ can be viewed as the sum of the mean number of circuits and a quantity which may be called margin. Thus, given any time $t > 0$, the margin is the capacity which must be built at $t_0 = 0$ in excess of the mean $m(t)$, in order to meet the demand, within the maximum held-order probability, **h**.

A significant quantity is the ratio of the margin to the mean in percent which will hereafter be referred to as the percent margin. Figure 6 shows a plot of this ratio as a function of time for various values of growth and no immigration. Turnover has been taken to be 0.3, the initial number of circuits 5, and the maximum probability of a held order 0.05. As may be seen, the percent margin increases with time, and generally less percent margin is needed for larger growth rates. The sensitivity of the percent margin with respect to turnover, for a growth of 0.1, is shown in Fig. 7. It may be seen that the larger the turnover, i.e., the larger the "activity" in the network, the more

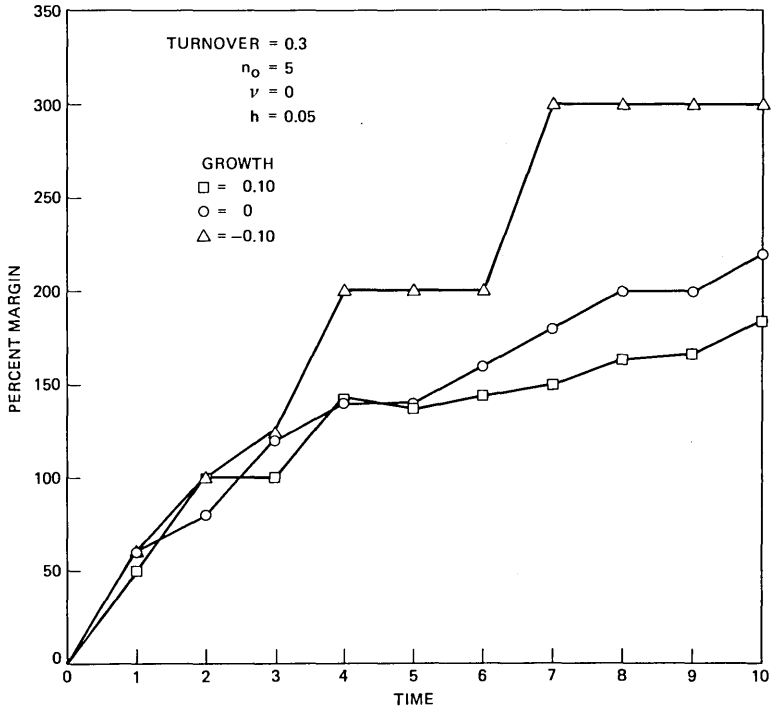


Fig. 6—Sensitivity of percent margin to growth (no immigration).

margin one has to build to provide the same maximum held-order probability.

The concept of margin has several applications, one of which is the determination of an appropriate capacity increment at each expansion. Given a minimum desired time, T, between expansions, it would be useful to determine the minimum capacity increment, c, which, if installed at time t, will exhaust in $[t, t + T]$ with a probability that is no larger than **h**, or equivalently, the increment which will last the interval T with a probability greater than or equal to $1 - \mathbf{h}$. Thus, the condition on c may be written as

$$\text{Prob}\{\mathcal{N}(t + \xi) \leq n_0 + c \,|\, \mathcal{N}(t) = n_0, \quad \forall\, \xi \in [0, T]\} \geq 1 - \mathbf{h}. \quad (53)$$

Since the λ and μ coefficients are time independent, the process is time homogeneous. Consequently, changing the origin of time does not affect the problem. Choosing it to be at t is equivalent to setting $t = 0$ in the above expression, and determination of c reduces to finding

$$\min\left\{c = 0, 1, \cdots \,\middle|\, \sum_{n=0}^{n_0+c} P_n(\xi) \geq 1 - \mathbf{h}, \quad \forall\, \xi \in [0, T]\right\}. \quad (54)$$

Tables may be set up permitting direct reading of the values of c corresponding to the growth, turnover, initial state parameters, and to the time interval T. Some typical results are plotted in Fig. 8.

For completeness, it must be mentioned that in problems of the type described above, questions about service-order queue disciplines must be entertained. A careful consideration of this aspect of the problem is beyond the scope of the present analysis, and the queue discipline implicitly followed has been Blocked Customers Held (BCH). See Appendix B.

5.3 Effect of randomness: Aggregation benefits

An inspection of Fig. 6 shows that the percent margin needed is large, for the particular case examined. Since the initial state considered is rather small ($n_0 = 5$), an interesting question is to find out whether the percent margin can be reduced by combining demand to form larger quantities, in the hope that the statistics of the aggregated

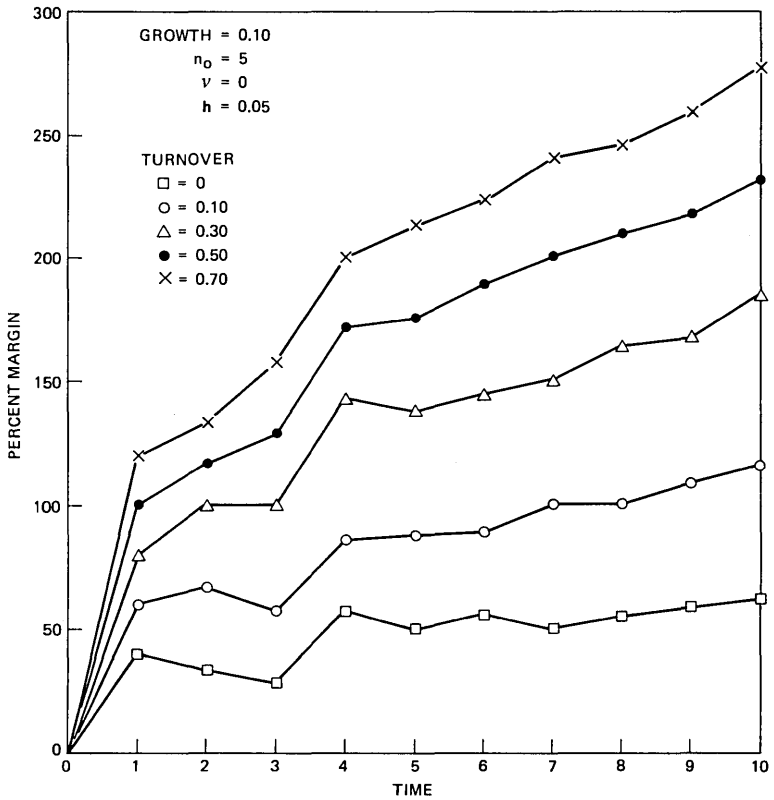


Fig. 7—Sensitivity of percent margin to turnover (no immigration).

process will be better behaved, i.e., less susceptible to fluctuations. Our studies have shown that benefits are indeed obtained by aggregation. Figure 9 illustrates the results. The time evolution of the percent margin is plotted as the initial state of the system is varied, for fixed values of growth, turnover, held-order probability, and for no immigration. Two important observations may be made. The first one is the fact that the percent margin decreases as the initial state of the system increases. An implication of this behavior, for example, is the following: Suppose demand between two points is being satisfied by two independent routes (with initial number of circuits $n_0^{(1)}$ and $n_0^{(2)}$, respectively). Benefits would be obtained by combining the two demands on one route (with initial number of circuits $n_0^{(1)} + n_0^{(2)}$) because the margin one would have to build in this case is less than in the nonaggregated case, the held-order probability being the same. The second observation is that the change in the percent margin with respect to n_0 is larger for small values of n_0. The implication is that the benefits will be especially significant when aggregating small demands.

It must be mentioned that the conclusions about aggregation benefits in the example given above were based on an examination of Fig. 9. The implicit assumption was that the combined process would obey the same birth and death equations as each single process, and

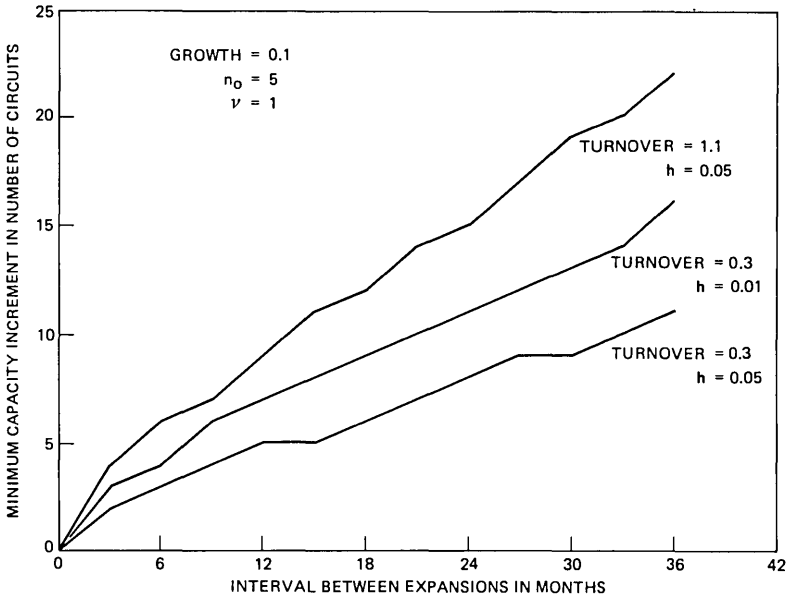


Fig. 8—Minimum capacity increments.

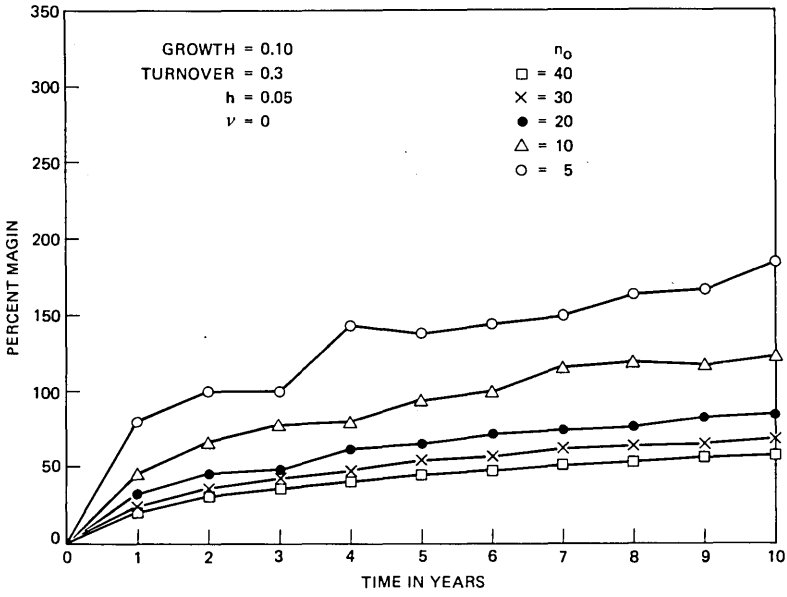


Fig. 9—Sensitivity of percent margin to initial demand (no immigration).

that it would be described by the same pair of transition rates, λ and μ. This assumption is only true if all the individual processes are of the same type, i.e., if they all have identical transition rates. It turns out, however, that the method for determining margin described in the previous sections may be extended with little additional effort to the general case in which there are M simultaneous processes characterized by a set of transition rates $\{\lambda^i, \mu^i\}$ for $i = 1, \cdots, M$. Since the processes are independent, the joint probability distribution is the product of the individual distributions,

$$P_{n_1,n_2,\ldots,n_M}(t) = P^{(1)}_{n_1}(t)P^{(2)}_{n_2}(t) \cdots P^{(M)}_{n_M}(t). \tag{55}$$

The probability of being in some level $\tilde{n}$, regardless of the composition of that level, may then be written

$$\tilde{P}_{\tilde{n}}(t) = \sum_{n_1}{}' \sum_{n_2}{}' \cdots \sum_{n_M}{}' P^{(1)}_{n_1}(t)P^{(2)}_{n_2}(t) \cdots P^{(M)}_{n_M}(t),$$
$$n_1 + n_2 + \cdots + n_M = \tilde{n}. \tag{56}$$

The sums are over all values of $n_1, n_2, \cdots, n_M$ such that $n_1 + n_2 + \cdots + n_M = \tilde{n}$. The primes over the summations are an indication of this restriction. The margin for the combined process may then be determined from the mean $\tilde{m}(t)$, and the quantity $\tilde{d}(t)$, analogous to that defined in Section 5.2. The mean for the combined process is

simply the sum of the individual means,

$$\tilde{m}(t) = m_1(t) + m_2(t) + \cdots + m_M(t), \tag{57}$$

and $\tilde{d}(t)$ may be obtained from an expression similar to eq. (52), namely,

$$\tilde{d}(t) = \min\left\{ d = 0, 1, \cdots \;\middle|\; \sum_{\tilde{n}=d+1}^{\infty} \tilde{P}_{\tilde{n}}(t) \leq \mathbf{h} \right\}. \tag{58}$$

5.4 Effect of immigration

Immigration affects the problem in several ways. First, it eliminates the absorbing state at $n = 0$, and consequently the probability of extinction, for all cases except the $\lambda < \mu$, $\beta \neq 0$ case. Furthermore, the $\lim_{t\to\infty} P_n(t) = 0$ $(n > 0)$, for all cases except the case mentioned above, for which the limit is nonzero. Finally, a nonzero value of β gives the model flexibility to represent linear growth (for $\lambda = \mu$) as well as exponential growth (for $\lambda \neq \mu$). It is interesting to note that for the $\lambda > \mu$ case, immigration actually reduces the percent margin for moderately large to large times, as seen in Fig. 10. This behavior is due to the fact that introducing immigration increases the mean (which tends to decrease the percent margin) faster than it increases the variance (which tends to increase the percent margin).

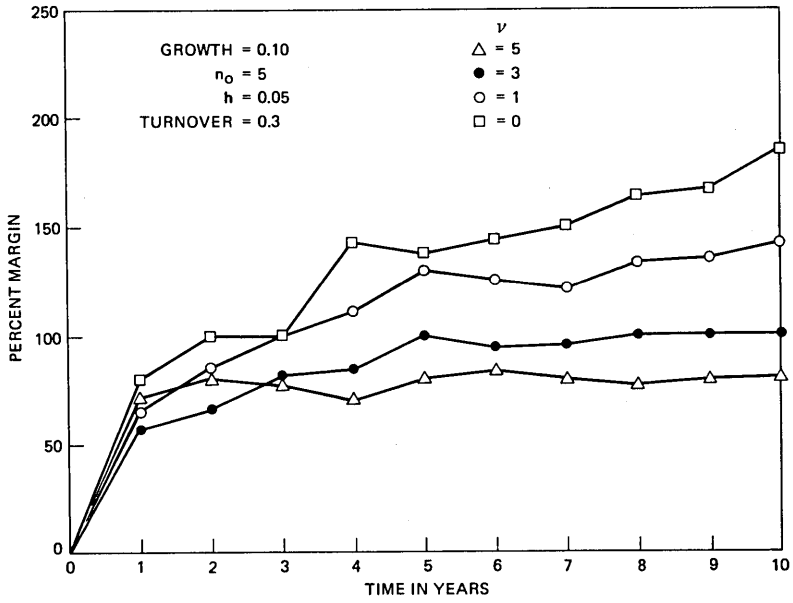


Fig. 10—Effect of immigration on percent margin.

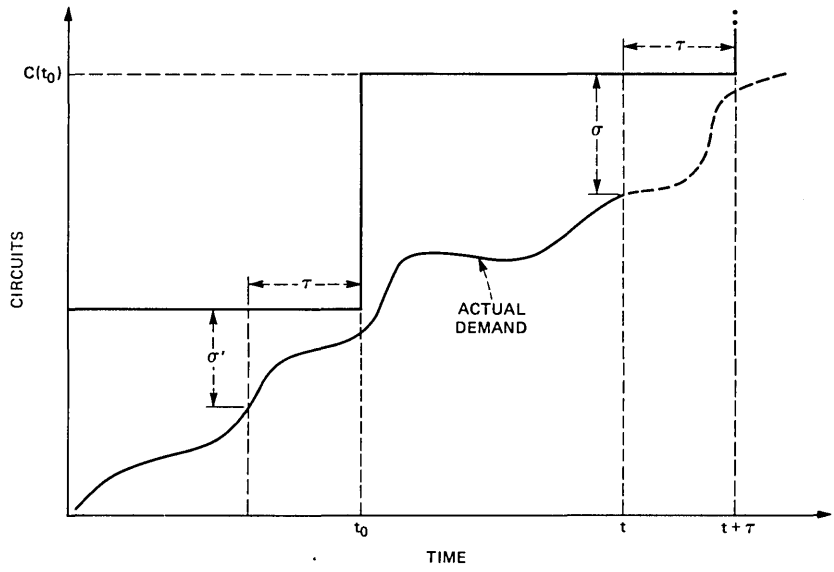


Fig. 11—Spare threshold.

5.5 Other applications: Spare threshold

In capacity expansion problems, a question that arises is when to expand. If facilities could be installed instantaneously, the expansion time would be simply the time at which remaining spare exhausted. However, there usually is a *lead time*, τ, between the moment facilities are ordered and the time they are actually installed. If existing spare can be monitored, it would be useful to determine *a priori* what the particular level of spare would be at the time when new facilities should be ordered. This information would then yield the order time, since as soon as that spare level is attained, it is time to order. This value of spare, or spare threshold, σ, is the amount of remaining capacity which will exhaust in $[t, t + \tau]$ with a probability that is no larger than **h**, or equivalently, the amount of capacity which will last the interval τ with a probability greater than or equal to $1 - \mathbf{h}$. If the last expansion occurred at t_0, providing a total capacity of $C(t_0)$ (see Fig. 11), the constraint on σ may be written as

$$\text{Prob}\{\mathcal{N}(t + \xi) \leq C(t_0) \mid \mathcal{N}(t) = C(t_0) - \sigma, \quad \forall\, \xi \in [0, \tau]\} \geq 1 - \mathbf{h}. \quad (59)$$

Time homogeneity of the process allows setting $t = 0$ in the above expression, as discussed in Section 5.2, and the determination of σ reduces to finding

$$\min\left\{\sigma = 0, 1, \cdots \;\middle|\; \sum_{n=0}^{C(t_0)} P_n(\xi) \geq 1 - \mathbf{h}, \quad \forall\, \xi \in [0, \tau]\right\}, \quad (60)$$

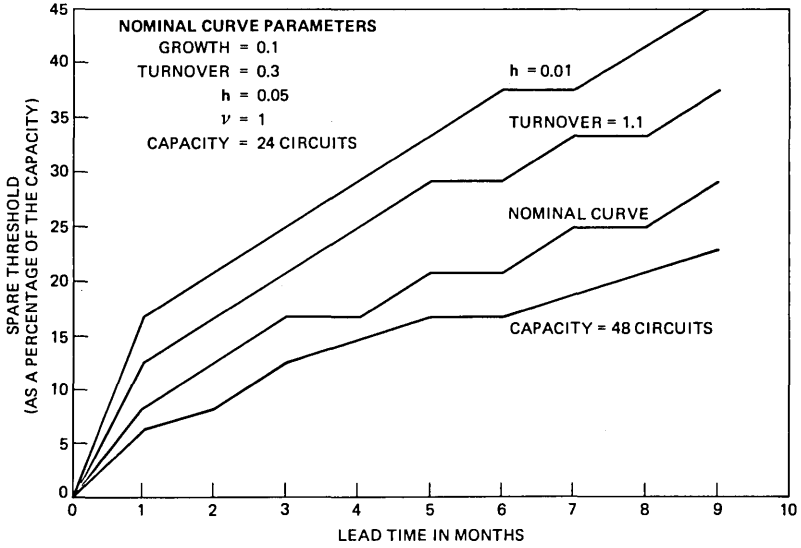


Fig. 12—Spare threshold as a function of lead time.

where σ appears implicitly in the initial condition used to evaluate the state probabilities $P_n(\xi)$. Again, tables may be set up permitting the direct reading of the values of σ corresponding to the growth, turnover, capacity levels, and the lead time, τ. Some typical results are plotted in Fig. 12, where σ is shown as a percentage of the total capacity.

VI. CONCLUSION

A summary of the main results has been given in Section II. Explicit solutions for the birth-death process in which the births and deaths are proportional to the state have been derived, and some of their applications to capacity expansion problems have been discussed. A method for determining margin has been described in detail for the case in which one birth-death process exists with exponential growth characterized by a pair of transition rates, λ and μ, and it was shown how to extend the method to cases in which there were several simultaneous processes.

In addition to its applications to capacity expansion problems, the proposed model is useful in assessing the potential of routing strategies for special services. It has already been shown, for example, that benefits are to be expected by aggregating small demands. These conclusions were based largely on service robustness considerations. To obtain more comprehensive results about routing strategies, it is clear that cost robustness considerations must be addressed as well.

VII. ACKNOWLEDGMENTS

The author would like to thank Bob Klessig for many valuable discussions without which this work would not have been possible.

The state probabilities were initially derived by applying the inverse Z transform to the generating function. The author is grateful to Ward Whitt for suggesting the easier approach described in Section III. The author would also like to thank Paul Burke for valuable discussions.

APPENDIX A

Expressions for b(t) and d(t) (Section 5.1)

As indicated in Section 2.1,

$$\lambda_n\delta + o(\delta)^* = \text{Probability of a birth in } [t, t+\delta] \text{ given that the system was in state } n \text{ at time } t,$$

$$\mu_n\delta + o(\delta) = \text{Probability of a death in } [t, t+\delta] \text{ given that the system was in state } n \text{ at time } t.$$

Letting $\mathcal{N}(t)$ be the random variable representing the number of circuits at time t, the total probability of a birth in $[t, t+\delta]$ may be written as

$$\begin{aligned}
&\text{Prob}\{\text{a birth in } [t, t+\delta]\} \\
&= \sum_{n=0}^{\infty} \text{Prob}\{\text{a birth in } [t, t+\delta] \text{ and } \mathcal{N}(t) = n\} \\
&= \sum_{n} \text{Prob}\{\text{a birth in } [t, t+\delta] \mid \mathcal{N}(t) = n\}\text{Prob}\{\mathcal{N}(t) = n\} \\
&= \sum_{n} [\lambda_n\delta + o(\delta)]P_n(t), \qquad (61)
\end{aligned}$$

where the certain event and Bayes' rule were successively used. Similarly, the total probability of a death in $[t, t+\delta]$ is $\sum_n [\mu_n\delta + o(\delta)]$ $P_n(t)$. With this information, a differential equation for $b(t)$ may be set up as follows:

$$\begin{aligned}
b(t+\delta) = b(t)&[\text{Probability no event } \textit{or} \text{ a death}] \\
&+ [b(t)+1][\text{Probability of a birth}] + o(\delta),
\end{aligned}$$

where $o(\delta)$ is the contribution of more than one birth.

$$\begin{aligned}
b(t+\delta) = b(t)&\Big[\Big\{1 - \sum_n (\lambda_n\delta + \mu_n\delta + o(\delta))P_n\Big\} \\
&+ \sum_n (\mu_n\delta + o(\delta))P_n\Big] \\
&+ [b(t)+1]\sum_n [\lambda_n\delta + o(\delta)]P_n(t) + o(\delta). \qquad (62)
\end{aligned}$$

---

* $o(\cdot)$: $R^1 \to R^1$ is such that $\lim_{\delta\to 0} \frac{o(\delta)}{\delta} = 0$.

In the limit $\delta \to 0$, eq. (62) becomes

$$\frac{d}{dt} b(t) = \sum_{n=0}^{\infty} \lambda_n P_n(t). \tag{63}$$

In a similar manner, the following differential equations for $d(t)$ and for $m(t)$ may be obtained:

$$\frac{d}{dt} d(t) = \sum_{n=0}^{\infty} \mu_n P_n(t), \tag{64}$$

$$\frac{d}{dt} m(t) = \sum_{n=0}^{\infty} (\lambda_n - \mu_n) P_n(t). \tag{65}$$

Since

$$m(t) = \sum_{n=0}^{\infty} n P_n(t),$$

one can use relations (3) to write eq. (65) as a differential equation for $m(t)$. Its solutions have already been given in eqs. (4) and (6). With knowledge of the mean, eqs. (63) and (64) may be solved for $b(t)$ and $d(t)$, respectively. The results are easily found to be

$$b(t) = \begin{cases} \dfrac{\lambda}{\Delta} n_{\text{eff}}(e^{\Delta t} - 1) - \dfrac{\beta\mu t}{\Delta} & (\lambda \neq \mu), \\ \dfrac{\lambda\beta t^2}{2} + (\lambda n_0 + \beta)t & (\lambda = \mu), \end{cases} \tag{66}$$

$$d(t) = \begin{cases} \dfrac{\mu}{\Delta} n_{\text{eff}}(e^{\Delta t} - 1) - \dfrac{\beta\mu t}{\Delta} & (\lambda \neq \mu), \\ \dfrac{\lambda\beta t^2}{2} + \lambda n_0 t & (\lambda = \mu), \end{cases} \tag{67}$$

where

$$n_{\text{eff}} = n_0 + \frac{\beta}{\Delta}$$

and

$$\Delta = \lambda - \mu.$$

APPENDIX B

Queue Discipline and Held-Order Probability

In the model described in this paper, the queue discipline followed is Blocked Customers Held (BCH). Let the random variable T denote the sojourn time of the customer, i.e., the total time he spends in the system, either waiting for service or being served. The assumption inherent in the BCH queue discipline is that the customer will spend

time T in the system, after which he will depart, regardless of whether he has been served or is still waiting for service. The choice of $\mu_n = n\,\mu$ implies that the sojourn times have a negative exponential distribution.

In special services, if the sojourn time distribution is in fact negative exponential, then the queue discipline used here should be correct. If, on the other hand, it is the service-time distribution that is negative exponential, then the BCH queue discipline assumed here may still be approximately correct if the average waiting time of a customer is much smaller than the average service time.

To compute the held-order probability, care must be given as to whether the held order is seen by an outside observer or by an arriving customer. For processes with Poisson input, it is well known that the distribution $P_n(t)$ seen by an outside observer is identical to the distribution $\pi_n(t)$ seen by an arriving customer (see Section 3.2 of Ref. 3), and hence the distinction is unimportant. For the process described in this paper, however, the distributions are different. Define

$$P_n(t) = \text{Prob}\{\mathcal{N}(t) = n\}, \tag{68}$$

$$\pi_n(t) = \text{Prob}\{\mathcal{N}(t) = n \mid \text{a customer arrives at } t^+\}. \tag{69}$$

Expression (69) is the probability that a customer who arrives at t finds n *other* customers being served or waiting to be served. Letting the event A refer to the arrival of a customer in the interval $(t,\ t + \delta]$, and using conditional probabilities, one may write $\pi_n(t)$ as the following limit, if it exists:

$$\pi_n(t) = \frac{\lim_{\delta \to 0} \text{Prob}\{\mathcal{N}(t) = n, A\}}{\lim_{\delta \to 0} \text{Prob}\{A\}}$$

$$= \lim_{\delta \to 0} \frac{\text{Prob}\{A \mid \mathcal{N}(t) = n\}\text{Prob}\{\mathcal{N}(t) = n\}}{\sum_{j=0}^{\infty} \text{Prob}\{A \mid \mathcal{N}(t) = j\}\text{Prob}\{\mathcal{N}(t) = j\}}.$$

Since $\text{Prob}\{A \mid \mathcal{N}(t) = n\} = \lambda_n \delta + o(\delta)$,* one obtains, with the help of (68),

$$\pi_n(t) = \frac{\lambda_n P_n(t)}{\sum_{j=0}^{\infty} \lambda_j P_j(t)}. \tag{70}$$

The held-order probability may thus be defined as

$$h(t) = \sum_{n=d+1}^{\infty} P_n(t) \quad \text{as seen by an outside observer} \tag{71}$$

* $o(\cdot)$: $R^1 \to R^1$ is such that $\lim_{\delta \to 0} \frac{o(\delta)}{\delta} = 0$.

or

$$h'(t) = \frac{\sum_{n=d}^{\infty} \lambda_n P_n(t)}{\sum_{j=0}^{\infty} \lambda_j P_j(t)} \quad \text{as seen by an arriving customer,} \qquad (72)$$

where d is the number of servers. Expression (72) is the conditional probability that if a customer were to arrive at t, he would find all the servers engaged. This quantity is known in congestion theory as the transient call-congestion function.[19] Expression (71) is the probability that at least one customer is waiting to be served. This quantity is similar but not identical to the transient time-congestion function, $S(t)$, which is the probability that all servers are busy at time t, and which may be written as [19]

$$S(t) = \sum_{n=d}^{\infty} P_n(t). \qquad (73)$$

For Poisson input ($\lambda_n = \lambda$), it may be shown that

$$h'(t) = S(t) > h(t).$$

For the Kendall process, the relationship is

$$h'(t) > S(t) > h(t).$$

For the Poisson input case, since $S(t)$ is equal to $h'(t)$, the time-congestion function may be used as a meaningful measure of the held orders. For the Kendall process, on the other hand, $S(t)$ does not describe the held orders as seen by either an arriving customer or an outside observer. Consequently, the time-congestion function is not believed to be a meaningful measure of the held orders. Throughout this paper, expression (71) was used for the held-order probability, although eq. (72) could have been used instead.

Both $h(t)$ and $h'(t)$ are instantaneous quantities. Since the Kendall process with $\lambda > \mu$ is not ergodic (i.e., space averaging is different than time averaging), a space average must be made when measuring either of these quantities. Thus, one cannot measure $h(t)$ or $h'(t)$ by examining one sample for a long enough time; rather, one needs an ensemble of samples. Because of these measurement difficulties, an open question remains as to whether this type of held-order probability is the best measure of the provided service, or whether other quantities such as the time average of $h(t)$ or $h'(t)$, or the duration of the held order might be more meaningful. Nevertheless, it is clear that eqs. (71) and (72) are some measure of the provided service and, as such, are useful when comparing different special services provisioning methods meant to provide the same level of service.

REFERENCES

1. W. Feller, *An Introduction to Probability Theory and Its Applications*, 3rd ed., Vol. I, New York: John Wiley, 1968, pp. 444–82.
2. K. R. Swaminathan, Bell Laboratories, private communication.
3. R. B. Cooper, *Introduction to Queueing Theory*, New York: Macmillan, 1972, pp. 62–103.
4. D. G. Kendall, "On the Generalized Birth-and-Death Process," Ann. Math. Stat., *19* (1948), pp. 1–15.
5. D. G. Kendall, "On Some Modes of Population Growth Leading to R. A. Fisher's Logarithmic Series Distribution," Biometrika, *35* (1948), pp. 6–15.
6. D. G. Kendall, "Stochastic Processes and Population Growth," J. Roy. Stat. Soc., *B11* (1949), pp. 230–65.
7. W. Ledermann and G. E. H. Reuter, "Spectral Theory for the Differential Equations of Simple Birth and Death Processes," Philos. Trans. Roy. Soc., London, *A246* (1954), pp. 321–69.
8. W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. II, New York: Wiley, 1968, pp. 479–81.
9. N. Arley and V. Borchsenius, "On the Theory of Infinite Systems of Differential Equations and Their Application to the Theory of Stochastic Processes and the Perturbation Theory of Quantum Mechanics," Acta Math., *76* (1945), pp. 261–322 (especially p. 299).
10. N. T. J. Bailey, *The Elements of Stochastic Processes*, New York: Wiley, 1964, pp. 84–105.
11. N. U. Prabhu, *Stochastic Processes*, New York: Macmillan, 1965, Chapter 4.
12. B. Wallstrom, "Congestion Studies in Telephone Systems with Overflow Facilities," Ericsson Tech., *22*, No. 3 (1966).
13. A. Y. Khintchine, *Mathematical Methods in the Theory of Queueing*, New York: Hafner, 1969.
14. D. L. Jagerman, "Nonstationary Blocking in Telephone Traffic," B.S.T.J., *54*, No. 3 (March 1975), pp. 625–61.
15. J. Freidenfelds, Bell Laboratories, private communication.
16. J. Freidenfelds, "Capacity Expansion when Demand is a Birth-Death Random Process," Oper. Res., *28*, No. 3, Part 2 (May–June 1980).
17. H. Luss and W. Whitt, Bell Laboratories, private communiction.
18. J. C. Lagarias, Bell Laboratories, private communication.
19. R. Syski, *Introduction to Congestion Theory in Telephone Systems*, Edinburgh, Great Britain: Oliver and Boyd, 1960, Chapter 5.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 1, January 1981
Printed in U.S.A.

High-Frequency Behavior of Waveguides with Finite Surface Impedances

By C. DRAGONE

(Manuscript received January 25, 1980)

We derive the modes inside a cylindrical waveguide of finite surface impedances, assuming the waveguide transverse dimensions are large compared to the wavelength λ. This paper restricts its consideration to the modes with $\beta \approx k$, where β is the propagation constant and $k = 2\pi/\lambda$. For these modes we show that asymptotically, for large values of k, the field ψ becomes infinitesimal (of the same order of $1/k$) at the boundary. Taking this into account, we obtain simple expressions for the asymptotic properties of ψ for large k. The theory applies to a variety of waveguides: corrugated waveguides, optical fibers, waveguides with smooth walls of lossy metal, and so on. An important property of the modes considered here is that their attenuation constant is very low, i.e., asymptotic to $1/k^2$ for large k. Thus, these modes are useful for long-distance communication. Another consequence of $\psi \to 0$ at the boundary is that for large k the distribution of ψ inside the boundary is essentially independent of the boundary parameters, i.e., independent of the surface impedances in the longitudinal and transverse directions. This consequence implies that the same radiation characteristics of the corrugated feed can be obtained using other structures and, therefore, construction can be simplified in many cases, with little sacrifice in performance. We also derive general expressions for ψ and the propagation constant β.

I. INTRODUCTION

It is known[1-8] that in certain waveguides the field becomes, under certain conditions, very small at the boundary. Consider, for instance, a corrugated waveguide of radius a and let λ be the free-space wavelength. This waveguide is characterized at the boundary by finite surface impedance Z_z in the longitudinal direction. The frequency dependence of Z_z, which is determined by the depth of the corruga-

tions, causes the transverse field distribution $\psi(x, y)$ of a mode to vary with the frequency $k = 2\pi/\lambda$. However, this frequency dependence virtually disappears (for all modes except surface waves) if the waveguide dimensions are large enough. In fact, one finds that $\psi(x, y)$ approaches for $ka \to \infty$ a frequency independent distribution that vanishes at the boundary.[5] This behavior is responsible for the low attenuation constant, for the excellent radiation characteristics, and the wide bandwidth of corrugated waveguides.[5] We show here that the same behavior also occurs, under quite general conditions, in a variety of uncorrugated waveguides.[1-21] Figures 1a and 1b show two examples, a dielectric waveguide[7,8,13-16] of general cross section and a hollow waveguide with metal walls coated by a dielectric layer.[4,17] Other examples can be obtained by modifying the boundary conditions in a variety of different ways. For instance, several dielectric layers may be used in Fig. 1b, or a metal grid of transverse wires may be placed at the boundary, as pointed out in Section II. Other examples are the waveguides of dielectric or lossy metal considered in Ref. 2. We now outline the main assumptions.

Consider a cylindrical waveguide with an interior region of homogeneous and isotopic material, as in Fig. 1c. Let Z and k be the wave impedance and propagation constant for a plane wave in the interior region, and let C denote the boundary. Then

$$Z = \sqrt{\frac{\mu}{\epsilon}}, \qquad k = \omega\sqrt{\epsilon\mu}. \tag{1}$$

Consider a mode with propagation constant β and let $\mathbf{E}_t$ denote the transverse component of the electric field,

$$\mathbf{E}_t = \psi(x, y)e^{-j\beta z}. \tag{2}$$

Let $2a$ be a characteristic dimension of the waveguide, for instance the width in the x direction as in Fig. 1c. We are concerned about the asymptotic behavior of $\psi(x, y)$ for large values of ka. Consideration will be restricted to the modes for which the propagation constant β approaches k, as $ka \to \infty$. Thus, we assume

$$\beta \to k \qquad \text{for} \quad ka \to \infty. \tag{3}$$

This excludes surface waves, as pointed out in the following section. Then, a property of the modes considered here is that $\mathbf{E}$ and $\mathbf{H}$ become transverse, in the limit as $ka \to \infty$,

$$\lim_{ka\to\infty} E_z = H_z = 0. \tag{4}$$

Another property is that asymptotically, per large ka, a set of linear relations exist at the boundary among the tangential components of $\mathbf{E}$

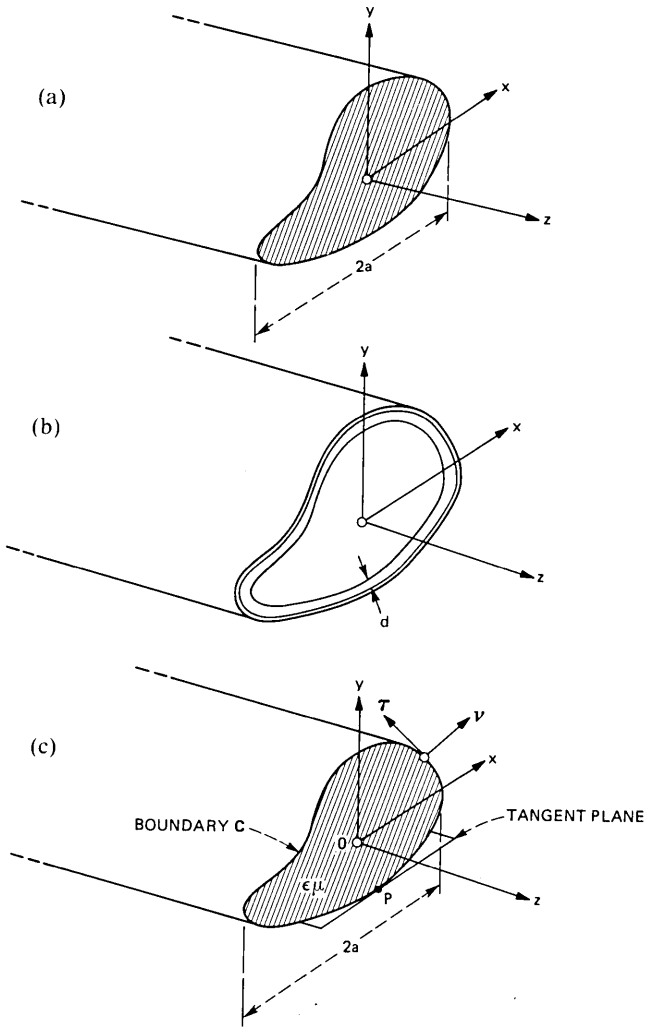


Fig. 1—Three examples of cylindrical waveguides: (a) a dielectric rod, (b) a waveguide with metal walls coated by a thin dielectric layer, and (c) a waveguide with boundary conditions shown in Fig. 2.

and **H**. It is convenient to write these relations in the form

$$\begin{bmatrix} ZH_\tau \\ E_\tau \end{bmatrix} = j[H] \begin{bmatrix} E_z \\ -ZH_z \end{bmatrix} \quad \text{on } C, \tag{5}$$

where $[H]$ is a 2×2 matrix and H_τ, E_τ denote the components of **H**, **E** in the direction of the unit vector τ in Fig. 1c. These relations

together with eq. (4) give at the boundary

$$\lim_{ka\to\infty} E_\tau = H_\tau = 0 \qquad \text{on } C, \tag{6}$$

provided the matrix $[H]$ does not diverge for $ka \to \infty$,

$$[H] \neq \infty \qquad \text{for} \quad ka \to \infty. \tag{7}$$

Throughout this article, we assume conditions (3), (5), and (7). Condition (5) is discussed in the following section, where it is pointed out that for most waveguides considered here, $[H]$ is a diagonal matrix. In this case it is convenient to define at the boundary surface impedances Z_z and Z_τ by writing

$$E_\tau = Z_\tau H_z, \qquad E_z = -Z_z H_\tau, \tag{8}$$

where Z_z is the longitudinal impedance, and Z_τ is the transverse impedance. Then condition (7) requires

$$Z_\tau,\ 1/Z_z \neq \infty \qquad \text{for} \quad ka \to \infty. \tag{9}$$

It is important to realize that this requirement is violated in a number of cases. It is violated in a hollow waveguide with metal walls of perfect conductivity, since then $Z_z = 0$. Furthermore, in a corrugated waveguide with corrugations of depth d, the longitudinal impedance Z_z is determined by kd, and there are certain frequencies for which $Z_z = 0$. A similar situation arises in Fig. 1b where both Z_τ and Z_z vary with kd. Throughout this article it will be assumed that the quantities

$$\frac{1}{ka}, \qquad \frac{1}{ka}\frac{Z_\tau}{Z}, \qquad \frac{1}{ka}\frac{Z}{Z_z}$$

are small. Therefore, the results will not apply in the vicinity of the above frequencies.

A direct consequence of condition (6) is that the boundary values of ψ vanish, in the limit as $ka \to \infty$,

$$\lim_{ka\to\infty} \psi(x, y) = 0 \qquad \text{on } C. \tag{10}$$

Another consequence is that ψ approaches a distribution ψ_∞ independent of ka, for $ka \to \infty$. For finite ka,

$$\psi = \psi_\infty + \delta\psi, \tag{11}$$

where $\delta\psi$ (but not ψ_∞) varies with ka and

$$\lim_{ka\to\infty} \delta\psi = 0. \tag{12}$$

Notice condition (10) implies that

$$\psi_\infty = 0 \qquad \text{on } C. \tag{13}$$

These results are of practical interest for several reasons. In the design of a feed,[6] it is desirable that the boundary values of ψ be small [as implied by eq. (10)] since these values determine radiation in the sidelobes due to edge diffraction at the aperture. Furthermore, for broadband applications, it is desirable that the frequency dependent part $\delta\psi$ of the aperture illumination be small, as implied by eq. (12). Finally, in a corrugated waveguide, or a waveguide with metal walls coated by dielectric layers, power is lost only at the walls and, therefore, it is determined by the boundary values of ψ. Since these boundary values vanish for $ka \to \infty$, the attenuation constant for the above waveguides for large ka is very small[1,2,4,18-21]; it is asymptotic to ka^{-2}. We shall see that in general ψ_∞ is independent of $[H]$ and, therefore, a variety of different waveguides, with different $[H]$ but the same boundary shape, give rise to the same ψ_∞. This explains the similarity, noted in Ref. 9, between the modes of a corrugated waveguide and those of an optical fiber, a dielectric lined waveguide,[4] and a hollow dielectric waveguide.[2] This similarity implies that essentially the same radiation characteristics of corrugated waveguides can also be obtained with a variety of other waveguides.

In the particular case of the optical fiber, some of our results are implied by the asymptotic expressions derived in Ref. 7. Exact solutions for the modes of the corrugated waveguide,[12] the optical fiber,[7,8] and the hollow waveguide of dielectric[2] are known for circular geometry. For a rectangular cross section, only approximate solutions[3,22] are known, except in special cases.[23] Exact solutions for the slab waveguide are given in Refs. 8 and 24. In all these cases one finds that condition (3) implies condition (10). Measurements of the radiation characteristics of a dielectric horn are described in Refs. 25 and 26. Some of the properties derived here apply also to propagation in stratified media.[27-29] The use of surface impedances to characterize a boundary is discussed in Ref. 30.

II. BOUNDARY CONDITIONS FOR $ka \to \infty$

We now derive and discuss eq. (5). Figure 1c shows a waveguide directed along the z axis and of general cross section in which ν is the outwardly directed normal and τ is a unit vector tangential to the boundary,

$$\tau = \mathbf{i}_z \times \nu. \tag{14}$$

The medium inside the boundary is assumed to be homogeneous and isotropic. Let C denote the closed contour of the boundary in the plane $z = 0$.

We are concerned with the properties of $\psi(x, y)$ in a waveguide of large transverse dimensions. Thus, consider a mode propagating in the

waveguide of Fig. 1b and suppose the width $2a$ is increased keeping the dielectric thickness d fixed. The resulting behavior of $\psi(x, y)$ as $ka \to \infty$ can be derived exactly in two cases, when (see Appendix C)

$$\frac{\partial \psi}{\partial y} = 0 \tag{15}$$

and when the boundary is a circle. In both cases one finds that for most of the modes $\beta \to k$, as $ka \to \infty$. For these modes, the normalized field amplitude $\psi(x, y)/\psi(0, 0)$ becomes infinitesimal at the boundary for $ka \to \infty$. For the remaining modes, those for which β does not approach k, just the opposite behavior takes place: The field becomes confined to the immediate vicinity of the walls, degenerating into a surface wave with propagation constant determined by the surface impedances of the walls. Here, consideration will be restricted to the modes satisfying condition (3). An important property of these modes is that asymptotically, for large ka, the surface impedances Z_τ and Z_z become independent of ka. In fact, if one writes

$$X = -j\frac{Z_\tau}{Z}, \qquad Y = -j\frac{Z}{Z_z}, \tag{16}$$

then in Fig. 1b

$$X \to \frac{1}{\sqrt{n^2 - 1}} \tan(\sqrt{n^2 - 1}\, kd), \tag{17}$$

$$Y \to -\frac{n^2}{\sqrt{n^2 - 1}} \frac{1}{\tan(\sqrt{n^2 - 1}\, kd)}, \tag{18}$$

as shown in Appendix C. Thus, Z_τ and Z_z depend only on the refractive index n and the thickness kd of the dielectric layer.

For a circular boundary, the above relations can be derived rigorously by expressing the field in terms of Bessel functions, and then making use of well-known expressions giving the asymptotic behavior of these functions for large arguments as in Ref. 5. They can also be derived by the following argument, which applies in general to a boundary of arbitrary shape (Fig. 1c). Consider the field in the vicinity of a boundary point P in Fig. 1c. Since ka is large, the curved boundary can be approximated locally by the tangent plane. Furthermore, since $\beta \simeq k$, the field is produced locally by a plane wave at *grazing incidence*. If one determines the law of reflection for such a plane wave, and takes into account that the plane of incidence is parallel to the z axis, one obtains[24,27] eqs. (16)–(18).

The above argument can be used to derive the asymptotic behavior of Z_z and Z_τ for a variety of other waveguides, illustrated in Fig. 2.* In case (d) the boundary is a smooth surface of lossy metal with surface

* Notice k_0 denotes k in free space.

resistance R_s. In (a) the metal surface is corrugated. In (b) the medium outside the boundary is dielectric with refractive index $n_2 < n_1$. This corresponds to the optical fiber of Fig. 1a. In (c), $n_2 > n_1$ as in Ref. 2, and, therefore, both Z_z and Z_τ are real, which implies power is lost through the boundary. Other boundaries of practical interest are obtained from Fig. 2 by placing a grid of thin wires tangent to τ on the boundary. This will cause $Z_\tau \simeq 0$ in all cases. For all these waveguides one finds, for large ka, that $[H]$ in eq. (5) is a diagonal matrix,

$$[H] = -\begin{vmatrix} Y & 0 \\ 0 & X \end{vmatrix}, \tag{19}$$

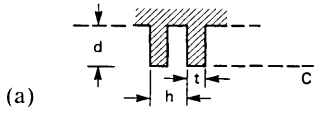
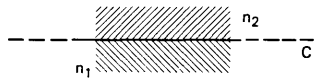
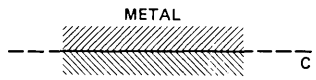
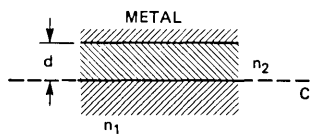
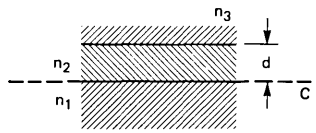
TYPE OF BOUNDARY	Z_z	Z_τ
(a) 	$jZ\left(1-\frac{t}{H}\right)\text{TANK}_0\, d$ $(h \ll \lambda)$	$Z_t = 0$
(b)  $(n_1 > n_2)$	$-j\frac{\sqrt{n_1^2-n_2^2}}{n_2}\frac{n_1}{n_2}Z$ $(Z = \sqrt{\mu/\epsilon})$	$j\frac{n_2}{\sqrt{n_1^2-n_2^2}}\frac{n_1}{n_2}Z$
(c)  $(n_1 < n_2)$	$r\frac{n_1}{n_2}Z$ $(r = \sqrt{n_2^2-n_1^2}/n_2)$	$\frac{1}{r}\frac{n_1}{n_2}Z$
(d) 	$(1+j)\,R_s$	$(1+j)\,R_s$
(e)  $(n_2 > n_1)$	$jr\frac{n_1}{n_2}ZT + R_s(1+T^2)$ $(T = \text{TANK}_0\sqrt{n_2^2-n_1^2}\, d,$ $\frac{R_s}{Z} \ll T, 1/T)$	$j\frac{1}{r}\frac{n_1}{n_2}ZT +$ $+ R_s(1+T^2)$
(f)  $n_3, n_2 > n_1$	$r\frac{n_1}{n_2}Z\frac{T'+jT}{1+jTT'}$ $\left(T' = \frac{\sqrt{n_3^2-n_1^2}}{\sqrt{n_2^2-n_1^2}}\frac{n_2^2}{n_3^2}\right)$	$\frac{1}{r}\frac{n_1}{n_2}Z\frac{T'+jT}{1+jTT'}$ $\left(T' = \frac{\sqrt{n_2^2-n_1^2}}{\sqrt{n_3^2-n_1^2}}\right)$

Fig. 2—Asymptotic values of the surface impedances Z_z and Z_τ for different boundary conditions.

and the coefficients X and Y are determined by the surface impedances given in Fig. 2. Notice there are cases where $[H]$ is not a diagonal matrix, as pointed out in Section VIII.

In the following sections we consider a waveguide with boundary conditions given by eq. (5) and derive simple expressions for the dependence of $\psi(x, y)$ upon the matrix $[H]$. These expressions are obtained neglecting terms of order higher than $1/ka$ and, therefore, they do not require an exact knowledge of $[H]$, but only of the asymptotic behavior of $[H]$, which can be determined as seen in this section. In Appendix A, a procedure for determining ψ to any desired accuracy is pointed out, but the procedure requires that $[H]$ be known accurately. Any desired accuracy for $[H]$ can be obtained by a procedure of successive approximations, but the resulting expressions are in general too complicated to be of practical interest.

Concerning the validity of the following derivation, it is important to realize that even though the expressions obtained for ψ will not satisfy the actual boundary conditions exactly, the errors will be small, of order two in $1/ka$. These errors imply the conditions satisfied at the boundary by the expressions in question can be obtained, from the actual conditions, by small perturbations, of order two in $1/ka$.

III. ASYMPTOTIC PROPERTIES OF ψ

We now determine the dependence of $\psi(x, y)$ upon the matrix $[H]$. To this purpose, it is convenient to assume that $[H]$ is independent of ka. The expressions obtained for $\psi(x, y)$ will depend on the coefficients of $[H]$. By substituting for these coefficients the expressions obtained in Section II, one obtains the dependence of ψ upon ka in general, for a waveguide with frequency dependent $[H]$.

Thus, consider a waveguide characterized by a given matrix $[H]$, and let

$$\sigma^2 = k^2 - \beta^2. \tag{20}$$

Let $\psi(x, y)$ and σ^2 be expanded in power series of l/ka,

$$\psi = \psi_\infty(x, y) + \sum_{i=1}^{\infty} \frac{1}{(ka)^i} \psi_i(x, y), \tag{21}$$

$$\sigma^2 = \sigma_\infty^2 \left(1 + \sum_{i=1}^{\infty} \frac{c_i}{(ka)^i}\right) \tag{22}$$

where the distributions ψ_∞, ψ_1, etc., are independent of ka; they are determined entirely by the shape of the boundary and the coefficients of [H].

Using eqs. (21) and (22) one can derive from eq. (5) for large ka (see Appendix A) a set of linear relations involving ψ and the normal

derivative $\partial\psi/\partial\nu$,

$$\begin{bmatrix} \psi_\nu \\ \psi_\tau \end{bmatrix} = \frac{1}{k}[H] \begin{bmatrix} \dfrac{\partial\psi_\nu}{\partial\nu} \\ \dfrac{\partial\psi_\tau}{\partial\nu} \end{bmatrix} + O[k^{-2}] \qquad \text{on } C. \tag{23}$$

Expressing ψ_ν and ψ_τ in terms of the x- y-components of ψ, we obtain

$$\begin{bmatrix} \psi_x \\ \psi_y \end{bmatrix} = \frac{1}{k}[A] \begin{bmatrix} \dfrac{\partial\psi_x}{\partial\nu} \\ \dfrac{\partial\psi_y}{\partial\nu} \end{bmatrix} + O[k^{-2}] \qquad \text{on } C, \tag{24}$$

where

$$[A] = \begin{bmatrix} \nu_x, & \nu_y \\ -\nu_y, & \nu_x \end{bmatrix} [H] \begin{bmatrix} \nu_x, & -\nu_y \\ \nu_y, & \nu_x \end{bmatrix}, \tag{25}$$

ν_x and ν_y being the direction cosines of $\boldsymbol{\nu}$.

3.1 Derivation of ψ_∞ and c_1

We now show that for each nondegenerate eigenvalue σ_∞ there are in general two modes, characterized by different values of c_1. For most waveguides, $[H]$ has the diagonal form (19), and in this case we shall see that ψ_∞ is linearly polarized. Furthermore, if the boundary has an axis of symmetry, then the polarization vector $\mathbf{i}$ of ψ_∞ is either parallel or orthogonal to the symmetry axis. Very simple expressions are obtained in this case for ψ_∞, c_1, ψ_1. More difficult is the treatment for degeneracy of order $N > 1$. Then, in order to determine ψ_∞, one must find the $2N$ latent roots of a certain characteristic equation.

The function ψ must satisfy the wave equation,

$$\nabla_t^2\psi + \sigma^2\psi = 0, \tag{26}$$

∇_t being the transverse part of ∇. Equation (24) implies that the boundary values of ψ vanish in the limit as $k \to \infty$.* It is, therefore, convenient to represent ψ in terms of the eigenfunctions f_1, f_2, etc., that satisfy the boundary condition

$$f_r = 0 \qquad \text{on } C. \tag{27}$$

Let u_r be the eigenvalue of f_r,

$$\nabla_t^2 f_r + u_r^2 f_r = 0. \tag{28}$$

From equations (21) and (22) for $k \to \infty$, one has

$$\psi \to \psi_\infty, \qquad \sigma \to \sigma_\infty, \tag{29}$$

* From now on the waveguide dimensions will be kept fixed, as k is increased.

and, therefore, ψ_∞ must satisfy the wave equation with σ replaced by σ_∞. Furthermore, from eq. (24),

$$\psi_\infty = 0 \qquad \text{on } C. \tag{30}$$

Therefore, σ_∞ must be one of the eigenvalues u_r. Let

$$\sigma_\infty = u_1 \tag{31}$$

and suppose there is degeneracy of order N, so that N distinct eigenfunctions $f_1, \cdots, f_N$ correspond to the same eigenvalue. Then

$$u_1 = \cdots = u_N \tag{32}$$

and ψ_∞ can be written in the form

$$\psi_\infty = \sum_{m=1}^{N} \alpha_{xm} f_m(x, y)\, \mathbf{i}_x + \sum_{m=1}^{N} \alpha_{ym} f_m(x, y)\, \mathbf{i}_y, \tag{33}$$

involving N eigenfunctions and $2N$ coefficients α_{xm}, α_{ym}. We now show that these coefficients cannot be chosen arbitrarily, but there are in general only $2N$ possible choices corresponding to $2N$ distinct modes.

3.2 Values of α_{xm}, α_{ym}

The values of ψ at any point (x', y') inside the boundary are related to the boundary values through the integral relation[24, 32]

$$\psi(x', y') = \int_C \psi(x, y) \frac{\partial G}{\partial \nu}\, dl, \tag{34}$$

where $G = G(x, y; x', y')$ is Green's function satisfying the equation

$$\nabla_t^2 G + \sigma^2 G = \delta(x - x')\delta(y - y') \tag{35}$$

and the boundary condition $G = 0$ when x, y is a point of C. Let G be represented in terms of the eigenfunctions f_r,

$$G = \sum_1^\infty \frac{1}{\sigma^2 - u_r^2} f_r(x, y) f_r(x', y'), \tag{36}$$

where it is assumed that f_r are properly normalized so that they are real functions satisfying

$$\iint_S f_r^2(x, y)\, dxdy = 1 \qquad (r = 1, 2, \cdots), \tag{37}$$

S being the region inside the boundary C. From eq. (36), G contains a component

$$G_\infty = \frac{1}{\sigma^2 - \sigma_\infty^2} \sum_1^N f_r(x, y) f_r(x', y'), \tag{38}$$

which diverges for $k \to \infty$, since $\sigma \to \sigma_\infty$. For large k,

$$G_\infty = ka \frac{1}{\sigma_\infty^2} \frac{1}{c_1} \sum_1^N f_r(x, y) f_r(x', y'). \tag{39}$$

The asymptotic behavior of eq. (34) for large k is now examined. Approximating G by G_∞, from eqs. (24), (25), and (34) one obtains the integral relation

$$\left.\psi_\infty\right] = \frac{1}{k} \int_C [A] \left.\frac{\partial \psi_\infty}{\partial \nu}\right] \frac{\partial G_\infty}{\partial \nu}\, dl, \tag{40}$$

where

$$\left.\psi_\infty\right] = \left.\begin{matrix} \psi_{\infty x} \\ \psi_{\infty y} \end{matrix}\right]. \tag{41}$$

Substituting eq. (39) in eq. (40), and taking into account eq. (33), one obtains for the coefficients α_{xm}, α_{ym} the characteristic equation

$$\left.\begin{matrix} \alpha_x] \\ \alpha_y] \end{matrix}\right] c_1 = \left[\begin{matrix} [I_{xx}] & [I_{xy}] \\ [I_{yx}] & [I_{yy}] \end{matrix}\right] \left.\begin{matrix} \alpha_x] \\ \alpha_y] \end{matrix}\right], \tag{42}$$

where

$$\alpha_x] = \left.\begin{matrix} \alpha_{x1} \\ \vdots \end{matrix}\right], \qquad \alpha_y] = \left.\begin{matrix} \alpha_{y1} \\ \vdots \end{matrix}\right], \tag{43}$$

and

$$(I_{xx})_{i,s} = \frac{a}{\sigma_\infty^2} \int_C A_{xx} \frac{\partial f_s}{\partial \nu} \frac{\partial f_i}{\partial \nu}\, dl, \tag{44}$$

and similarly for I_{xy}, I_{yx}, I_{yy} (replace A_{xx} with A_{xy}, A_{yx}, A_{yy}).

Equation (42) admits, in general , a total of $2N$ independent solutions $\alpha_1]$, $\alpha_2]$, $\cdots$, $\alpha_{2N}]$ for

$$\alpha] = \left.\begin{matrix} \alpha_x] \\ \alpha_y] \end{matrix}\right]. \tag{45}$$

Each solution is obtained by solving eq. (42) with c_1 set equal to one of the $2N$ latent roots $\lambda_1, \lambda_2, \cdots, \lambda_{2N}$ of the matrix

$$[I] = \left[\begin{matrix} [I_{xx}] & [I_{xy}] \\ [I_{yx}] & [I_{yy}] \end{matrix}\right]. \tag{46}$$

If the boundary is lossless, then one can verify that $[I]$ is Hermitian and its latent roots are *real.* In this case the $2N$ solutions are orthogonal,

$$\alpha_i]\alpha_s]_t^* = 0 \qquad \text{for} \quad i \neq s, \tag{47}$$

where $(\ )_t^*$ denotes the transpose conjugate.

Thus, to determine the coefficients α_{x1}, α_{y1}, etc., which specify ψ_∞, the $2N$ latent roots of the matrix $[I]$ must be determined. If the roots are all distinct, then they correspond to $2N$ distinct modes characterized by different c_1, i.e., by different propagation constant β. If there is degeneracy ($N > 1$), the expressions obtained from eq. (33) for ψ_∞ are quite complicated. Much simpler is the treatment for $N = 1$, since then only one eigenfunction $f_1(x, y)$ is involved. This is the most important case in practice, since it applies to the fundamental modes, which correspond to the lowest σ_∞ (see Appendix B).

3.3 Case N = 1

For $N = 1$, only one eigenfunction $f_1(x, y)$ corresponds to σ_∞ and eq. (33) reduces to

$$\psi_\infty = f_1(x, y)\mathbf{i}, \tag{48}$$

where $\mathbf{i}$ is a unit vector that determines the polarization of ψ_∞. If $\lambda_1 \neq \lambda_2$, then eq. (42) for $N = 1$ specifies two polarizations, corresponding to two modes with different propagation constants. To determine these two polarizations, let α_x and α_y be the direction cosines of $\mathbf{i}$. Then from eq. (42) with $N = 1$,

$$\begin{bmatrix} I_{xx} - c_1 & I_{xy} \\ I_{xy} & I_{yy} - c_1 \end{bmatrix} \begin{bmatrix} \alpha_x \\ \alpha_y \end{bmatrix} = 0, \tag{49}$$

where

$$I_{xx} = \frac{a}{\sigma_\infty^2} \int_C A_{xx} \left[\frac{\partial f_1}{\partial \nu}\right]^2 dl, \tag{50}$$

and similarly for I_{xy}, I_{yx}, I_{yy} (replace A_{xx} with A_{xy}, etc.). From eq. (49) one obtains for c_1 the characteristic equation

$$(I_{xx} - c_1)(I_{yy} - c_1) = I_{xy}I_{yx}, \tag{51}$$

whose solutions λ_1 and λ_2 determine for $\mathbf{i}$ two eigenvectors $\mathbf{i}_1$ and $\mathbf{i}_2$ with direction cosines specified by eq. (49). Notice $\mathbf{i}_1 = \mathbf{i}_2$ if $\lambda_1 \neq \lambda_2$. If the boundary is lossless and $\lambda_1 \neq \lambda_2$, then from eq. (47)

$$\mathbf{i}_1 \cdot \mathbf{i}_2^* = 0, \tag{52}$$

and, therefore, the two eigenvectors represent orthogonal polarizations.

For all the waveguides of Fig. 2, $[H]$ is given by eq. (19). Then

$$I_{xy} = I_{yxx} = -\frac{a}{\sigma_\infty^2} \int_C (Y - X)\nu_x\nu_y \left[\frac{\partial f_1}{\partial \nu}\right]^2 dl, \tag{53}$$

$$I_{xx} = -\frac{a}{\sigma_\infty^2} \int_C (Y\nu_x^2 + X\nu_y^2) \left[\frac{\partial f_1}{\partial \nu}\right]^2 dl, \tag{54}$$

and similarly for I_{yy} (interchange $x \rightleftarrows y$). Since in this case $[I]$ is a symmetric matrix,

$$\mathbf{i}_1 \cdot \mathbf{i}_2 = 0, \tag{55}$$

and if $\mathbf{i}_1$ is real (ψ_∞ is linearly polarized), then $\mathbf{i}_2$ is also real and orthogonal to $\mathbf{i}_1$.

We conclude by deriving a general condition that must be satisfied so that ψ_∞ is linearly polarized. Suppose $\mathbf{i}$ is real. Then the x axis can be oriented so that $\mathbf{i} = \mathbf{i}_x$. This implies $\alpha_y = 0$ and therefore $I_{yx} = 0$. We conclude that linear polarization is obtained only if it is possible to orient the x axis so that

$$\int_C (Y - X)\nu_x\nu_y \left[\frac{\partial f_1}{\partial \nu}\right]^2 dl = 0. \tag{56}$$

Notice then $I_{xy} = I_{yx} = 0$, and, therefore,

$$\mathbf{i}_1 = \mathbf{i}_x, \qquad \mathbf{i}_2 = \mathbf{i}_y. \tag{57}$$

The above requirement is always satisfied if the boundary is lossless (then X and Y are real) or if the values of X and Y are independent of position l on the boundary. It is also satisfied if the boundary has an axis of symmetry. In fact, let the symmetry axis be the x axis. Then X, Y and $\partial f_1/\partial \nu$ in eqs. (53) and (54) are even functions of y,

$$X(x, y) = X(x, -y), \qquad Y(x, y) = Y(x, -y), \tag{58}$$

whereas $\nu_x\nu_y$ is an odd function of y, which gives condition (56).

We thus conclude that in most cases of practical interest ψ_∞ is linearly polarized. This is of importance in the design of a feed for reflector antennas, to obtain good cross-polarization discrimination over a wide frequency range.

IV. PROPAGATION CONSTANT FOR $N = 1$

Assume the boundary has an axis of symmetry given by the x axis and let $N = 1$. Let $[H]$ be given by eq. (19), which applies to all waveguides of Fig. 2. Then, for the mode polarized in the x direction, the coefficient c_1 coincides with I_{xx} and it is given by eq. (54). Notice eq. (54) assumes that $f_1(x, y)$ is normalized as shown by eq. (37). If f_1 is not normalized, we must divide the right-hand side of eq. (54) by the left-hand side of eq. (37) with $r = 1$, then obtaining

$$c_1 = ja \frac{\displaystyle\int_C (Z_\tau \nu_y^2/Z + Z\nu_x^2/Z_z)\,(\partial f_1/\partial \nu)^2\, dl}{\displaystyle u_1^2 \iint f_1^2\, dxdy} \tag{59}$$

for $\mathbf{i} = \mathbf{i}_x$. For the other polarization $\mathbf{i} = \mathbf{i}_y$, interchange $\nu_y \rightleftarrows \nu_x$ in the above expression, which shows that the two polarizations are in general characterized by different propagation constants.

Once c_1 is known, the propagation constant β can be derived using eqs. (20) and (22) which for $\sigma_\infty = u_1$ give

$$\beta = k - \frac{1}{2}\frac{u_1^2}{k} - \frac{1}{2}\frac{u_1^2}{k^2 a} c_1 + \cdots, \tag{60}$$

where the dots indicate terms of order higher than two in $1/k$. If the medium inside the waveguide is lossless, k is real and the attenuation constant η is determined by the imaginary part of c_1. Then eqs. (59) and (60) give

$$\eta = -\mathrm{Im}(\beta) = \frac{\displaystyle\int_C (r\nu_y^2 + g\nu_x^2)(\partial f_1/\partial \nu)^2 \, dl}{2k^2 \displaystyle\iint f_1^2 \, dx\,dy}, \tag{61}$$

where r and g are the real parts of Z_τ/Z and Z/Z_z. This relation was used in Ref. 1 to determine the attenuation constant for a variety of waveguides of practical interest.

Using the above expressions one can straightforwardly calculate the dispersion and attenuation characteristics of any mode for large ka. In the special case of a hollow waveguide of dielectric with circular boundary, eqs. (59) and (60) give eq. (31) of Ref. 2.

Of greatest importance are the fundamental modes, which correspond to the lowest σ_∞. Then, for the circular boundary of Fig. 3,

$$f_1(x, y) = AJ_0(\sigma_\infty \rho), \tag{62}$$

where $\rho = \sqrt{x^2 + y^2}$, J_0 is the Bessel function of order zero, and $\sigma_\infty a$ is the first root of J_0,

$$\sigma_\infty a = 2.4048. \tag{63}$$

For the rectangular boundary of Fig. 3,

$$f_1(x, y) = A \cos\left(\frac{\pi}{2}\frac{x}{a}\right) \cos\left(\frac{\pi}{2}\frac{y}{b}\right). \tag{64}$$

If $f_1(x, y)$ is normalized [see eq. (37)], then

$$A = \begin{cases} \dfrac{1}{a\sqrt{\pi}} \dfrac{1}{J_1(\sigma_\infty a)} & \text{(circle)}, \\ \dfrac{1}{\sqrt{ab}} & \text{(rectangle)}. \end{cases} \tag{65}$$

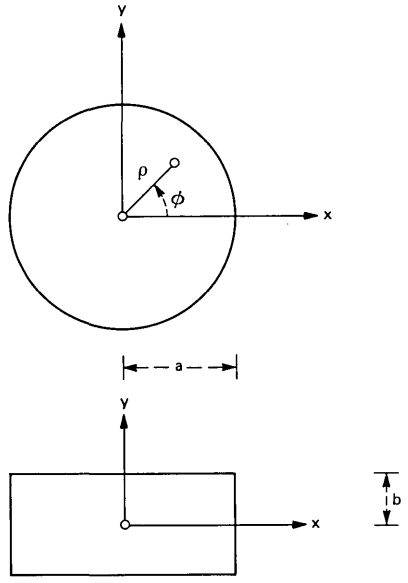


Fig. 3—Circular and rectangular boundaries.

V. THE DISTRIBUTION ψ_1

Once ψ_∞ is known, the boundary values of ψ can be calculated with error of order two in $1/k$ using eq. (23). If $N = 1$, $\mathbf{i} = \mathbf{i}_x$ and $[H]$ is given by eq. (19), then one obtains at the boundary

$$\psi \simeq -\frac{1}{k}\left[(X\nu_y^2 + Y\nu_x^2)\mathbf{i}_x - (X - Y)\nu_x\nu_y\mathbf{i}_y\right]\frac{\partial f_1}{\partial \nu} \qquad \text{on } C. \qquad (66)$$

To determine ψ inside the boundary we separate ψ into two parts, a component

$$f_1(x, y)\mathbf{i},$$

plus a component due to the other eigenfunctions f_2, f_3, etc. The latter component can be determined with error of order two in $1/k$ by substituting eq. (66) in the integral of eq. (34), with G replaced by $G - G_\infty$ for $\sigma = \sigma_\infty$, as shown in Appendix A. Concerning the former component, it is shown in Appendix A that if the boundary has an axis of symmetry, then $\mathbf{i}$ is independent of k, and, therefore, one can set $\mathbf{i} = \mathbf{i}_x$ for all values of k. If there is no symmetry, then in general $\mathbf{i}$ is a function of k and, to determine its dependence upon k, one must follow the same procedure used in this section to determine $\mathbf{i}$ for $k \to \infty$.

VI. APPLICATIONS

We now derive the fundamental modes of circular and rectangular boundaries with diagonal matrix $[H]$ given by eq. (19), which applies

to all the waveguides of Fig. 2. The surface parameters X and Y will be assumed to be independent of position on the boundary. Then for $\mathbf{i} = \mathbf{i}_x$,

$$\psi_\infty = f_1(x, y)\mathbf{i}_x, \tag{67}$$

giving ψ in the limit as $k \to \infty$. We now wish to obtain a better approximation for ψ.

For the rectangular boundary of Fig. 3b in eq. (66), one has $\nu_x \nu_y = 0$, and, therefore, ψ has the same polarization as ψ_∞, i.e.,

$$\psi \simeq \psi \mathbf{i}_x,$$

neglecting terms of order two in $1/k$. The boundary condition (23) can then be satisfied separating ψ into a product of two functions, $\psi_1(x)$ and $\psi_2(y)$, subject to the conditions

$$\psi_1 = -\frac{Y}{k}\frac{\partial \psi_1}{\partial \nu} \qquad \text{for} \quad x = \pm a, \tag{68}$$

$$\psi_2 = -\frac{X}{k}\frac{\partial \psi_2}{\partial \nu} \qquad \text{for} \quad y = \pm b, \tag{69}$$

whose solutions are well known.[26-31] For the fundamental mode one obtains

$$\psi = A \cos \alpha x \cos \gamma y, \tag{70}$$

where

$$\alpha a = \frac{\pi}{2}\left[1 - \frac{Y}{ka} + \cdots\right], \qquad \gamma b = \frac{\pi}{2}\left[1 - \frac{X}{kb} + \cdots\right]. \tag{71}$$

These results are closely related to expressions derived in Ref. 3 for a dielectric waveguide. Notice that for $b \to \infty$ the rectangular waveguide degenerates into two parallel plates placed at $x = \pm a$, in which case $\psi \to \psi_1(x)$ and the modes can be derived exactly as in Appendix C. Similarly, for $a \to \infty$ one obtains two plates at $y = \pm b$ and $\psi \to \psi_2(y)$. If both a and b are finite, then eq. (70) shows that ψ is simply the product of the two distributions $\psi_1(x)$ and $\psi_2(y)$, provided terms of order two in $1/k$ can be neglected.

Next consider the circular waveguide of Fig. 3. In this case it is convenient to introduce polar coordinates ρ, ϕ. Taking into account that X and Y are independent of ϕ, one obtains for the fundamental mode

$$\psi = A\left\{J_0(\sigma\rho)\mathbf{i}_x - \frac{X - Y}{4}\frac{(\sigma_\infty a)^2}{ka} J_2(\sigma\rho)(\cos 2\phi \mathbf{i}_x + \sin 2\phi \mathbf{i}_y) + \cdots\right\}, \tag{72}$$

where the dots indicate terms of order two in $1/k$, and from eqs. (22) and (59),

$$\sigma a = \sigma_\infty a\left[1 - \frac{X+Y}{2}\frac{1}{ka} + \cdots\right], \tag{73}$$

$\sigma_\infty a$ being given by eq. (63). These expressions were derived previously for $X = 0$ in Ref. 5 and therefore details of their derivation are not given here.

The above results for rectangular and circular waveguide are valid provided k is large enough so that the field at the boundary is small. From eq. (66) this requires

$$\frac{1}{ka} \ll 1, \qquad \frac{X}{ka} \ll 1, \qquad \frac{Y}{ka} \ll 1.$$

If the waveguide dimensions are large enough, the results apply even to an ordinary waveguide with metal walls of finite conductivity. To determine how large the dimensions must be, consider for instance copper at 100 GHz. Then $R_s/Z = 2.189 \times 10^{-4}$, and from Fig. 2d one obtains $|Y| = 514$ and $|X| = 5 \times 10^{-5}$. Therefore, the above inequalities require

$$2a \gg 1000\lambda,$$

which is too large a diameter for all practical purposes.

The above requirement is a consequence of the large value of Y for copper walls. By coating the walls with a thin dielectric layer, or by corrugating them, much lower values of Y can be obtained. Suppose for instance in Fig. 2e one chooses $T = 1$, $R_s \approx 0$, $n_1 = 1$, and $n_2^2 = 2$. Then, instead of the above requirement, one obtains $a \gg 0.318\lambda$, which is a much more realistic condition. The attenuation constant of such a waveguide, or of other waveguides realized using one of the structures of Fig. 2, can be determined straightforwardly using eq. (61).[1]

When a waveguide is used to illuminate a feed aperture, then at that aperture usually $ka \gg 1$. Then the aperture illumination is given accurately by eqs. (70) or (72) for rectangular and circular apertures. For other apertures, it can be determined as pointed out in Section V. By deriving the Fourier transform of eqs. (70) and (72) the far-field can be determined and thus its dependence on the aperture parameters X and Y can be obtained. These applications are discussed in Ref. 1.

Helical waveguide— A case where [H] is not a diagonal matrix

Consider one of the waveguides of Fig. 2, and let a helical wire be placed at the boundary, as in Fig. 4. Then one finds that $[H]$ is not a

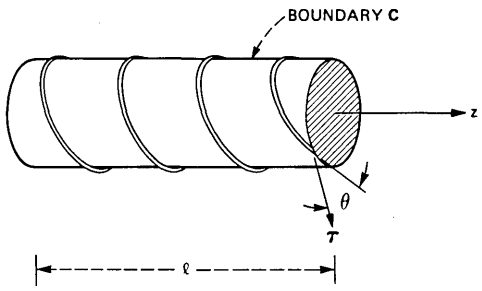


Fig. 4—Helix with pitch angle θ.

diagonal matrix,

$$[H] = \frac{-1}{\cos^2\theta - XY\sin^2\theta} \cdot \begin{bmatrix} Y, & j\sin\theta\cos\theta(1+XY) \\ -j\sin\theta\cos\theta(1+XY), & X \end{bmatrix}, \quad (74)$$

where X, Y denote the coefficients of $[H]$ for $\theta \to 0$, θ being the pitch angle between the wire and τ.

Consider a circular boundary. Then, we have seen that for $\theta = 0$ the fundamental modes have the same propagation constant, and are linearly polarized for $k \to \infty$. In this section we show that for $\theta \neq 0$ they become circularly polarized, and have different propagation constants β_1 and β_2. This implies the following. Suppose at the input of such a waveguide (Fig. 4) the fundamental modes are combined so as to obtain, to a good approximation, linearly polarized excitation. Then, one will not obtain, in general, linear polarization at the output, unless the difference between the two propagation constants is small enough so that

$$(\beta_2 - \beta_1)l \ll \frac{\pi}{8},$$

l being the waveguide length. In this section we derive β_2 and β_1. Helical waveguides are of importance for their simplicity of construction, as compared to corrugated waveguides, and for their excellent performance as hybrid-mode feeds, as shown recently by R. H. Turrin[33] whose work motivated the calculation of this section. The following results agree with Ref. 34.

Let consideration be restricted to the fundamental modes of a lossless boundary with $[H]$ independent of position on the boundary. Then, from eqs. (25) and (44)

$$I_{xx} = H_{11}v_{xx} - (H_{12} + H_{21})v_{xy} + H_{22}v_{yy}, \quad (75)$$

$$I_{xy} = (H_{11} - H_{22})v_{xy} + H_{12}v_{xx} - H_{21}v_{yy}, \quad (76)$$

and similarly for I_{yx} and I_{yy} (interchange $1 \rightleftarrows 2$, $x \rightleftarrows y$), where

$$\nu_{jk} = \frac{1}{\sigma_\infty^2} \int_C \nu_j \nu_k \left| \frac{\partial f_1}{\partial \nu} \right|^2 dl \qquad (j, k = x, y). \tag{77}$$

Since the boundary is assumed to be lossless,

$$(\mathrm{Im})(I_{xy}) = H_{12}(\nu_{xx} + \nu_{yy}), \tag{78}$$

and, therefore, $I_{xy} \neq 0$. We conclude that for $H_{12} \neq 0$ there is no degeneracy possible for the fundamental modes. For a circular boundary, using eqs. (74), (25), (50), (51), and (62), we obtain

$$\mathbf{i} = \frac{1}{\sqrt{2}} (\mathbf{i}_x \pm j i_y) \tag{79}$$

and

$$\sigma a = \sigma_\infty a \left[1 - \frac{1}{2(\cos^2\theta - XY \sin^2\theta)} \cdot \left(\frac{X + Y}{ka} \mp 2 \sin\theta \cos\theta \frac{1 + XY}{ka} \right) + \cdots \right], \tag{80}$$

with the plus sign of eq. (79) corresponding to the minus sign of eq. (80). The same expressions apply also to a square aperture with $a = b$. Thus, in both cases ψ_∞ is circularly polarized. If $XY + 1 > 0$, then the lower value of σ corresponds to a mode with polarization rotating in the sense of the helix in Fig. 4. The opposite is true for $XY + 1 < 0$.

VII. CONCLUSIONS

To summarize, we have shown for most of the modes inside a cylindrical waveguide of finite surface impedances that asymptotically, for large values of ka, the field ψ vanishes at the boundary. We have seen in Section III that for each eigenvalue u_r there are, in general, $2N$ modes, given for $k \to \infty$ by eq. (33). For the lowest eigenvalue one has $N = 1$, and for the corresponding two modes ψ_∞ is given by $f_1(x, y)\mathbf{i}$, where $\mathbf{i}$ is a unit vector. If the direction cosines of $\mathbf{i}$ are complex, then ψ_∞ is elliptically polarized. If $[H]$ is a diagonal matrix, as for the waveguides of Fig. 2, and the boundary has an axis of symmetry, then $\mathbf{i}$ is real, and one can always orient the x axis so that $\mathbf{i} = \mathbf{i}_x$. In this case the propagation constant β is given by eqs. (59) and (60), and using the procedure of Appendix A we can straightforwardly determine ψ, with error of order two in $1/k$, for any boundary shape. For rectangular and circular boundaries ψ is given by eqs. (70) and (72).

Of special importance are the fundamental modes, which correspond to the lowest eigenvalue σ_∞. These modes, treated in Section 3.2, are needed in reflector antennas to minimize cross-polarization and edge

illumination over the feed aperture. They are also needed for long distance waveguide or fiberguide communication. Our results, show that there is no need to corrugate the walls of a feed in order to obtain conditions (11) and (48) or to obtain the low attenuations calculated in Refs. 20 and 21. Furthermore, they imply that the low attenuations predicted in Ref. 2 for a hollow waveguide of dielectric, or for a waveguide with metal walls coated with dielectric,[4] can be achieved also using other waveguides. These applications are discussed in Ref. 1.

APPENDIX A

A.1 Derivation of eq. (23)

Taking into account that $\nabla \cdot \mathbf{E} = 0$, one can express E_z in terms on the transverse component given by eq. (2). One obtains

$$E_z = \frac{e^{-j\beta z}}{j\beta} \nabla_t \psi.$$

Using this relation and Maxwell's equation $-j\omega\mu\mathbf{H} = \nabla \times \mathbf{E}$ one can express H_z and $\mathbf{H}_t$ in terms of ψ. Substituting these expressions in eq. (5), we obtain the boundary condition

$$\begin{bmatrix} \psi_\nu \\ \psi_\tau \end{bmatrix} = \frac{1}{k} [H] \begin{bmatrix} \frac{k}{\beta} (\nabla_t \psi) \\ -\nabla_t \cdot (\mathbf{i}_z \times \psi) \end{bmatrix} + \begin{bmatrix} \frac{1}{\beta^2} \frac{\partial}{\partial \nu} (\nabla_t \cdot \psi) \\ 0 \end{bmatrix}. \tag{81}$$

Taking into account eqs. (21) and (22) for $k \to \infty$, we have

$$\psi \to 0, \qquad \text{on} \quad C.$$

This implies

$$\nabla_t \cdot \psi \to \frac{\partial \psi_\nu}{\partial \nu}, \qquad \nabla_t \times \psi \to \mathbf{i}_z \frac{\partial \psi_\tau}{\partial \nu}.$$

Taking into account these relations, from eq. (81) we obtain eq. (23) with error of order two in $1/k$.

A.2 Development of σ^2 and ψ in Asymptotic Series of $1/ka$

A general procedure for deriving the various terms c_r, ψ_r in eqs. (21) and (22) is now described, thus justifying these equations. Assume that the boundary has at least one line of symmetry, since this simplifies considerably the derivation, and it applies to most cases of practical interest. Also assume that $[H]$ has the diagonal form of eq. (19), and that a single eigenfunction $f_1(x, y)$ corresponds to σ_∞.

Since there is no degeneracy, ψ_∞ is given by eq. (48). Let the x axis coincide with the symmetry line. Then, the surface parameters X and

Y are even functions of y,

$$X(x, -y) = X(x, y) \qquad \text{and} \qquad Y(x, -y) = Y(x, y), \tag{82}$$

and we now show the modes can be divided in two groups, namely *even modes* satisfying

$$\psi_x(x, -y) = \psi_x(x, y), \qquad \psi_y(x, -y) = -\psi_y(x, y), \tag{83}$$

and *odd modes* satisfying

$$\psi_x(x, -y) = -\psi_x(x, y), \qquad \psi_y(x, -y) = \psi_y(x, y). \tag{84}$$

In fact, let

$$\psi = \psi_x(x, y)\mathbf{i}_x + \psi_y(x, y)\mathbf{i}_y$$

be a solution of the wave equation and of the boundary condition (81) with $[H]$ given by eq. (19). Then we wish to show that

$$\psi' = \psi_x(x, -y)\mathbf{i}_x - \psi_y(x, -y)\mathbf{i}_y$$

is also a solution. Notice that ψ' is the *image* of ψ with respect to the x axis, as shown in Fig. 5, where P and P' denote two corresponding points (x, y) and $(x, -y)$. One can verify that

$$(\nabla_t \times \psi')_{P'} = -(\nabla_t \times \psi)_P, \tag{85}$$

$$(\nabla_t \psi')_{P'} = (\nabla_t \psi)_P. \tag{86}$$

If P is a boundary point, and ν' and ν denote the normals to the boundary at P' and P, respectively, one has that ν' is the *image* of ν because the boundary is symmetrical. Taking all this into account one can verify that ψ' satisfies the boundary condition (81) with $[H]$ given by eq. (19) at P'. We conclude that if an arbitrary solution ψ is known, two independent solutions ψ_e and ψ_o can be obtained by the relations

$$\psi_e = \tfrac{1}{2}[\psi + \psi'], \qquad \psi_o = \tfrac{1}{2}[\psi - \psi']. \tag{87}$$

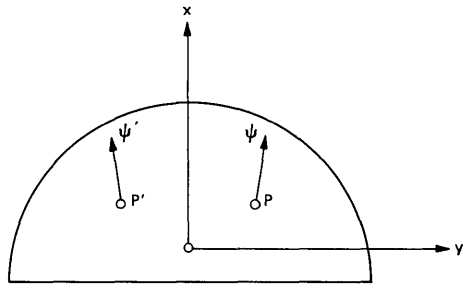


Fig. 5—For a symmetrical boundary, the boundary conditions are satisfied by both ψ and its mirror image ψ'.

One solution, ψ_e, is *even* and the other, ψ_o, is *odd.*

We now proceed to derive σ^2 and ψ. Let $\mathbf{i} = \mathbf{i}_x$ and assume $f_1(x, y)$ is even, i.e.,

$$f_1(x, y) = f_1(x, -y), \tag{88}$$

all other cases ($\mathbf{i} = \mathbf{i}_y$, or f_1 odd) being entirely analogous. Then ψ_x and ψ_y are, respectively, even and odd and it is convenient to separate G into three parts,

$$G = G_\infty + G_x + G_y, \tag{89}$$

where G_x, G_y denote, respectively, the even and odd parts of $G - G_\infty$.* Taking into account that ψ_x is even, from eqs. (34), (36), and (89) one obtains

$$\psi_x = A_1 f_1(x, y) + \int_C \psi_x \frac{\partial G_x}{\partial \nu}\, dl, \tag{90}$$

where

$$A_1 = \frac{1}{\sigma^2 - u_1^2} \int_C \psi_x \frac{\partial f_1}{\partial \nu}\, dl. \tag{91}$$

Similarly, for ψ_y,

$$\psi_y = \int_C \psi_y \frac{\partial G_y}{\partial \nu}\, dl. \tag{92}$$

In eq. (90), since ψ can be multiplied by an arbitrary constant, the coefficient A_1 can be chosen arbitrarily. We choose $A_1 = 1$, to be consistent with $\psi_\infty = f_1(x, y)\mathbf{i}_x$. Then eq. (91) gives

$$\sigma^2 - \sigma_\infty^2 = \int_C \psi_x \frac{\partial f_1}{\partial \nu}\, dl, \tag{93}$$

a basic relation that expresses σ^2 in terms of the boundary values of ψ_x. Expanding ψ_x in a power series of $1/ka$, from eq. (91) one obtains for the ith coefficient of σ^2,

$$c_i = \frac{1}{\sigma_\infty^2} \int_C \psi_{xi} \frac{\partial f_1}{\partial \nu}\, dl, \tag{94}$$

where ψ_{xi} is the x component of ψ_i in eq. (21). From eq. (66),

$$\psi_{x1} = -(X\nu_y^2 - Y\nu_x^2)a \frac{\partial f_1}{\partial \nu}, \tag{95}$$

---

* Thus G_y is obtained from eq. (36) considering only those terms for which f_r are odd, whereas for G_x, consideration is restricted to f_r even with $r > 1$.

and, therefore, for $i = 1$, eq. (94) gives eq. (59).

From eqs. (90) and (92), taking into account that at the boundary $\delta\psi = \psi - \psi_\infty = \psi$,

$$\delta\psi_x = \int_C \delta\psi_x \frac{\partial G_x}{\partial \nu}\, dl, \qquad \delta\psi_y = \int_C \delta\psi_y \frac{\partial G_y}{\partial \nu}\, dl. \tag{96}$$

Thus, using eq. (95), we obtain

$$\psi_{x1} = -\int_C (X\nu_y^2 + Y\nu_x^2) \frac{\partial f_1}{\partial \nu} \left(\frac{\partial G_x}{\partial \nu}\right)_\infty dl, \tag{97}$$

for the values of ψ_{x1} inside the boundary whereas for ψ_{y1}, using eq. (66), we obtain

$$\psi_{y1} = \int_C \nu_x \nu_y (x - Y) \frac{\partial f_1}{\partial \nu} \left(\frac{\partial G_y}{\partial \nu}\right)_\infty dl, \tag{98}$$

where $(\ )_\infty$ denotes the value for $\sigma = \sigma_\infty$. Equations (97) and (98) allow ψ and its derivatives to be determined with an error $O(1/k)$. Therefore, the right-hand side of eq. (81), which contains the factor $1/k$, can now be calculated with an error $O(1/k^2)$. The boundary values of ψ_{x2}, ψ_{y2} are then obtained from eq. (81).

Once these values are known, eq. (96) can be used to determine ψ_{x2} and ψ_{y2} inside the boundary. Equation (81) can then be used again to determine the boundary values of ψ_{x3} ψ_{y3}, whose values inside the boundary can then be calculated using eq. (96). By proceeding this way, we can successively calculate all ψ_{xi}, ψ_{yi}. Notice in eq. (96) that the kernels depend on σ^2. Therefore, to determine ψ_{xi}, ψ_{yi} for $i > 1$, we must first calculate the coefficients $c_1, \cdots, c_{i-1}$ using eq. (94). Once these coefficients are known, the kernels must be developed in power series of $1/k$, and then the first $i - 1$ terms in these series must be determined. These terms then allow eq. (96) to be used to determine ψ_{xi}, ψ_{yi}.

APPENDIX B

Let f_1 be one of the eigenfunctions satisfying the boundary condition $f_1 = 0$ and the wave equation

$$\nabla_i^2 f_1 + u_1^2 f_1 = 0,$$

and let u_1 be the lowest eigenvalue. This implies that if $g(x, y)$ is an arbitrary function with continuous derivatives, then

$$\iint_S g \nabla_i^2 g \, dxdy \leq u_1^2 \iint_S g^2 \, dxdy, \tag{99}$$

where the inequality sign applies to any $g(x, y)$ that is not an eigenfunction corresponding to the eigenvalue u_1.

Condition (99) is now used to show that f_1 cannot have nodal lines inside S. In fact, suppose f_1 has a nodal line inside S, and let

$$g = |f_1|.$$

Then, since $\nabla_t g$ is discontinuous across the nodal line, g cannot be an eigenfunction, and, therefore, condition (99) should give an inequality. To evaluate the integral

$$I = \iint g \nabla_t^2 g \, dxdy \tag{100}$$

in the immediate vicinity of the nodal line, where $\nabla_t^2 g$ diverges because of the discontinuity of $\nabla_t g$, write

$$g\nabla_t^2 g = \nabla_t[g\nabla_t g] - (\nabla_t g)^2$$

and notice that

$$g\nabla_t g$$

is continuous because $g = 0$ on the nodal line. It follows that $g\nabla_t^2 g$ does not diverge on the nodal line, and, therefore, its integral over a narrow strip containing the nodal line vanishes as the width of the strip goes to zero. Thus,

$$I = \iint_S f_1 \nabla_t^2 f_1 \, dxdy = u_1^2 \iint_S f_1^2 \, dxdy \tag{101}$$

and, therefore, condition (99) gives an equality, which implies f_1 cannot have nodal lines inside S.

It is now shown that f_1 is the only eigenfunction corresponding to u_1. In fact, if f_2 is another eigenfunction corresponding to u_1, then this must be true also for

$$f = f_1 + \alpha f_2,$$

where α is an arbitrary constant. But this is not possible, since one can always choose α causing f to have a nodal line inside S, and we have already seen that this violates condition (99).

APPENDIX C

Consider the modes propagating between two parallel planes orthogonal to the x axis and let $2a$ be the spacing of the two planes. Assume at the boundary one of the conditions of Fig. 2, and let the x axis be oriented in the direction of propagation, so that there is no

variation in the y direction. Then

$$\psi = \psi(x)\mathbf{i}, \tag{102}$$

where $\mathbf{i}$ is a unit vector. For the TM-modes, $\mathbf{i} = \mathbf{i}_x$, whereas for the TE-modes $\mathbf{i} = \mathbf{i}_y$. For the former case $E_\tau = H_z = 0$ and, therefore, ψ is independent of the transverse impedance Z_τ. In the latter case $H_\tau = E_z = 0$ and ψ is independent of Z_z. In either case one finds[24,27] that the surface impedances Z_τ and Z_z in eq. (9) can be determined straightforwardly. For two metal plates with dielectric coating of thickness d, for instance, Z_τ and Z_z are determined entirely by β, d, and the refractive index n of the dielectric,

$$X = \frac{k}{\sqrt{k^2n^2 - \beta^2}} \tan\left(d\sqrt{k^2n^2 - \beta^2}\right), \tag{103}$$

and

$$Y = -\frac{n^2k}{\sqrt{k^2n^2 - \beta^2}} \frac{1}{\tan\left(d\sqrt{k^2n^2 - \beta^2}\right)}, \tag{104}$$

which for $\beta \to k$ give eqs. (17) and (18). Analogous expressions are obtained for the other boundaries in Fig. 2.

Consider the even modes with $\mathbf{i} = \mathbf{i}_x$,

$$\psi = \psi\mathbf{i}_x = \cos\sigma x\mathbf{i}_x, \tag{105}$$

where from eq. (81) the wavenumber σ must satisfy[31]

$$\frac{Y}{ka} = \frac{1}{\sigma a \tan\sigma a}, \tag{106}$$

whose behavior for real values of σa is illustrated in Fig. 6a. If $Y \neq \infty$, then Y/ka vanishes for $ka \to \infty$ and, therefore, for most of the solutions of eq. (106),

$$\sigma a \to m\pi - \frac{\pi}{2}, \qquad m = 1, 2, \cdots, \tag{107}$$

as $ka \to \infty$. In addition to these solutions, for $Y < 0$ one mode exists for which σa is imaginary and[31]

$$\sigma a \to -j\frac{ka}{Y}, \qquad \text{as} \quad a \to \infty. \tag{108}$$

Equation (107) implies that the boundary values of ψ vanish for $ka \to \infty$, whereas eq. (108) implies that the mode is a surface wave whose

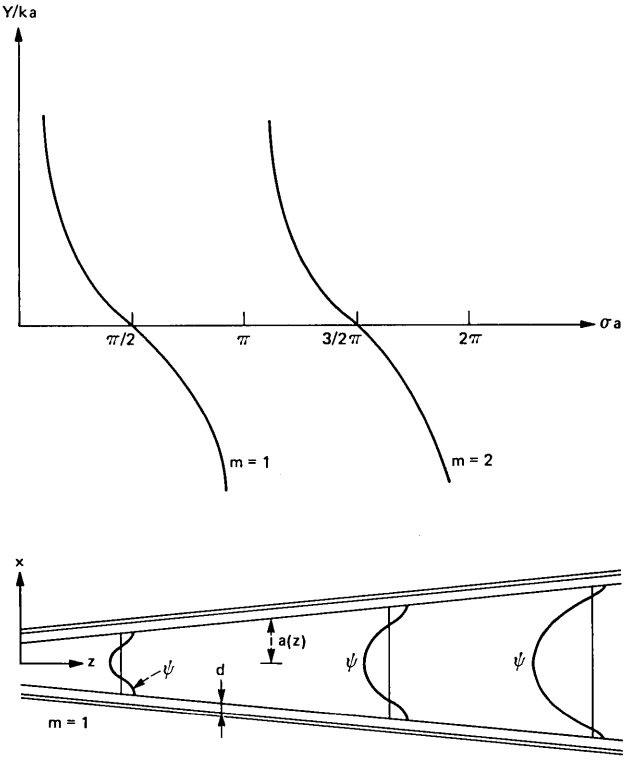


Fig. 6—Relationship between σa and Y/ka for the TM-modes of a parallel plate waveguide. Notice in the dielectric-lined waveguide the spacing $2a$ of the two plates increases with z. As a consequence, the field amplitude ψ at the boundary decreases with a, for the mode with $m = 1$.

amplitude in the vicinity of the walls is given by

$$\psi \simeq \frac{1}{2} \exp\left(+\frac{k}{|Y|} a\right) \exp\left(-\frac{k}{|Y|} |x - a|\right). \tag{109}$$

Notice eq. (107) implies

$$\beta \to k, \tag{110}$$

as $ka \to \infty$, whereas for the surface wave

$$\beta \to k \sqrt{1 + \frac{1}{Y^2}} \neq k. \tag{111}$$

To understand the significance of these results, consider two metal plates with dielectric coating as in Fig. 1a. Let the separation a of the two plates be gradually increased in the direction of propagation, as shown in Fig. 6, and let $a \to \infty$ as $z \to \infty$. Let the dielectric thickness

d be so small that initially, for $z = 0$,

$$\frac{Y}{ka} \simeq -\infty. \tag{112}$$

Then the modes for $z = 0$ are essentially those of an ordinary waveguide without dielectric coating,

$$\sigma a \simeq m\pi \qquad (m = 0, 1, 2, \cdots). \tag{113}$$

For $z > 0$, however, the magnitude of Y/ka decreases with z, and $Y/ka \to 0$ for $z \to \infty$. This implies, for all the modes with $m \neq 0$, conditions (107) and (110) and, therefore, the boundary values of ψ vanish for $z \to \infty$. For $m = 0$, on the other hand, σa is imaginary and initially $\sigma a \simeq 0$. This mode will degenerate for $z \to \infty$ into a surface wave with propagation constant determined by Y as shown by eq. (111). This is the only mode for which the field does not become infinitesimal at the walls for $z \to \infty$. For all the other modes the boundary values of ψ for large ka are given by

$$(-1)^{m+1} Y \frac{\sigma a}{ka}, \tag{114}$$

which vanishes for $ka \to \infty$.

The above considerations apply also to the TE-modes. In fact, it can be verified that if in eq. (2) E_t is replaced with ZH_t so that ψ represents the transverse component of ZH_t, then, for the even modes, ψ is still given by eqs. (105) and (106), provided Y is replaced with X. The behavior of the odd modes is entirely analogous; simply replace $\cos \sigma x$ with $\sin \sigma x$ in eq. (105), and $\tan$ with $-\operatorname{cotan}$ in eq. (106).

REFERENCES

1. C. Dragone, "Attenuation and Radiation Characteristics of the HE_{11}-Mode," IEEE Trans. MTT, 7 (July 1980), pp. 704–10.
2. E. A. J. Marcatili and R. E. Schmeltzer, "Hollow Metallic and Dielectric Waveguides for Long Distance Optical Transmission and Lasers," B.S.T.J., *43*, No. 4 (July 1964), pp. 1783–1809.
3. E. A. J. Marcatili, "Dielectric Rectangular Waveguide and Directional Coupler for Integrated Optics," B.S.T.J., *48*, No. 7 (September 1969), pp. 2071–2102.
4. J. W. Carlin and P. D'Apostino, "Normal Modes in Over-Moded Dielectric-Lined Circular Waveguide," B.S.T.J., Vol. *52*, No. 4 (April 1973), pp. 453–86.
5. C. Dragone, "Reflection, Transmission, and Mode Conversion in a Corrugated Feed," B.S.T.J., *56*, No. 6 (July–August 1977), pp. 835–67.
6. C. Dragone, "Characteristics of a Broadband Corrugated Feed: A Comparison Between Theory and Experiment," B.S.T.J., *56*, No. 6 (July–August 1977), pp. 869–88.
7. A. W. Snyder, "Asymptotic Expressions for Eigenfunctions and Eigenvalues of a Dielectric or Optical Waveguide," IEEE Trans. MTT, *17* (December 1969), pp. 1130–8.
8. D. Marcuse, *Theory of Dielectric Optical Waveguides*, New York: Academic, 1974.
9. P. J. B. Clarricoats, "Similarities in the Electromagnetic Behavior of Optical Waveguides and Corrugated Feeds," Electron. Lett., *6*, No. 6 (March 1970), pp. 178–80.
10. V. H. Rumsey, "Horn Antennas with Uniform Power Patterns Around their Axes,"

IEEE Trans. Antenna Propag., *AP-14*, No. 5 (September 1966), pp. 656–8.
11. H. C. Minnett and B. MacA. Thomas, "A Method of Synthesizing Radiation Patterns with Axial Symmetry," IEEE Trans. Antenna Propag., *AP-14*, No. 5 (September 1966), pp. 654–6.
12. P. J. B. Clarricoats and P. K. Saha, "Propagation and Radiation Behavior of Corrugated Feeds; Part 1—Corrugated Waveguide Feed," Proc. IEEE, *118*, No. 9 (September 1971), pp. 1167–76.
13. S. E. Miller, E. A. J. Marcatili, and Tingye Li, "Research Toward Optical-Fiber Transmission Systems, Part I: The Transmission Medium," Proc. IEEE, *61* (December 1973), pp. 1703–51.
14. D. Gloge, "Propagation Effects in Optical Fibers," IEEE Trans. MTT, *23* (January 1975), pp. 106–20.
15. D. Marcuse, *Light Transmission Optics*, Princeton, New Jersey: Van Nostrand, 1972.
16. N. S. Kapany and J. J. Burke, *Optical Waveguides*, New York: Academic, 1972.
17. H. G. Unger, "Lined Waveguide," B.S.T.J., *41*, No. 2 (March 1962), pp. 745–68.
18. J. W. Carlin and P. D'Apostino, "Low-Loss Modes in Dielectric Lined Waveguide," B.S.T.J., *50*, No. 5 (May–June 1971), pp. 1631–9.
19. J. W. Carlin, "A Relation for the Loss Characteristics of Circular Electric and Magnetic Modes in Dielectric Lined Waveguide," B.S.T.J., *50*, No. 5 (May–June 1971), pp. 1639–44.
20. P. J. B. Clarricoats and P. K. Saha, "Attenuation in Corrugated Circular Waveguide," Electron. Lett., *6* (1970), pp. 370–2.
21. P. J. Clarricoats, A. D. Oliver, and S. L. Chang, "Attenuation in Corrugated Circular Waveguide," Parts 1 and 2: Theory and Experiment, Proc. IEEE, *122*, No. 11 (November 1975), pp. 1173–83.
22. J. E. Goell, "A Circular-Harmonic Computer Analysis of Rectangular Dielectric Waveguides," B.S.T.J., *48* (1969) pp. 2133–60.
23. R. B. Dydbal, L. Peters, and W. H. Peake, "Rectangular Waveguide with Impedance Walls," IEEE Trans. MTT, *19*, No. 1 (January 1971), pp. 2–9.
24. R. E. Collin, *Field Theory of Guided Waves*, New York: McGraw-Hill, 1960.
25. N. Brooking, P. J. B. Clarricoats, and A. D. Oliver, "Radiation Patterns of Pyramidal Dielectric Waveguides," Electron. Lett., *10* (February 1974), pp. 33–4.
26. P. J. B. Clarricoats and C. E. R. C. Salemo, "Antennas Employing Conical Dielectric Horns," Proc. Inst. Elect. Eng., *120* (July 1973), pp. 741–9.
27. J. R. Wait, *Electromagnetic Waves in Stratified Media*, New York: Pergamon, 1970.
28. E. Bahar, "Propagation of VLF Radio Waves in a Model Earth-ionosphere Waveguide of Arbitrary Height and Finite Surface Impedance Boundaries: Theory and Experiment," Radio Sci. (New Series), *Vol. 1*, No. 8 (1966), pp. 925–38.
29. E. Bahar, "Generalized Scattering Matrix Equations for Waveguide Structures of Varying Surface Impedance Boundaries," Radio Sci., *2* (New Series), No. 3 (March 1967), pp. 287–97.
30. R. A. Waldron, "Theory of Guided Electromagnetic Waves," London: Van Nostrand, 1969.
31. V. V. Scherchenko, "Continuous Transitions in Open Waveguides," translated from Russian by P. Beckman, Boulder, Colorado: GOLEM Press, 1971.
32. A. Sommerfeld, *Partial Differential Equations in Physics*, New York: Academic, 1967.
33. R. H. Turrin, "A Helical-Wire Hybrid-Mode Conical-Horn Antenna," IREECON, Sydney, Australia, August 30, 1979.
34. S. Ghosh and G. P. Srivastava, "Corrugated Waveguide with Helically Continuous Corrugations," IEEE Trans. Antenna Propag. *AP-27*, No. 4 (July 1979), pp. 564–7.

Contributors to This Issue

Corrado Dragone, Laurea in E.E., 1961, Padua University (Italy); Libera Docenza, 1968, Ministero della Pubblica Istruzione (Italy); Bell Laboratories, 1961—. Mr. Dragone has been engaged in experimental and theoretical work on microwave antennas and solid-state power sources. He is currently concerned with problems involving electromagnetic wave propagation and microwave antennas.

John A. Morrison, B.Sc., 1952, King's College, University of London; Sc.M., 1954 and Ph.D., 1956, Brown University; Bell Laboratories, 1956—. Mr. Morrison has done research in various areas of applied mathematics and mathematical physics. He has recently been interested in queuing problems associated with data communications networks. He was a Visiting Professor of Mechanics at Lehigh University during the fall semester 1968. Member, American Mathematical Society, SIAM, IEEE, Sigma Xi.

Roger N. Nucho, B.S. (Physics), 1971, B.A. (Fine Arts), 1972, American University of Beirut; Ph.D. (Physics), 1977, Massachusetts Institute of Technology; Bell Laboratories, 1978—. Mr. Nucho was Research Associate at the University of Southern California in 1977–1978, during which time his main interest was the electronic structure of semiconductors. Since joining Bell Laboratories, he has been involved in designing special services facility networks which are insensitive with respect to forecast uncertainty. Member, Sigma Xi.

Donald R. Smith, A.B. (Physics), 1969, Cornell University; M.S. (Operations Research), 1974, Columbia University; Ph.D. (Operations Research), 1975, University of California, Berkeley; Bell Laboratories, 1980— . Before joining Bell Laboratories, Mr. Smith was employed at Adaptive Technology, Inc., 1970–1974, and was Assistant Professor in the Department of Industrial Engineering and Operations Research, Columbia University, 1975–1979. At Adaptive Technology, Mr. Smith

developed mathematical models for new techniques in statistical multiplexing. At Bell Laboratories he is in the Operations Research Center pursuing interests in applied stochastic processes, including queuing theory and reliability theory.

Ward Whitt, A.B. (Mathematics), 1964, Dartmouth College; Ph.D. (Operations Research), 1968, Cornell University; Stanford University, 1968–1969; Yale University, 1969–1977; Bell Laboratories, 1977—. At Yale University, from 1973–1977, Mr. Whitt was Associate Professor in the departments of Administrative Sciences and Statistics. At Bell Laboratories he supervises the Operations Research Analysis Group in the Operations Research Center. His work focuses on stochastic processes and stochastic models in operations research.

Yehuda Vardi, B.S. (Mathematics and Statistics), 1970, Hebrew University; M.S. (Operations Research), 1973, Technion; Ph.D. (Operations Research), 1977, Cornell University; Bell Laboratories, 1977—1980. At Bell Laboratories, Mr. Vardi worked on problems in operations research, financial management, statistical sequential analysis, and distribution theory and on estimation procedures in renewal processes.

Papers by Bell Laboratories Authors

PHYSICAL SCIENCES

Bounds on Spread-Spectrum Systems. H. E. Rowe, Electronics Conv Record (Nov 1979).

Carbon Films with Relatively High Conductivity. M. L. Kaplan, P. H. Schmidt, A. H. Chen, and W. M. Walsh, Jr., Appl Phys Lett, *36* (May 1980), pp 867–9.

Convection in a Porous Layer. P. G. Simpkins and P. A. Blythe, Int J Heat Mass Trans, *23* (June 1980), pp 881–7.

CVSD to LPC Conversion Using Noise Tolerant Analysis. J. D. Tomcik and J. L. Melsa, Proc IEEE Int'l Conf on Acoust, Speech & Sig Proc, *3* (Apr 1980), pp 719–24.

Coupled Modes with Random Propagation Constants. USIR North Amer Radio Sci Conf (June 1980), p 80.

Electron Acceptor Surface States Due to Oxygen Adsorption on Metal Phthalocyanine Films. S. C. Dahlberg, J Chem Phys, *72* (1980), pp 6706–11.

Magnetoelastic Interactions in Ionic π-Electron Systems: Magnetogyration. M. A. Bosch, M. E. Lines, and M. Labhart, Phys Rev Lett, *45* (July 1980), pp 140–3.

Optical Waveguides in $LiTaO_3$: Silver Lithium Ion Exchange. J. L. Jackel, Appl Optics, *19* (June 1980), p 1996.

Pyroelectric $Ba(NO_2)_2H_2O$: Room Temperature Crystal Structure. S. C. Abrahams, J. L. Bernstein, and R. Liminga, J Chem Phys, *72* (June 1980), pp 5857–62.

The Raman Excitation Spectra and Absorption Spectrum of a Metalloporphyrin in an Environment of Low Symmetry. J. A. Shelnutt, J Chem Phys, *72* (Apr. 1980), pp 3948–58.

Raman Spectra of Light-Coupling Prism and Gemstone Materials. J. E. Griffiths and K. Nassau, Appl Spectroscopy, *34* (1980), pp 395–9.

Rayleigh-Brillouin Scattering in Polymers. G. D. Patterson, Method of Exper Phys, *16* (1980), pp 170–204.

The Relation of Elastooptic and Electrostrictive Tensors. D. F. Nelson, Basic Optical Properties of Materials (May 1980), pp 209–12.

Surface Phenomena in Noble and Rare Earth Metals. G. K. Wertheim, Mat Sci Eng, *42* (1980), pp 85–90.

MATHEMATICS

Evolution of a Stable Profile for a Class of Nonlinear Diffusion Equations. III Slow Diffusion on the Line. J. G. Berryman, J Math Phys, *21* (June 1980), pp 1326–31.

Hamming Association Schemes and Codes on Spheres. S. P. Lloyd, SIAM J Math Anal, *11* (May 1980), pp 488–505.

COMPUTING

Computational Procedures for Markov Decision Processes. R. W. Henry, *19,* Annual Tech Sym Pathways to Sys Integrity (June 1980), pp 155–9.

A Mixed-Mode Simulator. V. D. Agrawal, A. K. Bose, P. Kozak, H. N. Nham, and E. Pacas-Skewes, 17th Design Auto Conf Proc (June 1980), pp 618–25.

A Standard Operating System Commanded Response Language. The ANSI X3H1 Effort. C. T. Schlegel, L. L. Frampton, and S. Mellor, Proc of IFIP Conf (1980), pp 83–99.

ENGINEERING

Acceleration Factors for IC Leakage Current in a Steam Environment. W. Weick, IEEE Trans on Reliability, *29* (June 1980), pp 109–14.

Digital Test Generation and Design for Testability. J. Grason and A. W. Nagle, 17th Design Auto Conf Proc (June 1980), pp 175–89.

The Measurement of the Effect of Photon Noise on Detection. J. Krauskopf and A. Reeves, Vision Res, *20* (1980), pp 193–6.

Multilocation Audiographic Conferencing. C. Stockbridge, Telecommun Policy, *4* (June 1980), pp 96–107.

Rerouting Stability in Virtual Circuit Data Networks. E. F. Wunderlich and T. S. Printis, Int'l Conf on Commun (June 1980), pp 13.5.1–.5.

Solar Cells. W. D. Johnston, Jr., New York: Marcel Dekker (1980).

Strengths and Diameter Variations of Fused Silica Fibers Prepared in Oxy-Hydrogen Flames. T. T. Wang and H. M. Zupko, *3* (1980), pp 73–87.

Triggerable Semiconductor Lasers and Light-Coupled Logic. J. A. Copeland, J Appl Phys, *51* (Apr 1980), pp 1919–21.

Triggerable Semiconductor Lasers. J. A. Copeland, S. M. Abbott, and W. S. Holden, J Quant Electronics, *16* (Apr 1980), pp 388–90

A Two-Chip CMOS-μ-Law Encoder/Decoder Set. M. R. Dwarakanath and D. G. Marsh, Conf Digest, *1* (June 1980), pp 11-3.1-.4.

A Welded Optical Fiber Signal Splitter. A. F. Judy and T. D. Mathis, Fiber & Integated Optics, *3* (1980), pp 63–71.

Contents, February 1981

THE BELL SYSTEM TECHNICAL JOURNAL is abstracted or indexed by *Abstract Journal in Earthquake Engineering, Applied Mechanics Review, Applied Science & Technology Index, Chemical Abstracts, Computer Abstracts, Current Contents/Engineering, Technology & Applied Sciences, Current Index to Statistics, Current Papers in Electrical & Electronic Engineering, Current Papers on Computers & Control, Electronics & Communications Abstracts Journal, The Engineering Index, International Aerospace Abstracts, Journal of Current Laser Abstracts, Language and Language Behavior Abstracts, Mathematical Reviews, Science Abstracts (Series A, Physics Abstracts; Series B, Electrical and Electronic Abstracts; and Series C, Computer & Control Abstracts), Science Citation Index, Sociological Abstracts, Social Welfare, Social Planning and Social Development,* and *Solid State Abstracts Journal.* Reproductions of the Journal by years are available in microform from University Microfilms, 300 N. Zeeb Road, Ann Arbor, Michigan 48106.


Bell System