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,

$$F_1(s, t) = \Delta^\nu \sum_{m=0}^{\infty} \binom{n_0 + \nu + m - 1}{m} (-cs)^m d^{-(m+n_0+\nu)}$$

$$= \sum_{m=0}^{\infty} s^m P_m^{(1)}(t), \tag{27}$$

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,

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

$$= \sum_{i=0}^{\min\{n_0, n\}} \left[\binom{n_0}{i} b^{n_0-i} a^i \right] \cdot \left[\Delta^\nu \binom{n_0 + \nu + n - i - 1}{n - i} (-c)^{n-i} d^{-(n_0+\nu+n-i)} \right], \tag{29}$$

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

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

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

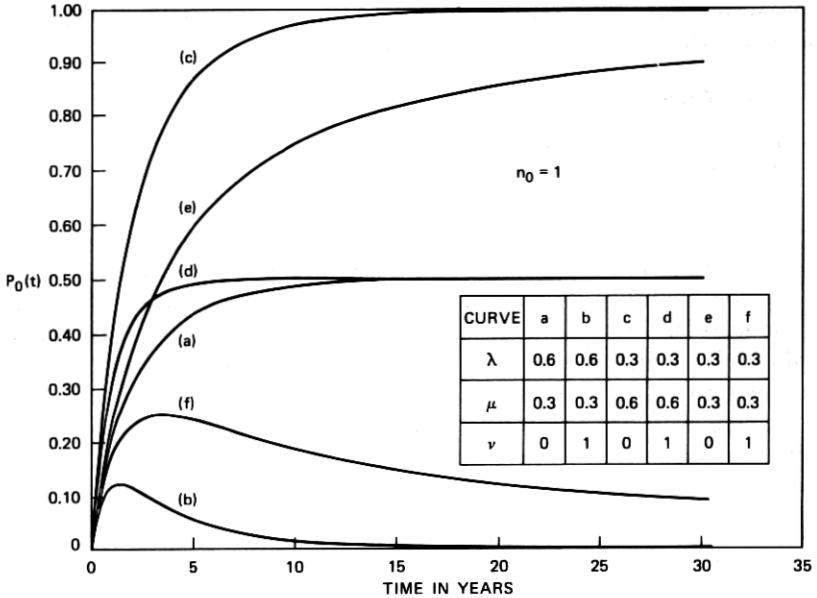

Fig. 1—Zero-state probability.

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.

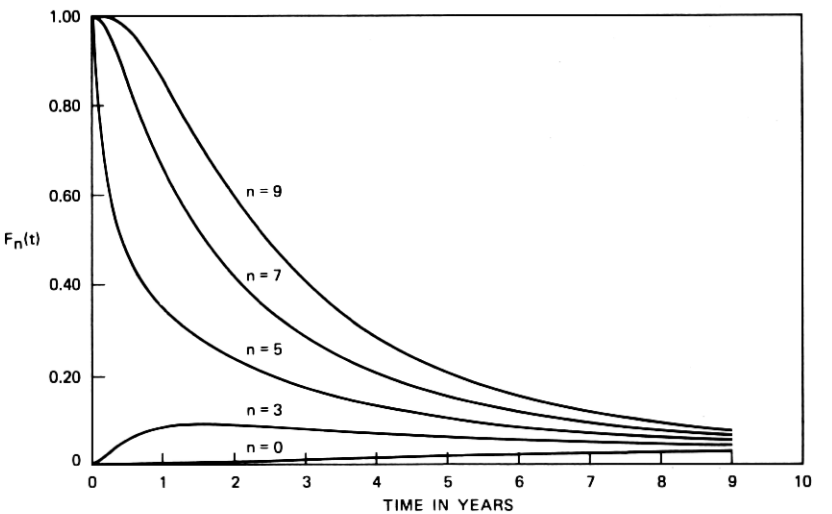

Fig. 2—Cumulative distribution without immigration ($\lambda > \mu$).

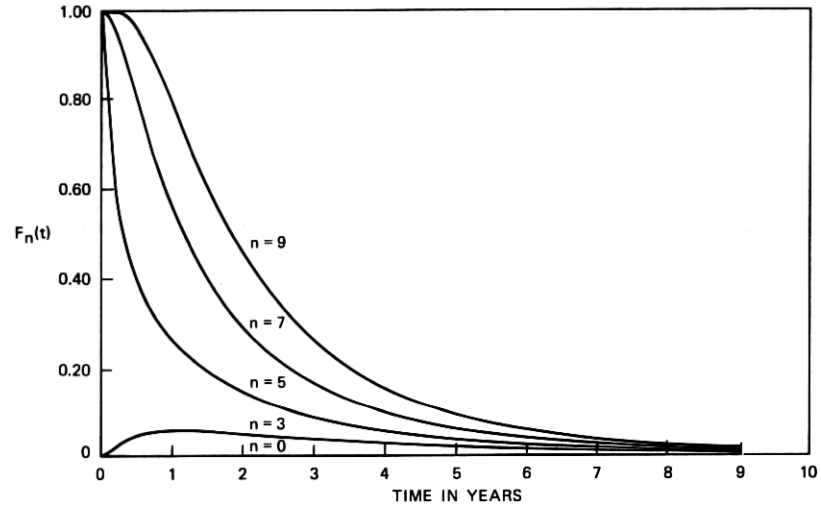

Fig. 3—Cumulative distribution with immigration ($\lambda > \mu$).

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

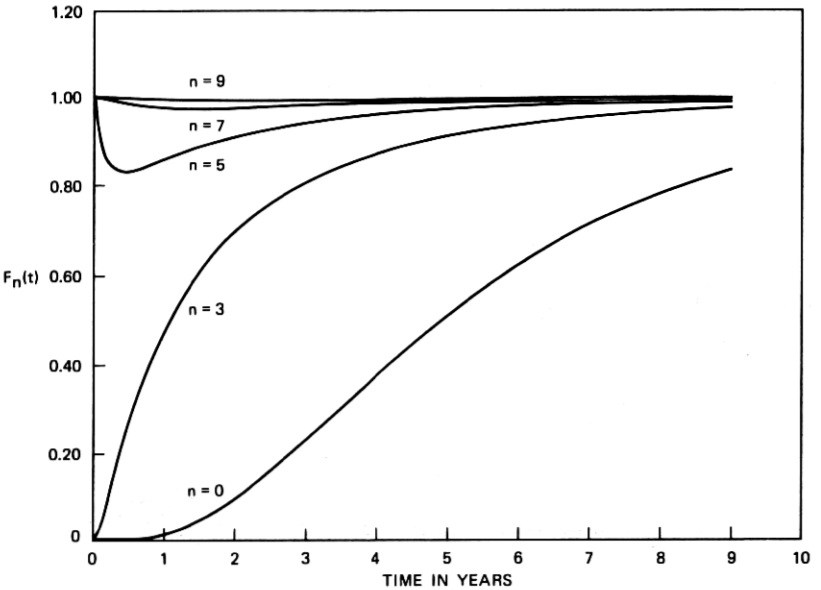

Fig. 4—Cumulative distribution without immigration ($\lambda < \mu$).

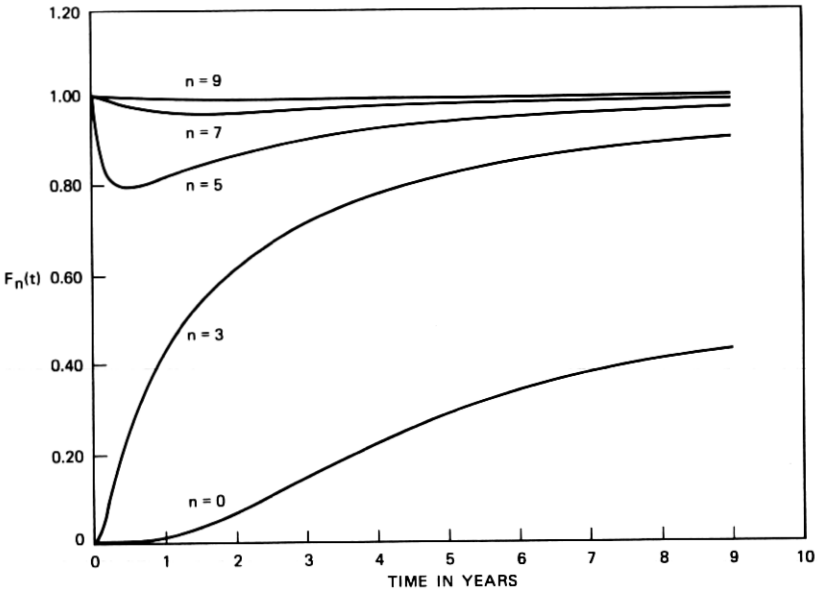

Fig. 5—Cumulative distribution with immigration ($\lambda < \mu$).

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 ($\lambda \neq \mu$, $\beta = 0$)

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

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

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

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

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

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

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

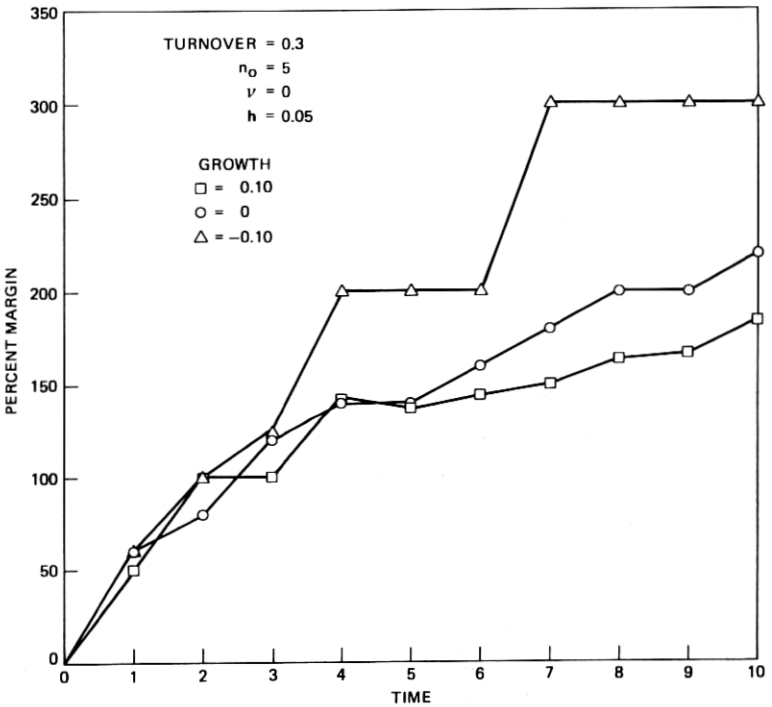

Fig. 6—Sensitivity of percent margin to growth (no immigration).

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

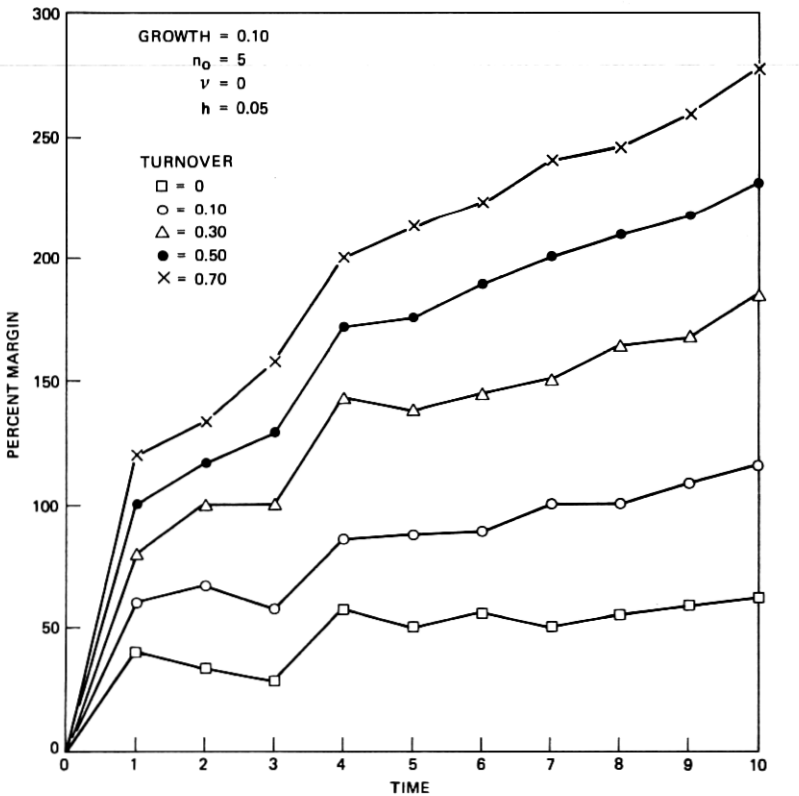

Fig. 7—Sensitivity of percent margin to turnover (no immigration).

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

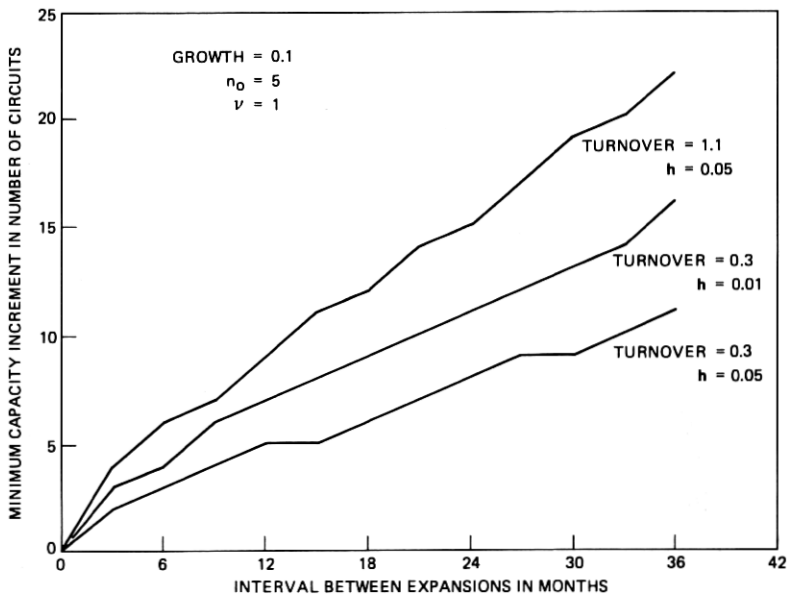

Fig. 8—Minimum capacity increments.

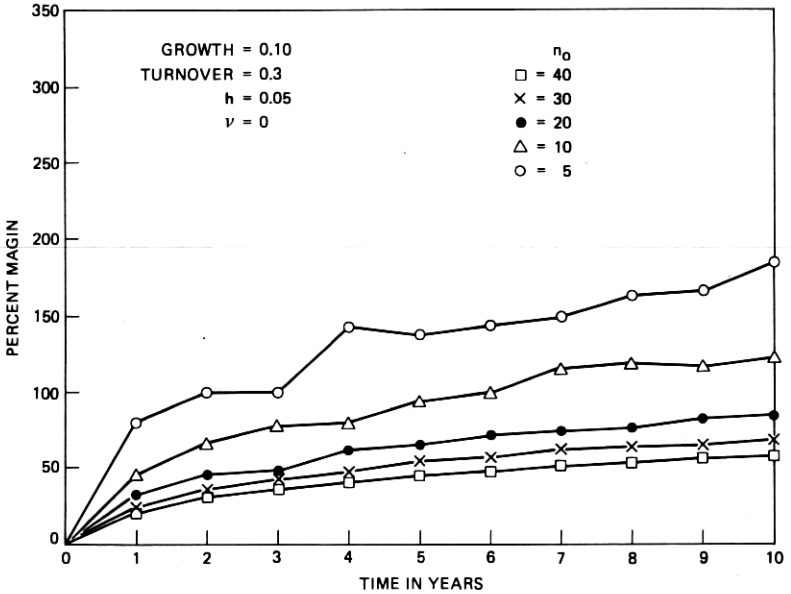

Fig. 9—Sensitivity of percent margin to initial demand (no immigration).

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,

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

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

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

$$n_1 + n_2 + \cdots + n_M = \tilde{n}.$$

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).

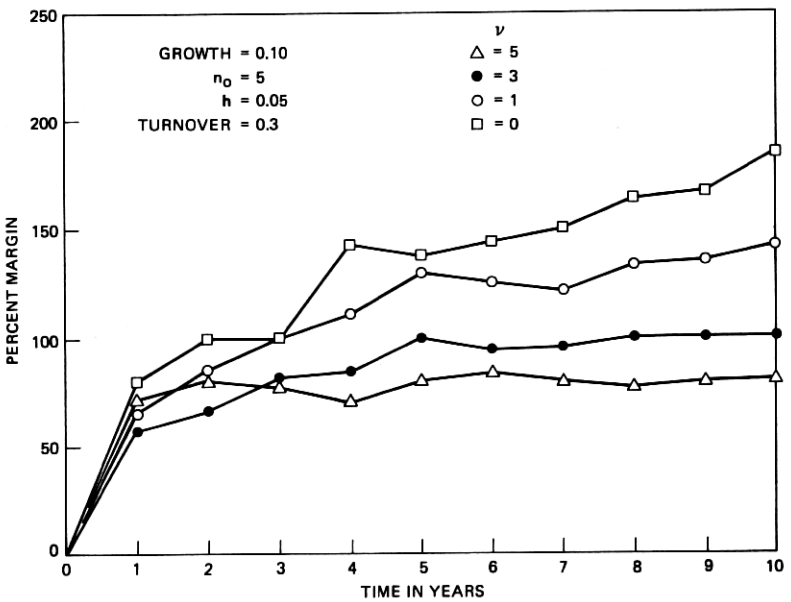

Fig. 10—Effect of immigration on percent margin.

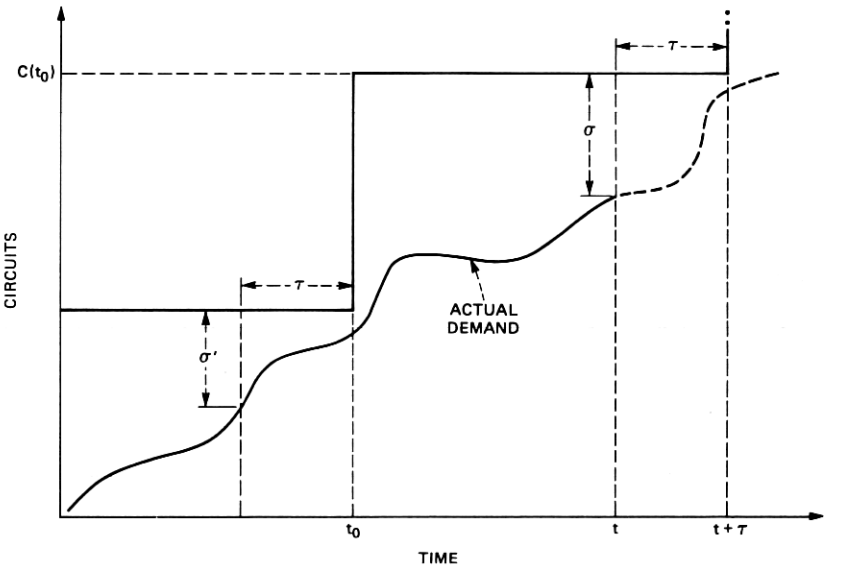

Fig. 11—Spare threshold.

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 $\mathbf{h}$, 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

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

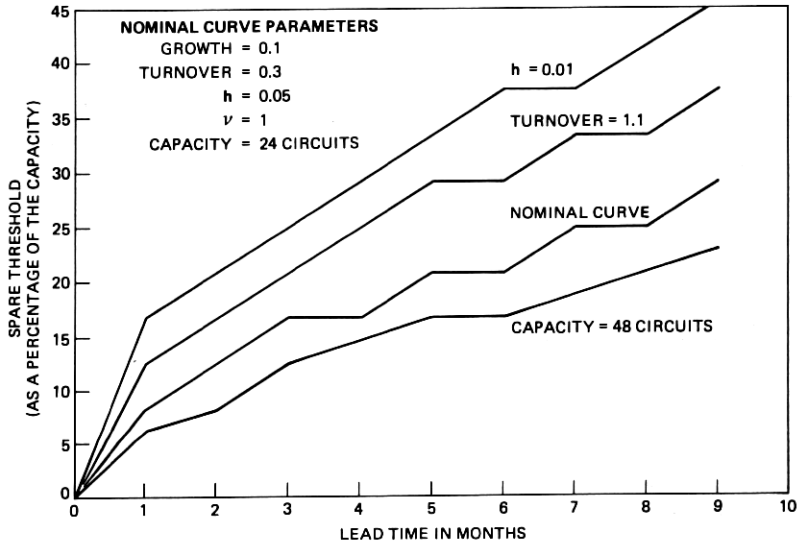

Fig. 12—Spare threshold as a function of lead time.

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 $\lim_{\delta\to 0} \dfrac{o(\delta)}{\delta} = 0$.

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.