THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 7, September 1975
Printed in U.S.A.

Variance of Load Measurements in Markovian Service Systems

By A. DESCLOUX

(Manuscript received November 25, 1974)

The load carried by a queuing system under equilibrium conditions is the average amount of server usage per unit of time. In telephony, this parameter is often evaluated by recording the number of busy servers at regular time intervals; these readings are then cumulated and their sum, after division by the number of observations, is an unbiased estimate of the carried load. The purpose of this paper is to derive exact formulas for the computation of the variance of this measurement in systems with arbitrary input and departure rates. The results obtained here thus apply to a wide class of teletraffic models which includes, in particular, the delay-and-loss systems with finite- or infinite-source inputs, exponential service times, and arbitrary defection rates from the queue. Problems related to computations are also considered, special attention being paid to the reduction of both computer time and storage when the number of states is large.

I. INTRODUCTION

Analysis of the stochastic behavior of traffic measurements is of considerable practical relevance, as it provides means for appraising field data as well as guidelines for selecting performance standards. Load measurements play a central role in this effort, and determination of their accuracy is therefore of particular interest. The present investigation yields an answer to this problem for a broad class of teletraffic models.

Whenever statistical equilibrium prevails (and it is assumed to throughout this paper), the load carried by a service system is the average amount of server usage per unit of time or, equivalently, the average number of busy servers at an arbitrary instant. In telephony, an estimate of this parameter is often obtained by "switch-counting." [1] This statistic, which is determined by recording the number of busy servers at regular intervals and then by taking the arithmetic mean of these discrete observations, is an unbiased estimate of the carried load.

The variance of this measurement, called hereafter the switch-count load to distinguish it from the estimate obtained by continuous observation, was first determined approximately by Palm[2] and Hayward[1] in the case of an infinite server group with Poisson input and exponential holding times. This result was later extended by Beneš,[3] who obtained the exact variance of the switch-count load for groups of finite sizes without waiting positions (loss systems). A further generalization to loss systems with recurrent input and exponential service is due to Neal and Kuczura.[4] Their formal analysis stops, however, with a derivation of the Laplace transform of the covariance function of the underlying carried-load process. From this point on, they proceed numerically, since explicit inversion of the transform appears to be difficult in general.

In this paper we are concerned with derivations of exact formulas for the variance of the switch-count load in finite systems with arbitrary state-dependent input and departure rates. The results presented here, therefore, fill a rather large gap, since they apply to a broad class of teletraffic models that includes, in particular, the (finite) delay systems with exponential holding-time distributions, arbitrary defection rates from the queue (if one is allowed to form) and either Poisson or quasi-random input (in the latter case, the traffic is generated by a finite number of sources that place demands for service at the same constant probability rate when free but that do not submit requests while being either served or waiting).

Let $N(t)$, the state of the system at time t, be defined as the number of busy devices at that instant (by device, we mean here either a server or a waiting position). Let c and d be, respectively, the number of servers and the number of devices.

Unless stated otherwise, we make the following assumptions:

(*i*) When $N(t) = n$ and $0 \leqq n < d$, the probability that a request originates during $(t, t + h)$, $h > 0$, is of the form $\lambda_n h + o(h)$, with $\lambda_n > 0$.

(*ii*) The requests which are submitted when all the devices are occupied are dismissed and, accordingly, λ_d is set equal to zero.

(*iii*) When $N(t) = n$ and $0 < n \leqq c$, the probability that a service time terminates during $(t, t + h)$ is of the form $\mu_n h + o(h)$, where $\mu_n > 0$.

(*iv*) When $N(t) = n > c$, the probability that either a service time terminates or a waiting request defects from the queue is of the form $\mu_n h + o(h)$ where $\mu_n > 0$ and $n \leqq d$.

(*v*) When a server becomes free, it is immediately reseized by one of the waiting requests if any are present in the system at that time.

Let $N_c(t)$ be the number of busy servers at time t and let $c \wedge n$ be the smaller of the two integers c and n. Then

$$N_c(t) = c \wedge N(t) = \begin{cases} N(t) & \text{if } N(t) \leqq c, \\ c & \text{if } N(t) > c, \end{cases}$$

and the switch-count load, $L_n(T)$, based on n observations (scans) made over $[0, T]$ at times $\tau, 2\tau, \cdots, n\tau$, is, by definition, equal to

$$n^{-1} \sum_{j=1}^{n} N_c(j\tau),$$

where $\tau \equiv T/n$.

Let $\text{Cov}\,[N_c(t_1), N_c(t_2)]$ be the covariance between $N_c(t_1)$ and $N_c(t_2)$. Under equilibrium conditions, this covariance depends only on $|t_1 - t_2|$ so that

$$\text{Cov}\,[N_c(t_1), N_c(t_2)] = \text{Cov}\,[N_c(0), N_c(|t_1 - t_2|)].$$

Hence, the variance of $L_n(\tau)$, cast in a form that will be convenient later, is given by the formula (Ref. 3, p. 137):

$$\text{Var}\, L_n(T) = n^{-2} \sum_{k=-n}^{n} (n - |k|) R_c(k\tau), \tag{1}$$

where

$$\begin{aligned} R_c(k\tau) &\equiv \text{Cov}\,[N_c(0), N_c(k\tau)] \\ &= \text{Cov}\,[N_c(0), N_c(|k|\tau)]. \end{aligned}$$

It is clear from (1) that the variance of the switch-count load is completely determined by the covariance function $R_c(\cdot)$ of the carried-load process $\{N_c(t), -\infty < t < \infty\}$, and therefore much of what follows is concerned with expressing $R_c(\cdot)$ in the most convenient form.

The covariance function can be stated at first in terms of the transition probabilities, and the resulting expression can then be reduced by taking the structural properties of the process into account. But alternate forms can also be obtained by making use of the fact that the conditional expectations, $E\{N_c(t) \mid N(0) = m\}$, $m = 0, 1, \cdots, d$, satisfy simple linear differential equations. The covariance formulas obtained by these diverse procedures exhibit distinct features that may be exploited in the computations. In all cases, however, $R_c(t)$ is expressed as a diagonal, positive-definite quadratic form which reveals that $R_c(\cdot)$ is completely monotonic.[5]

Expressions for the transition probabilities, the covariance function, and the variance of the switch-count load are derived in Sections II, III, and IV, respectively. The variance of load measurements based on

continuous observations can be found in Section IV. Extensions of the results of Sections II to IV to reversible Markov processes are also considered. Questions of a computational nature are dealt with in Section V, while Section VI is concerned with some asymptotic properties (d large) of the spectrum of the underlying process.

The formulas presented here have been programmed and used to explore the effects of parameter changes on the variance of the switch-count load. The result of that investigation will appear in another paper.

II. TRANSITION PROBABILITIES

In this section, we express the transition-probability matrix as a symmetric product of vectors and matrices. As becomes apparent later, this representation makes it possible to write the variance of the switch-count load in a way that greatly simplifies its evaluation.

Let $p_{mn}(t)$ be the probability of a transition from state m to state n in time t:

$$p_{mn}(t) \equiv \Pr[N(t) = n | N(0) = m], \qquad m, n = 0, 1, \cdots, d.$$

These transition probabilities satisfy the following system of differential equations:

$$\begin{aligned}
\frac{d}{dt} p_{m0} &= \mu_1 p_{m1} - \lambda_0 p_{m0}, \\
\frac{d}{dt} p_{mn} &= \mu_{n+1} p_{m,n+1} + \lambda_{n-1} p_{m,n-1} - (\lambda_n + \mu_n) p_{mn}, \qquad 1 \leqq n < d, \\
\frac{d}{dt} p_{md} &= \lambda_{d-1} p_{m,d-1} - \mu_d p_{md}.
\end{aligned} \tag{2}$$

Let

$$\mathbf{A}_d \equiv \begin{bmatrix}
-\lambda_0 & \lambda_0 & & & & & \\
\mu_1 & -(\lambda_1+\mu_1) & \lambda_1 & & & & \\
 & \mu_2 & -(\lambda_2+\mu_2) & \lambda_2 & & & \\
 & & \cdots & \cdots & \cdots & & \\
 & & & \mu_{d-1} & -(\lambda_{d-1}+\mu_{d-1}) & \lambda_{d-1} & \\
 & & & & \mu_d & -\mu_d &
\end{bmatrix}$$

and

$$\mathbf{P}_d(t) \equiv [p_{mn}(t)], \qquad m, n = 0, 1, \cdots, d,$$

be the transition-probability matrix. Capital and lower-case bold-face letters are used exclusively to designate matrices and vectors, respectively.

With this notation, the system of differential equations (2) becomes

$$\frac{d}{dt}\mathbf{P}_d(t) = \mathbf{P}_d(t)\cdot\mathbf{A}_d, \qquad t \geqq 0, \tag{3}$$

so that, for $k = 1, 2, \cdots$,

$$\frac{d^k}{dt^k}\mathbf{P}_d(t) = \frac{d^{k-1}}{dt^{k-1}}\mathbf{P}_d(t)\cdot\mathbf{A}_d, \qquad t \geqq 0. \tag{4}$$

It follows from our assumptions that if the system is in state m at time zero $[N(0) = m]$, then $\lim_{t\downarrow 0} p_{mm}(t) = 1$ and $\lim_{t\downarrow 0} p_{mn}(t) = 0$ for $n \neq m$. Hence, with $\mathbf{I}_d$ the identity matrix of order $d + 1$, the initial conditions take the following form:

$$\mathbf{P}_d(0) \equiv \lim_{t\downarrow 0}\mathbf{P}_d(t) = \mathbf{I}_d,$$

and by (3) and (4) we therefore have

$$\lim_{t\downarrow 0}\frac{d^k}{dt^k}\mathbf{P}_d(t) = \mathbf{A}_d^k. \tag{5}$$

The initial conditions state that $\mathbf{P}_d(\cdot)$ is right-continuous at $t = 0$ and imply that $\mathbf{P}_d(\cdot)$ is continuous for all $t > 0$. By (3) and (4), all the derivatives of $\mathbf{P}_d(\cdot)$ exist for $t > 0$, and by (5) they are also right-continuous at $t = 0$. An application of Taylor's theorem then yields (Ref. 6, pp. 240 ff.)

$$\mathbf{P}_d(t) = \exp(\mathbf{A}t) = \sum_{k=0}^{\infty}\frac{1}{k!}\mathbf{A}_d^k t^k, \qquad t \geqq 0. \tag{6}$$

The elements of $\mathbf{A}_d$ situated immediately either above or below the diagonal are all strictly positive and so $\mathbf{A}_d$ can be symmetrized. Indeed, let

$$\mathbf{D}_d \equiv \operatorname{diag}[\delta_0, \delta_1, \cdots, \delta_d]$$

with

$$\delta_0 = \zeta \quad \text{and} \quad \delta_m \equiv \zeta\left(\frac{\mu_1\mu_2\cdots\mu_m}{\lambda_0\lambda_1\cdots\lambda_{m-1}}\right)^{\frac{1}{2}} = \xi p_m^{-\frac{1}{2}}, \qquad m = 1, \cdots, d,$$

where (*i*) ζ is a nonvanishing but otherwise arbitrary constant, (*ii*) the p_m are the equilibrium state probabilities, and (*iii*) $\xi = \zeta p_0^{\frac{1}{2}}$. Without loss of generality, we can—and shall—set $\zeta = p_0^{\frac{1}{2}}$ so that $\xi = 1$ and

$$\mathbf{D}_d^{-1} = \operatorname{diag}[p_0^{\frac{1}{2}}, p_1^{\frac{1}{2}}, \cdots, p_d^{\frac{1}{2}}]. \tag{7}$$

It is easy to verify that

$$\mathbf{S}_d = \mathbf{D}_d^{-1}\cdot\mathbf{A}_d\cdot\mathbf{D}_d \tag{8}$$

is symmetric, its nonvanishing elements being

$$s_{mm} = -(\lambda_m + \mu_m), \qquad m = 0, 1, \cdots, d, \qquad (\lambda_d = 0),$$
$$s_{m,m+1} = s_{m+1,m} = (\lambda_m \mu_{m+1})^{\frac{1}{2}}, \qquad m = 0, 1, \cdots, d-1.$$

Hence, by (8) we have

$$\mathbf{A}_d^k = \mathbf{D}_d \cdot \mathbf{S}_d^k \cdot \mathbf{D}_d^{-1}, \qquad k = 0, 1, 2, \cdots, \tag{9}$$

and, by (6),

$$\mathbf{P}_d(t) = \mathbf{D}_d \cdot \exp(\mathbf{S}_d t) \cdot \mathbf{D}_d^{-1} = \sum_{k=0}^{\infty} \frac{1}{k!} (\mathbf{D}_d \cdot \mathbf{S}_d^k \cdot \mathbf{D}_d^{-1}) t^k. \tag{10}$$

The representation of $\mathbf{A}_d$ in terms of the symmetric tridiagonal matrix $\mathbf{S}_d$ entails substantial formal simplification of the final results. And it is also particularly convenient computationally, since the determination of the characteristic values of $\mathbf{A}_d$ (which are needed for an exact solution) is best carried out after symmetrization.

The matrices $\mathbf{A}_d$ and $\mathbf{S}_d$ clearly have the same characteristic values, $r_0, r_1, \cdots, r_d$. But $\mathbf{S}_d$ is symmetric and is therefore unitarily similar to the diagonal matrix

$$\mathbf{C}_d \equiv \operatorname{diag}[r_0, r_1, \cdots, r_d].$$

This means that an orthogonal matrix $\mathbf{B}_d$ exists such that

$$\mathbf{S}_d = \mathbf{B}_d' \cdot \mathbf{C}_d \cdot \mathbf{B}_d, \qquad \mathbf{B}_d \cdot \mathbf{B}_d' = \mathbf{B}_d' \cdot \mathbf{B}_d = \mathbf{I}_d, \tag{11}$$

where $\mathbf{B}_d'$ is the transpose of $\mathbf{B}_d$.

But $\mathbf{S}_d$ is also tridiagonal, and its off-diagonal elements never vanish. Hence, $\mathbf{S}_d$ is nonderogatory and its characteristic values are necessarily distinct (Ref. 7, p. 26). The elements in the nth column of $\mathbf{B}_d'$ are then the components of the (uniquely defined) normalized characteristic vector associated with the nth characteristic value $r_n (n = 0, 1, \cdots, d)$.

We now substitute (11) into (10). This yields

$$\mathbf{P}_d(t) = \sum_{k=0}^{\infty} \frac{1}{k!} (\mathbf{D}_d \cdot \mathbf{B}_d' \cdot \mathbf{C}_d^k \cdot \mathbf{B}_d \cdot \mathbf{D}_d^{-1}) t^k,$$

so that

$$\begin{aligned} \mathbf{P}_d(t) &= \mathbf{D}_d \cdot \mathbf{B}_d' \cdot \exp(\mathbf{C}_d \cdot t) \cdot \mathbf{B}_d \cdot \mathbf{D}_d^{-1} \\ &= \mathbf{D}_d \cdot \mathbf{B}_d' \cdot \operatorname{diag}[e^{r_0 t}, e^{r_1 t}, \cdots, e^{r_d t}] \cdot \mathbf{B}_d \cdot \mathbf{D}_d^{-1}. \end{aligned} \tag{12}$$

We note now that all the row sums of $\mathbf{A}_d$ vanish and one of the characteristic roots, r_0, say, must therefore be equal to zero. Furthermore, known extremal properties of the characteristic values can be used to show that $r_1, r_2, \cdots, r_d$ are negative. It is also readily seen that

$$\mathbf{p}_d^{(\frac{1}{2})} \equiv (p_0^{\frac{1}{2}}, p_1^{\frac{1}{2}}, \cdots, p_d^{\frac{1}{2}})'$$

is the characteristic vector of $\mathbf{S}_d$ that corresponds to the vanishing

characteristic root r_0. Indeed, let $\mathbf{e}_d$ and $\mathbf{o}_d$ be the $(d+1)$ dimensional (column) vectors whose components are all equal to 1 and 0, respectively. Then, since $\mathbf{A}_d \cdot \mathbf{e}_d = \mathbf{o}_d$, we have, by (8) and (7),

$$\mathbf{D}_d \cdot \mathbf{S}_d \cdot \mathbf{D}_d^{-1} \cdot \mathbf{e}_d = \mathbf{D}_d \cdot \mathbf{S}_d \cdot \mathbf{p}_d^{(\frac{1}{2})} = \mathbf{o}_d.$$

But none of the diagonal elements of $\mathbf{D}_d$ vanishes and the relation

$$\mathbf{D}_d \cdot \mathbf{S}_d \cdot \mathbf{p}_d^{(\frac{1}{2})} = \mathbf{o}_d$$

can hold if and only if $\mathbf{S}_d \cdot \mathbf{p}_d^{(\frac{1}{2})} = \mathbf{o}_d$. Thus, $\mathbf{p}_d^{(\frac{1}{2})}$ is the characteristic vector associated with $r_0(=0)$, a fact that may be of relevance in the computations, as a comparison of $\mathbf{p}_d^{(\frac{1}{2})}$ with $\mathbf{D}_d^{-1}$ provides an accuracy check for the method used to determine the characteristic vectors.

In the derivation of formula (12), advantage was taken of the fact that the transition-rate matrix $\mathbf{A}_d$ is symmetrizable. It is worth noting that this relatively simple expression for $\mathbf{P}_d$ is a consequence of this property, and therefore holds for all (and actually only for) reversible Markovian processes with finite state spaces. Indeed, by definition, the class of these processes—which includes those of the birth-and-death type—is fully characterized by the following conditions (Refs. 8 and 9):

$$p_m p_{mn}(t) = p_n p_{nm}(t), \qquad m, n = 0, 1, \cdots, d, \tag{13}$$

or, equivalently, by the single relation:

$$\mathbf{D}_d^{-2} \cdot \mathbf{P}_d = \mathbf{P}_d' \cdot \mathbf{D}_d^{-2}. \tag{14}$$

Hence (12), written in terms of $\mathbf{S}_d$, implies that

$$\begin{aligned} \mathbf{D}_d^{-2} \cdot \mathbf{P}_d &= \mathbf{D}_d^{-1} \cdot \exp(\mathbf{S}_d t) \cdot \mathbf{D}_d^{-1} \\ &= (\mathbf{D}_d^{-1} \cdot \exp(\mathbf{S}_d t) \cdot \mathbf{D}_d^{-1})' = \mathbf{P}_d' \cdot \mathbf{D}_d^{-2}, \end{aligned}$$

and (14) is therefore satisfied.

Conversely, we show next that (14) is a sufficient condition for (12) to hold.

Pre- and post-multiplication of (14) by $\mathbf{D}_d$ yield

$$\mathbf{D}_d^{-1} \cdot \mathbf{P}_d \cdot \mathbf{D}_d = \mathbf{D}_d \cdot \mathbf{P}_d' \cdot \mathbf{D}_d^{-1}. \tag{15}$$

Substituting the expansion of $\mathbf{P}_d$ as given by (6) into (15), and performing the multiplications by $\mathbf{D}_d$ and $\mathbf{D}_d^{-1}$ under the summation sign (which is clearly legitimate), we obtain:

$$\sum_{k=0}^{\infty} \frac{1}{k!} (\mathbf{D}_d^{-1} \cdot \mathbf{A}_d^k \cdot \mathbf{D}_d) t^k = \sum_{k=0}^{\infty} \frac{1}{k!} (\mathbf{D}_d \cdot (\mathbf{A}_d^k)' \cdot \mathbf{D}_k^{-1}) t^k, \qquad t \geqq 0.$$

However, this relation cannot be satisfied unless

$$\mathbf{D}_d^{-1} \cdot \mathbf{A}_d \cdot \mathbf{D}_d = \mathbf{D}_d \cdot \mathbf{A}_d' \cdot \mathbf{D}_d^{-1},$$

so that, by transposition,

$$(\mathbf{D}_d^{-1} \cdot \mathbf{A}_d \cdot \mathbf{D}_d)' = \mathbf{D}_d^{-1} \cdot \mathbf{A}_d \cdot \mathbf{D}_d.$$

This means that $\mathbf{A}_d$ is symmetrizable by pre- and post-multiplication by $\mathbf{D}_d^{-1}$ and $\mathbf{D}_d$, respectively, and (12) then follows as shown earlier.

Under the assumption that the process is reversible and that all the states communicate with each other[10] (i.e., $p_{mn}(t) > 0$, $m, n = 0, \cdots, d$, $t > 0$), the characteristic roots of $\mathbf{A}_d$ are necessarily simple. (Note that $\mathbf{A}_d$, and hence $\mathbf{S}_d = \mathbf{D}_d^{-1} \cdot \mathbf{A}_d \cdot \mathbf{D}_d$, need no longer be tridiagonal.) This can be proved as follows.

The matrix $\mathbf{S}_d$ is symmetric and can therefore be tridiagonalized by a method from Householder (Ref. 7, pp. 152, 153, 290–293, and 343). According to this procedure, the tridiagonalization of $\mathbf{S}_d$ is achieved by successive right and left multiplications by symmetric orthogonal matrices, $\mathbf{U}_1, \mathbf{U}_2, \cdots, \mathbf{U}_{d-1}$, of the form

$$\mathbf{U}_r = \mathbf{I}_d - 2\mathbf{w}_r \cdot \mathbf{w}_r',$$

where $\mathbf{w}_r$ is a suitably chosen $d + 1$ dimensional (column) vector whose first r components are zero. (All the $\mathbf{U}_r$ are of order $(d + 1)$ and $\mathbf{U}_r^2 = \mathbf{I}_d$, $r = 1, 2, \cdots, d - 1$.) A direct application of the results derived in Ref. 7, above, shows that $\mathbf{S}_d$ admits of the following representation:

$$\mathbf{S}_d = \mathbf{U}_1 \cdot \mathbf{U}_2 \cdot \cdots \cdot \mathbf{U}_{d-1} \cdot \mathbf{T}_d \cdot \mathbf{U}_{d-1} \cdot \cdots \cdot \mathbf{U}_2 \cdot \mathbf{U}_1,$$

where $\mathbf{T}_d$ is a symmetric tridiagonal matrix of order $d + 1$.

Let θ_{ij} be the elements of $\mathbf{T}_d$.

We are now faced with two possibilities. Either $\theta_{i,i+1} = \theta_{i+1,i} \neq 0$ for $i = 0, 1, \cdots, d - 1$, or there is an index $j (< d)$ such that $\theta_{j,j+1} = \theta_{j+1,j} = 0$ so that

$$\mathbf{T}_d = \begin{bmatrix} \mathbf{T}_j & 0 \\ 0 & \mathbf{T}_{d-j-1} \end{bmatrix}. \tag{16}$$

In the first instance, all the characteristic roots of $\mathbf{S}_d$, and hence of $\mathbf{A}_d$, are distinct (Ref. 7, p. 26). To complete the proof, it is therefore sufficient to show that the second contingency cannot occur when all states communicate with each other. To this end, we proceed indirectly. We assume that (16) is satisfied for some $j < d$ and show that some states then do not communicate with others.

When (16) holds for $j < d$, we have, for any $k \geqq 0$,

$$\begin{aligned} \mathbf{A}_d^k &= \mathbf{D}_d \cdot \mathbf{S}_d^k \cdot \mathbf{D}_d^{-1} \\ &= \mathbf{D}_d \cdot \mathbf{U}_1 \cdot \cdots \cdot \mathbf{U}_{d-1} \cdot \begin{bmatrix} \mathbf{T}_j^k & 0 \\ 0 & \mathbf{T}_{d-j-1}^k \end{bmatrix} \cdot \mathbf{U}_{d-1} \cdot \cdots \cdot \mathbf{U}_1 \cdot \mathbf{D}_d^{-1}. \end{aligned} \tag{17}$$

The first row in $\mathbf{D}_d$ is $(p_0^{-\frac{1}{2}}, 0, \cdots, 0)$ and the elements in the first column and first row of each of the $\mathbf{U}_r$'s are zero except for their first component, which is always equal to 1. Hence,

$$(1, 0, \cdots, 0)\cdot\mathbf{D}_d\cdot\mathbf{U}_1\cdot\cdots\cdot\mathbf{U}_{d-1} = (p_0^{-\frac{1}{2}}, 0, \cdots, 0). \tag{18}$$

Similarly, since the $\mathbf{U}_r$'s are symmetric, we have

$$\mathbf{U}_{d-1}\cdot\cdots\cdot\mathbf{U}_1\cdot\mathbf{D}_d^{-1}\cdot(1, 0, \cdots, 0)' = (p_0^{\frac{1}{2}}, 0, \cdots, 0)'. \tag{19}$$

Hence, by (17) to (19),

$$\begin{aligned} P_{00}(t) &= (1, 0, \cdots, 0)\left(\sum_{k=0}^{\infty}\frac{1}{k!}\mathbf{A}_d^k\cdot t^k\right)(1, 0, \cdots, 0)' \\ &= \sum_{k=0}^{\infty}\frac{1}{k!}\theta_0^{(k)}t^k, \end{aligned} \tag{20}$$

where $\theta_0^{(k)}$ is the element belonging to the first row and first column of $\mathbf{T}_j^k$.

Let a_{mn} and s_{mn}, $m, n = 0, 1, \cdots, d$, be the elements of $\mathbf{A}_d$ and $\mathbf{S}_d$, respectively. Under the present assumptions, $a_{mn}\cdot a_{nm} \geqq 0$ and $s_{mn} = (a_{mn}\cdot a_{nm})^{\frac{1}{2}}$, $m, n = 0, 1, \cdots, d$, $m \neq n$. The elements in the first r rows and columns of $\mathbf{S}_d$ are therefore uniquely determined by the elements in the first r rows and columns of $\mathbf{A}_d$. Similarly, the vector $\mathbf{w}_r$ depends only on the components, s_{mn}, of $\mathbf{S}_d$ for which either $m \leqq r-1$ and $n = r-1, \cdots, d$ or, by symmetry, $n \leqq r-1$ and $m = r-1, \cdots, d$ (Ref. 7, pp. 290 ff). Consequently, the elements of $\mathbf{T}_j^k$ (which are all obtained after $j-1$ steps) depend only on the elements of the first $j+1$ rows and columns of $\mathbf{A}_d$. This implies that the transition probability $P_{00}(t)$, as given by (20), is independent of the rates a_{mn}, $m, n > j$. However, the process being Markovian, this can only be true if $P_{0m}(t) = 0$ for $m > j$ which means (since, by assumption, $j < d$) that state 0 does not communicate with states $j+1, \cdots, d$, as was to be proved.

III. COVARIANCE FUNCTION

3.1 First version

The covariance function of the carried-load process is, by definition,

$$\begin{aligned} R_c(t) &\equiv \sum_{m,n=0}^{d}(c \wedge n)\cdot(c \wedge m)p_n p_{nm}(t) - M_{c1}^2 \\ &= \sum_{m,n=0}^{d}(c \wedge n)\cdot(c \wedge m)p_n[p_{nm}(t) - p_m], \end{aligned}$$

where

$$M_{c1} \equiv EN_c(0) = \sum_{n=0}^{d}(c \wedge n)p_n.$$

However, if γ is an arbitrary constant, the covariance of the process $\{N_c(t) + \gamma, t \geqq 0\}$ is also R_c. Hence, with the notation

$$\rho_n = (c \wedge n) + \gamma, \qquad n = 0, 1, \cdots, d,$$

we also have

$$R_c(t) = \sum_{m,n=0}^{d} \rho_n \rho_m p_n [p_{nm}(t) - p_m]. \tag{21}$$

Let

$$\hat{\mathbf{P}}_d \equiv \begin{bmatrix} p_0 & p_1 & \cdots & p_d \\ p_0 & p_1 & & p_d \\ \vdots & \vdots & & \vdots \\ p_0 & p_1 & \cdots & p_d \end{bmatrix}$$

and

$$\mathbf{G}_d(t) \equiv [p_{nm}(t) - p_m].$$

The matrix $\hat{\mathbf{P}}_d$ can be obtained by letting $t \to \infty$ in (13). Hence,

$$\hat{\mathbf{P}}_d = \mathbf{D}_d \cdot \mathbf{B}_d' \cdot \text{diag}\,[1, 0, 0, \cdots, 0] \cdot \mathbf{B}_d \cdot \mathbf{D}_d^{-1}$$

and

$$\begin{aligned} \mathbf{G}_d(t) &= \mathbf{P}_d(t) - \hat{\mathbf{P}}_d \\ &= \mathbf{D}_d \cdot \mathbf{B}_d' \cdot \text{diag}\,[0, e^{r_1 t}, \cdots, e^{r_d t}] \cdot \mathbf{B}_d \cdot \mathbf{D}_d^{-1}. \end{aligned}$$

We now introduce two auxiliary row vectors:

$$\mathbf{r}_d' \equiv (\rho_0, \rho_1, \cdots, \rho_d), \qquad \mathbf{s}_d' \equiv (p_0\rho_0, p_1\rho_1, \cdots, p_d\rho_d).$$

Then the coefficient of $e^{r_i t}$ in the linear form

$$\mathbf{s}_d' \cdot \mathbf{G}_d(t) \cdot \mathbf{r}_d = \mathbf{s}_d' \cdot \mathbf{D}_d \cdot \mathbf{B}_d' \cdot \text{diag}\,[0, e^{r_1 t}, \cdots, e^{r_d t}] \cdot \mathbf{B}_d \cdot \mathbf{D}_d^{-1} \mathbf{r}_d$$

is the same as the coefficient of $e^{r_i t}$ in (21), and we may conclude that

$$R_c(t) = \mathbf{s}_d' \cdot \mathbf{D}_d \cdot \mathbf{B}_d' \cdot \text{diag}\,[0, e^{r_1 t}, \cdots, e^{r_d t}] \cdot \mathbf{B}_d \cdot \mathbf{D}_d^{-1} \cdot \mathbf{r}_d.$$

With the notation

$$\mathbf{q}_d' \equiv (\rho_0 p_0^{\frac{1}{2}}, \cdots, \rho_d p_d^{\frac{1}{2}}),$$

we have

$$\mathbf{q}_d = \mathbf{s}_d' \cdot \mathbf{D}_d \qquad \text{and} \qquad \mathbf{q}_d = \mathbf{D}_d^{-1} \cdot \mathbf{r}_d,$$

so that

$$R_c(t) = \mathbf{q}_d' \cdot \mathbf{B}_d \cdot \text{diag}\,[0, e^{r_1 t}, \cdots, e^{r_d t}] \cdot \mathbf{B}_d \cdot \mathbf{q}_d \tag{22}$$

or, alternatively,

$$R_c(t) = \sum_{i=1}^{d} b_i^2 e^{r_i t} \tag{23}$$

with b_i the ith component of the row vector $\mathbf{q}_d' \cdot \mathbf{B}_d'$. This last expression shows that the coefficient of $e^{r_i t}$ in either (22) or (23) is necessarily

nonnegative. Furthermore, since all the r_i's ($i = 1, \cdots, d$) are negative, we also have

$$(-1)^k \frac{d^k}{dt^k} R_c(t) \geqq 0, \qquad t \geqq 0, \qquad k = 0, 1, \cdots,$$

and R_c is therefore completely monotonic over $[0, \infty)$.[5]

If we now set $\gamma = -c$, the last $d - c$ components of $\mathbf{q}_d$ are equal to zero, and so the determination of R_c by means of (22) necessitates only the computation of the first c components of the characteristic vectors. Formula (22) is therefore often well-suited to the case of delay systems. But unless the number of waiting positions exceeds the number of servers, greater reduction of the computations can be achieved by means of the formulas derived below.

In the preceding derivation, the ρ's are independent of the arrival and departure rates, and the formulas of this subsection therefore hold for arbitrary, reversible, Markov processes with finite state spaces. In contrast, the results of the next subsection are restricted to birth-and-death processes.

3.2 Alternative forms

Multiplying the nth equation in (2) by $(c \wedge n)$ and then summing with respect to $n (0 \leqq n \leqq d)$, we obtain, after rearranging and canceling terms,

$$\sum_{n=0}^{d} (c \wedge n) \cdot \frac{d}{dt} p_{mn} = \sum_{n=0}^{c-1} \lambda_n p_{mn} - \sum_{n=1}^{c} \mu_n p_{mn}. \tag{24}$$

But

$$\sum_{n=0}^{d} (c \wedge n) p_{mn}(t) = E\{N_c(t) \,|\, N(0) = m\},$$

so that, by (24),

$$\frac{d}{dt} E\{N_c(t) \,|\, N(0) = m\} = \sum_{n=0}^{c-1} \lambda_n p_{mn} - \sum_{n=1}^{c} \mu_n p_{mn}. \tag{25}$$

Adding and subtracting $\kappa E\{N_c(t) \,|\, N(0) = m\}$ on the right-hand side of (25), we obtain

$$\frac{d}{dt} E\{N_c(t) \,|\, N(0) = m\} = \kappa E\{N_c(t) \,|\, N(0) = m\} + \sum_{n=0}^{d} \rho_n^*(\kappa) p_{mn}, \qquad m = 0, 1, \cdots, d, \tag{26}$$

where

$$\rho_n^*(\kappa) = \begin{cases} (\lambda_n - \mu_n - \kappa n) & \text{if } n = 0, 1, \cdots, c-1, \quad (\mu_0 = 0), \\ -(\kappa c + \mu_c) & \text{if } n = c, \\ -\kappa c & \text{if } n = c+1, \cdots, d. \end{cases} \tag{27}$$

In the preceding formulas, κ is an arbitrary real number that may be positive, negative, or null. We see later that the covariance formulas can occasionally be simplified by appropriate choices of κ.

Taking the initial conditions,

$$E\{N_c(0)\,|\,N(0) = m\} = (c \wedge m), \qquad 0 \leqq m \leqq d,$$

into account, the solution of (26) is

$$E\{N_c(t)\,|\,N(0) = m\} = (c \wedge m)e^{\kappa t} + \int_0^t e^{\kappa(t-u)} \left[\sum_{n=0}^{d} \rho_n^*(\kappa) p_{mn}(u) \right] \cdot du \quad (28)$$

so that

$$\begin{aligned} R_c(t) &\equiv \sum_{m=0}^{d} (c \wedge m) p_m E\{N_c(t)\,|\,N(0) = m\} - M_{c1}^2 \\ &= M_{c2}e^{\kappa t} - M_{c1}^2 \\ &\quad + \int_0^t e^{\kappa(t-u)} \sum_{m=0}^{d} (c \wedge m) p_m \left[\sum_{n=0}^{d} \rho_n^*(\kappa) p_{mn}(u) \right] \cdot du, \end{aligned}$$

where $M_{c2} \equiv EN_c^2(0)$. By means of (13), the preceding relation can be expressed in a much more convenient form:

$$\begin{aligned} R_c(t) &\equiv M_{c2}e^{\kappa t} - M_{c1}^2 \\ &\quad + \int_0^t e^{\kappa(t-u)} \sum_{n=0}^{d} \rho_n^*(\kappa) p_n \left[\sum_{m=0}^{d} (c \wedge m) p_{nm}(u) \right] \cdot du \\ &= M_{c2}e^{\kappa t} - M_{c1}^2 \\ &\quad + \int_0^t e^{\kappa(t-u)} \sum_{n=0}^{d} \rho_n^*(\kappa) p_n E\{N_c(u)\,|\,N(0) = n\} \cdot du. \end{aligned} \quad (29)$$

Next, substituting (28) into (29), we obtain

$$\begin{aligned} R_c(t) &= M_{c2}e^{\kappa t} - M_{c1}^2 + te^{\kappa t} \sum_{n=0}^{d} (c \wedge n) \cdot \rho_n^*(\kappa) \cdot p_n \\ &\quad + \int_0^t e^{\kappa(t-u)} \sum_{n,m=0}^{d} \rho_m^*(\kappa) \cdot p_n \int_0^u e^{\kappa(u-v)} p_{nm}(v) \cdot dv \cdot du. \end{aligned}$$

Let R_c^* be the Laplace transform of R_c and p_{nm}^* that of p_{nm}. The preceding relation then yields

$$\begin{aligned} R_c^*(s) &= \frac{M_{c2}}{s - \kappa} - \frac{M_{c1}^2}{s} + \frac{1}{(s - \kappa)^2} \sum_{n=0}^{d} (c \wedge n) \cdot \rho_n^*(\kappa) \cdot p_n \\ &\quad + \sum_{n,m=0}^{d} \rho_n^*(\kappa) \cdot \rho_m^*(\kappa) \cdot p_n \frac{p_{nm}^*(s)}{(s - \kappa)^2}. \end{aligned} \quad (30)$$

We know, however, that $p_{nm}(t)$ is of the form

$$p_{nm}(t) = p_m + \sum_{i=1}^{d} \gamma_{nmi} e^{r_i t},$$

so that

$$R_c(t) = \sum_{m,n=0}^{d} (c \wedge n)\cdot(c \wedge m)p_n \sum_{i=1}^{d} \gamma_{nmi} e^{r_i t}.$$

This implies that the only poles of R_c^* are r_i, $i = 1, \cdots, d$, and that $\lim_{t\to\infty} R_c(t) = 0$. Taking these two facts into consideration, we see at once that (30) reduces to

$$R_c^*(s) = \sum_{n,m=0}^{d} \rho_n^*(\kappa)\cdot\rho_m^*(\kappa)\cdot p_n \sum_{i=1}^{d} \frac{\gamma_{nmi}}{(r_i - \kappa)^2}\cdot\frac{1}{s - r_i},$$

provided $\kappa \neq r_i$, $i = 1, 2, \cdots, d$. And referring back to the derivation of (22), it is readily seen that

$$R_c(t) = [\mathbf{q}_d^*(\kappa)]'\cdot\mathbf{B}_d'\cdot\operatorname{diag}\left[0, \frac{e^{r_1 t}}{(r_1 - \kappa)^2}, \cdots, \frac{e^{r_d t}}{(r_d - \kappa)^2}\right]\cdot\mathbf{B}_d\cdot\mathbf{q}_d^*(\kappa), \quad \kappa \neq r_i, \quad i = 1, \cdots, d, \tag{31}$$

where

$$[\mathbf{q}_d^*(\kappa)]' \equiv [\rho_0^*(\kappa)\cdot p_0^{\frac{1}{2}}, \cdots, \rho_d^*(\kappa)\cdot p_d^{\frac{1}{2}}].$$

The modifications needed when κ is equal to one of the characteristic roots are immediate. Let

$$\operatorname{diag}^{(j)}[a_0, a_1, \cdots, a_d]$$

be the diagonal matrix obtained by setting the jth diagonal element of $\operatorname{diag}[a_0, a_1, \cdots, a_d]$ equal to zero. Then if $\kappa = r_j$ we must have, with some as-yet-undetermined constant a and σ_c^2 the variance of $N_c(0)$,

$$R_c(t) = (\sigma_c^2 + a)e^{\kappa t} + [\mathbf{q}_d^*(\kappa)]'\cdot\mathbf{B}_d'\cdot\operatorname{diag}^{(j)}\left[0, \frac{e^{r_i t}}{(r_i - \kappa)^2}, i = 1, \cdots, d\right]\cdot\mathbf{B}_d\cdot\mathbf{q}_d^*(\kappa).$$

But $R_c(0) = \sigma_c^2$, so that

$$a = -[\mathbf{q}_d^*(\kappa)]'\cdot\mathbf{B}_d'\cdot\operatorname{diag}^{(j)}[0, (r_i - \kappa)^{-2}, i = 1, \cdots, d]\cdot\mathbf{B}_d\cdot\mathbf{q}_d^*(\kappa).$$

Hence,

$$R_c(t) = \sigma_c^2 e^{\kappa t} + [\mathbf{q}_d^*(\kappa)]'\cdot\mathbf{B}_d'\cdot\operatorname{diag}^{(j)}\left[0, \frac{e^{r_i t} - e^{\kappa t}}{(r_i - \kappa)^2}, i = 1, \cdots, d\right]\cdot\mathbf{B}_d\cdot\mathbf{q}_d^*(\kappa). \tag{32}$$

It should be noted that (32) is valid even if $\kappa \neq r_i$, $i = 1, \cdots, d$, and that it should be used in the computations [rather than (31)] whenever κ is "close" to one of the characteristic roots so as to avoid accuracy losses (see below). Since κ is arbitrary, one could always choose it so that it is not "close" to any of the characteristic roots. But, as shown next, it is often preferable to select it in such a way as to reduce the amount of computation, and this, in turn, may dictate

the use of (32). As we have seen,

$$R_c(t) = [\mathbf{q}_d^*(\kappa)]' \cdot \mathbf{B}_d' \cdot \text{diag}\left[0, \frac{e^{r_1 t}}{(r_1 - \kappa)^2}, \cdots, \frac{e^{r_d t}}{(r_d - \kappa)^2}\right] \cdot \mathbf{B}_d \cdot \mathbf{q}_d^*(\kappa).$$

But, as noted previously, $\mathbf{p}_d^{(\frac{1}{2})}$ is the characteristic vector associated with the vanishing root (r_0), so that

$$(\mathbf{p}_d^{(\frac{1}{2})})' \cdot \mathbf{B}_d' \cdot \text{diag}\left[0, \frac{e^{r_1 t}}{(r_1 - \kappa)^2}, \cdots, \frac{e^{r_d t}}{(r_d - \kappa)^2}\right] \cdot \mathbf{B}_d \cdot \mathbf{p}_d^{(\frac{1}{2})} = 0. \quad (33)$$

Consequently, (31) remains valid for all $[\mathbf{q}_d^*(\kappa)]'$ of the form

$$\{[\rho_0^*(\kappa) + \gamma] \cdot p_0^{\frac{1}{2}}, \cdots, [\rho_d^*(\kappa) + \gamma] \cdot p_d^{\frac{1}{2}}\},$$

where γ is an arbitrary constant. [The same remark, of course, also applies to (32).]

We are therefore always at liberty to add the same constant to all the $\rho_i^*(\kappa)$'s. Under some circumstances, this degree of freedom, together with the one provided by the introduction of κ, can be used to reduce the dimension of $\mathbf{B}'$ and $\mathbf{B}$: entire rows in $\mathbf{B}'$ and the corresponding columns in $\mathbf{B}$ can be set equal to zero without affecting the computation either of (31) or (32), or of the variance of the switch-count load. It is relevant to note here that this reduction would be largely illusory were it not for the fact that the normalized components of any of the characteristic vectors can be obtained without having to compute other components of that vector (see below).

According to the result of Section 3.1, the covariance can always be cast in a form that involves only the first c components of the characteristic vectors. But when the input and departure rates for $0 \leqq n < c$ are linear in n, the covariance can also be expressed in terms of the last $d - c + 1$ components of these vectors. Indeed, the rates are then of the form

$$\lambda_n = \lambda n + \lambda',$$
$$\mu_n = \mu n + \mu', \qquad n = 0, 1, \cdots, c - 1,$$

so that

$$\lambda_n - \mu_n = (\lambda - \mu)n + (\lambda' - \mu').$$

Hence, with $\kappa = \lambda - \mu$ and $\gamma = \mu' - \lambda'$, (27) yields

$$\rho_n^*(\lambda - \mu) = \begin{cases} 0 & \text{if } n = 0, 1, \cdots, c - 1, \\ (\mu - \lambda)c - \mu_c + \mu' - \lambda' & \text{if } n = c, \\ (\mu - \lambda)c + \mu' - \lambda' & \text{if } n = c + 1, \cdots, d. \end{cases}$$

For the random (Poisson) and the quasi-random inputs, the ρ's take the

following simple form whenever the service time is exponential with mean 1.

(*i*) Random input ($\lambda_n = \alpha$, $n = 0, 1, \cdots$):

$$\rho_n^*(-1) = \begin{cases} 0 & \text{if } n = 0, 1, \cdots, c-1, \\ -\alpha & \text{if } n = c, \\ c - \alpha & \text{if } n = c+1, \cdots, d. \end{cases}$$

(*ii*) Quasi-random input [N sources, $\lambda_n = (N - n)\lambda$, $n = 0, 1, \cdots, N$]:

$$\rho_n^*[-(1+\lambda)] = \begin{cases} 0 & \text{if } n = 0, 1, \cdots, c-1, \\ (c - N)\lambda & \text{if } n = c, \\ c + (c - N)\lambda & \text{if } n = c+1, \cdots, d. \end{cases}$$

From the preceding developments, we see that the ρ's can be chosen in such a way that the number of components of the characteristic vectors needed to express R_c is the smaller of the two integers c and $d - c + 1$. In particular, in the case of loss systems, only the $(c+1)$st component of each vector is needed.

The parameters κ and γ can also be chosen so that only the first $c + 1$ components of the characteristic vectors actually enter in the expression of R_c. This will be the case if we set $\kappa = \mu_{c-1} - \lambda_{c-1}$ and $\gamma = c(\mu_{c-1} - \lambda_{c-1})$.

In Ref. 3, the derivation of the covariance function for loss systems [$d = c$, $N_c(t) \equiv N(t)$] with Poisson input and exponential service time makes use of the differential equations

$$\frac{d}{dt} E\{N(t) \mid N(0) = m\} = -E\{N(t) \mid N(0) = m\} + \alpha[1 - p_{mc}(t)], \qquad m = 0, 1, \cdots, c.$$

These equations appear here as that particular instance of (26) for which $\kappa = -1$, $\lambda_n = \alpha$, $n = 0, 1, \cdots, c-1$, and $\mu_n = n$, $n = 1, \cdots, c$. Note also that now $\sum_{n=0}^{c-1} p_{mn}(t) = 1 - p_{mc}(t)$. But we stress that, in Ref. 3, the determination of the covariance relies on known recurrence relations between the so-called "sigma" functions (Ref. 3, pp. 129 and 143 ff.); the more general problem considered in the present paper is not as readily amenable to such a treatment because of the greater complexity of the expressions that would now have to be used instead of the sigma functions. As we have seen, however, relatively simple formulas for R_c can be obtained without extensive algebraic developments as long as the underlying process is reversible.

IV. VARIANCE OF THE SWITCH-COUNT LOAD

The variance of the switch-count load is now readily obtained. Depending on which expression we select for R_c, we have either

(*i*) Var $L_n(T)$

$$= n^{-2} \sum_{k=-n}^{n} (n - |k|) R_c(k\tau)$$

$$= n^{-2} \cdot \mathbf{q}_d' \cdot \mathbf{B}_d' \cdot \text{diag} \left[0, \sum_{k=-n}^{n} (n - |k|) e^{r_i |k| t}, i = 1, \cdots, d \right] \cdot \mathbf{B}_d \cdot \mathbf{q}_d^*(\kappa), \quad (34)$$

with

$$\mathbf{q} = (\rho_0 p_0^{\frac{1}{2}}, \cdots, \rho_{c-1} p_{c-1}^{\frac{1}{2}}, 0, \cdots, 0), \text{ or}$$

(*ii*) Var $L_n(T) = n^{-2} [\mathbf{q}_d^*(\kappa)]' \cdot \mathbf{B}_d'$

$$\cdot \text{diag} \left[0, \sum_{k=-n}^{n} (n - |k|) \frac{e^{r_i |k| t}}{(r_i - \kappa)^2}, i = 1, \cdots, d \right] \cdot \mathbf{B}_d \cdot \mathbf{q}_d^*(\kappa), \quad (35)$$

with $\kappa \neq r_i$, $i = 1, \cdots, d$, or

(*iii*) Var $L_n(T) = n^{-2} \{ \sigma_c^2 - [\mathbf{q}_d^*(\kappa)]' \cdot \mathbf{B}_d$

$$\cdot \text{diag}^{(j)} [0, (r_i - \kappa)^{-2}, i = 1, \cdots, d] \cdot \mathbf{B}_d \cdot \mathbf{q}_d \}$$

$$\cdot \sum_{k=-n}^{n} (n - |k|) e^{\kappa |k| \tau} + n^{-2} [\mathbf{q}_d^*(\kappa)]' \cdot \mathbf{B}_d'$$

$$\cdot \text{diag}^{(j)} \left[0, \sum_{k=-n}^{n} (n - |k|) \frac{e^{r_i |k| \tau}}{(r_i - \kappa)^2}, i = 1, \cdots, d \right] \cdot \mathbf{B}_d \cdot \mathbf{q}_d^*(\kappa), \quad (36)$$

where $\kappa \neq r_i$ for $i \neq j$.

We now make use of the following identity (Ref. 3, p. 137):

$$\sum_{n=-k}^{k} (n - |k|) e^{-2|k|u} = n \cdot \coth u - \frac{1 - e^{-2nu}}{2} \cdot \text{csch}^2 u.$$

By means of this relation, (34) to (36) can also be written as

$$\text{Var } L_n(T) = n^{-1} \cdot \mathbf{q}_d' \cdot \mathbf{B}_d' \cdot \text{diag} \left[0, \coth \left(\frac{-\tau r_i}{2} \right) - \frac{1 - e^{n\tau r_i}}{2n} \cdot \text{csch}^2 \left(\frac{-\tau r_i}{2} \right), i = 1, \cdots, d \right] \cdot \mathbf{B}_d \mathbf{q}_d; \quad (34a)$$

$$\text{Var } L_n(T) = n^{-1} \cdot [\mathbf{q}_d^*(\kappa)]' \cdot \mathbf{B}_d' \cdot \text{diag}\left[0, \frac{1}{(r_i - \kappa)^2}\left\{\coth\left(\frac{-\tau r_i}{2}\right) - \frac{1 - e^{n\tau r_i}}{2n} \cdot \text{csch}^2\left(\frac{-\tau r_i}{2}\right)\right\}, i = 1, \cdots, d\right] \cdot \mathbf{B}_d \cdot \mathbf{q}_d^*(\kappa), \quad \kappa \neq r_i, \quad i = 1, \cdots, d; \quad (35a)$$

and

$$\text{Var } L_n(T) = n^{-1}\{\sigma_c^2 - [\mathbf{q}_d^*(\kappa)]' \cdot \mathbf{B}_d' \cdot \text{diag}^{(j)}[0, (r_i - \kappa)^{-2}, i = 1, \cdots, d] \cdot \mathbf{B}_d \cdot \mathbf{q}_d^*(\kappa)\} \cdot \left\{\coth\left(\frac{-\tau\kappa}{2}\right) - \frac{1 - e^{n\tau\kappa}}{2n} \cdot \text{csch}^2\left(\frac{-\tau\kappa}{2}\right)\right\} + n^{-1} \cdot [\mathbf{q}_d^*(\kappa)]' \cdot \mathbf{B}_d' \cdot \text{diag}^{(j)}\left[0, \frac{1}{(r_i - \kappa)^2}\left\{\coth\left(\frac{-\tau r_i}{2}\right) - \frac{1 - e^{n\tau r_i}}{2n} \cdot \text{csch}^2\left(\frac{-\tau r_i}{2}\right)\right\}, i = 1, \cdots, d\right] \cdot \mathbf{B}_d \cdot \mathbf{q}_d^*(\kappa), \quad (36a)$$

where $\kappa \neq r_i$ for $i \neq j$.

Let $\text{Var } L_\infty(T) \equiv \lim_{n\to\infty} \text{Var } L_n(T)$ be the variance of the load measurement obtained by continuous observation of the number of busy servers. If we replace τ by T/n in (34a) to (36a) and then let n tend to infinity while keeping T fixed, we obtain the following formulas:

$$\text{Var } L_\infty(T) = -\frac{2}{T} \cdot \mathbf{q}_d' \cdot \mathbf{B}_d' \cdot \text{diag}\left[0, \frac{1}{r_i}\left(1 + \frac{1 - e^{r_i T}}{T r_i}\right), i = 1, \cdots, d\right] \cdot \mathbf{B}_d \cdot \mathbf{q}_d, \quad (34b)$$

$$\text{Var } L_\infty(T) = -\frac{2}{T} [\mathbf{q}_d^*(\kappa)]' \cdot \mathbf{B}_d' \cdot \text{diag}\left[0, \frac{1}{r_i(r_i - \kappa)^2}\left(1 + \frac{1 - e^{r_i T}}{T r_i}\right), i = 1, \cdots, d\right] \cdot \mathbf{B}_d \cdot \mathbf{q}_d^*(\kappa), \quad \kappa \neq r_i, \quad i = 1, \cdots, d, \quad (35b)$$

and

$$\text{Var } L_\infty(T) = \frac{2}{T}\{\sigma_c^2 - [\mathbf{q}_d^*(\kappa)]' \cdot \mathbf{B}_d' \cdot \text{diag}^{(j)}[0, (r_i - \kappa)^{-2}, i = 1, \cdots, d] \cdot \mathbf{B}_d \cdot \mathbf{q}_d^*(\kappa)\} \frac{1}{\kappa}\left(1 + \frac{1 - e^{\kappa T}}{T\kappa}\right) - \frac{2}{T}[\mathbf{q}_d^*(\kappa)]' \cdot \mathbf{B}_d' \cdot \text{diag}^{(j)}\left[0, \frac{1}{r_i(r_i - \kappa)^2}\left(1 + \frac{1 - e^{r_i T}}{T r_i}\right), i = 1, \cdots, d\right] \cdot \mathbf{B}_d \cdot \mathbf{q}_d^*(\kappa), \quad \kappa \neq r_i \quad \text{for} \quad i \neq j. \quad (36b)$$

We note that the formula for the variance of sums of dependent random variables makes it possible to compute the covariance between load measurements performed over distinct time intervals. Indeed, consider for instance a sequence of load measurements over the intervals $(0, T]$, $(T, 2T]$, $(2T, 3T]$, $\cdots$. Let $L_n^{(i)}(T)$ be the switch-count load over the ith interval ($i = 0, 1, \cdots$), $S_n(t) = n^2 L_n(t)$, and $\Gamma_n^{(i)} \equiv \text{Cov}\,[L_n^{(0)}(T), L_n^{(i)}(T)]$. Then

$$\text{Var}\, S_{n(k+1)}[(k+1)T] = (k+1)\,\text{Var}\, S_n(T) + 2\sum_{i=1}^{k}(k+1-i)n^2\Gamma_n^{(i)}(T)$$

and

$$\Gamma_n^{(k)}(T) = \frac{(k+1)^2}{2}\,\text{Var}\, L_{n(k+1)}[(k+1)T] - \frac{k+1}{2}\,\text{Var}\, L_n(T) - \sum_{i=1}^{k-1}\Gamma_n^{(i)}(T).$$

The preceding formulas may be used to determine the $\Gamma_n^{(k)}(T)$ recurrently. But the results of such computations shall be exact only if, for some choice of the time origin, all the scanning instants are multiples of τ.

We conclude this section with the remark that the variance formulas (34), (34a), and (34b) are valid for arbitrary reversible Markov processes with finite state spaces.

V. NUMERICAL CONSIDERATIONS

The exact variance formulas of the preceding section are very well suited to electronic computation and are easily programmed since, apart from straightforward evaluations of hyperbolic functions and simple products of matrices and vectors, they only involve the determination of characteristic values and vectors for which powerful subroutines are readily available. The fact that $\mathbf{S}_d$ is symmetric and tridiagonal (or reducible to tridiagonal form by an orthogonal similarity transformation) allows us to use the subprogram TQL2, which is particularly efficient under the present circumstances (Ref. 11, pp. 227–240). Without going into details, we mention here only that this subprogram is based on the so-called QR-algorithm and relies on the construction of a sequence of symmetric tridiagonal matrices, $\mathbf{S}_d^{(n)}$, $n = 1, 2, \cdots$, unitarily similar to $\mathbf{S}_d$, which converges to diag $[0, r_1, r_2, \cdots, r_d]$. At the nth iteration $\mathbf{S}_d^{(n)}$ is expressed as a product of an

orthogonal matrix $\mathbf{Q}_d^{(n)}$ and a lower-triangular matrix $\mathbf{L}_d^{(n)}$:

$$\mathbf{S}_d^{(n)} = \mathbf{Q}_d^{(n)} \cdot \mathbf{L}_d^{(n)}.$$

In Ref. 11, this decomposition is carried out by the Givens' triangularization in which the $\mathbf{Q}_d^{(n)}$'s are expressed as products of simple plane rotations (Ref. 7, pp. 239–240).

The $(n + 1)$st iterate of $\mathbf{S}_d$ is then given by

$$\mathbf{S}_d^{(n+1)} = \mathbf{L}_d^{(n)} \cdot \mathbf{Q}_d^{(n)},$$

whose unitary similarity to $\mathbf{S}_d^{(n)}$ (and hence to $\mathbf{S}_d$) follows from the relation

$$\mathbf{L}_d^{(n)} \cdot \mathbf{Q}_d^{(n)} = (\mathbf{Q}_d^{(n)})' \cdot \mathbf{S}_d^{(n)} \cdot \mathbf{Q}_d^{(n)}.$$

This method avoids the numerical difficulties frequently associated with the computation of the zeros of the characteristic polynomial.

As shown in Ref. 11, p. 228,

$$\mathbf{Q}_d^{(n)} = \mathbf{R}_d^{(n)} \cdot \cdots \cdot \mathbf{R}_d^{(1)}, \tag{37}$$

where $\mathbf{R}_d^{(i)}$ is of the form:

$$\begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & C_i & S_i & & & \\ & & -S_i & C_i & & & \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{bmatrix} \leftarrow \text{row } i. \tag{38}$$

According to Ref. 11, p. 231, the matrix $\mathbf{B}_d'$ (whose columns are the characteristic vectors of $\mathbf{S}_d$) is given "almost to working accuracy" by

$$\mathbf{Q}_d^{(n)} \cdot \cdots \cdot \mathbf{Q}_d^{(1)}, \tag{39}$$

where n ($\leqq 30$) is the number of iterations needed for the (numerical) symmetrization of $\mathbf{S}_d$. Taking (37) to (39) into account, it is then readily seen that the elements of $\mathbf{B}_d'$ can be determined row by row which, as we have remarked earlier, is a desirable feature in the present context.

Computations have been carried out for systems having as many as 400 devices (and hence transition-rate matrices of order 401) to determine the numerical accuracy of the approach described in the preceding sections. Checks were performed by comparing the value of Var $L_1(T)$ obtained by means of (35a) or (36a) with the corresponding σ_c^2, which can be calculated directly and independently from the equilibrium state probabilities. These two quantities, which are theoretically equal,

turned out in all cases to agree to at least 10 decimal places with the greatest difference occurring when d was largest. Hence, our procedure indeed yields very accurate results for the type of systems that are likely to occur in practice. But when d is large, the storage requirements and the amount of computations become critical. It is therefore always important to select κ and γ in such a way as to minimize the number of $\mathbf{B}'$ rows that actually enter into the computations. (It follows from earlier remarks that this number, for proper choice of κ and γ, never exceeds the integral part of $(d + 1)/2$.) Further reduction can also be achieved by excluding the states whose probabilities of occurrences are so small that neglecting them will not materially affect the final results. In this connection, we make the following remarks.

The variance of the switch-count load is perturbed by at most

$$[\rho_j^*(\kappa)]^2 \cdot p_j \cdot \sigma_c^2$$

if p_j is set equal to zero in the particular formula used to evaluate Var $L_n(T)$. Hence, since

$$\text{Var } L_n(T) \geqq \sigma_c^2/n,$$

we always have the following upper bound for the relative error, ϵ_j, induced by setting p_j equal to zero:

$$\epsilon_j \leqq [\rho_j^*(\kappa)]^2 \cdot p_j \cdot n, \qquad j = 0, 1, \cdots, d.$$

For a given relative accuracy of Var $L_n(T)$, these inequalities make it possible to determine ahead of time whether some components of the characteristic vectors can be "safely" eliminated from the computations. In large systems, the gains achieved by such a reduction may be quite substantial, as either low occupation states [$N(t)$ small] and/or high occupation states [$N(t)$ large] have then frequently very small probabilities of occurrences.

Computations could be arranged to determine only those characteristic roots that are required to reach a given degree of accuracy [plus those needed to compute $\mathbf{B}_d \cdot \mathbf{q}_d^*(\kappa)$]. This is rather readily achieved in loss systems with Poisson input and exponential service times since, in this case, the coefficients b_i^2 of

$$\coth\left(\frac{-\tau r_i}{2}\right) - \frac{1 - e^{n\tau r_i}}{2n} \cdot \text{csch}^2\left(\frac{-\tau r_i}{2}\right)$$

in the variance formulas of Section 3.1 are then monotonically decreasing as $|r_i|$ increases:

$$b_i^2 < b_j^2 \quad \text{if} \quad |r_i| > |r_j|, \qquad i, j = 1, \cdots, d.$$

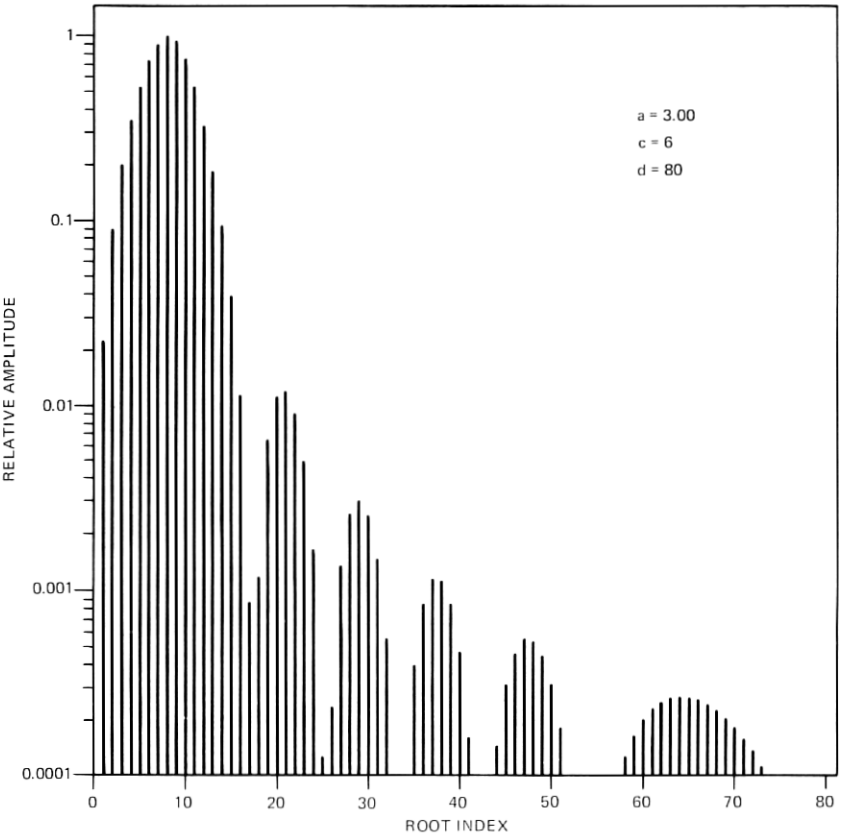

Fig. 1—Spectral measure of the carried load process.

But, in general, one would encounter an additional difficulty, namely, that the b_i^2's do not have the monotonicity property alluded to above and may actually fluctuate widely. This is illustrated in Fig. 1, where the roots are assumed to be indexed in order of increasing magnitude and the ordinates are the corresponding b_i^2's, normalized in such a way that $\max_i b_i^2 = 1$.

The computations should be based on (35a)–(36a) or on (35b)–(36b) in the case of continuous measurements—as these formulas provide us with all the flexibility needed to cut down both storage space and computation time. When choosing between (35a) and (36a) or between (35b) and (36b), one should keep in mind that, for κ close to r_i, the difference $r_i - \kappa$ may not be determinable with enough precision to allow accurate computation of Var $L_n(T)$. This is shown in Table I where $\kappa = -1$ and r_1 is the root of smallest positive absolute value.

Table I — Loss system, 80 servers, Poisson input, exponential service

Offered Load in Erlangs	$1 + r_1$	σ_c^2	Var $L_1(T)$	
			Formula (36a)	Formula (35a)
10	1.09×10^{-13}	10.000000	10.000000	0.034614
20	-9.90×10^{-13}	20.000000	20.000000	0.013173
30	-1.14×10^{-11}	30.000000	30.000000	20.891365
40	-3.38×10^{-6}	39.999986	39.999986	39.999926

(Note that the last two columns of this table should be equal and that errors of the same magnitude would arise if one were to use (18) of Ref. 3.) In all our computations, we have made use of (35a) and (36a) whenever $|\kappa - r_i| < 10^{-4}$ for some i. This bound for $|\kappa - r_i|$ is both large enough to ward off appreciable accuracy losses and small enough, under prevailing conditions, to be satisfied by only one root.

VI. REMARKS ON INFINITE SYSTEMS

It is known that infinite systems can be regarded as limits of finite ones,[12] and it is therefore of practical interest to have information concerning the spacing of the characteristic values as the dimension, d, of these approximating systems becomes large. Indeed, as d increases, computational difficulties may arise because of a lack of separation between these roots. Such problems would certainly come up sooner or later if the spectrum of $\mathbf{A} \equiv \lim_{d\to\infty}\mathbf{A}_d$ happens to be dense over some interval as, for instance, in the case of a single-server queue with Poisson input, exponential service time, and unlimited waiting room (Ref. 12, pp. 365–366). Infinite systems with well-separated roots do, of course, also occur. As an example of this type, we mention the systems with an infinite number of servers, Poisson input, and exponential service which often provide useful idealizations. (In this case, as is well known, the nonvanishing characteristic roots are the negative integers, $-1, -2, -3, \cdots$.) Other examples of systems with discrete spectra are given in Ref. 12, where sufficient conditions for this to occur are discussed in some details; but in all these instances the λ_n and μ_n both increase as n^ν for some $\nu > 0$. This condition is unlikely to be satisfied in queuing systems; generally, in this particular area, the arrival and the departure rates remain bounded:

$$0 \leqq \lambda_n \leqq \Lambda < \infty, \qquad 0 \leqq \mu_n \leqq M < \infty, \qquad n = 0, 1, \cdots. \tag{40}$$

As briefly described below, these inequalities imply the existence of definite bounds for the spectrum of $\mathbf{A}$.

Consider an infinite system, and let $\mathbf{A}$ be its (infinite) transition-rate matrix. Let $\mathbf{A}_d$ be the matrix obtained by retaining only the elements belonging to the first $(d+1)$ rows and columns of $\mathbf{A}$ and then setting λ_d equal to 0. Let $r_{d0}(=0) > r_{d1} > \cdots$ be the characteristic roots of $\mathbf{A}_d$. Then, under conditions (40) it can be shown that, for any $k \geqq 0$:

(i) $\quad |r_{dk}| < \Lambda + M \quad$ for d sufficiently large,

(ii) $\quad |r_{d,d-k}| < 2(\Lambda + M) \quad$ for $d \geqq k$.

Either of these two inequalities implies that the characteristic roots do not remain separated as $d \to \infty$ whenever (40) is satisfied. Under the more stringent requirements that (40) holds and that

$$\lim_{n\to\infty} \lambda_n = \Lambda, \qquad \lim_{n\to\infty} \mu_n = M,$$

more precise statements can be made, namely, that, for all k's and d's,

$$|r_{dk}| < (\sqrt{\Lambda} + \sqrt{M})^2$$

and that the spectrum of $\mathbf{A}$ always comprises a closed interval, viz.,

$$\Omega = [-(\sqrt{\Lambda} + \sqrt{M})^2, -(\sqrt{\Lambda} - \sqrt{M})^2].$$

(In addition to Ω, the spectrum of $\mathbf{A}$ may also include a finite number of roots in $[-(\sqrt{\Lambda} - \sqrt{M})^2, 0]$.) But it turns out (as will be shown elsewhere) that, as d increases, the characteristic roots of $\mathbf{A}_d$ fill Ω rather "evenly"; furthermore, for practical accuracy levels, large values of d are needed only when the length of Ω tends to be relatively large (a circumstance corroborated by extensive computations). Hence, within the present framework, it appears that root-spacing is not likely to be critical except in the improbable event that extreme precision is required.

VII. ACKNOWLEDGMENT

I wish to thank Peter Businger for very helpful suggestions.

REFERENCES

1. W. S. Hayward, Jr., "The Reliability of Telephone Traffic Load Measurements by Switch Counts," B.S.T.J., *31*, No. 2 (March 1952), pp. 357–377.
2. C. Palm, "Accuracy of Measurements in Determining Traffic Volumes by the Scanning Method," Tekniska Medelanden fran Kungl. Telegrafstyrelen, No. 7–9, 1941.
3. V. E. Beneš, "The Covariance Function of a Simple Trunk Group, with Applications to Traffic Measurement," B.S.T.J., *40*, No. 1 (January 1961), pp. 117–148.
4. S. R. Neal and A. Kuczura, "A Theory of Traffic-Measurement Errors for Loss Systems with Renewal Input," B.S.T.J., *52*, No. 6 (July–August 1973), pp. 967–990.
5. D. V. Widder, *The Laplace Transform*, Princeton: Princeton University Press, 1946.

6. J. L. Doob, *Stochastic Processes*, New York: John Wiley, 1953.
7. J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford: Clarendon Press, 1965.
8. E. Reich, "Waiting Times When Queues are in Tandem," Ann. Math. Stat., *28*, No. 3 (1957), pp. 768–773.
9. S. Karlin and J. McGregor, "The Classification of Birth and Death Processes," Trans. Amer. Math. Soc., *86*, 1957, pp. 366–400.
10. K. L. Chung, *Markov Chains*, second edition, New York: Springer-Verlag, 1967.
11. H. Bowdler, R. S. Martin, C. Reinsch, and J. H. Wilkinson, "The QR and QL Algorithms for Symmetric Matrices," Contribution II/3, pp. 227–240, in *Handbook for Automatic Computation, Vol. II, Linear Algebra*, edited by J. H. Wilkinson and C. Reinsch, New York, Heidelberg, Berlin: Springer-Verlag, 1971.
12. W. Ledermann and G. E. H. Reuter, "Spectral Theory for the Differential Equations of Simple Birth and Death Processes," Phil. Trans. Roy. Soc. of London, Series A, *246*, 1954, pp. 321–369.