A Class of FIFO Queues Arising in Computer Systems
Author(s): Edward G. Coffman Jr. and Micha Hofri
Source: *Operations Research*, Vol. 26, No. 5, Operations Research/Computer Science Interface (Sep. - Oct., 1978), pp. 864-880
Published by: INFORMS
Stable URL: http://www.jstor.org/stable/170081
Accessed: 11/04/2010 06:19

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/page/info/about/policies/terms.jsp. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/action/showPublisher?publisherCode=informs.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

INFORMS is collaborating with JSTOR to digitize, preserve and extend access to *Operations Research.*

http://www.jstor.org

OPERATIONS RESEARCH
Vol. 26, No. 5, September-October 1978

0030-364X/78/2605-0864 $01.25
© 1978 Operations Research Society of America

A Class of FIFO Queues Arising in Computer Systems

EDWARD G. COFFMAN, JR.
University of California, Santa Barbara, California
MICHA HOFRI
Technion—Israel Institute of Technology, Haifa, Israel

(Received October 1976; accepted January 1978)

We model secondary memory devices as single-server queuing systems. The non-random access to data within these devices is explicitly accounted for as "set-up" times. Requests are typed by the location of the desired record. No distinction is made between "read" and "write" requests. Each request is assumed to be satisfiable from one location on the device (e.g., a single directory search may result in a number of distinct requests). Requests arrive according to a homogeneous Poisson process. The types of successive requests form a first-order Markov chain, which is an approximation of reality. Alternative computational procedures and closed expressions are given for queue length, waiting times, and device utilization. We present some specializations to disks and drums. Only FIFO service is considered.

IMPORTANT congestion points in general-purpose computer systems frequently occur through interactions with secondary storage devices. In such cases the efficiency with which information is exchanged between these devices and primary (e.g., core) storage determines the system's maximum throughput or work rate. Secondary storage units such as magnetic drums, disks, bubble memories, and tapes (whether singly or within libraries) have the characteristic feature that the total service time of a read or write request depends on the location addressed by the request previously served. Of course, it is precisely this property that specifies the manner in which these devices fail to be random access, as are primary storage devices.

Our purpose is to present and analyze a mathematical model that will explicitly take into account the above characteristic of non-random access devices. Since the difficulty arises mainly from the unpredictable arrivals of requests, it is natural that a stochastic model is required for a realistic presentation of the salient features of these systems.

A specific goal will be to provide a FIFO (first-in-first-out) service queuing analysis of secondary storage devices sufficiently general to embrace the detailed structure of a large majority of existing systems.

The parameters of the mathematical model will include a stationary, discrete probability distribution describing the patterns by which requests address information on secondary storage devices (successive addresses are allowed as well to form a first-order Markov chain). Such patterns normally influence system performance, and they are determined by the mechanism that allocates specific storage locations to records (units of information). Thus, in the calculation of conventional performance measures we shall also briefly consider the essentially combinatorial problem of determining the influence of different record allocations.

The first two sections present a general model and its analysis. The remainder of the paper specializes the results to certain common secondary storage devices and discusses alternative computational methods.

1. THE MATHEMATICAL MODEL

We will model the devices discussed above as a single-server facility as

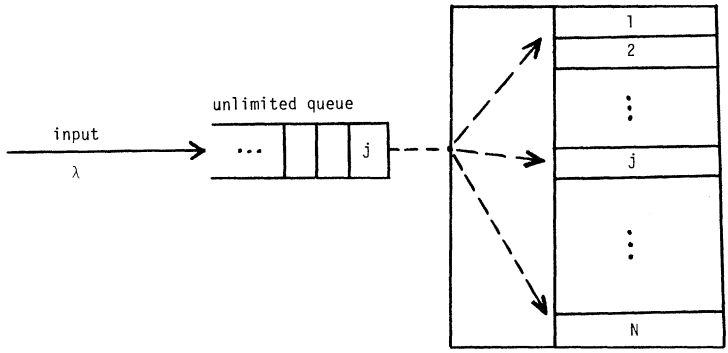

Figure 1. Secondary memory as a service facility with a waiting queue.

illustrated in Figure 1. Incoming requests are immediately inducted into service when the facility is idle. Arrivals at a busy facility wait for service, and there is no limit to the number of such requests that may wait at any given time. All service periods contain an initial period of set-up delay, possibly of zero length. The selection from the queue for service, at the termination of a service period, is done without prior knowledge of the requested service times. During all of our analysis we consider selection procedures that provide service in the order of arrival (FIFO), but some of the results admit more general regimes.

Requests are of N types, simply called type 1 through N. The probability that an arriving request is of type j, given that the preceding one was of type i, is p_{ij} and is otherwise independent of the state of the system and its history. These "transition probabilities" form a matrix P with an invariant probability vector we denote $\tilde{p}$. (We are interested only in situations where P is irreducible and all its states are recurrent.) The

matrix with all its rows equal to $\tilde{p}$ will be denoted by $\tilde{P}$. A request of type j, which is immediately preceded by a request of type i, requests service with duration S_{ij} drawn from a distribution $F_{ij}(\cdot)$, independently of the other descriptors of the state of the system. This service period is generally the sum of two components, $S_{ij}=T_{ij}+K_j$, where T_{ij}, the set-up time, is the time it takes the service facility to switch over from a state of having finished the service of a type i request to the beginning of service for a type j request. The quantity K_j depends on the request type, does not depend on the state of the system, and usually represents the actual transmission time of the information. In most of the applications toward which this paper is directed, the variables S_{ij} are in fact constants.

In some situations we find it expedient to distinguish the service rendered to a request that starts a busy period (i.e., it finds upon arrival an idle system). Invariably, it is the set-up time T_{ij} that is affected, and its value under these circumstances will be denoted T^0_{ij}. Associated with T^0_{ij} is a service duration S^0_{ij}, but the value of K_j is not changed. The arrival process of requests is assumed Poisson, with rate λ, homogeneous in time and independent of the state of the system.

We shall be interested primarily in steady-state behavior. We observe the system at the epochs of departure of requests. Since arrivals and departures happen singly, the distribution of the states of the system at these epochs is the same as at the arrival epochs, and also equal to the so-called "long-term" distribution. We let X_n denote the number of requests in the system immediately following the departure of the nth request, the one in service included. η_n denotes the waiting time of the nth request, which terminates at the beginning of the nth set-up time. We let S be the random variable denoting general service time, $F(\cdot)$ the corresponding distribution, and $\mathscr{L}(\cdot)$ its LST.

2. ANALYSIS OF THE MODEL

The major difference between the model we investigate here and standard queuing models is the dependence between successive services. Depending on the type of device and its operating procedures, this relationship may even extend across an intervening idle period. Special cases of our model can be treated as applications of Skinner's model [12] (with a loss of structure severe enough to preclude its use for most of the devices for which our model is intended). For an example, see Fuller and Baskett [4] for approximate analyses of FIFO paging drums. A queuing model with similar structure—the main difference being that no distinction is made between a general service and one that starts a busy period—was treated in detail by Neuts [10].

We begin with an analysis that is independent of the order of arrivals. Then we proceed to evaluate the waiting times for a FIFO queue.

System capacity. As one usually finds in queuing systems, the input rates that the facility can sustain must be less than $\lambda_{\max}=1/E(S)$, where $E(S)=\Sigma_i\Sigma_j\tilde{p}_i p_{ij}E(S_{ij})$. This statement will not be proved explicitly here. We note the occurrence of the corresponding discontinuity point in numerical calculations.

Queue length. We observe that $(X_n, J_n; n=1, 2, \ldots)$, where J_n is the type of the nth departing request, is an aperiodic, irreducible and, for low enough input rates, recurrent Markov chain (MC). We proceed first to evaluate the probability generating function (pgf) of the steady-state distribution of the number in system. This will turn out to entail most of the complexity of the analysis that we require.

We define for $1\leqq i\leqq N$, $x\geqq 0$, $p_i(x)=\lim_{n\to\infty}P(X_n=x|J_n=i)$ where, as usual, the vertical bar is to be read as "given that . . .," and $G_i(z)=\sum_{x=0}^{\infty}p_i(x)z^x$. The dynamics of our MC are embodied in the matrix P and the relation $X_{n+1}=X_n-U_n+Y_{n+1}$, where U_n is 0 when $X_n=0$ and is 1 otherwise, and where Y_{n+1} is the number of arrivals during the service of the $(n+1)$st request. We proceed in a standard way to obtain directly

$$\begin{aligned}P(X_{n+1}=x|J_{n+1}=j)P(J_{n+1}=j)=\textstyle\sum_{i=1}^{N}P(J_n=i)\{P(X_n=0|J_n=i)\cdot\\ P(Y_{n+1}=x, J_{n+1}=j|X_n=0, J_n=i)+\textstyle\sum_{r=1}^{\infty}P(X_n=r|J_n=i)\cdot\\ P(Y_{n+1}=x-r+1, J_{n+1}=j|X_n=r, J_n=i)\}; \quad x\geqq 0,\ 1\leqq j\leqq N.\end{aligned} \tag{1}$$

The distribution of Y_{n+1} is now derived. It obviously depends on the duration of service of the $(n+1)$st request. As mentioned above, we distinguish between a departure followed by an idle period (with a subsequent service distributed according to $F_{ij}^0(\cdot)$) and a departure for which the next service commences immediately (and is distributed according to $F_{ij}(\cdot)$); the set-up duration may be different in the two cases.

Hence

$$P(Y_{n+1}=x|X_n=0, J_n=i, J_{n+1}=j)=\int_{s=0}^{\infty}(\exp(-\lambda s)(\lambda s)^x/x!)dF_{ij}^0(s) \tag{2}$$

and

$$P(Y_{n+1}=x|X_n>0, J_n=i, J_{n+1}=j)=\int_{s=0}^{\infty}(\exp(-\lambda s)(\lambda s)^x/x!)dF_{ij}(s). \tag{3}$$

We substitute (2) and (3) properly deconditoned from J_{n+1} in (1), multiply by z^x and sum over all values of x. Since the MC is recurrent, we may drop the subscripts n and $n+1$ to obtain the limiting equation

$$G_j(z)=\textstyle\sum_{i=1}^{N}\tilde{p}_i p_{ij}\{\pi_i\sum_{x=0}^{\infty}\int_{s=0}^{\infty}(\exp(-\lambda s)(\lambda sz)^x/x!)dF_{ij}^0(s)$$

$$+\sum_{r=1}^{\infty} z^r P(X=r|J=i)\sum_{x=t-1}^{\infty} z^{-1}\int_{s=0}^{\infty} \cdot (exp)(-\lambda s)(\lambda s z)^{x-r+1}/(x-r+1)!)dF_{ij}(s)\}/\tilde{p}_j$$
$$=\sum_{i=1}^{N}\{\pi_i L_{ij}^0(a)+z^{-1}(G_i(z)-\pi_i)L_{ij}(a)\} \qquad 1\leqq j\leqq N, \tag{4}$$

where $\pi_i=P(X=0|J=i)$, $a=\lambda(1-z)$, $\mathscr{L}_{ij}$ (respectively $\mathscr{L}_{ij}^0$) is the Laplace-Stieltjes transform of F_{ij} (respectively F_{ij}^0) and $L_{ij}(L_{ij}^0)$ is given by $\tilde{p}_i p_{ij}\mathscr{L}_{ij}/\tilde{p}_j(\tilde{p}_i p_{ij}\mathscr{L}_{ij}^0/\tilde{p}_j)$. The various changes of order of summation are allowed since all the sums are trivially absolutely convergent.

The N equations can be written in a more convenient and compact matrix form:

$$A(z)\vec{G}(z)=B(z)\vec{\pi}, \tag{5}$$

where $\vec{\pi}$ and $\vec{G}(z)$ are the obvious vectors and

$$\begin{aligned} A_{ij}(z)&=z\delta_{ij}-L_{ji}(a),\\ B_{ij}(z)&=zL_{ji}^0(a)-L_{ji}(a), \end{aligned} \qquad 1\leqq i,j\leqq N \tag{6}$$

where δ_{ij} is 1 if $i=j$ and is 0 otherwise. Equation (5) has the formal solution

$$\vec{G}(z)=A^{-1}(z)B(z)\vec{\pi}. \tag{7}$$

The unknown boundary probabilities π_i now have to be deduced. First we have

$$\vec{G}(1)=\vec{1}. \tag{8}$$

Second, letting $C(z)$ be the adjoint matrix of $A(z)$, and hence $C(z)A(z)=|A(z)|I$, then we must have

$$C(\zeta)B(\zeta)\vec{\pi}=\vec{0} \tag{9}$$

at all points ζ, $|\zeta|\leqq 1$, which are solutions of

$$|A(z)|=0. \tag{10}$$

Each of the equations in the system (5) is homogeneous, and thus (9) has to be supplemented by an equation that is inhomogeneous. Equation (8) does not give this directly, and we obtain it by noting that if π is the probability an incoming request finds an empty system, then balance equations yield

$$\pi=\sum_{i=1}^{N}\tilde{p}_i\pi_i \tag{11}$$

$$1-\pi=\lambda\sum_{i,j}\tilde{p}_i p_{ij}\pi_i[E(S_{ij}^0)-E(S_{ij})]+\lambda\sum_{i,j}\tilde{p}_i p_{ij}E(S_{ij}). \tag{12}$$

Equations (11) and (12) can now be combined to yield the necessary addendum to (9),

$$\sum_i \pi_i \tilde{p}_i \{1+\lambda \sum_j p_{ij}[E(S_{ij}^0)-E(S_{ij})]\} = 1-\lambda E(S);\ E(S) \equiv \sum_i \sum_j \tilde{p}_i p_{ij} E(S_{ij}).$$

An alternative method to compute the π_i is given in Section 4.

We present here an important result concerning those points at which $|A(z)|$ vanishes.

THEOREM 1. *The determinant of $A(z)$ vanishes at $z=1$ and at precisely $N-1$ points that satisfy $|z|<1$.*

Proof. The first claim is immediate by substitution and using $\tilde{p}P=\tilde{p}$. The second claim can be proved as follows:

Let $a_i(z)$ denote the N eigenvalues of the matrix $L(a)$, $|z| \leqq 1$. They need not necessarily be distinct, but in such a case we "perturb" the matrix P to separate them and invoke continuity arguments to assure that the number of roots of (10) stays the same.[1] We will assume they are distinct. Then (10) can be rewritten as

$$\prod_{i=1}^{N} (z-a_i(z)) = 0. \tag{13}$$

Since the matrix $L^T(a)$ is term by term strictly smaller (for $z \neq 1$) in absolute value than $L^T(0)$, a stochastic matrix (with spectral radius 1), all of $a_i(z)$ satisfy $|a_i(z)|<1$ (*see* [5], *vol. II, p.* 57). Rouché's theorem can now be applied to each of the factors of (13), to the effect that it has a single root in the open unit disk $|z|<1$ (except for that factor where $a_i(z)=1$).

We note a phenomenon that is interesting for its numerical implications: When successive request types are independent (i.e. $p_{ij}=p_j=\tilde{p}_j$) and $\lambda=0$, we have $|A(z)|=(z-1)z^{N-1}$, and thus it has one simple zero at $z=1$ and one of multiplicity $N-1$ at $z=0$. When λ increases continuously from zero to its operational value, the roots of (13) (which consists of continuous functions only) also move continuously in the z-plane. Writing L_{ij} as a power series in λ, $L_{ij}(a)=\sum_{k=0}^{r} b_{ijk} a^k + o(\lambda^r)$, we obtain $|A(z)|$ as a polynomial in z of degree N, with a simple zero at $z=1$ and a zero at $z=0$ of multiplicity $N-r-1 (r<N-1)$. This is obtained by using $0(\lambda)$ as an approximation for the other roots. The Lévy-Desplanques theorem assures us that for $|z|=1$, only $z=1$ is a root of the determinant[2]; since the roots departed continuously, they perforce are somewhere in the unit disk. For small values of λ we may expect them to have roots that are very close together, hard to separate and accurately evaluate. The method described in Section 4 is superior in such circumstances.

[1] What is not necessarily preserved is the strict inequality $|\zeta|<1$. It may happen that a root $\zeta \neq 1$ will have $|\zeta|=1$.

[2] The theorem states, in one of its versions, that if a matrix C satisfies the condition $|C_{ii}| > \Sigma_{j \neq i} |C_{ij}|$, for all i, then it is non-singular. For the matrix $A(z)$, where $|z|=1$, this condition reduces, using the inequality $|a-b| \geqq |a|-|b|$, to the requirement $|L_{ij}[\lambda(1-z)]|<1$, which is true when $\lambda>0$ and $\mathrm{Re}(1-z)>0$ [8].

To obtain the expected number in system $e_i (1 \leqq i \leqq N)$, we differentiate (7) at $z=1$. After some cancellations we obtain the set of relations.

$$e_i = (d/dz) G_i(z)\,|_{z=1} = \delta_1^{-2} (\{\delta_2 C_0 B_1 + \delta_1 C_1 B_1 + \delta_1 C_0 B_2\} \vec{\pi}_i), \qquad (14)$$

where the following derivatives are all with respect to z and are evaluated at $z=1$: $\delta_1 = |A(z)|'$, $\delta_2 = \frac{1}{2}|A(z)|''$, $C_0 = C(1)$, $C_1 = C'(z)$, $B_1 = -B'(z)$ and $B_2 = \frac{1}{2}B''(z)$. The values of these quantities, in terms of the model parameters, are given in the appendix. The overall mean queue size is then given by $\sum_{i=1}^{N} \tilde{p}_i e_i$.

Waiting time. We consider now the waiting time in a linear (FIFO) queue. Let η_n denote the waiting time of the nth request. As in any single-server linear queue,

$$\eta_{n+1} = [\eta_n + S_n - t_n]^+, \qquad n = 0,1,\cdots \qquad (15)$$

except that here the request types have to be incorporated into the calculation. S_n is the service duration of the nth request, and t_n is the time between its arrival and that of the $(n+1)$st. $t_n \sim \exp(\lambda)$, independently of the other variables.

Define

$$W_{ij}^n(x) = P(\eta_n \leqq x \,|\, \eta_0, J_0 = i, J_n = j)$$

$$w_{ij}^n = W_{ij}^n(0) \qquad (16)$$

$$\tilde{W}_{ij}^n(s) = \int_0^\infty \exp(-sx)\, d_x W_{ij}^n(x) = w_{ij}^n + \int_{0+}^\infty \exp(-sx)\, d_x W_{ij}^n(x).$$

Using the dependence structure of η_n and S_n, we obtain from (15), with some manipulations,

$$(\lambda - s)\,\tilde{W}_{ij}^{n+1}(s) = -s w_{ij}^{n+1} + \lambda \sum_{k=1}^{N} \{\tilde{W}_{ik}^n(s) L_{kj}(s) - w_{ik}^n [L_{kj}(s) - L_{kj}^0(s)]\}. \qquad (17)$$

Taking the limit $n \to \infty$ and assuming stationarity as before, (17) goes over to

$$C(s)\,\vec{W}(s) = D(s)\vec{w}, \qquad (18)$$

where

$$C_{ij}(s) = (\lambda - s)\delta_{ij} - \lambda L_{ji}(s) \text{ or } C(s) = (\lambda - s) I - \lambda L^T(s)$$

$$D_{ij}(s) = -s\delta_{ij} + \lambda [L_{ji}^0(s) - L_{ji}(s)], \text{ or } D(s) = \qquad (19)$$

$$-sI + \lambda [L^{0T}(s) - L^T(s)]$$

and w_j, the limit of w_{ij}^n is independent of i and equal to π_j. Note that $L^T(0)$ is a stochastic matrix with the invariant vector $\vec{p}$.

Equation (18) is of interest to us mainly as the starting point for the evaluation of the expected conditional waiting times

$$v_i=E(\eta|J=i)=-(d/ds)\,W_i(s)|_{s=0}. \tag{20}$$

We present a method of calculation. Higher moments can be calculated by continuation of the procedure below.

Differentiating (18) at $s=0$ yields

$$\lambda(I-L^T(0))\vec{v}=[-I+L^{0T'}(0)-\lambda L^{T'}(0)]\vec{\pi}=[\lambda L^{T'}(0)+I]\vec{e}, \tag{21}$$

where $\vec{e}$ is the all-ones N-vector.

Equation (21) is a singular set of equations for $\vec{v}$. We obtain a more convenient form by adding $\lambda\vec{P}\vec{v}$, which can be written as $\lambda(\vec{v}\cdot\vec{p})\vec{e}$, to both sides and get

$$\begin{aligned}\lambda\vec{v}=(I-L^T(0)+\vec{P})^{-1}\{&[\lambda\sigma-\lambda\sigma^0-I]\vec{\pi}\\ &+[I-\lambda\sigma]\vec{e}+\lambda(\vec{p}\cdot\vec{v})\vec{e}\},\ \sigma=-L^{T'}(0);\ \sigma^0=-L^{0T'}(0).\end{aligned} \tag{22}$$

Note that $(I-L^T(0)+\vec{P})^{-1}\vec{e}=\vec{e}$, $\vec{p}(I-L^T(0)+\vec{P})^{-1}=\vec{p}$.

The RHS of (22) is known, except the last term, $\vec{p}\cdot\vec{v}$, which we now determine. To this avail we need the Frobenius' eigenvalue of $L^T(s)$, $a(s)$, and its right and left eigenvectors, $\vec{\alpha}(s)$ and $\vec{\beta}(s)$, respectively.
Thus

$$[L^T(s)-a(s)I]\vec{\alpha}(s)=\vec{\beta}(s)[L^T(s)-a(s)I]=0 \tag{23}$$

and we may also stipulate, in addition, $\vec{\alpha}(s)\cdot\vec{\beta}(s)=\vec{\beta}(s)\cdot\vec{e}=1$. From these and (23) we immediately get $\vec{\alpha}(0)=\vec{e}$, $\vec{\beta}(0)=\vec{p}$, $a(0)=1$. As in Neuts [11] we obtain $a'(0)=-\rho/\lambda$, $\vec{\beta}'(0)=\vec{p}(\rho/\lambda I-\sigma)(I-L^T(0)+\vec{P}^{-1}$, $a''(0)=\vec{p}L^{T''}(0)\vec{e}-2\rho^2/\lambda^2+2\vec{p}\sigma(I-L^T(0)+\vec{P})^{-1}\sigma\vec{e}$. Now, multiplying (18) on the left by $\vec{\beta}(s)$ yields $\vec{\beta}(s)\vec{W}(s)=\vec{\beta}(s)[\lambda L^{0T}(s)-(s+\lambda a(s))I]\vec{\pi}/(\lambda-s-\lambda a(s))$. Differentiating by s and letting $s\to 0$ result in

$$\begin{aligned}\vec{p}\cdot\vec{v}=\{&[2(1-\rho)\vec{\beta}'(0)-\lambda a''(0)\vec{p}](\lambda\sigma-\lambda\sigma^0-I)\\ &-\lambda\vec{p}[L^{T''}(0)-L^{0T''}(0)]\}\,\vec{\pi}/2(1-\rho)^2.\end{aligned}$$

Thus, the expected waiting times can be directly calculated.

3. SPECIALIZATIONS

In this section we examine specific secondary storage devices, applying the results of previous sections. A specialization consists of specifying T_{ij}, T^0_{ij}, and K_i and their relations with the device parameters.

Drum-like devices. We consider a drum that comprises N logical sectors. The number of tracks is left unspecified. The time required for the ith sector to pass under the read heads is a constant δ_i; thus the set-up time, called rotational latency here, is given by

$$t_{ij}=\begin{cases}\sum_{k=i+1}^{j-1}\delta_k & \leqq i<j\leqq n\\ R-\sum_{k=j}^{i}\delta_k & 1\leqq j\leqq i\leqq n\end{cases}\qquad R=\sum_{k=1}^{N}\delta_k \tag{24}$$

We assume that the physical motion of the drum is the only element that creates delays; i.e., electronic switching times are entirely neglected. A similar device, with a slightly simpler distance structure, is magnetic bubble loop memory [9]. In terms of record structure our drum is midway between a "paging drum" and a "file drum" [4]. In our discussion of the drum we also specialize the input process by assuming that the types of successive requests are independent (and drawn from a distribution $\{p_i\}$). This is a reasonable assumption for a drum, which is normally the shared device *par excellence* in a system.

An interesting question in the design, and hence in the analysis, of such devices is the dependence of service capacity, or delays, on the pattern of use. For the drum as modeled here, it is well known that when requests are processed continuously, which would be the case in our model when the system is overloaded, the average rotational latency, T, depends on the distribution $\{p_i\}$ of relative frequencies but not on the manner in which the corresponding records are arranged around the circumference. Indeed, from (24) we easily have

$$t_{ij}+t_{ji}=\begin{cases} R-\delta_i-\delta_j & j\neq i \\ 2R-2\delta_i & j=i. \end{cases}$$

Hence

$$\begin{aligned} E(T_{\text{sat}}) &= \textstyle\sum_{i=1}^{N}\sum_{j=1}^{N} p_i p_j t_{ij} = \tfrac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} p_i p_j (t_{ij}+t_{ji}) \\ &= R(1+\textstyle\sum_{i=1}^{N} p_i^2)/2 - \sum_{i=1}^{N} p_i\delta_i \end{aligned} \tag{25}$$

for which the claimed invariance manifestly holds.

We shall show that this property is not retained when idle periods intervene. As in Section 2 we distinguish between a set-up within a busy period (T) and one that follows an idle period (T^0). From (25) we have

$$E(T)=(R/2)(1+\textstyle\sum p_i^2)-\sum p_i\delta_i. \tag{26}$$

To compute $E(T^0)$ consider the following sequence of events, on which we condition our calculation. A request for sector j is completed (and "departs"); no other request is queued for service; a request for sector i arrives and finds the head over sector M, at a distance D from its termination. Thus

$$T^0_{j,M,i,D}=D+t_{Mi}. \tag{27}$$

The duration τ between the departure of the request for the jth sector and the arrival of the new one is distributed exponentially with parameter λ. We may write

$$P(M=m,\ D=x)dx=\textstyle\sum_{k=0}^{\infty} P(\tau=kR+t_{jm}+\delta_m-x)dx,$$

where k is the number of complete revolutions the drum made between the departure and arrival. Using the distribution of τ we readily obtain

$$P(M=m,D=x)dx=\lambda \exp\{-\lambda(t_{jm}+\delta_m-x)\}/(1-\exp(-R\lambda))dx,\ 0\leqq x<\delta_m;1\leq m\leq N. \tag{28}$$

Using this result in (27) we have

$$E(T^0)=\sum_i \sum_j p_i p_j \sum_m \int_{x=0}^{\delta m} \lambda(x+t_{mi}) \exp\{-\lambda(t_{jm}+\delta_m-x)\}\, dx.$$

After integration and rather massive cancellations, one obtains

$$E(T^0)=R(1+\sum p_i^2)/2-1/\lambda - \sum_i p_i\delta_i+R\sum_i \sum_j p_i p_j \exp(-\lambda t_{ji})/(1-\exp(-R\lambda)). \tag{29}$$

Looking at the last term of (29) we see that $E(T^0)$ clearly depends, as claimed, on the relative arrangement of the records.

Remarks

1. We assumed here that sector lenghts δ_i may differ. In paging systems this is not necessarily the case. Nevertheless, the dependence expressed in (29) is maintained then as well.

2. Although we did not address ourselves to the problem of finding the optimal arrangement of the records on the drum (i.e., the relative placement of δ_i), which minimizes $E(T^0)$, this problem is of some theoretical interest. We digress here briefly to present a variation on the last model, where the nature of the problem is more evident.

In this variation all sectors have equal length, time is discrete, and arrivals may occur only at those evenly spaced epochs when an intersector boundary arrives at a read head. On this time scale, inter-arrival times are distributed geometrically, with a parameter we denote by α; this is the discrete analog of the exponential distribution. $E(T)$ is still given by (25), with no change, but when we come to evaluate $E(T^0)$ and examine (27), we see that D has no counterpart because of the discretization of arrival times.

If the calculation of $E(T^0)$ is carried to conclusion, we obtain instead of (29)

$$E(T^0, \text{discrete}) =(1-\alpha)/(1-\alpha^N) \sum_i \sum_j p_i p_j \sum_m t_{mi}\alpha^{t_{jm}}. \tag{30}$$

Consider the sum in (30). This is a polynomial of degree $N-1$ in α. The coefficient of α^r contains various terms that do not depend on the relative order of the records and the term $\beta_r=N\sum_{i=1}^{N}\sum_{j=i+1}^{i+r+1} p_i p_j$, which does (the indices j are calculated modulo N). Thus, the minimization of $E(T^0)$ here requires solution of $\min \sum_{r=0}^{N-1} \beta_r \alpha^r$.

3. Obviously, when the traffic intensity increases, idle times become rarer, and the relative arrangement is thus *least* important just when capacity is most critical. We note that this result does not justify ran-

domly placing records on a drum since this policy would affect the values of p_i as well (through the aggregating effects of tracks). As is apparent from (26), (29), and (30), the p_i do have considerable influence on $E(T)$, not just $E(T^0)$.

The maximum traffic intensity that the drum can handle under this regime is immediately given by (26) and bounded by

$$\lambda_{\max}=1/E(S)=1/[E(T)+\sum_i p_i\delta_i]=2/R(1+\sum_i p_i^2).$$

This result shows the way to obtain the distribution $\{p_i\}$ that results in efficient operation of the system. (Remember that the p_i are determined by the records that are placed in each sector, and normally some choice can be exercised in this respect.) To this end we only have to find the vector $\{p_i\}$ that minimizes $f=\sum_i p_i^2$, subject to $\sum p_j=1$. Since f and the constraint are convex and $p_i=1/N$ is an extremum point, that point must be a global minimum of f. Thus $\lambda_{\max}=2N/R(N+1)$, the best performance the system can exhibit. We remark that this optimization problem is "hard" (in fact, *NP*-complete [7]), as verified in [1] and [2]. Thus, one must expect an essentially enumerative search for that partition of the set of records such that f is minimized. (See [1] and [2] for analyses of a simple but very efficient heuristic.)

Finally, we look at the system of equations in (5) and their interpretation in the geometry and dynamics of the drum. We note first that $L_{ji}(a)=p_j\exp\{-a(t_{ji}+\delta_i)\}$ since the service time is constant. The transform of S_{ji}^0 is calculated in a way similar to that producing (29), and we obtain at some labor

$$\begin{aligned}L_{ji}^0(a)&=[p_j\exp(-a\delta_i)/z(1-\exp(-\lambda R))]\textstyle\sum_{m=1}^{N}\{\exp[-\lambda t_{jm}-a(\delta_m+t_{mi})]\\&\quad-\exp[-\lambda t_{jm}-\lambda\delta_m-at_{mi}]\}\\&=[p_j\exp(-a\delta_i)/z(1-\exp(-\lambda R))]\exp(-\lambda t_{ji})\{\exp(-aR)\\&\quad-1+(1-\exp(-\lambda R)\exp(\lambda z t_{ji})\}.\end{aligned}$$

Thus, we have from (6)

$$A_{ij}(z)=z\delta_{ij}-p_j\exp\{-a(t_{ji}+\delta_i)\} \tag{31}$$

and

$$B_{ij}(z)=[p_j(\exp(-aR)-1)/(1-\exp(-\lambda R))]\exp(-a\delta_i-\lambda t_{ji}).$$

In order to use (9) the roots of the equation $|A(z)|=0$ are required. The following result is instrumental in obtaining an efficient solution.

THEOREM 2. *The determinant of* $A(z)$ *(in (31)) can be expressed as*

$$|A(z)|=(b-z^N)/(q-1) \tag{32}$$

where $q=\exp(-aR)$, $b=q\prod_{j=1}^{N}[z-p_j(q-1)]$.

Proof. Consider $|A(z)|$ as an Nth degree polynomial expression in the term linear in z, with the exponentials regarded as coefficients. At the N values of z given by $z_j=p_j(q-1)$ (when we treat q as an explicit coefficient of z rather than display its functional dependence), the determinant can be easily evaluated, and we obtain $-z_j^N/(q-1)$. The right-hand side of (32) is a polynomial expression of degree N that correctly interpolates $|A(z)|$ at the $N+1$ points[3] $z=z_j$ and $z=1$ and is therefore the unique interpolating polynomial expression of degree N.

This is perhaps a somewhat curious result, since the roots of this equation turn out not to depend to any extent on lengths of individual records (sectors), but merely on their frequency of use. We have no intuitive explanation for this phenomenon.

The equation $|A(z)|=0$ can now be easily solved numerically, and our experience with a straightforward Newton-Raphson iteration procedure demonstrated very fast convergence and good resolution between the roots (we only looked inside the unit disk).

Disk-like devices. We consider now the characteristics of a disk pack (or cartridge), with N cylinders (tracks), and a single arm carrying the read heads. For the purposes of our analysis this is functionally identical with any device where the setup time T_{ij} is merely a function of $|i-j|$, such as magnetic bubble or shift register storage devices. The following will be in disk terminology.

It is customary to consider the set-up time in disks as composed of two parts: seek time, the duration required for the arm to move between cylinders, and a rotational latency similar to the drum.

As we consider here a primitive request-queue management technique, we also limit all explicit calculations concerning disks in two ways:

Rotational latency is eliminated by the method of reading, which is to transmit one whole track per request (the portion of the disk passing under a read head during one full revolution). The desired record is subsequently located in memory and perhaps pieced together from two portions. The latter situation occurs when the requested record was under the read head when the seek terminated and transmission started.

In these devices (in contrast to the situation in scanning disks) the arm does not react "on the fly" to changes of destination, but rather maintains a "busy" status until a desired seek is terminated and the arm is stopped; only then can a new seek be initiated. Comprehensive discussions of these delays can be found in [3] and [13].

Although the set-up times T_{ij} of bi-directional tapes conform with the above characterization, we exclude them from this discussion on both practical and analytical grounds. First, rotational latency is of course absent here; also, the tape system can usually handle changes of desti-

[3] Obviously $|A(1)|=0$ since $\Sigma_{p_j}=1$; a single application of L'Hospital's rule establishes that $z=1$ is a zero of the right-hand side of (32).

nation "on the fly" in a much simpler way than in a disk system. (Thus, FIFO is a less natural operating technique for tapes than it is for disks. Even there, however, low processor speeds may require a FIFO regime.) Analytically, we find the dependence structure between successive services even more involved than the model presented in Section 1: T_{ij} depends on the boundary of record i where its reading terminated, and this, in turn, involves the even earlier record. We note, though, that if the idle-period policy were of the type denoted by (a) in the following, one randomization on the identity of that preceding record is enough to properly define the necessary variables.

Unlike the drum, the behavior of the system when no requests are pending may have different modes. The more common ones are (in disk terminology)

(a) The arm remains in place, at the cylinder last used.

(b) The arm is directed to move to a predetermined "rest place," cylinder r.

These modes determine the distributions of the respective S^0_{ij}. In case (a) it is clear that $S^0_{ij} \sim S_{ij}$. In case (b) let $f(i, j)$ be the time taken to travel from cylinder i to cylinder j when no intervening cylinders are read. Normally, this is the same as the set-up time T_{ij} and we assume so in the following. The set-up time T^0_{ij} succeeding an idle period is then given by

$$T^0_{ij|s} = \begin{cases} f(r, j) & s \geqq f(i, r) \\ f(i, r) - s + f(r, j) & s < f(i, r) \end{cases}$$

where s is the length of the idle period. Since s is exponentially distributed, we immediately find $E_s(T^0_{ij}) = f(i, r) + f(r, j) - [1 - \exp(-\lambda f(i, r))]/\lambda$. The expected duration of this delay is calculated as follows. Note first that $p_i^0 \equiv P$ (cylinder i was just read|an idle period just started) $= \pi_i p_i / \sum_k \pi_k p_k$ where π_i, as defined earlier, is the probability that the request queue is empty following the completion of service from cylinder i. Thus we obtain

$$E(T^0) = \sum_i \sum_j p_i^0 p_j E_s(T^0_{ij})$$

and

$$E(T^0) = \sum_j p_j f(r, j) + \sum_i \pi_i p_i f(i, r) \exp(-\lambda f(i, r)) / (\lambda \sum_k \pi_k p_k) - 1/\lambda.$$

The value of $E(T^0)$ does not influence the overall service capacity of the system. It is a factor in its response when not fully loaded. In fact, it becomes more important as the load becomes lighter.

Unlike the drum, which is a constant speed system, we have here important acceleration and deceleration effects. An approximation that holds for a rather large subset of available disks is

$$T_{ij} = f(i, j) = \begin{cases} 0 & i = j \\ A + B|i - j & i \neq j, \end{cases}$$

where A "summarizes" the effects of the changes of speed of motion of the arm and B corresponds to movement in constant speed. (The approximation is not very good for short distances and quite acceptable when a sizeable portion of the disk radius has to be traversed.) This completes the specification, so that the procedures of Section 2 can be applied. (Nothing comparable to Theorem 2 was found here, however.)

4. ALTERNATIVE PROCEDURE TO CALCULATE BOUNDARY PROBABILITIES

In this section we describe a general method to evaluate the boundary probabilities π_j defined below (4). The analytical method given in Section 2 which depends on Theorem 1 is devoid of probabilistic-physical content. This makes any numerical idiosyncracies occurring in its implementation hard to interpret, and thus instabilities are not easy to move, even for moderate values of N. We present an approach parallel to the one in [11][4]; here all the steps and interim results have intuitive meaning, and thus error control is materially simplified. Excepting values of λ close to $\lambda_{\max}$, this method would also be cheaper than the method of Section 2.

We call upon a familiar result: In a recurrent Markov chain, the invariant probability of a state (its steady-state probability) is equal to the inverse of its recurrence time [6, p. 195].

Consider then the embedded chain, formed of the N states $(0, j)$, obtained at departure epochs. Its steady-state probabilities were called $p(0, j)=\tilde{p}_j\pi_j$, and its recurrence times are given by $\sum_{i=1}^{N} l_i\mu_i^*/l_j$, where $\vec{l}$ is the invariant probability vector of the matrix L, defined as follows: $L_{ij}(k, x)$=Prob (a busy period that starts at $(0, i)$ terminates at $(0, j)$, following k services and requiring up to x) and $L=\int_{x=0}^{\infty}\sum_{k=1}^{\infty}dL(k, x)$. The quantity μ_i^* is the expected number of service completions in a busy period that started in state $(0, i)$. If $L(z, s)$ is the LST-pgf of $L(k, x)$, we have $\vec{\mu}^*=(\partial L(z, s)/\partial z|_{z=1,s=0})\vec{e}$. We proceed to derive $\vec{l}$ and $\vec{\mu}^*$. Define now the matrices $\bar{c}(\cdot)$ and $\bar{c}^0(\cdot)$, $\bar{c}_{ij}(x)=p_{ij}F_{ij}(x)$ and $\bar{c}_{ij}^0(x)=p_{ij}F_{ij}^0(x)$, and first-passage measure $\bar{G}_{ij}(k, x)$=Prob (A first transition of the system from a state (n, i) to a state where $X=n-1$, will be to the state $(n-1, j)$, will involve k services and terminate within x). The interpretations of $\bar{c}(\cdot)$ and $\bar{c}^0(\cdot)$ are obvious. $\bar{G}(\cdot,\cdot)$ is also called a "down level-crossing" distribution.

We further define the matrices $\bar{c}_\nu(x)=\int_{t=0}^{x} p(\nu, t)d\bar{c}(t)$, $\nu\geqq 0$, and $\bar{c}_\nu^0(x)=\int_{t=0}^{x} p(\nu, t)d\bar{c}^0(t)$, $\nu\geqq 0$, where $p(\nu, t)$ is the probability of exactly ν arrivals within t.

Forming the LST's of $\bar{c}_\nu(\cdot)$, $\bar{c}_\nu^0(\cdot)$ and $\bar{G}(\cdot,\cdot)$ (the latter is also a pgf), denoted respectively by $c_\nu(s)$, $c_\nu^0(s)$, and $G(z, s)$ we have, following [11] from renewal considerations,

[4] The only essential difference between our model and the problem treated in [11], is that we must ascribe an extraordinary distribution to the service duration that initiates a busy cycle. We note, however, that in [11] batch arrivals are treated.

$$G(z, s)=\sum_{\nu=0}^{\infty} z c_\nu(s) G^\nu(z, s) \quad |z|\leqq 1,\ \mathrm{Re}(s)\geqq 0. \tag{33}$$

For $z=1$, $s=0$, $G(1, 0)\equiv G$ is a stochastic matrix, with invariant probabilty vector $\vec{g}$, from which we form $\tilde{G}$, as $\tilde{P}$ was defined.

The same consideration that led to (33) can be applied to $L(z, s)$, and we obtain $L(z, s)=\sum_{\nu=0}^{\infty} z c_\nu^0(s) G^\nu(z, s)$, $|z|\leqq 1$, $\mathrm{Re}(s)\geqq 0$.

Equation (33), at $z=1$, $s=0$ can be iteratively solved,[5] whence L and $\vec{l}$ are obtained quite painlessly. L is the probability transition matrix of our embedded chain.

We still need $\vec{\mu}^*$. To this avail we note the following result, given in [11]:

$$\vec{\mu}\equiv(\partial G/\partial z|_{z=1,s=0})\vec{e}=(I-G+\tilde{G})[I-P+\tilde{G}+\lambda\mathrm{diag}(\vec{\sigma})\tilde{G}]^{-1}\vec{e}=(I-G+\tilde{G})\vec{a},$$

where $\sigma_i=\sum_j p_{ij}E(S_{ij})$. Thus,

$$\vec{\mu}^*=(\partial L/\partial z|_{z=1,s=0})\vec{e}=[\sum_{\nu=0}^{\infty} c_\nu^0(0) G^\nu+\sum_{\nu=0}^{\infty} c_\nu^0 \sum_{i=0}^{\nu-1} G^i M_z G^{\nu-i-1}]\vec{e},$$

where $M_z=\partial G(z, s)/\partial z$, $z=1$, $s=0$. Since G is stochastic we get

$$\begin{aligned}\vec{\mu}^* &= 1+\sum_{\nu=0}^{\infty} c_\nu^0(0) \sum_{i=0}^{\nu-1} G^i\vec{\mu}=1+\sum_{\nu=0}^{\infty} c_\nu^0 \sum_{i=0}^{\nu-1} G^i(I-G+\tilde{G})\vec{a}\\ &= 1+[P-L+\lambda\mathrm{diag}(\vec{\sigma}^0)\tilde{G}](I-G+\tilde{G})^{-1}\vec{\mu},\end{aligned}$$

which can be readily calculated.

Reference 11 contains further results that are of interest and can be applied—mutatis mutandis—to our model. Expressions for mean queue lengths were derived, to be used as a check on (14). As they are rather involved, we do not present them here. We note, however, that the two procedures pose numerical problems of entirely different nature and the investigation of their respective behavior modes, particularly in extreme situations (very light or very heavy traffic, large N, etc.), is of great interest.

5. DISCUSSION

We have shown in the preceding sections a method of analyzing system models that, although they are simple to describe in queuing-theoretical terminology, display features that render standard methods ineffective in tackling them. The factor that particularly exacerbates the work is the dependence between successive services; put another way, the time required to service a set of requests depends on the way we order them. Rarely, if ever, will FIFO prove the most efficient service method, although we can very well imagine situations where its simplicity of implementation would outweigh other considerations.

[5] E.g., by the sequence $G_1=(I-c_1)^{-1}$, $G_{k+1}=(I-c_1)^{-1}\{c_0+\sum_{\nu=2}^{\infty} c_\nu G_k^\nu\}$, $k\geqq 1$, which converges quite well, normally.

In contrast with the foregoing analysis we mention a rather prevalent approach to the same situation that is often found in the literature (cf. [14] for a recent example). The approach we refer to consists of evaluating a distribution function for the duration required to service a request by averaging over request types [essentially, writing $P(S\leqq t)=\sum_{i,j}\tilde{p}_i p_{ij} P(S_{ij}\leqq t)$], and substituting the result within formulas derived in the standard analysis of an $M/G/1$ queuing system, which explicitly assumes (and uses) independence between successive services. This can often lead to gross misestimation of the evaluated quantities.

APPENDIX

Defining $\sigma_{ij}=E(S_{ij})$, $\sigma_{ij}^{(2)}=E(S_{ij}^2)$ (similarly with superscript, 0), and letting $\tilde{M}_{ij}$ be $\tilde{p}_i p_{ij} M_{ij}/\tilde{p}_j$ where M is any of σ, σ^0, $\sigma^{(2)}$, $\sigma^{0(2)}$, I we find $B_1{}^T=\lambda\tilde{\sigma}-\lambda\tilde{\sigma}^0-\tilde{I}$, $B_2{}^T=\tfrac{1}{2}[\lambda^2\tilde{\sigma}^{0(2)}+2\lambda\tilde{\sigma}^0-\lambda^2\tilde{\sigma}^{(2)}]$. The quantities δ_1, δ_2 and the matrices C_0 and C_1 have generally to be directly evaluated by differentiating $|A(z)|$ and the relation $|A(z)|I=A(z)C(z)$. For the special case $P=\tilde{P}$ (independent references) closed expressions can be readily found:

$$\delta_1=1-\rho;\ \delta_2=N-1-\lambda\sum_j p_j\sigma_{jj}-(N-2)\rho-\tfrac{1}{2}\lambda^2E(S^2)+\lambda^2\sum_j\sum_{k>j}p_jp_k$$
$$(\sigma_{jj}\sigma_{kk}-\sigma_{jk}\sigma_{kj})-\lambda^2\sum_j\sum_{k>j}\sum_{l>k}p_jp_kp_l\sigma_{(jkl)\sigma},$$

where

$$\sigma_{(jkl)\sigma}=\sigma_{jj}\sigma_{kk}+\sigma_{jj}\sigma_{ll}+\sigma_{kk}\sigma_{ll}+\sigma_{kj}\sigma_{lk}$$
$$+\sigma_{kj}\sigma_{jl}+\sigma_{lk}\sigma_{jl}+\sigma_{jk}\sigma_{lj}+\sigma_{lj}\sigma_{kl}$$
$$+\sigma_{jk}\sigma_{kl}-\sigma_{kl}\sigma_{lk}-\sigma_{kl}\sigma_{jj}-\sigma_{lk}\sigma_{jj}$$
$$-\sigma_{jl}\sigma_{kk}-\sigma_{jl}\sigma_{lj}-\sigma_{kk}\sigma_{lj}-\sigma_{ll}\sigma_{jk}$$
$$-\sigma_{ll}\sigma_{kj}-\sigma_{jk}\sigma_{kj}.$$

$$C_{0_{i,j}}=p_j,$$

$$C_{1_{i,i}}=-1-(N-2)p_i+\rho+\lambda p_i\sum_j p_j(\sigma_{jj}-\sigma_{ij}-\sigma_{ji}),$$

$$C_{1_{i,j}}=-p_j[N-2+\lambda(p_i+p_j)\sigma_{ji}-\lambda\textstyle\sum_{ki,j}p_k(\sigma_{kk}-\sigma_{jk}-\sigma_{ki})],\ i\neq j.$$

REFERENCES

1. A. K. CHANDRA AND C. K. WONG, "Worst-Case Analysis of a Placement Algorithm Related to Storage Allocation," *Siam J. Comput.* **4,** 249–263 (1975).
2. R. A. CODY AND E. G. COFFMAN, JR., "Record Allocation for Minimizing Expected Retrieval Costs on Drum-Like Storate Devices," *J. Assoc. Comput. Mach.* **23,** 103–115 (1976).
3. H. FRANK, "Analysis and Optimation of Disk Storage Devices for Time-Sharing Systems." *J. Assoc. Comput. Mach.* **16,** 602–620 (1969).

4. S. H. FULLER AND F. BASKETT, "An Analysis of Drum Storage Units," *J. Assoc. Comput. Mach.* **22,** 83–105 (1975).
5. F. R. GANTMACHER, *The Theory of Matrices,* Chelsea, N.Y., 1959.
6. J. J. HUNTER, "On the Moments of Markov Renewal Processes," *Adv. Appl. Prob.* **1,** 188–210 (1969).
7. R. M. KARP, "Reducibility among Combinatorial Problems," *Complexity of Computer Computation,* pp. 85–104, R. E. Miller and J. W. Thatcher (eds.), Plenum Press, N.Y., 1972.
8. M. MARCUS AND H. MINC, *A Survey of Matrix Theory and Matrix Inequalities,* Allyn and Bacon, Inc., Boston, 1964.
9. D. MITRA, "Some Aspects of Hierarchical Memory Systems, *J. Assoc. Comput. Mach.* **21,** 54–65 (1974).
10. M. F. NEUTS, "The Single Server Queue with Poisson Input and Semi-Markov Service Times," *J. Appl. Prob.* **3,** 202–230 (1966).
11. M. F. NEUTS, "Some Explicit Formulas for the Steady-State Behavior of the Queue with Semi-Markovian Service Times," Dept. of Statistics, Purdue University, Mimeograph Series, No. 444, 1976.
12. C. E. SKINNER, "Priority Queueing Systems with Server Walking Time," *Opns. Res.* **15,** 278–285 (1967).
13. S. J. WATERS, "Estimating Magnetic Disk Seeks," *Comp. J.* **18,** 12–17 (1975).
14. N. C. WILHELM, "An Anomaly in Disk Scheduling," *Comm. ACM* **19,** 13–17 (1976).