Probability in the Engineering and Informational Sciences, 6, 1992, 217–239. Printed in the U.S.A.

GATED, EXHAUSTIVE, PARALLEL SERVICE

SID BROWNE
Graduate School of Business Administration
Columbia University
New York, New York 10027

E. G. COFFMAN, JR., E. N. GILBERT, AND PAUL E. WRIGHT
AT&T Bell Laboratories
Murray Hill, New Jersey 07974

We analyze gated, exhaustive service of an infinite-server system with vacations. Customers enter a queue in a Poisson stream. The servers, working in parallel, serve customers in stages. A stage begins with all customers transferred from the queue to the servers (the gate opens). The servers then begin serving these customers, all simultaneously. The stage ends when their services are completed. Service is exhaustive because the servers must again examine the queue to see if any new customers arrived during the last stage. If there are any, a new stage begins. If there are none, the servers move on to other work. The time spent away from the queue is called *vacation time*. The queue may represent a node or station in a data transmission network and the servers may be communication channels.

We analyze the equilibrium behavior of the number of requests served during a stage for general service and vacation time distributions. This analysis leads to the solution of a Fredholm integral equation of the second kind. We find conditions under which the system is stable and compute bounds on performance metrics of interest. Approximate techniques are introduced and tested. Finally, an extension to polling systems is studied.

1. INTRODUCTION

In the problem to be studied, customers enter an infinite-server system in a Poisson stream at rate λ per unit time. The servers, working in parallel, give gated

 0269-9648/92 $5.00 + .00

and exhaustive service to the queue. Service is *gated* because it is performed in stages. A *stage* begins with all customers transferred from the queue to the servers (the gate opens). The servers then begin serving these customers, all simultaneously. The stage ends when their services are completed. Service is *exhaustive* because the servers must again examine the queue to see if any new customers arrived during the last stage. If any did, a new stage begins. If not, the servers move on to other work (secondary jobs), giving the queue time to accumulate customers. The time servers spend away will be called *vacation time*. Vacations alternate with busy periods of the queue.

The probability that k customers arrive to the queue during a vacation will be called ν_k. No generality is lost by taking $\nu_0 = 0$. For if the servers return from vacation and find the queue still empty, they merely extend the vacation another round. We assume that vacation times and busy periods are independent. That assumption is a great convenience but it usually holds only approximately. Thus, if secondary jobs to be done during vacations arrive at random during busy periods, long vacations tend to follow long busy periods. The independence assumption is satisfied in the important special case where there are no vacation jobs and the servers spend vacations idly awaiting the next arrival to the queue. This case corresponds to $\nu_1 = 1$ and is covered in Sections 3 and 8.

Customers have independent, random service times with a common distribution function $F(t) = \Pr\{\textit{service time} \leq t\}$. With λ, $\{\nu_k\}$, and $F(t)$ given, the equations of Sections 2 and 3 will determine the probability distributions of the duration of a stage and the number it serves.

The servers can work simultaneously on arbitrarily large numbers of customers. If k customers are in the queue when the gate opens, the stage lasts just long enough to serve k customers in parallel; i.e., the stage lasts for the longest of k service times. In a real application, there is some number $M < \infty$ of servers. The queue may represent a node or station in a data transmission network and the servers may be communication channels, for example. Models with M finite appear to be very difficult to analyze. However, the model with $M = \infty$ can provide insight into the number of servers needed for efficient operation.

Another application is task-oriented parallel simulation, in which tasks are repeatedly removed from a queue by processors working in parallel. Vacations correspond to idle periods of the server. Gating occurs when processors must synchronize after each stage of a computation. In parallel simulations, synchronization points are often necessary to guarantee correctness of the computation. Iterative solution of algebraic systems is another example where this occurs. Another example where synchronization is required is *bounded-lag* distributed computation (see Lubachevsky [3] for details). In the case of distributed computation, it is not unusual to design a program as if there were one logical processor per task. In general, each real processor performs the tasks of many logical processors, with synchronization points required after each set of tasks.

It is then of interest to determine how many processors are active in each stage of the computation.

Let k_n denote the number of customers served in the nth stage; k_n constitutes a Markov chain. Sections 2 and 3 analyze the transient and equilibrium behavior of this chain for general service and vacation time distributions. Section 3 reduces the analysis of stationary distributions to the study of a Fredholm integral equation of the second kind.

Servers working in parallel remove most of the instability that arises in single-server queuing with large λ. For instance, suppose service times are distributed only over a finite interval $[0, T]$; i.e., $F(T) = 1$. Then all stages last for time T at most. For large λ, each stage merely serves a number of customers that is approximately Poisson distributed with mean λT (stages following a vacation have another distribution but these stages are rare). If arbitrarily large service times are possible, then it is no longer obvious that the queue will not grow. However, Section 4 will show stability as long as the service times and numbers of arrivals during vacations have finite means and variances.

Section 5 derives bounds on the mean duration of a stage; their accuracy improves in heavy traffic, i.e., for λ large. Section 6 examines the waiting time of a typical customer and the number of other customers he or she finds waiting when he or she arrives. Section 7 specializes $F(t)$ to a step function to obtain a numerical approximation applicable to arbitrary $F(t)$. Section 8 applies the approximation to the case where $\nu_1 = 1$; i.e., there are no vacation jobs.

Polling systems are an interesting extension of the gated infinite-server queue with vacations. The literature on polling appears to be limited to single-server models (see Takagi [4] for an extensive survey). In an infinite-server polling system, the single queue discussed to this point is just one of several that the servers visit in an endless cycle. The times that the servers spend at the other queues belong to vacations. In these systems, vacations and busy periods are not independent. Polling systems with exhaustive service can have long mean waiting times. This is illustrated in Section 9, where a two-station infinite-server polling system is analyzed for several different service policies.

2. STATE PROBABILITIES

The number k_n served in stage n determines distribution functions for the duration of stage n, the number of arrivals during stage n, and hence the number k_{n+1} served in the next stage. Define the state probability $p_k^{(n)} = \Pr\{k_n = k\}$, $k \geq 1$. The duration of the nth stage and its distribution function will be denoted, respectively, by t_n and $H^{(n)}(t) = \Pr\{t_n \leq t\}$. The transition equations relating $p_k^{(n+1)}$ to $p_j^{(n)}$ are most conveniently expressed in terms of the generating functions

$$P^{(n)}(x) = \sum_{k=1}^{\infty} p_k^{(n)} x^k, \qquad V(x) = \sum_{k=1}^{\infty} \nu_k x^k, \tag{2.1}$$

and Laplace transforms

$$G^{(n)}(s) = \int_0^\infty e^{-st}\, dH^{(n)}(t). \tag{2.2}$$

With k_n given, t_n is the maximum of k_n service times. Then

$$\Pr\{t_n \le t \,|\, k_n = k\} = F(t)^k \tag{2.3}$$

and

$$H^{(n)}(t) = \sum_{k=1}^{\infty} p_k^{(n)} F(t)^k = P^{(n)}(F(t)). \tag{2.4}$$

Also, with $t_n = t$ given, k_{n+1} is determined from the number z of arrivals in time t, a Poisson-distributed random variable with mean λt. If $z \neq 0$, then $k_{n+1} = z$. If $z = 0$, then k_{n+1} has the distribution ν_k of the number of arrivals during a vacation. It follows that

$$\Pr\{k_{n+1} = k \,|\, t_n = t\} = e^{-\lambda t}\frac{(\lambda t)^k}{k!} + e^{-\lambda t}\nu_k, \qquad k = 1,2,\ldots, \tag{2.5}$$

$$p_k^{(n+1)} = \int_0^\infty \left\{\frac{(\lambda t)^k}{k!} + \nu_k\right\} e^{-\lambda t}\, dH^{(n)}(t), \qquad k = 1,2,\ldots, \tag{2.6}$$

and

$$\begin{aligned} P^{(n+1)}(x) &= \int_0^\infty \{e^{\lambda t x} - 1 + V(x)\} e^{-\lambda t}\, dH^{(n)}(t) \\ &= G^{(n)}(\lambda - \lambda x) + \{V(x) - 1\} G^{(n)}(\lambda). \end{aligned} \tag{2.7}$$

The Laplace transforms $G^{(n)}$ in Eq. (2.7) can be expressed in terms of $P^{(n)}$ by using Eq. (2.4),

$$G^{(n)}(s) = \int_0^\infty e^{-st}\, dP^{(n)}(F(t)). \tag{2.8}$$

Then

$$P^{(n+1)}(x) = \int_0^\infty [e^{-\lambda(1-x)t} + \{V(x) - 1\} e^{-\lambda t}]\, dP^{(n)}(F(t)). \tag{2.9}$$

When Eq. (2.9) is expanded in powers of x, equating coefficients of x^k gives $p_k^{(n+1)}$ as a linear function of $p_j^{(n)}$. Starting with any given initial state distribution p_k^0, one may iterate Eq. (2.9) to calculate the state probabilities $p_k^{(1)}$, $p_k^{(2)}, \ldots$ of the stages that follow.

The transition equations can also be expressed in terms of $H^{(n)}(t)$ alone. With the change of variable $x = F(y)$ in Eq. (2.7), we obtain a recurrence for

$H^{(n)}(y) = P^{(n)}(F(y))$, expressed in terms of $U(y) = V(F(y))$. A simpler form, obtained by substituting $\bar{F}(y) = 1 - F(y)$, $\bar{U}(y) = 1 - U(y)$, and $\bar{H}^{(n)}(t) = 1 - H^{(n)}(t)$, is

$$\bar{H}^{(n+1)}(y) = 1 + \int_0^\infty \{e^{-\lambda \bar{F}(y)t} - \bar{U}(y)e^{-\lambda t}\}\, d\bar{H}^{(n)}(t). \tag{2.10}$$

Integration by parts, using $\bar{H}^{(n)}(0) = 1$, produces

$$\bar{H}^{(n+1)}(y) = \bar{U}(y) + \int_0^T K(y,t)\bar{H}^{(n)}(t)\, dt \tag{2.11}$$

with

$$K(y,t) = \lambda\{\bar{F}(y)e^{-\lambda \bar{F}(y)t} - \bar{U}(y)e^{-\lambda t}\}, \tag{2.12}$$

and the service times are distributed over $[0, T]$; T may be infinity. Equation (2.11) will be used in the next section to study stationary distributions, obtained as n becomes large.

3. STATIONARY DISTRIBUTIONS

In statistical equilibrium, the probability distributions $p_k^{(n)}$ and $H^{(n)}(t)$ are the stationary distributions p_k and $H(t)$, independent of n. To study them we use the equations of Section 2 with all superscripts (n) deleted. (This convention is applied throughout.) The stationary form of Eq. (2.9) is

$$P(x) = \int_0^\infty e^{-\lambda(1-x)t}\, dP(F(t)) + \{V(x) - 1\}\int_0^\infty e^{-\lambda t}\, dP(F(t)), \tag{3.1}$$

now an integral equation for $P(x)$. For the moment, we assume that Eq. (3.1) has a unique solution. We will return to this question later in this section and in Section 4.

A more convenient form of Eq. (3.1) is sometimes obtained by changing the variable from t to $F = F(t)$. Then t in Eq. (3.1) becomes the inverse function $t = t(F)$ (such that $F(t(F)) = F$) and Eq. (3.1) becomes

$$P(x) = \int_0^1 e^{-\lambda(1-x)t(F)}\, dP(F) + \{V(x) - 1\}\int_0^1 e^{-\lambda t(F)}\, dP(F). \tag{3.2}$$

For example, if service times are uniformly distributed over $[0, T]$, $F(t) = t/T$, $t(F) = TF$, and Eq. (3.2) is

$$P(x) = \int_0^1 e^{-\lambda(1-x)TF}\, dP(F) + \{V(x) - 1\}\int_0^1 e^{-\lambda TF}\, dP(F). \tag{3.3}$$

If service times are exponentially distributed with mean τ, $F(t) = 1 - e^{-t/\tau}$, and Eq. (3.2) is

$$P(x) = \int_0^1 (1-F)^{\lambda r(1-x)}\, dP(F) + \{V(x) - 1\} \int_0^1 (1-F)^{\lambda r}\, dP(F). \tag{3.4}$$

By expanding $P(x)$ and $P(F)$ in Eq. (3.2) and equating coefficients of x^k, one obtains an infinite system of linear equations for the state probabilities

$$p_k = \sum_{j=1}^{\infty} j p_j \int_0^1 e^{-\lambda t(F)} \left\{ \frac{[\lambda t(F)]^k}{k!} + \nu_k \right\} F^{j-1}\, dF, \qquad k = 1,2,\ldots. \tag{3.5}$$

When service times are uniform or exponential, the coefficients in Eq. (3.5) are elementary (although unpleasant) integrals, but the solution requires numerical methods.

The stationary form of Eq. (2.11) is a Fredholm integral equation for $\bar{H}(t)$. It may be written compactly as $(\mathbf{I} - \mathbf{K})\bar{H} = \bar{U}$, where $\mathbf{K}$ is the integral operator in Eq. (2.11) with the kernel $K(y,t)$, and $\mathbf{I}$ is the identity operator. Then Eq. (2.11) suggests the formal solution

$$\bar{H} = (\mathbf{I} - \mathbf{K})^{-1} \bar{U} = \sum_{r=0}^{\infty} \mathbf{K}^r \bar{U}. \tag{3.6}$$

Equation (3.6) is the familiar Neumann series, which can also be obtained by solving the integral equation by successive approximation. Indeed, Eq. (2.11) has the solution

$$\bar{H}^{(n)} = (\mathbf{I} + \mathbf{K} + \ldots + \mathbf{K}^{n-1})\bar{U} + \mathbf{K}^n \bar{H}^{(0)}. \tag{3.7}$$

Then $\bar{H}^{(n)}$ will approach Eq. (3.6) for large n if the Neumann series converges and the term $\mathbf{K}^n \bar{H}^{(0)}$ of Eq. (3.7) tends to zero. That will happen if $\mathbf{K}$ has a sufficiently small norm $\|\mathbf{K}\|$, i.e., if

$$\|\mathbf{K}\| \equiv \sup_y \int_0^T |K(y,t)|\, dt < 1$$

(see Kress [2]). Moreover, when the sequence $\bar{H}^{(n)}(t)$ converges, so does the sequence of generating functions $P^{(n)}(x)$. The no-vacation case, $\nu = 1$, gives a simple example.

THEOREM 3.1: *Suppose $V(x) = x$ and service times have a finite bound T; i.e., $F(T) = 1$. Then the integral equation $\bar{H} = \bar{U} + \mathbf{K}\bar{H}$ has a unique solution, given by a convergent Neumann series (Eq. (3.7)). Moreover, $H^{(n)}(t)$ approaches $H(t)$ as $n \to \infty$ for any choice of $H^{(0)}(t)$.*

PROOF: It suffices to prove $\|\mathbf{K}\| < 1$. With $V(x) = x$, we have $\bar{U}(y) = \bar{F}(y)$. But then Eq. (2.12) shows that $K(y,t) > 0$ so $\int_0^T |K(y,t)|\, dt \le 1 - e^{-\lambda \bar{F}(y)T} \le 1 - e^{-\lambda T}$. ∎

In the next section, we examine stability for general $V(x)$ and unbounded service times.

4. STABILITY

When λ is large, an instability of the following sort may be imagined. A typical stage n has a large number, about λt_n, of arrivals. At least one of them requires an abnormally long service time, so $t_{n+1} > t_n$. But then stage $n+1$ must serve even more customers, and so on. Is it possible for k_n to grow with n and eventually remain forever above any preassigned finite bound? Of course, as Section 1 has already mentioned, this kind of instability can only be imagined for service time distributions $F(t)$ with long tails.

THEOREM 4.1: *Suppose service times have a finite mean τ and variance σ^2. Also suppose the number of requests arriving during a vacation has a finite mean ν and variance ω^2. Then $k_n \le 8A$ holds for infinitely many n with probability* 1, *where*

$$A = \lambda\tau + \lambda^2\tau^2 + \lambda^2\sigma^2 + \nu^2 + \omega^2. \tag{4.1}$$

Thus, k_n cannot grow and remain arbitrarily large. Indeed, whenever $k_n \le 8A$, there is a fixed, small but positive probability that no customers arrive during stage n. Then there is probability 1 that the servers go on vacation infinitely often.

PROOF: The proof begins with Eq. (2.5), from which

$$\Pr\{k_{n+1} = l \,|\, k_n = k\} = \int_0^\infty e^{-\lambda t}\left\{\frac{(\lambda t)^l}{l!} + \nu_l\right\} dF(t)^k$$

and

$$\begin{aligned}\mathbf{E}\{k_{n+1}^2 \,|\, k_n = k\} &= \int_0^\infty \sum_{l=1}^\infty \left\{\frac{l^2(\lambda t)^l}{l!} + l^2\nu_l\right\} e^{-\lambda t}\, dF(t)^k \\ &= \int_0^\infty (\lambda t + \lambda^2 t^2)\, dF(t)^k + (\nu^2 + \omega^2)\int_0^\infty e^{-\lambda t}\, dF(t)^k.\end{aligned}$$

The second integral is at most 1. Because $F(t)^k$ is the distribution function for the maximum of k independent service times $t_1, t_2, \ldots, t_k$, the first integral is an expectation

$$\mathbf{E}\max_{1\le i\le k}\{\lambda t_i + \lambda^2 t_i^2\} \le \mathbf{E}\sum_{i=1}^k (\lambda t_i + \lambda^2 t_i^2) = k\lambda\tau + k\lambda^2(\tau^2 + \sigma^2).$$

Thus, for $k = 1,2,\ldots,$

$$\mathbf{E}\{k_{n+1}^2 | k_n = k\} \le k\{\lambda\tau + \lambda^2\tau^2 + \lambda^2\sigma^2\} + \nu^2 + \omega^2 \le kA. \tag{4.2}$$

Apply Chebyshev's inequality to Eq. (4.2) to get $\Pr\{k_{n+1} \ge l | k_n = k\} \le kA/l^2$. Then also

$$\Pr\{k_{n+1} \ge l | k_n \le k\} \le \sum_{j=1}^{k} \Pr\{k_n = j | k_n \le k\} \frac{jA}{l^2}$$

$$\le \sum_{j=1}^{k} \Pr\{k_n = j | k_n \le k\} \frac{kA}{l^2} = \frac{kA}{l^2}.$$

With $l = k/2$,

$$\Pr\left\{k_{n+1} < \frac{k}{2} \middle| k_n = k\right\} \ge 1 - \frac{4A}{k}$$

and

$$\Pr\left\{k_{n+1} < \frac{k}{2} \middle| k_n \le k\right\} \ge 1 - \frac{4A}{k}.$$

Now consider a stage that serves a large number of customers, say $k_0 = k$, where $2^{r+2}A < k \le 2^{r+3}A$, and r is a positive integer. The bounds established above show that the event $E = \{k_1 < 2^{-1}k, k_2 < 2^{-2}k, \ldots, k_r < 2^{-r}k\}$ has conditional probability

$$\Pr\{E | k_0 = k\} \ge \left(1 - \frac{4A}{k}\right)\left(1 - \frac{8A}{k}\right)\cdots\left(1 - \frac{2^{r+1}A}{k}\right)$$

$$\ge (1 - 2^{-r})(1 - 2^{-r+1})\cdots(1 - 2^{-1})$$

$$\ge \prod_{i=1}^{\infty} (1 - 2^{-i}) = .288788\ldots.$$

When event E occurs, $k_r < 2^{-r}k \le 8A$. Thus, whenever k_n grows to a high value, $k_n = k > 8A$, it returns to a value below $8A$ within $\log_2(k/A)$ stages, with probability at least .288788. From this it follows that each fluctuation to a large value of k_n is followed, with probability 1, by a return below $8A$ eventually. Then $k_n \le 8A$ infinitely often, as claimed. ■

5. BOUNDS

Because vacations follow stages in which no customers arrive, the state probabilities give information about the rate at which the servers leave on vacation. The stationary probability that a randomly chosen stage is followed by a vacation is

$$\Pr\{\textit{vacation}\} = \int_0^\infty e^{-\lambda t}\, dH(t) = G(\lambda) \tag{5.1}$$

by Eq. (2.8). Because a fraction $G(\lambda)$ of the stages precede a vacation, a busy period contains a mean number, $1/G(\lambda)$ of stages.

A stage requires mean time

$$T_S = \int_0^\infty t\, dH(t). \tag{5.2}$$

A relation between T_S and K_S, the expected number served in a stage in statistical equilibrium, follows from Eq. (2.6):

$$K_S = \lambda T_S + \mathbf{E}(e^{-\lambda t})\nu = \lambda T_S + G(\lambda)\nu < \lambda T_S + \nu, \tag{5.3}$$

where $\nu = \sum_{k=1}^\infty k\nu_k$ is the mean number of arrivals during a vacation, and t has the stationary distribution of stage times. Bounds on T_S and K_S in terms of ν alone can be derived as follows. Let $T_S(k)$ denote the mean duration of a stage serving k customers; i.e., $T_S(k)$ is the expected maximum of k service times. Note that

$$T_S(k) = \mathbf{E}\max\{t_1, \ldots, t_k\} \le \mathbf{E}(t_1 + \cdots + t_k) = kT_S(1), \tag{5.4}$$

so the means $T_S(k)$ are finite if the mean service time $T_S(1)$ is. Also $T_S(k)$ is a concave function so long as $T_S(1)$ is finite. For,

$$T_S(k) = \int_0^\infty \{1 - F(t)^k\}\, dt, \tag{5.5}$$

and so, for $k > 1$,

$$\begin{aligned} &T_S(k+1) - T_S(k) - [T_S(k) - T_S(k-1)] \\ &\quad = T_S(k+1) - 2T_S(k) + T_S(k-1) \\ &\quad = -\int_0^\infty \{F(t)^{k+1} - 2F(t)^k + F(t)^{k-1}\}\, dt \\ &\quad = -\int_0^\infty F(t)^{k-1}[1 - F(t)]^2\, dt < 0. \end{aligned}$$

By linear interpolation, or otherwise, we may now extend $T_S(k)$ to a concave function $T_S(x)$ of a real variable x. Then

$$T_S = \sum_{k=1}^\infty p_k T_S(k) \le T_S\left(\sum_{k=1}^\infty kp_k\right) = T_S(K_S). \tag{5.6}$$

Substitution into Eq. (5.3) puts an implicit bound on K_S,

$$K_S < \lambda T_S(K_S) + \nu. \tag{5.7}$$

For service times uniformly distributed on [0,1],

$$T_S(k) = \frac{k}{k+1} \tag{5.8}$$

and, with $T_S(x) = x/(x+1)$, Eq. (5.7) becomes

$$K_S < \frac{\lambda + \nu - 1 + \sqrt{(\lambda + \nu - 1)^2 + 4\nu}}{2}. \tag{5.9}$$

For exponential service times with mean 1,

$$T_S(k) = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{k} \tag{5.10}$$

and Eq. (5.7) gives a bound approximating the solution of $K_S = \lambda \log K_S + \nu$. Numerical results in Section 8 will test the accuracy of the bound Eq. (5.7).

6. WAITING TIMES

The waiting time W of a customer is the time that elapses between his arrival and the start of his service. A related quantity is the number C of other customers that the given customer finds waiting when he arrives. This section examines the state of the system seen by a random arriving customer and finds the probability distribution for C. The mean waiting time then follows immediately from the mean of C.

Stages that serve k customers serve a fraction kp_k/K_S of all customers. Then a random customer has probability kp_k/K_S of being served along with $k-1$ others in the same stage. These k customers all arrived in the preceding stage or vacation, and the random customer is equally likely to have been the first, second, ..., or kth in order of arrival. Thus the random customer found c others waiting with probability

$$\Pr\{C = c\} = \sum_{k=c+1}^{\infty} \frac{kp_k}{K_S} \cdot \frac{1}{k} = \frac{1}{K_S} \sum_{k=c+1}^{\infty} p_k. \tag{6.1}$$

As a check, Eq. (6.1) with $c = 0$ states that a random customer has probability $1/K_S$ of being the first arrival among those served in the same stage.

The mean number waiting is

$$\mathbf{E}[C] = \sum_{c=0}^{\infty} c \Pr\{C = c\} = \sum_{c=0}^{\infty} c \frac{1}{K_S} \sum_{k=c+1}^{\infty} p_k = \sum_{k=1} \frac{k(k-1)}{2K_S} p_k. \tag{6.2}$$

By PASTA [5], $\mathbf{E}[C]$ is the time-averaged mean number of customers in the queue. Now Eq. (6.2) supplies $E[C]$ in Little's formula for the mean waiting time [1, p. 17]:

$$\mathbf{E}[C] = \lambda \mathbf{E}[W]. \tag{6.3}$$

The distribution function for W requires more details about the vacation mechanism than just the $\{\nu_r\}$. Two different models of how vacations occur will help to make that clear.

Random number—At the start of a vacation, a random number $r = 1,2,\ldots$ is chosen with probability ν_r. The vacation then lasts until just after the rth customer arrives, however long that takes.

Random duration—At the start of a vacation a random duration t is chosen with probability distribution $\Phi(t)$. After t units of time have elapsed, if any customers have arrived, the vacation then ends. If not, the vacation is extended another random time. These extensions continue, if necessary, until some customers do arrive.

The two vacation mechanisms give rise to the same vacation arrival distribution if

$$\nu_r = \int_0^\infty \frac{(\lambda t)^r}{r!\,(e^{\lambda t} - 1)}\, d\Phi(t), \qquad r = 1,2,\ldots;$$

i.e.,

$$V(x) = \int_0^\infty \frac{e^{\lambda t x} - 1}{e^{\lambda t} - 1}\, d\Phi(t).$$

Nevertheless, they produce different waiting-time distributions. Random-number vacations give zero waiting time to each customer who is the last to arrive before the start of a busy period. On the other hand, in general, random-duration vacations give no customers zero waiting time.

7. APPROXIMATIONS

To approximate Eq. (3.2) by a finite system one could set $p_{J+1} = p_{J+2} = \cdots = 0$ for some large integer J and solve the J equations (Eq. (3.2)) with $k = 1,2,\ldots,J$, for $p_1,\ldots,p_J$. Another way to get a finite problem might be to approximate the system itself by a loss system having a large finite number M of servers. Whenever the servers find more than M customers in the queue they serve only M customers and make the rest leave without service. This loss system is again a Markov process but with only M states. Equation (2.4) still holds. However, Eq. (2.5) now holds only for $k = 1,2,\ldots,M-1$. For $k = M$,

$$\Pr\{k_{m+1} = M \mid t_m = t\} = e^{-\lambda t} \sum_{m=M}^{\infty} \left\{ \frac{(\lambda t)^m}{m!} + \nu_m \right\}. \tag{7.1}$$

Then Eq. (2.7) and the equations that follow from it require some modifications. For most $F(t)$, approximating $F(t)$ by a step function with a finite number of steps seems the most attractive way to obtain a finite problem. The results for step functions are given below.

Let $F(t)$ be a step function with steps at times $T_1, T_2, \ldots, T_N$,

$$F(t) = \begin{cases} q_0 = 0, & t < T_1, \\ q_1, & T_1 \le t < T_2, \\ \vdots & \vdots \\ q_{N-1}, & T_{N-1} \le t < T_N, \\ q_N = 1, & T_N \le t < \infty. \end{cases} \tag{7.2}$$

Equation (2.9) becomes

$$P(x) = \sum_{r=1}^{N} \{P(q_r) - P(q_{r-1})\} e^{-\lambda T_r} \{e^{\lambda T_r x} + V(x) - 1\}. \tag{7.3}$$

The unknowns $P(q_1), \ldots, P(q_{N-1})$ in Eq. (7.3) may be found by successively replacing x by $q_1, q_2, \ldots, q_{N-1}$ to get a linear system that can be solved for the $P(q_r)$. Since there are only $N-1$ unknowns, the solution is easy when $F(t)$ has few steps. Once the $P(q_r)$ are known, expanding $P(x)$ gives

$$p_k = \sum_{r=1}^{N} \{P(q_r) - P(q_{r-1})\} e^{-\lambda T_r} \left\{ \frac{(\lambda T_r)^k}{k!} + \nu_k \right\}, \qquad k \ge 1. \tag{7.4}$$

As a check, consider Eq. (7.4) with $N = 1$ step of height $q_1 = 1$ at $T_1 = T$ (all service times are exactly T). In Eq. (7.4), $P(q_0) = P(0) = 0$ and $P(q_1) = P(1) = 1$. Then

$$p_k = e^{-\lambda T} \left\{ \frac{(\lambda T)^k}{k!} + \nu_k \right\}, \qquad k \ge 1, \tag{7.5}$$

which is indeed the probability that a stage, lasting for duration T, is followed by a stage that serves k arrivals.

With $N = 2$ steps, $q_1 = q$, and $q_2 = 1$, one must find $P(q_1)$ by substituting $x = q$ in Eq. (7.3). Then Eq. (7.4) holds with $P(q_0) = 0$, $P(q_2) = 1$, and

$$P(q_1) = P(q) = \frac{e^{-\lambda T_2(1-q)} + (V(q) - 1)e^{-\lambda T_2}}{1 - e^{-\lambda T_1(1-q)} + e^{-\lambda T_2(1-q)} - (V(q) - 1)(e^{-\lambda T_1} - e^{-\lambda T_2})}. \tag{7.6}$$

For larger values of $N, P(q_1), \ldots, P(q_{N-1})$ must be obtained by solving the $N-1$ equations obtained from Eq. (7.3). For example, the linear system for the case $V(x) = x$, studied in the next section, can be written

$$\sum_{j=1}^{N-1} (c_{ij} - \delta_{ij}) P(q_i) = e^{-\lambda(1-q_i)T_N} - (1 - q_i)e^{-\lambda T_N}, \qquad 1 \le i \le N-1, \tag{7.7}$$

where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise, and where

$$c_{ij} = e^{-\lambda(1-q_i)T_j} - (1-q_i)e^{-\lambda T_j} - [e^{-\lambda(1-q_i)T_{j+1}} - (1-q_i)e^{-\lambda T_{j+1}}],$$
$$1 \le i \le N-1,\ 1 \le j \le N.$$

Although the solution is more complicated than Eq. (7.6), there is still a simple interpretation for Eq. (7.4). From Eq. (2.4), $H(T_r) = P(F(T_r)) = P(q_r)$ is the probability that a stage has duration T_r or less. Then $P(q_r) - P(q_{r-1})$ in Eq. (7.4) is the probability that a stage has duration exactly T_r. The other factor in the rth term of Eq. (7.4) has the same interpretation as that in Eq. (7.5).

8. NUMERICAL RESULTS

This section examines the system with $V(x) = x$. Discrete approximations are used for the uniform and exponential distributions of service times. For the discrete approximation to $F(t) = t$, $0 \le t \le 1$, put

$$q_i = \frac{i}{N}, \qquad T_i = \frac{2i-1}{2N}, \qquad 1 \le i \le N,$$

which preserves the mean $\frac{1}{2}$. With the $P(q_i)$ determined from Eq. (7.7), the distribution $\{p_k\}$ is obtained from Eq. (7.4).

Table 1 shows the mean K_S and standard deviation σ_S of the number of requests served in a stage, for several values of λ and the discretization parameter N. As can be seen, the results strongly suggest that convergence in N to the results for the continuous model is fast, especially at low to moderate values of λ. The quick convergence of K_S and σ_k to the parameters of a Poisson distribution with mean $\lambda(1 - 1/2N)$ is apparent, as λ becomes large (the discretized

TABLE 1. Moments of Number Served with Uniform $F(t)$[a]

	$N = 5$		$N = 10$		$N = 15$	
λ	K_S	σ_S	K_S	σ_S	K_S	σ_S
.5	1.0372	.2128	1.0375	.2138	1.0375	.2139
1.0	1.1386	.4383	1.1397	.4406	1.1399	.4410
2.0	1.5067	.9048	1.5120	.9113	1.5129	.9125
10.0	1.5304	3.1170	8.7594	3.1774	8.8048	3.1898
(10.0)	(9.0)	(3.0)	(9.5)	(3.0822)	(9.6667)	(3.1091)
20.0	17.8831	4.2856	18.6293	4.4174	18.7967	4.4506
(20.0)	(18.0)	(4.2426)	(19.0)	(4.3589)	(19.3333)	(4.3970)
40.0	35.9940	6.0035	37.9074	6.1876	38.4435	6.2521
(40.0)	(36.0)	(6.0)	(38.0)	(6.1644)	(38.6667)	(6.2182)

[a]Numbers in parentheses refer to the Poisson distribution with mean $\lambda(1 - 1/2N)$.

distribution is supported over an interval of length $1 - 1/2N$). For the original continuous model, the Poisson distribution with mean λ is a good estimate for moderately large λ ($\lambda > 20$ say); with these values of λ the expected number served in a stage is sufficiently large that the expected maximum service time is very close to the maximum possible service time, 1.

We also test the tightness of the upper bound $\hat{K}_S(\lambda)$ on K_S obtained from Eq. (5.7). The bound Eq. (5.9) is

$$K_S \le \hat{K}_S(\lambda) = \frac{\lambda + \sqrt{\lambda^2 + 4}}{2}.$$

The bounds $\hat{K}_S(0.5) = 1.2808$, $\hat{K}_S(20) = 20.0499$, and $\hat{K}_S(40) = 40.025$ compare favorably with K_S presented in Table 1.

To approximate the exponential distribution $F(t) = 1 - e^{-t}$, set

$$T_i = \frac{i}{2}, \quad 1 \le i \le N,$$

$$q_i = 1 - e^{-(T_i + 1/4)}, \qquad 1 \le i \le N - 1, \qquad q_N = 1.$$

The approximation, shown in Figure 1, is rather coarse for small t. N and λ should be moderately large for good accuracy. For small λ, more points T_i are probably needed in the region out to the knee of the curve but fewer beyond the

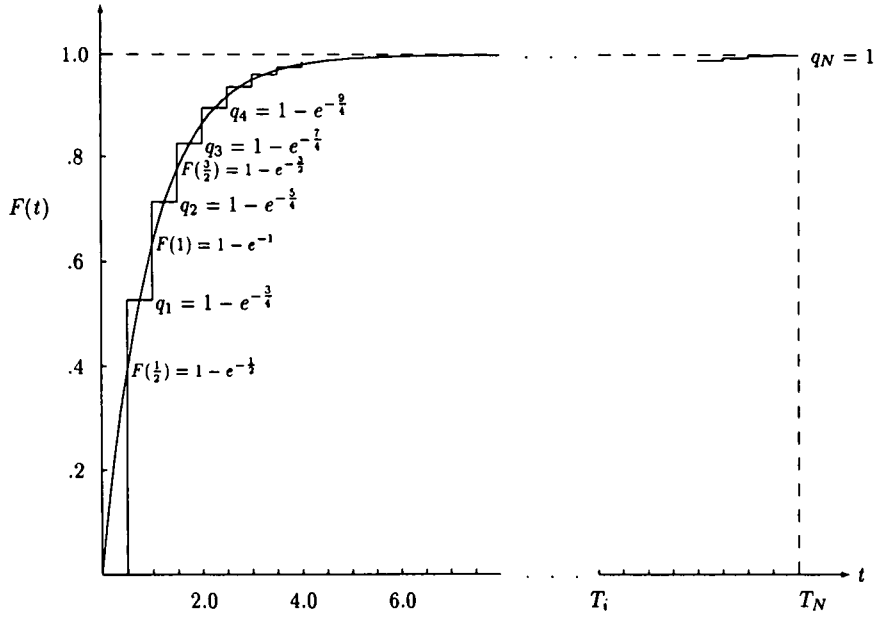

FIGURE 1. Step function approximation of $F(t) = 1 - e^{-t}$.

knee. Table 2 gives the results K_S and σ_K for several values of λ and N. Clearly, for given λ and a desired accuracy, the exponential distribution needs a finer discretization (larger N) than the uniform distribution. Numerical values of $\hat{K}_S(\lambda)$ from Eqs. (5.7) and (5.10) are $\hat{K}_S(2) = 6$, $\hat{K}_S(4) = 14$, $\hat{K}_S(6) = 24$, $\hat{K}_S(8) = 34$, and $\hat{K}_S(10) = 45$, which Table 2 shows to be reasonably accurate for large λ.

The distributions $\{p_k\}$ for $N = 20$ are plotted in Figure 2 with smooth interpolating curves to bring out the change in shape as the parameter λ increases.

Some features of heavy traffic behavior are already apparent, even for the modest values of λ shown. The distribution $\{p_k\}$ is fairly peaked. The value $K_{\max}$ of k, where p_k is maximized, is shown in Table 3. The corresponding modal value $p_{\max} = \sup_k p_k$ is also shown. A rough calculation gives the following estimate $\hat{K}$ for the value of $K_{\max} = K_{\max}(\lambda)$. For a stage with $k \gg 1$ jobs,

$$\Pr[\textit{stage length} \le \log k + c] = \left(1 - \frac{e^{-c}}{k}\right)^k \sim \exp(-e^{-c}),$$

which rises sharply from 0 for $c \ll 0$ to 1 for $c \gg 0$, with most of the transition occurring in the neighborhood of $c = 0$. Thus for $k \gg 1$, stages with k customers usually persist for times near $\log k$. Then the next stage contains very nearly $\lambda \log k$ jobs. In heavy traffic equilibrium therefore, we expect a stage to contain close to $\hat{K}$ jobs where $\lambda \log \hat{K} = \hat{K}$. $\hat{K}$ is also shown in Table 3. It agrees rather well with $K_{\max}$ even for λ as low as 4.

The solution of Eq. (7.7) gives $N - 1$ points of the stationary distribution of stage duration, $H(T_i) = P(q_i)$, $1 \le i \le N - 1$. These distributions are shown in Table 4.

9. POLLING WITH CONSTANT SERVICE

The gated infinite-server system in a general polling setting is difficult to analyze, owing to the dependence between the vacation times and busy periods at

Table 2. Moments of Number Served with Exponential $F(t)$

	$N = 5$		$N = 10$		$N = 15$	
λ	K_S	σ_S	K_S	σ_S	K_S	σ_S
2.0	3.7620	2.9372	3.8198	3.0727	3.8244	3.0899
4.0	11.3296	5.8100	11.6708	6.3833	11.6983	6.4660
6.0	20.4504	7.7787	21.3445	9.0762	21.4191	9.2834
8.0	30.0177	9.3898	31.7244	11.6204	31.8714	12.0063
10.0	40.8716	10.7525	42.6433	14.0770	43.8756	14.4278

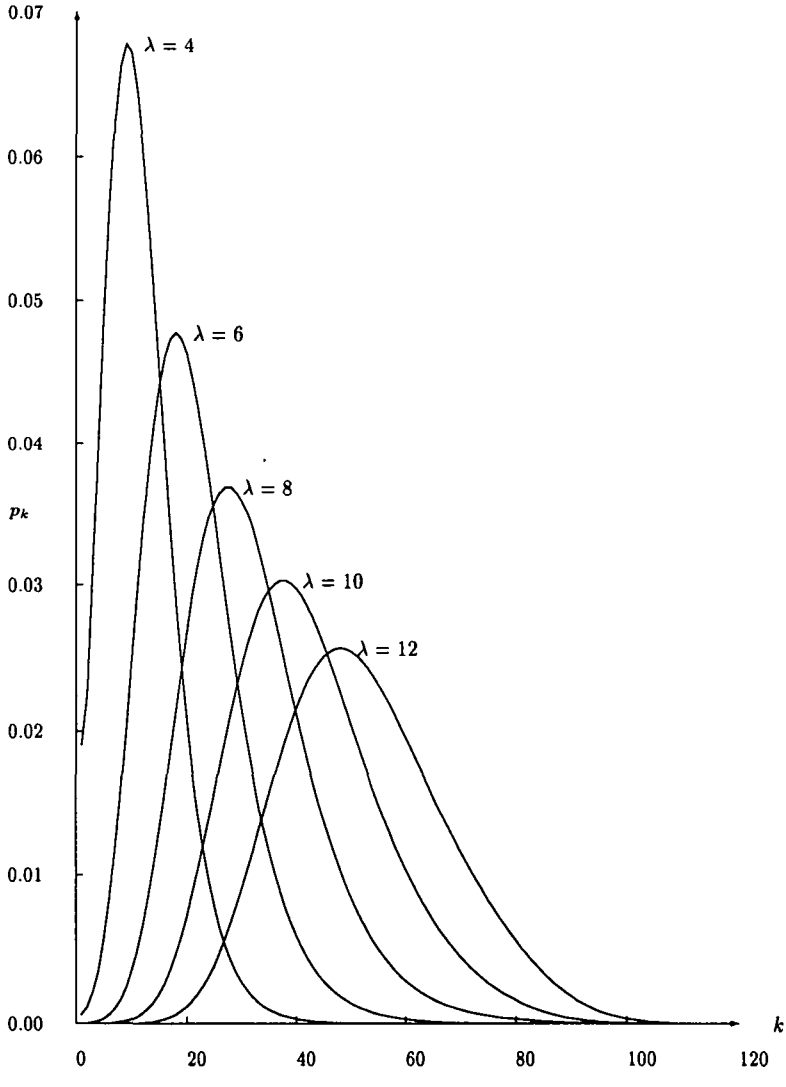

FIGURE 2. Distribution of $\{p_k\}$ for unit exponential $F(t)$, $N = 20$.

any one of the stations. However, the case of two identical stations and constant service times T is simple enough to analyze completely, as shown later. Besides illustrating the difficulty of the general problem, this section shows that different polling rules can dramatically affect waiting times. The vacation-time distribution is no longer prescribed but will be deduced analytically.

Three new service rules will be compared with exhaustive service. These rules choose the next queue to be served as follows:

TABLE 3. Mode of $\{p_k\}$ for Unit Exponential Distribution $F(t)$, $N = 20$

λ	K_{max}	$\hat{K}$	p_{max}
4.0	9	8.6132	.067763
6.0	18	16.9989	.047710
8.0	27	26.0935	.036948
10.0	38	35.7715	.030260
12.0	48	45.9238	.025684

- Rule 1 chooses one of the occupied queues at random.
- Rule 2 chooses the other queue if it is occupied.
- Rule 3 chooses the queue with the customer who has waited longest.

The basis of comparison will be the mean wait for service of a typical customer. Mean waiting times W_0, W_1, W_2, and W_3 for exhaustive service and service by rules 1, 2, and 3, respectively, will be derived. The calculations will show that

TABLE 4. $H(t)$ for Unit Exponential Distribution, with $N = 20$

	λ				
t	2.0	4.0	6.0	8.0	10.0
.5	.2265	.0287	.0020	.0001	.0000
1.0	.3913	.0834	.0119	.0015	.0002
1.5	.5506	.1876	.0494	.0123	.0030
2.0	.6868	.3364	.1399	.0559	.0217
2.5	.7915	.5006	.2853	.1578	.0856
3.0	.8657	.6490	.4557	.3133	.2122
3.5	.9154	.7656	.6146	.4864	.3810
4.0	.9474	.8488	.7414	.6417	.5515
4.5	.9677	.9047	.8328	.7621	.6943
5.0	.9802	.9408	.8945	.8473	.8003
5.5	.9879	.9636	.9344	.9040	.8732
6.0	.9927	.9777	.9596	.9405	.9208
6.5	.9955	.9864	.9753	.9634	.9511
7.0	.9973	.9917	.9849	.9776	.9700
7.5	.9984	.9950	.9908	.9864	.9817
8.0	.9990	.9970	.9944	.9917	.9889
8.5	.9994	.9981	.9966	.9950	.9932
9.0	.9996	.9989	.9979	.9969	.9959
9.5	.9998	.9993	.9988	.9981	.9975

$W_3 < W_2 < W_1 < W_0$. Little's theorem [1, p. 17] then shows that the mean queue lengths for the four service rules are also ordered in the same way.

All stages now have a duration equal to the constant service time T, regardless of the number served; the system can be analyzed as a Markov chain with five states called (0,0), (1*,0), (0,1*), (1*,1), and (1,1*). The numbers in each pair show whether a queue is empty (0) or occupied (1), the asterisk denoting which queue is being served. A discrete Markov chain is formed by the states at the beginnings of new stages or vacations (most vacations are stages for some queue, but (0,0) represents a vacation for both). The transition probabilities depend on the service rule but the main parameter will be $q = e^{-\lambda T}$ (the probability that a queue has no arrivals during a service time). We set $p = 1 - q$.

All four service rules treat the two queues alike, a symmetry that simplifies the equilibrium state probabilities. In steady state, the probability of being in either of the states (1*,1) or (1,1*) is the same and denoted by a. Similarly, the probabilities of (1*,0) and (0,1*) are the same and denoted by b. Finally, state (0,0) has probability

$$c = 1 - 2a - 2b. \tag{9.1}$$

The probabilities c, $2b$, and $2a$ of 0, 1, or 2 queues being occupied satisfy the same transition equations

$$\begin{aligned} c &= 2q^2 b \\ a &= pa + p^2 b \end{aligned} \tag{9.2}$$

for all four service rules. Then, Eq. (9.1) shows that

$$\begin{aligned} a &= \frac{p^2}{2(1 - q + q^2 + q^3)} \\ b &= \frac{q}{2(1 - q + q^2 + q^3)} \\ c &= \frac{2q^3}{2(1 - q + q^2 + q^3)}. \end{aligned} \tag{9.3}$$

9.1. Waiting Times

The elapsed time between most transitions equals the service time T. However, transitions from state (0,0) are different, occurring after a mean wait $1/2\lambda$ for the next arrival to either queue. The mean time between transitions is then

$$\begin{aligned} \tau &= 2aT + 2bT + \frac{c}{2\lambda} \\ &= \frac{2\lambda(1 - q + q^2)T + q^3}{2\lambda(1 - q + q^2 + q^3)}. \end{aligned} \tag{9.4}$$

Now consider a customer arriving at random, say to the first queue. His or her mean wait W_i $(i = 0,1,2,3)$ will be a sum

$$W_i = \sum_S R(S) w_i(S), \tag{9.5}$$

where the sum is over the five states S. $R(S)$ denotes the probability that the last transition before the customer arrived was to state S and $w_i(S)$ is the customer's expected wait, given S. Note that the state the customer finds at his or her time of arrival may differ from S because of earlier arrivals.

Because the states persist for different lengths of time, $R(S)$ is not just the probability of S at a transition (which is given by Eq. (9.3)). Instead,

$$\begin{aligned} R(0,0) &= \frac{c}{2\lambda\tau} \\ R(1^*,0) = R(0,1^*) &= \frac{bT}{\tau} \\ R(1^*,1) = R(1,1^*) &= \frac{aT}{\tau}. \end{aligned} \tag{9.6}$$

In general, the customer's conditional expected wait $w_i(S)$ depends on the service rule (indexed by i), although $w_i(0,0) = 0$ for all i. For other states S, the waiting time is a sum $t_0 + LT$, where t_0 is the wait for the current stage to end and L is the number of extra stages the job waits before service begins.

With exhaustive service, $L = 0$ if the customer finds the first queue being served so

$$w_0(1^*,0) = w_0(1^*,1) = \mathbf{E}(t_0) = \frac{T}{2}.$$

If the second queue is being served, the customer must also wait for a stage in which the second queue has no arrivals. Then L has a geometric distribution with mean p/q:

$$w_0(0,1^*) = w_0(1,1^*) = \frac{T}{2} + \frac{Tp}{q}.$$

Now that $w_0(S)$ is known, Eqs. (9.4), (9.5), and (9.6) combine to give the mean wait W_0. Similar arguments give W_1, W_2, and W_3 from

$$w_1(1^*,0) = w_1(0,1^*) = w_1(1,1^*) = \frac{T(3-q)}{2(1+q)},$$

$$w_1(1^*,1) = \frac{T(3+q)}{2(1+q)};$$

$$w_2(0,1^*) = w_2(1,1^*) = \frac{T}{2},$$

$$w_2(1^*,0) = \frac{T(3-q)}{2};$$

$$w_3(1,1^*) = \frac{T}{2},$$

$$w_3(1^*,1) = \frac{3T}{2},$$

$$w_3(0,1^*) = w_3(1^*,0) = T - \frac{(1-q^2)}{4\lambda}.$$

Formulas for the W_i are

$$\begin{aligned} W_0 &= \frac{T}{2q}\left(1 - \frac{q^3}{d}\right) \\ W_1 &= \frac{\lambda T^2(3p + 2q^2)}{(1+q)d} \\ W_2 &= \frac{\lambda T^2(2p + q^2)}{d} \\ W_3 &= T\left(1 - \frac{q + q^2}{2d}\right), \end{aligned} \tag{9.7}$$

where

$$d = 2\lambda T(1 - q + q^2) + q^3.$$

Only the argument for $w_3(0,1^*)$ and $w_3(1^*,0)$ merits comment. The arrival time t, measured from the last transition, is uniformly distributed over the interval $[0,T]$. If the customer is the first to arrive to either queue, then his service begins after the wait $t_0 = T - t$. That event has probability $e^{-2\lambda t}$. Otherwise, another customer arrived first and is equally likely to be in either queue. Then the waiting time is $T - t$ or $2T - t$, each with probability $\{1 - \exp(-2\lambda t)\}/2$. The mean wait for these states is then found by averaging over t.

Table 5 illustrates numerical waiting times from Eq. (9.7). The factor $1/q$ in W_0 grows exponentially with the arrival rate and makes W_0 huge, for large values of λT. With Rules 2 and 3, no customer waits longer than $2T$. Their mean waits, and even the mean wait with Rule 1, remain bounded.

9.2. Vacation Distribution

Previous sections regard the vacation distribution $\{v_k\}$ as prescribed. But in polling applications, the service time distribution $F(t)$ and the arrival rate de-

TABLE 5. Mean Waiting Times for Constant Service Polling Disciplines

λT	W_0/T	W_1/T	W_2/T	W_3/T
0.2	0.2339	0.2330	0.2322	0.1629
0.5	0.6375	0.6058	0.5865	0.5051
1.0	1.3164	0.9997	0.8832	0.8412
2.0	3.6919	1.3111	0.9889	0.9783
4.0	27.2991	1.4729	0.9998	0.9988
6.0	201.7144	1.4963	1.0000	0.9999
8.0	1490.4790	1.4995	1.0000	1.0000

termine $\{\nu_k\}$ automatically. The same five-state Markov chain that entered into the derivation of mean waiting times will now be used to obtain $\{\nu_k\}$ for exhaustive polling of two queues with constant service time. We consider a vacation that follows a busy period of the first queue.

The busy period ended in a stage that began in state (1*,1) or (1*,0). Since stages that begin in these two states all have probability q of being the last stage of the period, busy periods end in these two kinds of stages with probabilities $a' = a/(a+b)$ and $b' = b/(a+b)$. As Section 1 remarked, the end of the busy period has an influence on the vacation. Thus, the vacation could be a period of service to neither queue if the state at the start of the final stage was (1*,0), but not if it was (1*,1). $V(x)$ will be a suitably weighted average of generating functions for states (1*,1) and (1*,0).

States during the vacation may be (1,1*), (0,1*), or (0,0). The vacation ends with a transition to (1*,0). Consider the number of arrivals to queue 1 during the rest of the vacation, starting just after a transition to (1,1*). Let $A(x)$ be the generating function for the probability distribution of that number of arrivals. Also let $B(x)$ and $C(x)$ be similar generating functions for (0,1*) and (0,0). Equations for $A(x)$, $B(x)$, and $C(x)$ follow from an analysis of transitions. A stage starting at (1,1*) adds a Poisson-distributed number of arrivals to queue 1 and ends with a transition to (1,1*) or (1*,0) (end of vacation) with probabilities p and q, respectively. Thus,

$$A(x) = (pA(x) + q)e^{\lambda T(x-1)}.$$

In the same way,

$$B(x) = pq(e^{\lambda Tx} - 1)A(x) + pqB(x) + q^2C(x) + q^2(e^{\lambda Tx} - 1)$$

and

$$C(x) = [B(x) + x]/2.$$

These equations may be solved for $A(x)$, $B(x)$, and $C(x)$.

If the stage before the vacation began started in state $(1^*,1)$, an event of probability a', then the vacation started in state $(0,1^*)$. Otherwise, with probability b', the last stage started in state $(1^*,0)$; the vacation then began in state $(0,1^*)$ with probability p or in state $(0,0)$ with probability q. The number of arrivals during the vacation is distributed with the generating function

$$\begin{aligned} V(x) &= a'B(x) + b'[pB(x) + qC(x)] \\ &= \frac{q^2}{1-q+q^2}\left\{\frac{e^{\lambda Tx}-1}{1-pqe^{\lambda Tx}} + x\right\}. \end{aligned} \tag{9.8}$$

Expanding Eq. (9.8) in powers of x and matching coefficients gives an explicit expression for v_k, but it is an infinite series. For computing purposes, a more convenient recurrence is obtained by multiplying both sides of Eq. (9.8) by the denominator $1 - pqe^{\lambda Tx}$ before expanding:

$$\begin{aligned} v_0 &= 0, \\ v_1 &= \frac{q^2}{1-pq}\left(1 + \frac{\lambda T}{1-pq}\right), \\ v_k &= \frac{pq}{1-pq}\sum_{j=1}^{k-1}\frac{(\lambda T)^{k-j}}{(k-j)!}v_j \\ &\quad + \frac{q^2}{(1-pq)^2}\left\{\frac{(\lambda T)^k}{k!} - pq\frac{(\lambda T)^{k-1}}{(k-1)!}\right\}, \qquad k = 2,3,\ldots. \end{aligned} \tag{9.9}$$

The mean number of vacation arrivals is

$$v = V'(1) = \frac{\lambda T}{q} + \frac{q^2}{1-pq}. \tag{9.10}$$

One use for $\{v_k\}$ is to predict how much storage space the queue will need. Of course, the storage must hold the Poisson-distributed number that arrives during a service stage. But at high arrival rates the queue tends to be much longer after a vacation. The number served in a stage has the probability distribution in Eq. (7.5). Using Eqs. (7.5) and (9.9) with $\lambda T = 1$, one finds that the queue length exceeds 3 after 2% of the stages but exceeds 10 after 2% of the vacations. For $\lambda T = 2$, the queue lengths at the 2% level become 5 and 56.

The queuing system of Sections 2–8 differed from the present polling systems in having vacations independent of busy periods. Just how much difference that assumption makes can now be tested by comparing the mean waiting time W_0 for exhaustive polling with the mean waiting time $\mathbf{E}(W)$ of Section 6, with $V(x)$ being the vacation distribution in Eq. (9.8). The mean wait obtained

from Eq. (6.3) was $\mathbf{E}[W] = P''(1)/(2\lambda K_S)$. With constant service times, and ν given by Eq. (9.10), Eq. (5.3) becomes

$$K_S = \lambda T_S + \nu G(\lambda) = \lambda T + \nu q = 2\lambda T + \frac{q^3}{1 - pq},$$

while Eq. (3.1) reduces to

$$P(x) = e^{-\lambda T(1-x)} + [V(x) - 1]e^{-\lambda T}.$$

After substituting Eq. (9.8) for the vacation distribution $V(x)$ and differentiating the expression for $P(x)$, the result is

$$\mathbf{E}[W] = \frac{(1 - pq)\lambda T^2}{q[2\lambda T(1 - pq) + q^3]}. \tag{9.11}$$

But now, rearranging terms shows that $\mathbf{E}[W]$ is exactly W_0 of Eq. (9.7).

References

1. Kleinrock, L. (1975). *Queueing systems*, Vol. 1. New York: Wiley.
2. Kress, R. (1989). *Linear integral equations*, Vol. 82. *Applied Mathematical Sciences*. New York: Springer-Verlag.
3. Lubachevsky, B.D. (1989). Stability of the bounded-lag distributed discrete event simulation. In B. Unger & R. Fujimoto (eds.), *Distributed simulation 1989*. San Diego: Society for Computer Simulation International, pp. 100-107.
4. Takagi, H. (1986). *Analysis of polling systems*. Cambridge, MA: MIT Press.
5. Wolff, R. (1982). Poisson arrivals see time averages. *Operations Research* 30: 232-233.