

Queues Served in Cyclic Order

By R. B. COOPER and G. MURRAY*

(Manuscript received September 30, 1968)

We study two models of a system of queues served in cyclic order by a single server. In each model, the i th queue is characterized by general service time distribution function $H_i(\cdot)$ and Poisson input with parameter λ_i .

In the exhaustive service model, the server continues to serve a particular queue until the server becomes idle and there are no units waiting in that queue; at this time the server advances to and immediately starts service on the next nonempty queue in the cyclic order.

The gating model differs from the exhaustive service model in that when the server advances to a nonempty queue, a gate closes behind the waiting units. Only those units waiting in front of the gate are served during this cycle, with the service of subsequent arrivals deferred to the next cycle.

We find expressions for the mean number of units in a queue at the instant it starts service, the mean cycle time, and the Laplace-Stieltjes transform of the cycle time distribution function.

I. INTRODUCTION

We consider a system of queues served in cyclic order by a single server. The i th queue is characterized by general service time distribution function $H_i(\cdot)$ and Poisson input with parameter λ_i .

We study two variations of this model. In the first, called the exhaustive service model, the process begins with the arrival of a unit at some queue, say A , when the system is otherwise empty. The server begins on this unit immediately, and continues to serve queue A until for the first time the server becomes idle and there are no units waiting in queue A . The server then looks at the next queue in the cyclic order, queue $A + 1$, and serves those units, if any, that have accumulated during the serving period of queue A . The server continues to serve queue $A + 1$ until for the first time the server

* The RAND Corporation.

becomes idle and there are no units waiting in queue $A + 1$. The process continues in this manner, with the queues being served in cyclic order, until for the first time the system becomes completely empty. The process is then re-initiated by the arrival of the next unit. No time is required to switch from one queue to the next.

The second variation, called the gating model, differs from the first in the following way: When the server moves to a queue with at least one waiting unit, the server accepts only those units that were waiting when the server arrived, deferring service of all subsequent arriving units until the next cycle. That is, in the gating model, at the instant the server advances to a nonempty queue a gate closes behind the waiting units, and only those units waiting in front of the gate are served during that cycle.

The exhaustive service model is analyzed in detail. We obtain the generating function for the joint probability distribution of the number of units in each queue at an instant at which the server finishes serving any one of the queues. We then obtain expressions for the mean number of units in a queue at the instant it starts service, the mean cycle time, and, in a form suitable for numerical computation, the Laplace-Stieltjes transform of the cycle time distribution function.

Finally, we note that the equations describing the gating model differ only trivially from those of the exhaustive service model, and that the same method of solution applies to each.

Systems in which a single server is shared among several queues are common. For example, in the No. 1 Electronic Switching System the central control spends much of its time polling various hoppers and performing work requests that it finds in these hoppers. Similarly, in a time-shared computer system the users have access through teletypewriters to a central computer which is shared among them. The cyclic queueing models studied here are of a type which may be useful in the analyses of these and similar problems.

The exhaustive service model for the special case of two queues has been studied by L. Takács,¹ B. Avi-Itzhak, W. L. Maxwell, and L. W. Miller,^{2,3} and M. F. Neuts and M. Yadin.⁴ Avi-Itzhak *et al* used an argument based on the properties of mean values, to obtain an expression for the mean waiting time suffered by a unit in either queue. Takács, by a more direct argument, utilizing the Markov chain imbedded at the epochs of service completion, obtained the corresponding Laplace-Stieltjes transform and formulas for the waiting time moments assuming service in order of arrival. Neuts and Yadin obtained waiting time results for the transient case.

The more general two-queue model in which the time required to switch from one queue to the next has some arbitrary distribution function is also studied in Ref. 3, and has been investigated in addition by M. Eisenberg⁵ and J. S. Sykes.⁶ M. A. Leibowitz^{7,8} has studied a multiqueue model similar to the gating model studied here. A nonprobabilistic approach to cyclic queueing problems has been used by J. B. Kruskal.⁹

In the present paper we use the imbedded Markov chain approach but, as with Neuts and Yadin, our chain is imbedded at the instants at which the server completes serving a queue, rather than at the set of all instants of service completion used by Takács. Whereas Takács and Neuts and Yadin obtained waiting time results, our analysis yields cycle time results. The mathematical analyses characterizing the three approaches share some common ground, although the differences, especially those arising from our consideration of an arbitrary number of queues, are significant.

Also, in a recent nontechnical article on queues by Leibowitz,¹⁰ the present problem is offered as a prime example of an important, difficult, unsolved queueing problem.

II. PRELIMINARIES

In the analysis of the exhaustive service model, we take the number of queues to be $N + 1 \geq 2$. Units arrive at the i th queue according to the Poisson process with rate λ_i ; that is, the probability $Q_i(k; t)$ that k units arrive at the i th queue in an interval of length t is

$$Q_i(k; t) = \frac{(\lambda_i t)^k}{k!} \exp(-\lambda_i t) \quad (k = 0, 1, 2, \dots; i = 0, 1, \dots, N).$$

The length of time required to serve a unit from queue i has distribution function $H_i(\cdot)$ with mean h_i ($i = 0, 1, \dots, N$).

In the analysis of the exhaustive service model, we shall use the concept of busy period, discussed at length by Takács¹¹. For the ordinary single-server queue, the busy period is defined as the length of time from the instant a unit enters a previously empty system until the next instant at which the system is completely empty. Both the distribution function of the busy period and its Laplace-Stieltjes transform are known explicitly for the $M/G/1$ queue. In particular, the $M/G/1$ queue with arrival rate λ and mean service time h has a busy period with mean $b = h/(1 - \lambda h)$ if $\lambda h < 1$ and $b = \infty$ if $\lambda h \geq 1$.

Consider now the $M/G/1$ queue with j waiting units; define the

j -busy period as the length of time from the instant at which service starts on the first of the j units until the next instant at which the system is completely empty. (When $j = 1$, the j -busy period and the busy period are identical.) Each of the j units, which together generate a j -busy period, individually generates a 1-busy period. Thus (as Takács shows) the distribution function of the j -busy period is the j -fold convolution with itself of the distribution function of the 1-busy period.

Denote by $B_i(\cdot)$ the distribution function of a 1-busy period for queue i , by $\beta_i(\cdot)$ its Laplace-Stieltjes transform, and by $b_i = h_i/(1 - \lambda_i h_i)$ its mean. Let $B_i^{*j}(\cdot)$ be the j -fold convolution of $B_i(\cdot)$ with itself, $B_i^{*1}(\cdot) = B_i(\cdot)$. Then a j -busy period for the i th queue has distribution function $B_i^{*j}(\cdot)$ and Laplace-Stieltjes transform $(\beta_i(\cdot))^j$.

III. FORMULATION OF IMBEDDED MARKOV CHAIN STATE EQUATIONS FOR THE EXHAUSTIVE SERVICE MODEL

There are $N + 1 \geq 2$ queues. Suppose that the system is idle and a unit arrives at some queue at epoch τ_0 . The server immediately commences service at that queue, and continues to serve units at that queue until the first instant τ_1 at which that queue becomes empty. If the system is not empty at τ_1 , the server advances to the next queue in the cyclic order. The server immediately commences work at this queue until the instant τ_2 at which this queue becomes empty (where $\tau_2 = \tau_1$ if the server finds the queue empty), and continues on in this manner until for the first time, τ_n say, the server finishes serving a queue and there are no units waiting anywhere in the system. The process terminates at τ_n and is reinitiated by the next arrival.

Thus the process generates a set of points $\tau_0, \tau_1, \dots, \tau_n$, where τ_0 is the arrival instant of a unit at some queue in the previously empty system, and τ_n is the first instant at which the system becomes completely empty again. The next arrival, at epoch τ_0' say, reinitiates the process, and a new set of points, $\tau_0', \tau_1', \dots, \tau_n'$ is generated. We call the points τ_1, \dots, τ_n (and τ_1', \dots, τ_n') switch points.

Note that τ_0 is not a switch point, whereas τ_n is a switch point. Successive switch points may occur simultaneously in time, but are nevertheless considered distinct. Thus, with each switch point is associated a queue, namely, that queue at which the server has just completed its visit.

When the server finishes serving a queue and finds the system completely empty, a switch point associated with that queue is recorded. The next switch point is recorded when the server leaves the queue at which the process is reinitiated, and is associated with that queue.

Let $(i; n_1, \dots, n_N)$ denote the state of the system at an arbitrary switch point, where i is the index of the associated queue, and n_k is the number of units waiting in queue $i + k$ ($k = 1, \dots, N$). [For simplicity, no special notation will be used to denote arithmetic mod $(N + 1)$.] Let the state $(i; n_1, \dots, n_N)$ have probability $P_i(n_1, \dots, n_N)$; that is, $P_i(n_1, \dots, n_N)$ is the joint probability that at a switch point, the server has just completed a visit to queue i ($i = 0, 1, \dots, N$) and n_1 units are waiting in queue $i + 1$, n_2 units in queue $i + 2$, \dots , and n_N units in queue $i + N$.

The state $(i; n_1, \dots, n_N)$ can occur through the following exhaustive and mutually exclusive contingencies:

- (i) The server leaves queue $i - 1$ and finds $j \geq 1$ units waiting for service in queue i , where it thus spends a length of time equal to a j -busy period.
- (ii) The server leaves queue $i - 1$ and finds $j = 0$ units waiting for service in queue i , but at least one unit waiting for service somewhere else in the system, so that the server then "passes through" queue i in zero time. [That is, the state $(i; n_1, \dots, n_{N-1}, 0)$ necessarily follows the state $(i - 1; 0, n_1, \dots, n_{N-1})$ where at least one of the $n_k \neq 0$ ($k = 1, \dots, N - 1$).]
- (iii) The server leaves some queue and finds no units waiting anywhere in the system. With probability λ_i/λ ($\lambda = \lambda_0 + \dots + \lambda_N$) the next arrival (which reinitiates the process) occurs at queue i , where the server then spends a 1-busy period.

These considerations lead directly to the imbedded (at the switch points) Markov chain probability state equations:

$$\begin{aligned}
 &P_i(n_1, \dots, n_N) \\
 &= \sum_{j=1}^{\infty} \sum_{k_1=0}^{n_1} \dots \sum_{k_{N-1}=0}^{n_{N-1}} P_{i-1}(j, k_1, \dots, k_{N-1}) \\
 &\quad \cdot \int_0^{\infty} \prod_{m=1}^{N-1} Q_{i+m}(n_m - k_m; t) Q_{i+N}(n_N; t) dB_i^{*i}(t) \\
 &\quad + P_{i-1}(0, n_1, \dots, n_{N-1}) \left(1 - \delta\left(\sum_{m=1}^{N-1} n_m\right) \right) \delta(n_N) \\
 &\quad + \frac{\lambda_i}{\lambda} \sum_{k=0}^N P_k(0, \dots, 0) \int_0^{\infty} \prod_{m=1}^N Q_{i+m}(n_m; t) dB_i(t) \\
 &\quad \cdot \left[\delta(x) = \begin{cases} 1 & \text{if } x = 0; \\ 0 & \text{if } x \neq 0 \end{cases}; \quad i = 0, 1, \dots, N \right]. \quad (1)
 \end{aligned}$$

[Throughout the analysis, arithmetic mod $(N + 1)$ in subscripts will not be specially denoted.] Assuming it exists, the distribution $\{P_i(n_1, \dots, n_N)\}$ is uniquely determined by (1) and the normalization equation

$$\sum_{i=0}^N \sum_{n_1=0}^{\infty} \cdots \sum_{n_N=0}^{\infty} P_i(n_1, \dots, n_N) = 1. \quad (2)$$

(Intuitively, one would expect a unique stationary distribution to exist when $\sum_{i=0}^N \lambda_i h_i < 1$.)

IV. FUNCTIONAL EQUATIONS FOR GENERATING FUNCTIONS

We define the probability generating functions $g_i(x_1, \dots, x_N)$:

$$g_i(x_1, \dots, x_N) = \sum_{n_1=0}^{\infty} \cdots \sum_{n_N=0}^{\infty} P_i(n_1, \dots, n_N) x_1^{n_1} \cdots x_N^{n_N} \quad (i = 0, 1, \dots, N). \quad (3)$$

Substitution of (1) into (3) yields, after some rearrangement,

$$\begin{aligned} g_i(x_1, \dots, x_N) &= \sum_{j=1}^{\infty} \sum_{k_1=0}^{\infty} \cdots \sum_{k_{N-1}=0}^{\infty} P_{i-1}(j, k_1, \dots, k_{N-1}) x_1^{k_1} \cdots x_{N-1}^{k_{N-1}} \\ &\quad \cdot \int_0^{\infty} \exp\left(-t \sum_{m=1}^N \lambda_{i+m}(1 - x_m)\right) dB_i^{*i}(t) \\ &\quad + \sum_{n_1=0}^{\infty} \cdots \sum_{n_N=0}^{\infty} \left(1 - \delta\left(\sum_{m=1}^{N-1} n_m\right)\right) \\ &\quad \cdot \delta(n_N) P_{i-1}(0, n_1, \dots, n_{N-1}) x_1^{n_1} \cdots x_N^{n_N} \\ &\quad + \frac{\lambda_i}{\lambda} \sum_{k=0}^N P_k(0, \dots, 0) \int_0^{\infty} \exp\left(-t \sum_{m=1}^N \lambda_{i+m}(1 - x_m)\right) dB_i(t) \end{aligned} \quad (i = 0, 1, \dots, N). \quad (4)$$

The integrals on the right side of (4) are recognized as the Laplace-Stieltjes transform $(\beta_i(\cdot))^j$ of the j -busy period distribution function with argument $\sum_{m=1}^N \lambda_{i+m}(1 - x_m)$. Hence (4) yields the set of simultaneous functional equations

$$\begin{aligned} g_i(x_1, \dots, x_N) &= g_{i-1}\left(\beta_i\left(\sum_{m=1}^N \lambda_{i+m}(1 - x_m)\right), x_1, \dots, x_{N-1}\right) \\ &\quad + \frac{\lambda_i}{\lambda} \beta_i\left(\sum_{m=1}^N \lambda_{i+m}(1 - x_m)\right) \sum_{k=0}^N P_k(0, \dots, 0) \\ &\quad - P_{i-1}(0, \dots, 0) \quad (i = 0, 1, \dots, N). \end{aligned} \quad (5)$$

V. SOLUTION OF THE FUNCTIONAL EQUATIONS

For notational convenience, we define the nesting operator Ξ for any sequence of functions $\{f_k(\cdot)\}$ for which it is meaningful:

$$\Xi_{k=0}^n f_k(x) = f_n(\cdots(f_2(f_1(f_0(x))))\cdots).$$

We shall denote by \mathbf{x} the vector with components x_1, \dots, x_N , and by $\mathbf{0}$ the vector with all components zero. Both vectors and vector-valued functions will be denoted by boldface type, and square brackets will be used to enclose vector arguments of vector-valued functions. Finally, we will denote by $\phi(\mathbf{v})$ the first component of a vector \mathbf{v} .

Define the vector functions

$$\mathbf{Z}_i[x_1, \dots, x_N] = \left[\beta_i \left(\sum_{m=1}^N \lambda_{i+m} (1 - x_m) \right), x_1, \dots, x_{N-1} \right] \\ (i = 0, 1, \dots, N) \quad (6)$$

so that (5) can be rewritten

$$g_i(\mathbf{x}) = g_{i-1}(\mathbf{Z}_i[\mathbf{x}]) + \frac{\lambda_i}{\lambda} \phi(\mathbf{Z}_i[\mathbf{x}]) \sum_{k=0}^N P_k(\mathbf{0}) - P_{i-1}(\mathbf{0}) \\ (i = 0, 1, \dots, N). \quad (7)$$

Iterating $\nu - 1$ times on i in (7) we obtain

$$g_i(\mathbf{x}) = g_{i-\nu} \left(\Xi_{k=0}^{\nu-1} \mathbf{Z}_{i-k}[\mathbf{x}] \right) + \frac{1}{\lambda} \sum_{k=0}^N P_k(\mathbf{0}) \sum_{m=0}^{\nu-1} \lambda_{i-m} \phi \left(\Xi_{k=0}^m \mathbf{Z}_{i-k}[\mathbf{x}] \right) \\ - \sum_{m=1}^{\nu} P_{i-m}(\mathbf{0}) \quad (i = 0, 1, \dots, N; \nu = 1, 2, \dots). \quad (8)$$

In particular, when $\nu = N + 1$ (8) can be written

$$g_i \left(\Xi_{k=0}^N \mathbf{Z}_{i-k}[\mathbf{x}] \right) - g_i(\mathbf{x}) = P(\mathbf{0}) \left(1 - \frac{1}{\lambda} \sum_{m=0}^N \lambda_{i-m} \phi \left(\Xi_{k=0}^m \mathbf{Z}_{i-k}[\mathbf{x}] \right) \right) \\ (i = 0, 1, \dots, N) \quad (9)$$

where we have set

$$P(\mathbf{0}) = \sum_{k=0}^N P_k(\mathbf{0}). \quad (10)$$

We shall now solve (9) by extending a method devised by M. F. Neuts¹² for the solution of a related equation in one variable.

Define the iteration procedure

$$\mathbf{V}_i^{(j)}[\mathbf{x}] = \sum_{k=0}^N \mathbf{Z}_{i-k}[\mathbf{V}_i^{(j-1)}[\mathbf{x}]]$$

$$(i = 0, 1, \dots, N; j = 1, 2, \dots; \mathbf{V}_i^{(0)}[\mathbf{x}] = \mathbf{x}). \quad (11)$$

Using (11) in (9) gives

$$g_i(\mathbf{V}_i^{(j)}[\mathbf{x}]) - g_i(\mathbf{V}_i^{(j-1)}[\mathbf{x}])$$

$$= P(0) \left(1 - \frac{1}{\lambda} \sum_{m=0}^N \lambda_{i-m} \phi \left(\sum_{k=0}^m \mathbf{Z}_{i-k}[\mathbf{V}_i^{(j-1)}[\mathbf{x}]] \right) \right)$$

$$(i = 0, 1, \dots, N; j = 1, 2, \dots). \quad (12)$$

Adding equations (12) for $j = 1, 2, \dots, n$ yields

$$g_i(\mathbf{V}_i^{(n)}[\mathbf{x}]) - g_i(\mathbf{x}) = P(0) \sum_{j=1}^n \left(1 - \frac{1}{\lambda} \sum_{m=0}^N \lambda_{i-m} \phi \left(\sum_{k=0}^m \mathbf{Z}_{i-k}[\mathbf{V}_i^{(j)}[\mathbf{x}]] \right) \right)$$

$$(i = 0, 1, \dots, N; n = 1, 2, \dots). \quad (13)$$

Now let $n \rightarrow \infty$ in (13). We will show in the next section that

$$\lim_{n \rightarrow \infty} \mathbf{V}_i^{(n)}[\mathbf{x}] = \mathbf{1} \quad (x_1 \leq 1, \dots, x_N \leq 1; i = 0, 1, \dots, N) \quad (14)$$

where $\mathbf{1} = [1, 1, \dots, 1]$, so that (13) becomes

$$g_i(\mathbf{1}) - g_i(\mathbf{x}) = P(0) \sum_{j=1}^{\infty} \left(1 - \frac{1}{\lambda} \sum_{m=0}^N \lambda_{i-m} \phi \left(\sum_{k=0}^m \mathbf{Z}_{i-k}[\mathbf{V}_i^{(j)}[\mathbf{x}]] \right) \right)$$

$$(i = 0, 1, \dots, N). \quad (15)$$

Notice that

$$\sum_{i=0}^N g_i(\mathbf{1}) = 1 \quad (16)$$

and

$$\sum_{i=0}^N g_i(\mathbf{0}) = P(0) \quad (17)$$

so that upon setting $\mathbf{x} = \mathbf{0}$ in (15) and adding for $i = 0, 1, \dots, N$ we obtain

$$P(0) = \left(1 + \sum_{i=0}^N A_i(\mathbf{0}) \right)^{-1} \quad (18)$$

where

$$A_i(\mathbf{x}) = \sum_{j=0}^{\infty} \left(1 - \frac{1}{\lambda} \sum_{m=0}^N \lambda_{i-m} \phi \left(\sum_{k=0}^m \mathbf{Z}_{i-k} [\mathbf{V}_i^{(j)}(\mathbf{x})] \right) \right) \quad (i = 0, 1, \dots, N). \quad (19)$$

(We remark that $P(0) \neq 1 - \sum_{i=0}^N \lambda_i h_i$ because the set of switch points is not an arbitrary subset of the set of all points at which units leave the server.)

It remains to calculate $g_i(1)$. Physically, $g_i(1)$ is the probability that at the instant the server leaves some queue, that queue is queue i . This event occurs if

(i) the last time the server left a queue the system was empty, and the next arrival occurred at queue i , or

(ii) the last time the server left a queue the system was not empty, and the queue was queue $i-1$.

Event (i) has probability $(\lambda_i/\lambda)P(0)$; event (ii) has probability $g_{i-1}(1) - g_{i-1}(0)$. Hence

$$g_i(1) = \frac{\lambda_i}{\lambda} P(0) + (g_{i-1}(1) - g_{i-1}(0)) \quad (i = 0, 1, \dots, N). \quad (20)$$

[Equation (20) can also be obtained directly from (7) with $\mathbf{x} = 1$.] But the difference $(g_{i-1}(1) - g_{i-1}(0))$ can be evaluated from (15) with $\mathbf{x} = 0$. Hence

$$g_i(1) = \frac{\lambda_i}{\lambda} P(0) + P(0) A_{i-1}(0) \quad (i = 0, 1, \dots, N) \quad (21)$$

so that (15) can be rewritten

$$g_i(\mathbf{x}) = \frac{\frac{\lambda_i}{\lambda} + A_{i-1}(0) - A_i(\mathbf{x})}{1 + \sum_{j=0}^N A_j(0)} \quad (i = 0, 1, \dots, N). \quad (22)$$

The quantities on the right side of (22) are completely specified; the set of simultaneous functional equations (5) has been solved in the sense that $g_i(\mathbf{x})$ may be calculated for any $\mathbf{x} \leq 1$.

VI. PROOF OF CONVERGENCE

We wish to prove statement (14):

$$\lim_{n \rightarrow \infty} \mathbf{V}_i^{(n)}(\mathbf{x}) = \mathbf{1} \quad (x_1 \leq 1, \dots, x_N \leq 1; i = 0, 1, \dots, N).$$

Note first that $\mathbf{V}_i^{(n)}[\mathbf{x}]$ is a vector whose $(N + 1 - m)$ th element ($m = 1, 2, \dots, N$) is

$$\phi\left(\sum_{k=0}^m \mathbf{Z}_{i-k}[\mathbf{V}_i^{(n-1)}[\mathbf{x}]]\right).$$

Therefore, we need show only that

$$\lim_{n \rightarrow \infty} \phi\left(\sum_{k=0}^m \mathbf{Z}_{i-k}[\mathbf{V}_i^{(n)}[\mathbf{x}]]\right) = 1$$

$$(i = 0, 1, \dots, N; m = 1, 2, \dots, N; x_1 \leq 1, \dots, x_N \leq 1). \quad (23)$$

From the definition (11) it is clear that the sequence $\{\mathbf{V}_i^{(n)}[\mathbf{x}]\}$ is bounded as $n \rightarrow \infty$ for $\mathbf{x} \leq 1$, and therefore the sequence $\{g_i(\mathbf{V}_i^{(n)}[\mathbf{x}])\}$ is bounded as $n \rightarrow \infty$ for $\mathbf{x} \leq 1$. Also,

$$0 \leq \phi\left(\sum_{k=0}^m \mathbf{Z}_{i-k}[\mathbf{V}_i^{(n)}[\mathbf{x}]]\right) \leq 1 \quad (n > 1, \mathbf{x} \leq 1). \quad (24)$$

We now turn our attention to equation (13). From (24) we see that the right side of (13) increases monotonically with n for $\mathbf{x} \leq 1$, and therefore the sequence $\{g_i(\mathbf{V}_i^{(n)}[\mathbf{x}])\}$ increases monotonically with n for $\mathbf{x} \leq 1$. Thus the sequence $\{g_i(\mathbf{V}_i^{(n)}[\mathbf{x}])\}$ is monotonically increasing and bounded for $\mathbf{x} \leq 1$, and therefore has a limit. Hence the left side of (13) has a limit, which implies that the series of nonnegative terms on the right side of (13) converges. This in turn implies that

$$\lim_{n \rightarrow \infty} \frac{1}{\lambda} \sum_{m=0}^N \lambda_{i-m} \phi\left(\sum_{k=0}^m \mathbf{Z}_{i-k}[\mathbf{V}_i^{(n)}[\mathbf{x}]]\right) = 1 \quad (\mathbf{x} \leq 1). \quad (25)$$

Statements (24) and (25) together imply (23), completing the proof.

VII. MEAN NUMBERS OF WAITING UNITS

Denote by $\bar{n}_i(k)$ the mean number of units waiting in queue $i + k$ when the server leaves queue i ($i = 0, 1, \dots, N; k = 0, 1, \dots, N; \bar{n}_i(0) = 0$). For convenience, let $g_i(1)\bar{n}_i(k) = \bar{m}_i(k)$ and $\bar{m}_i(1) = \bar{m}_i$. Then $\sum_{i=0}^N \bar{m}_i$ is the mean number of waiting units found by the server in the next queue in cyclic order at a switch point, and $\sum_{i=0}^N \bar{m}_i(k - i)$ is the mean number of waiting units in queue k at a switch point. We shall evaluate $\bar{m}_i(k)$ ($i = 0, 1, \dots, N; k = 1, 2, \dots, N$).

We first note that $\bar{m}_i(k)$ is given by

$$\bar{m}_i(k) = \frac{\partial}{\partial x_k} g_i(x_1, \dots, x_N) \Big|_{x_1 = \dots = x_N = 1}$$

$$(i = 0, 1, \dots, N; k = 1, 2, \dots, N) \quad (26)$$

and the mean 1-busy period $b_i = h_i / (1 - \lambda_i h_i)$ generated by a unit in queue i is given by

$$b_i = -\frac{d}{ds} \beta_i(s) \Big|_{s=0} \quad (i = 0, 1, \dots, N). \quad (27)$$

Differentiating through (5) we obtain

$$\begin{aligned} & \frac{\partial}{\partial x_k} g_i(x_1, \dots, x_N) \\ &= \frac{\partial}{\partial x_k} \beta_i \left(\sum_{m=1}^N \lambda_{i+m} (1 - x_m) \right) \frac{\partial}{\partial \beta_i} g_{i-1}(\beta_i, x_1, \dots, x_{N-1}) \\ &+ (1 - \delta(N-k)) \frac{\partial}{\partial x_k} g_{i-1}(\beta_i, x_1, \dots, x_{N-1}) \\ &+ \frac{\lambda_i}{\lambda} P(0) \frac{\partial}{\partial x_k} \beta_i \left(\sum_{m=1}^N \lambda_{i+m} (1 - x_m) \right) \\ & \quad (i = 0, 1, \dots, N; k = 1, 2, \dots, N) \end{aligned} \quad (28)$$

which upon setting $x_1 = \dots = x_N = 1$ gives the two-dimensional set of linear equations

$$\begin{aligned} \bar{m}_i(k) &= \lambda_{i+k} b_i \bar{m}_{i-1} + \frac{\lambda_i}{\lambda} P(0) \lambda_{i+k} b_i + (1 - \delta(N-k)) \bar{m}_{i-1}(k+1) \\ & \quad (i = 0, 1, \dots, N; k = 1, 2, \dots, N). \end{aligned} \quad (29)$$

For each i , (29) can be solved successively starting with $k = N$ and working backward:

$$\begin{aligned} \bar{m}_i(N-j) &= \lambda_{i+N-j} \sum_{\nu=1}^{j+1} b_{i+1-\nu} \bar{m}_{i-\nu} + \lambda^{-1} P(0) \lambda_{i+N-j} \sum_{\nu=1}^{j+1} \lambda_{i+1-\nu} b_{i+1-\nu} \\ & \quad (i = 0, 1, \dots, N; j = 0, 1, \dots, N-1). \end{aligned} \quad (30)$$

In particular, when $j = N-1$ equation (30) can be written

$$\begin{aligned} \bar{m}_i &= \lambda_{i+1} \sum_{\nu=i+1}^{i+N} b_{\nu+1} \bar{m}_\nu + \lambda^{-1} P(0) \lambda_{i+1} \sum_{\nu=i+1}^{i+N} \lambda_{\nu+1} b_{\nu+1} \\ & \quad (i = 0, 1, \dots, N). \end{aligned} \quad (31)$$

When $\lambda_{i+1} b_{i+1} \bar{m}_i$ is added to both sides of the i th equation of the set (31) we have after rearrangement

$$\begin{aligned} \bar{m}_i (1 + \lambda_{i+1} b_{i+1}) \lambda_{i+1}^{-1} - \lambda^{-1} P(0) \sum_{\nu=i+1}^{i+N} \lambda_{\nu+1} b_{\nu+1} &= \sum_{\nu=i}^{i+N} b_{\nu+1} \bar{m}_\nu \\ & \quad (i = 0, 1, \dots, N). \end{aligned} \quad (32)$$

The sum on the right side of (32) is a constant independent of the value of the index i . Hence

$$\begin{aligned} \bar{m}_i(1 + \lambda_{i+1}b_{i+1})\lambda_{i+1}^{-1} - \lambda^{-1}P(0) \sum_{\nu=i+1}^{i+N} \lambda_{\nu+1}b_{\nu+1} \\ = \bar{m}_{i+j}(1 + \lambda_{i+j+1}b_{i+j+1})\lambda_{i+j+1}^{-1} - \lambda^{-1}P(0) \sum_{\nu=i+j+1}^{i+j+N} \lambda_{\nu+1}b_{\nu+1} \end{aligned}$$

$$(i = 0, 1, \dots, N; j = 0, 1, \dots, N). \quad (33)$$

Combining (33) and (31) yields

$$\bar{m}_i = \frac{\lambda_{i+1}}{\lambda} P(0) \frac{\rho - \rho_{i+1}}{1 - \rho} \quad (i = 0, 1, \dots, N) \quad (34)$$

where we define $\rho_i = \lambda_i h_i$ and $\rho = \sum_{i=0}^N \rho_i$. Note that for (34) to be meaningful we must have $\rho < 1$. The $\{\bar{m}_i(k)\}$ can now be calculated from equations (34) and (30).

VIII. LAPLACE-STIELTJES TRANSFORM OF CYCLE TIME DISTRIBUTION FUNCTION

Consider the set of switch points associated with queue i , and append to this set every switch point associated with queue $i - 1$ at which the server finds the system completely empty. Call the elements of this augmented set the record points associated with queue i .

We define the partial cycle time for queue i as the elapsed time between a switch point associated with queue $i - 1$ and the temporally preceding record point associated with queue i . Denote by $\hat{G}_i(\cdot)$ the distribution function of the partial cycle time for the i th queue, and by $\hat{\gamma}_i(\cdot)$ its Laplace-Stieltjes transform.

Since queue i is necessarily empty at an associated record point, all of the units waiting for service in queue i at a switch point of queue $i - 1$ must have arrived during the preceding partial cycle time. Let $P_{i-1}(j)$ be the conditional probability that $j \geq 0$ units will be waiting for service in queue i , given that a switch point associated with queue $i - 1$ has just occurred. Then the distribution function $\hat{G}_i(\cdot)$ of the partial cycle time for the i th queue and the distribution $\{P_{i-1}(j)\}$ of the number of units that arrive (according to the Poisson process with rate λ_i) during the partial cycle time are related as follows:

$$P_{i-1}(j) = \int_0^\infty \frac{(\lambda_i t)^j}{j!} \exp(-\lambda_i t) d\hat{G}_i(t)$$

$$(i = 0, 1, \dots, N; j = 0, 1, \dots). \quad (35)$$

Notice also that the distribution $\{P_{i-1}(j)\}$ has probability generating function

$$\sum_{j=0}^{\infty} P_{i-1}(j)x^j = \frac{g_{i-1}(x, 1, \dots, 1)}{g_{i-1}(1)} \quad (i = 0, 1, \dots, N). \quad (36)$$

Substitution of (35) into (36) yields, for the Laplace-Stieltjes transform $\hat{\gamma}_i(\cdot)$ of the partial cycle time distribution function for queue i ,

$$\hat{\gamma}_i(s) = \frac{g_{i-1}\left(\frac{\lambda_i - s}{\lambda_i}, 1, \dots, 1\right)}{g_{i-1}(1)} \quad (i = 0, 1, \dots, N). \quad (37)$$

We define the (full) cycle time for queue i as the partial cycle time plus the time required to serve those units, if any, waiting in queue i when the server finishes queue $i - 1$. (Notice that in order to be counted as a cycle for queue i , a time interval must contain a partial cycle ending at a switch point at queue $i - 1$.) Denote by $G_i(\cdot)$ the distribution function of the cycle time for the i th queue, and by $\gamma_i(\cdot)$ its Laplace-Stieltjes transform.

The cycle time distribution function $G_i(\cdot)$ is related to the partial cycle time distribution function $\hat{G}_i(\cdot)$ as follows:

$$G_i(t) = \int_0^t \sum_{j=0}^{\infty} \frac{(\lambda_i \xi)^j}{j!} \exp(-\lambda_i \xi) B_i^{*j}(t - \xi) d\hat{G}_i(\xi) \\ (B_i^{*0}(\cdot) = 1; i = 0, 1, \dots, N). \quad (38)$$

Taking Laplace-Stieltjes transforms throughout (38) we obtain

$$\gamma_i(s) = \hat{\gamma}_i(\lambda_i + s - \lambda_i \beta_i(s)) \quad (i = 0, 1, \dots, N). \quad (39)$$

Hence we have for the Laplace-Stieltjes transform $\gamma_i(s)$ of the cycle time distribution function for the i th queue

$$\gamma_i(s) = \frac{g_{i-1}\left(\frac{\lambda_i \beta_i(s) - s}{\lambda_i}, 1, \dots, 1\right)}{g_{i-1}(1)} \quad (i = 0, 1, \dots, N). \quad (40)$$

By differentiating through (40) we obtain for the mean cycle time \bar{t}_i the intuitively obvious result

$$\bar{t}_i = (b_i + \lambda_i^{-1})\bar{n}_{i-1} \quad (i = 0, 1, \dots, N). \quad (41)$$

IX. THE GATING MODEL

Consider now a system of $N \geq 1$ cyclic queues described by the gating model of Section I. Define $P_i(n_1, \dots, n_N)$ as the joint prob-

ability that at the instant the server leaves a queue, that queue is queue i ($i = 0, 1, \dots, N - 1$) and n_1 units are waiting in queue $i + 1$, n_2 units in queue $i + 2$, \dots , and n_N units are waiting in queue i (that is, n_N units arrived at queue i after the closing of the gate). Denote by $H_i^{*j}(\cdot)$ the j -fold convolution with itself of the service time distribution function $H_i(\cdot)$. Then

$$\begin{aligned} P_i(n_1, \dots, n_N) &= \sum_{j=1}^{\infty} \sum_{k_1=0}^{n_1} \dots \sum_{k_{N-1}=0}^{n_{N-1}} P_{i-1}(j, k_1, \dots, k_{N-1}) \int_0^{\infty} \prod_{m=1}^{N-1} Q_{i+m}(n_m - k_m; t) \\ &\quad \cdot Q_{i+N}(n_N; t) dH_i^{*j}(t) + P_{i-1}(0, n_1, \dots, n_{N-1}) \left(1 - \delta \left(\sum_{m=1}^{N-1} n_m \right) \right) \\ &\quad \cdot \delta(n_N) + \frac{\lambda_i}{\lambda} \sum_{k=0}^{N-1} P_k(0, \dots, 0) \int_0^{\infty} \prod_{m=1}^N Q_{i+m}(n_m; t) dH_i(t) \\ &\quad (i = 0, 1, \dots, N - 1) \end{aligned} \quad (42)$$

where $Q_i = Q_{i+N}$.

Equation (42), for the N -queue gating model, is only trivially different from (1), which describes the $(N + 1)$ - queue exhaustive service model. The analogue of (5) is

$$\begin{aligned} g_i(x_1, \dots, x_N) &= g_{i-1} \left(\eta_i \left(\sum_{m=1}^N \lambda_{i+m} (1 - x_m) \right), x_1, \dots, x_{N-1} \right) \\ &\quad + \frac{\lambda_i}{\lambda} \eta_i \left(\sum_{m=1}^N \lambda_{i+m} (1 - x_m) \right) \sum_{k=0}^{N-1} P_k(0, \dots, 0) \\ &\quad - P_{i-1}(0, \dots, 0) \quad (i = 0, 1, \dots, N - 1) \end{aligned} \quad (43)$$

where $\eta_i(\cdot)$ is the Laplace-Stieltjes transform of the distribution function $H_i(\cdot)$, and $g_i(x_1, \dots, x_N)$ is now the generating function for the gating model state probabilities. The solution of (43) follows that given for (5), and a complete analysis may now be carried out in a manner similar to that employed for the exhaustive service model. (We remark in passing that the equations originally considered by Neuts are those of the gating model with $N = 1$.)

X. SUMMARY

Two models of a system of queues served in cyclic order by a single server have been presented. One of these, the exhaustive service model, has been analyzed in detail. This model is described by the imbedded Markov chain probability state equations (1), from which a set

of functional equations (5) for the probability generating functions are derived. The functional equations are solved with the help of a generalization of an iteration procedure used by Neuts. The equations (5) are then used to obtain explicit expressions for various mean values, such as the mean number of units found waiting by the server in the i th queue, given by equation (34), and the mean cycle time, given by equation (41). The Laplace-Stieltjes transform of the cycle time distribution function is given, in a form suitable for numerical computation (and hence numerical inversion), by equation (40).

It is then shown that the gating model is described by state equations only trivially different from those of the exhaustive service model. It is now easy to adapt the methods and results of the detailed analysis of the exhaustive service model to a similar analysis of the gating model.

It is noteworthy that all results are expressed directly in terms of the single state probability $P(0)$ and the relevant generating functions, so that there is no need to evaluate the individual state probabilities. The calculations are thus reduced to the iteration algorithm, which may be suited to digital computer solution.

REFERENCES

1. Takács, L., "Two Queues Attended by a Single Server," *Operations Research*, 16, No. 3 (May-June 1968), pp. 639-650.
2. Avi-Itzhak, B., Maxwell, W. L., and Miller, L. W., "Queuing With Alternating Priorities," *Operations Research*, 13, No. 2 (March-April 1965), pp. 306-318.
3. Conway, R. W., Maxwell, W. L., and Miller, L. W., *Theory of Scheduling*, New York: Addison-Wesley, 1967.
4. Neuts, M. F. and Yadin, M., "The Transient Behavior of the Queue with Alternating Priorities, with special reference to the Waitingtimes," Mimeo Series No. 136, Dept. of Statistics, Purdue University (January 1968).
5. Eisenberg, M., "Multi-Queues With Changeover Times," MIT Doctoral Dissertation, (September 1967).
6. Sykes, J. S., unpublished work.
7. Leibowitz, M. A., "An Approximate Method for Treating a Class of Multi-queue Problems," *IBM J. Research Development*, 5, No. 3 (July 1961), pp. 204-209.
8. Saaty, T. L., *Elements of Queueing Theory*, New York: McGraw-Hill, 1961, pp. 298-301.
9. Kruskal, J. B., unpublished work.
10. Leibowitz, M. A., "Queues," *Scientific American*, 219, No. 2 (August 1968), pp. 96-103.
11. Takács, L., "Introduction to the Theory of Queues," New York: Oxford University Press, 1962, pp. 32, 57-65.
12. Neuts, M. F., "The Queues With Poisson Input and General Service Times, Treated as a Branching Process," Purdue University, (September 1966), unpublished paper distributed by the Clearinghouse for Federal Scientific and Technical Information, Dept. of Commerce, AD640483.

