AT&T Technical Journal
Vol. 64, No. 5, May–June 1985
Printed in U.S.A.

Application of Decomposition Principle in M/G/1 Vacation Model to Two Continuum Cyclic Queueing Models—Especially Token-Ring LANs

By S. W. FUHRMANN* and R. B. COOPER†

(Manuscript received November 15, 1984)

We apply a recent decomposition result of Fuhrmann and Cooper for the M/G/1 queue with server vacations to obtain mean waiting times for the following two cyclic queueing models: The server scans at a constant velocity (1) serving work as it is encountered, or (2) collecting work that it serves at the end of each cycle. Model 1 describes token-ring polling in certain computer-communication networks; Model 2 has been used to describe mail pickup and delivery systems.

I. INTRODUCTION AND SUMMARY

Cyclic queueing models, in which a single server switches back and forth among a (large) number n of queues, have been studied by many authors. These studies were motivated largely by the need to describe the performance of electronic telephone-switching systems. Recent technological developments in computer-communication networks (local area networks, or LANs) have generated renewed interest in these models. In the present paper we consider two *continuum cyclic queueing models*, i.e., models where $n \to \infty$ while the total arrival rate remains fixed. Model 1 describes the behavior of certain token-ring LANs, while Model 2 has been used to describe mail pickup and delivery systems.

* AT&T Bell Laboratories. † Florida Atlantic University, Boca Raton, Florida.

Exact analytic models for finite n tend to be very complicated (see, for example, Refs. 1 through 8). One of the earliest techniques for the analysis of these complicated multiqueue models was to define the ordinary single-server *vacation model*, in which the server periodically leaves the queue and takes a "vacation"; this vacation model is then "connected" to the model of n queues served in cyclic order by interpreting the vacation as the time interval from when the server leaves a particular queue until its return to that queue after cycling through the other $n - 1$ queues.[2]

The basis for the analysis in the present paper is to relate the server vacations to the cyclic queueing model in an entirely different manner, and then to apply a new stochastic decomposition result of Fuhrmann and Cooper[9] for the M/G/1 vacation model. (As noted, in the present paper we are interested in models where n is infinite. Another paper[10] uses a related method to analyze certain cyclic queueing models where n is finite.)

In both of our continuum models, the server scans (or polls) at a constant velocity along a closed path. Customers arrive according to a Poisson process (with rate λ) in time, and are uniformly and independently distributed in space along the scanning path. Service times have distribution function $H(\cdot)$, with mean τ and variance σ^2, and are independent of the arrival process and each other. In Model 1, the server stops scanning and serves customers as they are encountered along the scanning path. In Model 2, the server *collects* customers (there is no time expended in collecting a customer) as they are encountered along the scanning path; when the server reaches a unique point on the path called the *origin*, the server stops and serves all the customers it has collected since last leaving the origin. Model 2 is closely related to a model studied by Nahmias and Rothkopf,[11] which we shall refer to as Model 3, in which customers are served at the origin and are then randomly (uniformly) redistributed (delivered) over the scanning path on the server's next cycle. This is in contrast to Models 1 and 2, where each customer departs from the system as soon as his service is completed.

Model 1 provides a good description of a large, symmetric, token-ring LAN. In such a network, a number n of terminals (devices, work stations) are interconnected in either a physical or logical ring structure. The terminals' access to the transmission medium is controlled by a "token" (a signal) that circulates around the ring. A terminal gains access to the medium by seizing the circulating token as it goes by. It retains the token while it is transmitting, thereby preventing other terminals from simultaneously accessing the medium; it then releases the token to circulate around the ring, enabling another terminal to gain access to the transmission medium. (For a general

description of LANs and token-passing protocols, see, for example, Refs. 12 through 14.)

One identifies the scan time c as the time required for the server to poll all the terminals once (equivalently, the time required for the token to cycle once around the LAN ring when no terminals are waiting to transmit). If the number n of terminals is large, and the terminals submit statistically identical loads (i.e., each terminal is characterized by the same arrival rate and distribution of service times), then there is a good correspondence between Model 1 and the LAN.

Model 3 has been studied by Nahmias and Rothkopf,[11] who used it to describe a delivery system in which a clerk (the server) traverses (scans) at a constant velocity a route along which letters (customers) are generated (arrive) randomly in space and time. As the clerk travels along the route, he picks up the letters that have been generated since his last traversal, and he delivers the letters (to locations distributed uniformly along the route) that were previously picked up and sorted. When the clerk reaches the end of the route he sorts (serves) the letters he has just picked up; then he again traverses the route, delivering the letters that have just been sorted, and picking up the new letters that have been generated since his last traversal of the route. This process is repeated indefinitely.

For each model, the equilibrium *cycle time* T_j, defined as the time (during equilibrium) between successive visits in Model j by the server to any given point along the scanning path, has the same mean value $\bar{T} = E(T_j)$, given by

$$\bar{T} = \frac{c}{1 - \rho} \qquad (\rho < 1), \tag{1}$$

where c is the (constant) length of time the server spends scanning during each cycle (which is the time to complete a scan cycle when there is no work to be done), and $\rho(=\lambda\tau)$ is the server utilization. [Equation (1) is easily derived by the following argument, given by Kuehn,[7] for a very general model of n queues served in cyclic order by a single server: The mean cycle time $\bar{T}$ is the sum of the (constant) time c spent scanning and the mean time s spent serving per cycle; that is, $\bar{T} = c + s$. Clearly, $s = \rho\bar{T}$, and (1) follows. Note that $\bar{T}$ does not depend on the form of the service-time distribution function, but only on its mean value; also, the parameter n does not appear explicitly.]

Our main results are these: Let $W_j (j = 1, 2)$ be the equilibrium *waiting time* (time from request for service until start of service) in Model j, and let W_0 be the equilibrium waiting time in the corresponding M/G/1 queue (Model 0). Then,

$$E(W_1) = \frac{1}{2}\bar{T} + E(W_0), \tag{2}$$

and

$$E(W_2) = \bar{T} + E(W_0), \tag{3}$$

where $\bar{T}$ is given by (1), and $E(W_0)$ is given by the celebrated Pollaczek-Khintchine formula [see, for example, Ref. 15, p. 217, eq. (8.39)]:

$$E(W_0) = \frac{\rho\tau}{2(1-\rho)}\left(1 + \frac{\sigma^2}{\tau^2}\right). \tag{4}$$

The simplicity of (2) and (3) and their similarity are quite remarkable.

We also define $S_j (j = 0, 1, 2)$ to be the equilibrium *sojourn time* (waiting time plus service time) in Model j. Since

$$E(S_j) = E(W_j) + \tau \qquad (j = 0, 1, 2),$$

eqs. (2) and (3) are equivalent to the following two equations:

$$E(S_1) = \frac{1}{2}\bar{T} + E(S_0) \tag{5}$$

and

$$E(S_2) = \bar{T} + E(S_0). \tag{6}$$

For Model 3, we define the equilibrium *delivery time* D_3 as the elapsed time between the generation of a letter and its delivery to its destination. We will show that $E(D_3)$ is given by the following formula, again remarkable in its simplicity:

$$E(D_3) = \tau + \frac{3}{2}\bar{T} + \frac{1+2\rho}{1+\rho}E(W_0). \tag{7}$$

The special case of (7) when $\sigma^2 = 0$ (i.e., when the time required to sort a letter is constant) was found (in a different form, by a more complicated argument) by Nahmias and Rothkopf.[11] The general result (7) was found also by Shanthikumar,[16] using level-crossing analysis.

The well-known textbook by Tanenbaum[12] discusses a model of a token-ring LAN with an arbitrary number n of terminals that, for n infinite, coincides with our Model 1. Tanenbaum gives eq. (1) and then states (p. 310) that the mean waiting time is "about half" the mean cycle time. It is interesting to note that, for the case of n infinite, eq. (2) shows that Tanenbaum's approximation (i.e., $E(W_1) = \bar{T}/2$) underestimates the correct value by exactly $E(W_0)$, an amount that can be considerable, being essentially proportional to σ^2 and inversely proportional to $1 - \rho$. [The reason that $\bar{T}/2$ underesti-

mates the correct value is a manifestation of the phenomenon of length biasing, i.e., the cycle to which an arbitrary customer (the *test customer*) arrives is stochastically longer than an arbitrary cycle. In particular, if T_1^* is the length of the cycle during which the test customer arrives, then $E(T_1^*) = E(T_1) + V(T_1)/E(T_1)$, where $V(T_1)$ is the variance of the cycle times in Model 1 (see, for example, Ref. 15, pp. 200–6). Since $E(W_1) = E(T_1^*)/2$, it follows from this observation and eq. (2) that $V(T_1) = 2\bar{T}E(W_0)$. A practical implication of this observation is that the mean waiting time $E(W_1)$ can be estimated using measurements of cycle times only.]

Several authors (see Refs. 4, 17, 18, and 19), using arguments more complicated than ours, have obtained results for related models with different queue disciplines (e.g., exhaustive service or gated service) and a finite number n of terminals; our result (2) can be obtained from their results when $n \to \infty$. Coffman and Gilbert[20] have analyzed Model 1 for the case of constant service times and, for this special case, derived a number of explicit distributional results, such as the distribution of waiting times.

In Section II we state the M/G/1 decomposition result alluded to earlier. In Section III we apply this decomposition result to obtain eqs. (2), (3), (5), and (6). For completeness, in Section IV we quickly derive eq. (7) by directly comparing the mean delays in Models 2 and 3.

II. A STOCHASTIC DECOMPOSITION RESULT

At all times the server is either scanning or is serving customers. The basis for the analysis of this paper is to interpret the time intervals when the server is scanning to be vacations, and then to invoke Proposition 3 of Fuhrmann and Cooper.[9] We define

$\psi_j(\cdot)$ = the p.g.f. (probability generating function) of the equilibrium distribution of the number of the customers present in Model j ($j = 1, 2$) at an arbitrary point in time;

$\chi_j(\cdot)$ = the p.g.f. of the equilibrium distribution of the number of customers present in Model j ($j = 1, 2$) at an arbitrary point in time, *given* that the server is scanning (on vacation); and

$\pi(\cdot)$ = the p.g.f. of the equilibrium distribution of the number of customers present in the corresponding M/G/1 queue at an arbitrary point in time (or, equivalently, just after a service completion epoch).

Thus, $\pi(\cdot)$ is given by a well-known formula [see, for example, eq. (8.12), p. 210, Ref. 15]. It follows directly from Proposition 3 of Fuhrmann and Cooper[9] that

$$\psi_j(z) = \chi_j(z)\pi(z) \qquad (j = 1, 2). \tag{8}$$

In terms of mean values,

$$\psi_j'(1) = \chi_j'(1) + \pi'(1) \qquad (j = 1, 2). \tag{9}$$

In Section III we show that it is a simple matter to find $\chi_j'(1)$ for $j = 1, 2$. Since (by Little's theorem) $\pi'(1) = \lambda E(S_0)$ and $\psi_j'(1) = \lambda E(S_j)$, eq. (9) yields eqs. (5) and (6) or, equivalently, (2) and (3).

III. MEAN WAITING TIMES: MODELS 1 AND 2

In this section we derive formulas (2) and (3) for the mean waiting times for Models 1 and 2. To do this, first note that (for either model) during each customer's service time, k new customers arrive to the system with probability

$$p_k = \int_0^\infty \frac{(\lambda t)^k}{k!} e^{-\lambda t} dH(t) \qquad (k = 0, 1, 2, \cdots). \tag{10}$$

We now define two auxiliary models, Auxiliary Model 1 and Auxiliary Model 2. Auxiliary Model 1 is defined in exactly in the same way as Model 1 except for the following aspect: Now, whenever the server encounters a customer, the customer is served in zero time and departs from the system; coincidental with his departure, however, a batch of k new customers joins the system (distributed along the scanning path in a uniform and independent manner) with probability p_k, given by (10). Thus, while the lengths of all service times have been collapsed to zero, the number of customers in the system just after a service completion epoch is stochastically the same for both Model 1 and Auxiliary Model 1. (This is true in a distributional sense. Or, if we go to the trouble to define Model 1 and Auxiliary Model 1 on the same sample space, we can make this statement true on every sample path.) This leads to the following conclusion: If we define A_1 to be the mean number of customers present in Auxiliary Model 1, then

$$\chi_1'(1) = A_1. \tag{11}$$

To calculate A_1, let λ_1^* and S_1^* be the arrival rate and sojourn time in Auxiliary Model 1. Then, by Little's theorem,

$$A_1 = \lambda_1^* E(S_1^*), \tag{12}$$

where, clearly,

$$E(S_1^*) = \frac{c}{2}. \tag{13}$$

To calculate λ_1^*, let N be the total number of customers (including himself) generated by a *Poisson* arrival in Auxiliary Model 1; then $\lambda_1^* = \lambda E(N)$. Now observe that the average number of new customers

generated when a customer is served is $\lambda\tau = \rho$; and each of these customers will generate, on average, $E(N)$ additional customers. Hence, $E(N) = 1 + \rho E(N)$; that is, $E(N) = (1 - \rho)^{-1}$, and therefore

$$\lambda_1^* = \frac{\lambda}{1 - \rho}. \tag{14}$$

[Note that $E(N)$ is precisely the mean number of customers served during an M/G/1 busy period. Observe also that (14) follows immediately from the requirement that the mean number of arrivals per cycle be the same in Auxiliary Model 1 as in the original Model 1: $\lambda_1^* c = \lambda \bar{T}$.] Equations (11) through (14) yield $\chi_1'(1) = \lambda c/2(1 - \rho)$; in light of (1), we have

$$\chi_1'(1) = \frac{\lambda \bar{T}}{2}. \tag{15}$$

This completes the calculation of (9) for Model 1, from which the main result (2) follows.

We now define Auxiliary Model 2 in a completely analogous manner, that is, in exactly the same way as Model 2, except that now when a customer is served (at the origin), he is served in zero time and is instantaneously replaced by a batch of k customers with probability p_k, given by (10). We define A_2 to be the mean number of customers present in Auxiliary Model 2. By the same argument used earlier,

$$\chi_2'(1) = A_2; \tag{16}$$

and the same argument that was used to derive eq. (15) for Model 1 applies. Hence, combining the equations that are analogous to (12) and (14), we have

$$A_2 = \frac{\lambda}{1 - \rho} E(S_2^*). \tag{17}$$

But, in contrast with (13), the mean sojourn time of a customer in Auxiliary Model 2 is exactly the cycle time c,

$$E(S_2^*) = c. \tag{18}$$

Therefore, the analogue of (15) is

$$\chi_2'(1) = \lambda \bar{T}, \tag{19}$$

and the main result (3) follows.

IV. MEAN DELIVERY TIME: MODEL 3

For completeness, we now derive eq. (7). This is accomplished by directly comparing the mean delays in Models 2 and 3. Recall that in

these models, all customers are served at the origin. For either model, consider an arbitrary customer (the test customer) and define

y = the mean number of customers in the test customer's batch;
y_b = the mean number of customers in the test customer's batch that are served before the test customer; and
y_a = the mean number of customers in the test customer's batch that are served after the test customer.

Then

$$y = y_b + y_a + 1 \tag{20}$$

and, by symmetry,

$$y_b = y_a. \tag{21}$$

Now let L_2 denote the number of customers present in Model 2, excluding the test customer, when the test customer enters service. Then

$$E(L_2) = \frac{\lambda c}{2} + y_b\rho + y_a. \tag{22}$$

The term $\lambda c/2$ equals the mean number of customers who arrived during the last scan, but behind the server. (These customers will be collected on the server's next cycle.) The term $y_b\rho(=\lambda y_b\tau)$ equals the mean number of customers who arrived during the service times of the customers (in the test customer's batch) who were served before the test customer. Finally, the term y_a equals the mean number of customers that have not yet been served.

Now observe, on the other hand, that L_2 has the same distribution as the number of customers present in Model 2 at an arbitrary point in time, excluding the customer being served (if any). [This is true because departures see the same distribution of customers that arrivals see (see Ref. 15, p. 187) and the arrivals see time averages (see Ref. 21).] Hence, by Little's theorem,

$$E(L_2) = \lambda E(W_2). \tag{23}$$

Combining (21), (22), and (23) yields

$$\lambda E(W_2) = \frac{\lambda c}{2} + y_a(1 + \rho). \tag{24}$$

Equations (3) and (24) determine y_a. Now observe that, clearly,

$$E(D_3) = E(W_2) + \tau + y_a\tau + \frac{c}{2}. \tag{25}$$

Equations (3), (24), and (25) now yield eq. (7) after some straightforward algebra.

REFERENCES

1. R. B. Cooper and G. Murray, "Queues Served in Cyclic Order," B.S.T.J., *48*, No. 3 (March 1969), pp. 675–89.
2. R. B. Cooper, "Queues Served in Cyclic Order: Waiting Times," B.S.T.J., *49*, No. 3 (March 1970), pp. 399–413.
3. M. Eisenberg, "Queues With Periodic Service and Changeover Times," Oper. Res., *20*, No. 2 (March–April 1972), pp. 440–51.
4. O. Hashida, "Analysis of Multiqueue," Review of the Electrical Communication Laboratories, N. T. T. Public Corp., *20*, Nos. 3–4 (March–April 1972), pp. 189–99.
5. A. G. Konheim and B. Meister, "Waiting Lines and Times in a System With Polling," J. Ass. Comput. Mach., *21*, No. 3 (July 1974), pp. 470–90.
6. S. Halfin, "An Approximate Method for Calculating Delays for a Family of Cyclic Type Queues," B.S.T.J., *54*, No. 10 (December 1975), pp. 1733–54.
7. P. J. Kuehn, "Multiqueue Systems With Nonexhaustive Cyclic Service," B.S.T.J., *58*, No. 3 (March 1979), pp. 671–98.
8. G. B. Swartz, "Polling in a Loop System," J. Ass. Comput. Mach., *27*, No. 1 (January 1980), pp. 42–59.
9. S. W. Fuhrmann and R. B. Cooper, "Stochastic Decompositions in the M/G/1 Queue With Generalized Vacations," Oper. Res., *33*, 1985.
10. S. W. Fuhrmann, private communication.
11. S. Nahmias and M. H. Rothkopf, "Stochastic Models of Internal Mail Delivery Systems," Manage. Sci., *30*, No. 9 (September 1984), pp. 1113–20.
12. A. S. Tanenbaum, *Computer Networks*, Englewood Cliffs: Prentice-Hall, 1981.
13. W. Bux, "Performance Issues in Local-Area Networks," Research Report RZ 1268 (45715), IBM Zurich Research Laboratory, November 1983.
14. H. Takagi and L. Kleinrock, "Analysis of Polling Systems," Technical Report TR87–0002, IBM Japan Science Institute, Chiyoda-ku, Tokyo, 1985.
15. R. B. Cooper, *Introduction to Queueing Theory*, second edition, New York: North-Holland (Elsevier), 1981.
16. J. G. Shanthikumar, private communication.
17. W. Bux, "Local-Area Subnetworks: A Performance Comparison," IEEE Trans. Commun., *COM-29*, No. 10 (October 1981), pp. 1465–73.
18. L. F. M. de Moraes and I. Rubin, "Analysis and Comparison of Message Queueing Delays in Token-Rings and Token-Buses Local Area Networks," Proc. IEEE Int. Conf. Commun., Amsterdam, May 14–17, 1984, pp. 130–5.
19. H. Takagi, "Mean Message Waiting Time in N Symmetric Queues With Nonexhaustive Cyclic Service Discipline," IBM Japan Science Institute, Chiyoda-ku, Tokyo, 1984.
20. E. G. Coffman and E. N. Gilbert, private communication.
21. R. W. Wolff, "Poisson Arrivals See Time Averages," Oper. Res., *30*, No. 2 (March–April 1982), pp. 223–31.

AUTHORS

Robert B. Cooper, B.S., 1961, Stevens Institute of Technology; M.S. (Systems Engineering and Operations Research), 1962, and Ph.D. (Electrical Engineering), 1968, University of Pennsylvania; AT&T Bell Laboratories, 1961–1969; Georgia Institute of Technology, 1969–1976; University of Michigan, 1975; New Mexico Institute of Mining and Technology, 1976–1978; Florida Atlantic University, 1978—. Mr. Cooper's primary interest is queueing theory and its application in telecommunications and computer science.

Steve W. Fuhrmann, B.A. (Mathematics), 1970, University of Montana; Ph.D. (Statistics), 1975, Purdue University; Rutgers University, 1975–1977; AT&T Bell Laboratories, 1977—. Mr. Fuhrmann's interests and work focus on stochastic processes and stochastic models for the performance evaluation of computer-communication systems.