A Queueing System in Which Customers Require a Random Number of Servers

LINDA GREEN
Columbia University, New York, New York
(Received September 1978; accepted March 1980)

We consider a multiserver queueing system in which customers request service from a random number of identical servers. In contrast to batch arrival queues, customers cannot begin service until all required servers are available. Servers assigned to the same customer may free separately. For this model, we derive the steady-state distribution for waiting time, the distribution of busy servers, and other important measures. Sufficient conditions for the existence of a steady-state distribution are also obtained.

IN CERTAIN STOCHASTIC service systems, it is sometimes necessary to provide simultaneous service from more than one server in order to perform the requested task. If customers require the same number of identical servers, and these servers start and finish service simultaneously, the system is equivalent to one which provides a single server per customer. However, in many of these systems, the number, and sometimes the type(s) of servers needed, varies from customer to customer.

This paper examines a multiserver queueing system in which customers require a random number of identical servers who must start serving together, but who may leave their assigned customer separately. This system is one member of a class of queueing systems which allow a random number of servers per customer. The most crucial characteristic of these systems is that a customer cannot begin service until all required servers are available. This characteristic has two important implications:

1. These systems are not members of the class of batch arrival models. Although an arrival who requests i servers can be thought of as a batch of i customers who each need 1 server, in a batch arrival system these customers may enter service one at a time.

2. Servers may be idle even when there are customers waiting to enter service.

Queues which fall into this category are found in many contexts. In computer systems, buffers and other temporary storage devices are used for programs and data of varying dimension. A loss system situation of

Operations Research
Vol. 28, No. 6, November-December 1980
0030-364X/80/2806-1335 $01.25
© 1980 Operations Research Society of America

this type was studied by Kaufman (1977). Communications systems provide various examples. Gimpelson (1965) examined a system in which a single wide-band facility is used to carry traffic of different bandwidths, and Wolman (1972) studied a problem in which data traffic is directed to two or more destinations (and cannot be transmitted until the required receivers are free). In both of these analyses, simplifying assumptions had to be made and/or numerical techniques employed to get approximate solutions.

Other applications include loading docks, where the number of people needed to lift an item varies according to the size and weight, and maintenance systems in which component failures are interdependent and occur in batches of varying size. Many other examples can be found; see Green (1978) for a general discussion of this class of queues, as well as results for various models.

Many of these applications have the characteristic of joint service—servers assigned to the same customer freeing together. Situations in which the model under examination provides a more accurate representation include:

1. *Firefighting.* The necessary number of fire engines and/or other equipment varies with the intensity of the fire. Although individual units may begin working as soon as they arrive, the major firefighting effort usually cannot begin until all required units are present. As the fire is brought under control, some units will be free to leave the scene of the fire.

2. *Jury selection.* Before a trial can begin, a jury panel of specified size (determined by the judge according to the type of trial) must be available from the jury pool. Most of the impaneled jurors will be released one at a time after questioning by the judge and lawyers.

3. *Emergency surgery.* The number of surgeons and/or other medical personnel required to begin an operation varies according to the type of surgery and the severity of the patient's condition. However, at various stages of the surgery, medical personnel who are no longer required will become free.

These applications are intended to be illustrative. The model was not chosen because it fits any one application exactly, but because it is tractable and preserves the most significant characteristic of such appliations—randomness in the number of required servers. It is also of interest as an approximation to the joint service queuing system (see Green [1978]).

Since there is the possibility in these systems of servers being idle when a queue exists, it is of interest to consider alternative service order disciplines to FIFO that may use some of these servers sooner. In Green (1980), such disciplines are considered for various models and, in many

cases, found to perform "better" than FIFO with respect to one or more measures of efficiency.

The operating characteristic of this system of primary interest is the steady-state waiting time in queue, denoted by W. In this paper, the distribution for W (as well as other important measures) is derived by noticing that the redefinition of certain random variables results in an $M/G/1$ queue being embedded in the more complex system. The distribution of busy servers is also derived for the case when a queue exists, as well as when it does not. This is of particular interest since, as noted previously, all servers are not necessarily busy when there is a queue. Sufficient conditions for the existence of a steady-state solution are presented in the last section of the paper.

1. THE MODEL

We consider a multiserver queueing system in which arriving customers request service from a random number of servers and cannot begin service unless at least that number of servers is free. In particular, the system consists of s identical and independent servers with completion times that are exponentially distributed with mean $1/\mu$. Customers arrive according to a Poisson process with rate λ and each of them requests simultaneous service from i servers with probability c_i, $1 \leq i \leq s$. The number of servers requested by successive customers is independent. Without loss of generality, we assume $c_0 = 0$. Customers enter service in their order of arrival (FIFO) and leave the system only after all servers requested have finished service. Since individual server completion times are independent once work is begun, servers associated with the same customer do not end service together. Therefore, customers are not in service for an exponentially distributed amount of time. A customer's service time, B_s, is distributed as the maximum of a random number of exponentially distributed random variables. Thus if $B_s(t)$ is the distribution function for B_s,

$$B_s(t) = \sum_{i=1}^{s} (1 - e^{-\mu t})^i c_i.$$

2. DEFINITIONS AND NOTATION

Before proceeding with the analysis, we define some random variables. See Figure 1 for illustrations.

The *queueing period*, generically denoted by Q, is defined as the period of time beginning when a customer arrives at an empty queue and must wait for service, and ending when the queue next becomes empty. Similarly, the *non-queue period*, $\bar{Q}$, is defined as beginning when the preceding queueing period ends, and ending when a queue next forms.

Let $\{t_n, n = 1, 2, \cdots\}$ be defined as the times when the customers in a queueing period enter service and define $B_{n+1} = t_{n+1} - t_n$, $n \geq 1$. We call B_n the *interservice time* of the nth customer in the queueing period. In Section 3 it is shown that the B_n are independent and identically distributed (i.i.d.) random variables. Therefore we can drop the subscript and denote an interservice time generically by B.

Since B is defined only for customers who join an existing queue when they enter the system, define the *initial delay* random variable, D, as the delay in entering service encountered by a customer who initiates a

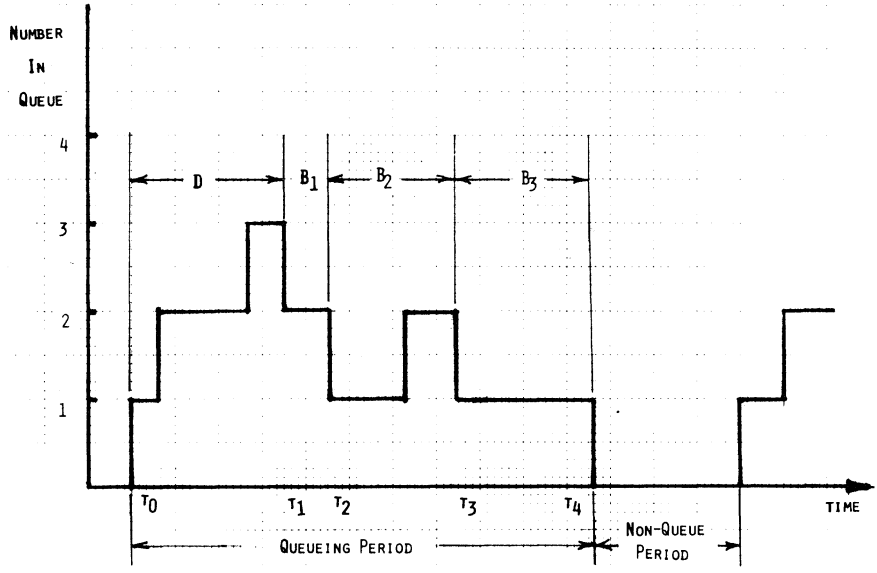

Figure 1

queueing period. If a customer arrives at an empty queue at time t_0 and enters service at t_1, then $D = t_1 - t_0$.

Derivations for the distributions and expected values of B and D can be found in the Appendix.

Unless otherwise noted $X(t)$ will denote the distribution function for the random variable X and $E(X)$ will be its expected value. $\tilde{X}(s)$ will denote the Laplace-Stieltjes Transform (LST) defined as

$$\tilde{X}(s) = \int_0^\infty e^{-st}\, dX(t) = E(e^{-sX}).$$

3. PRELIMINARY RESULTS

The analyses which appear in the following sections rely upon the results presented below.

PROPOSITION 1. *All s servers are busy whenever a customer enters service during a queueing period.*

Proof. Consider a customer who begins service during a queueing period and needs i servers. Since servers free one at a time, this customer cannot enter service simultaneously with any other customer and will occupy the first queue position for a positive amount of time. Therefore, he enters service at that epoch when the number of idle servers changes from $i - 1$ to i, causing all s servers to be busy.

COROLLARY 1. *The sequence of interservice times $\{B_n, n > 1\}$ are i.i.d. random variables.*

Proof. B_n is the time it takes for the nth customer of the queueing period to enter service, measured from the time he becomes first in the queue. Since, from Proposition 1, all servers will be busy when this customer becomes first in line, the time it takes for him to enter service depends only on the number of servers he requires and the time it takes for each server to free up. Since these random variables are i.i.d. for each customer, the result follows.

We define the sum of the non-queue period, $\bar{Q}$, and the queueing period, Q, to be the *queueing cycle*. Since, by Proposition 1, all s servers are busy at the start of each non-queue period, we have the following result.

COROLLARY 2. *The sequence of queueing cycles forms a renewal process and the queue-length process is regenerative.*

4. WAITING TIME ANALYSIS

The LST expression for the equilibrium waiting time in queue is derived by noticing that an $M/G/1$ queue is embedded within the structure of the original queue. Since the waiting time distribution for this $M/G/1$ queue is known, and the relationship between the waiting time in the original system and the waiting time in the embedded system can be written as a simple algebraic expression, the desired result is obtained. The LST is then easily inverted to arrive at the waiting time distribution.

Let p_q be the steady-state probability that there exists a queue, and let p_d be the probability that a customer arriving to an empty queue experiences a delay before entering service (i.e., initiates a queueing

period). These probabilities will be derived in Sections 5 and 6. If W is the steady-state waiting time in queue, we have:

THEOREM 1.

$$\tilde{W}(s) = (1 - p_q)\cdot(1 - p_d) + (1 - p_q)p_d\tilde{D}(s) + p_q[1 - \tilde{D}(s)][1 - \lambda E(B)]/([s - \lambda + \lambda\tilde{B}(s)]E(D))\cdot\tilde{B}(s). \tag{1}$$

Proof. Consider a customer who initiates or arrives during a queueing period. Define a new queueing system such that the total time this customer spends in the new system is equal to the time he actually spends waiting in queue. That is, ignore the actual service function and assume that when this customer enters service in the original system, he is leaving the constructed system. The "service" time for this constructed system is defined as the time spent occupying the first queue position in the original system. Since this system has only one "server" and Poisson arrivals, it is an $M/G/1$ queue with a busy period equivalent to the queueing period of the original system. Service times are distributed as the interservice random variable B except for those customers who initiate a queueing period, whose "service" time is distributed as D. Thus we have constructed an $M/G/1$ queue in which the first customer in a busy period has exceptional service; such a queue was studied by Welch (1964). Let Ω be the waiting time in queue for the constructed sytem. Using Welch's result for the LST of the waiting time in queue,

$$\tilde{\Omega}(s) = [1 - \lambda E(B)][\lambda(\tilde{D}(s) - \tilde{B}(s)) - s]/([1 - \lambda(E(B) - E(D))][\lambda - s - \lambda\tilde{B}(s)]). \tag{2}$$

In terms of the original system, Ω is the time it takes for a customer who joins or initiates a queue to become first in the queue. The LST expression for the wait in queue is obtained as follows:

$$W|\text{enter in non-queueing period} = \begin{cases} 0 & \text{if customer has no delay} \\ D & \text{if customer is delayed} \end{cases}$$

and

$$W|\text{enter in queueing period} = \Omega|\text{enter in queueing period} + B. \tag{3}$$

From Welch's (1964) Theorem 2, π_0, the steady-state probability of the constructed system being empty is given by $\pi_0 = (1 - \lambda E(B))/(1 - \lambda[E(B) - E(D)])$. Thus, (2) can be rewritten as

$$\tilde{W}(s) = \pi_0 + (1 - \pi_0)\cdot[1 - \tilde{D}(s)][1 - \lambda E(B)]/([s - \lambda + \lambda\tilde{B}(s)]E(D)) \tag{4}$$

and clearly

$$\tilde{\Omega}(s|\text{enter in queueing period}) = [1-\tilde{D}(s)][1-\lambda E(B)]/([s-\lambda+\lambda\tilde{B}(s)]E(D)). \tag{5}$$

Using (3) and (5), the result is obtained.

Let B_e and D_e be the equilibrium excess (residual life) random variables for the renewal processes with distribution B and D, respectively. That is,

$$B_e(t) = \int_0^t [1-B(u)]du/E(B), \qquad D_e(t) = \int_0^t [1-D(u)]du/E(D).$$

COROLLARY 3.

$$W(t) = (1-p_q)(1-p_d) + (1-p_q)p_d D(t) + p_q[1-\lambda E(B)]\textstyle\sum_{n=0}^{\infty}[\lambda E(B)]^n (B_e^{(n)} * D_e * B)(t)$$

where $B_e^{(n)}$ is the n-fold convolution of B_e and $*$ is the convolution operator.

Proof.

$$[1-\tilde{D}(s)]\tilde{B}(s)/(s-\lambda+\lambda\tilde{B}(s)) = [1-\tilde{D}(s)]\tilde{B}(s)/(sE(D))E(D)/(1-\lambda E(B) \cdot [1-\tilde{B}(s)]/(sE(B)). \tag{6}$$

Since $[1-\tilde{B}(s)]/sE(B)$ is the LST of the equilibrium excess random variable B_e, and $\lambda E(B) < 1$ in order to ensure a steady-state solution (see Section 7), we have from (1) and (6),

$$\tilde{W}(s) = (1-p_q)(1-p_d) + (1-p_q)p_d\tilde{D}(s) + p_q[1-\lambda E(B)]\tilde{D}_e(s)B(s)\textstyle\sum_{n=0}^{\infty}[\lambda E(B)\tilde{B}_e(s)]^n$$

and the result follows.

As in the case of the $M/G/1$ queue, the generating function $Q(z)$ for the distribution of the number of customers in queue is easily obtained from the LST of the waiting time in queue. The proof is the same as for the $M/G/1$ queue and will be omitted.

COROLLARY 4.

$$Q(z) = \tilde{W}[\lambda(1-z)].$$

5. QUEUEING CYCLE ANALYSIS

In order to evaluate transform equation (1) for the waiting time in queue, we must derive p_q, the stationary probability that there exists a queue. Since by Corollary 2 the sequence of queueing cycles forms a regenerative process, we can write (see Ross [1970], Chapter 5)

$$p_q = E(Q)/[E(\bar{Q}) + E(Q)] \tag{7}$$

and thus it suffices to obtain $E(Q)$ and $E(\bar{Q})$.

THEOREM 2.

$$E(Q) = E(D)/[1 - \lambda E(B)]. \tag{8}$$

Proof. From the proof of Theorem 1, it is clear that the queueing period can be viewed as the busy period of an $M/G/1$ queue with exceptional first service. Thus by direct application of the result for the expected length of the busy period for this type of system, as found in Takács (1962), we obtain the result.

To obtain $E(\bar{Q})$, let time 0 be defined as an epoch at which the queue first becomes empty. At time 0 there are s servers busy and no one waiting for service. Define the indicator function

$$I_Q(t) = \begin{cases} 0 & \text{if no queue at } t \\ 1 & \text{otherwise} \end{cases} \tag{9}$$

and let

$$\bar{N}_B(t) = \begin{cases} \text{number of busy servers at } t & \text{if } I_Q(t) = 0 \\ s + 1 & \text{if } I_Q(t) = 1. \end{cases} \tag{10}$$

The length of the non-queue period which starts at time 0 is the first passage time of $\bar{N}_B(t)$ to state $s + 1$.

Consider the Markov chain embedded in the process $\{\bar{N}_B(t), t \geq 0\}$ at arrival and server completion epochs with absorbing state $s + 1$. Using standard algebraic methods (see Kemeny and Snell [1960] Chapter 3), we obtain the matrix (V_{ij}) of the expected number of visits to transient state j starting in transient state i before absorption from

$$V = (I - T)^{-1} \tag{11}$$

where T is the matrix of transient states,

$$T = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ s \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & & 1 & & 2 & \cdots & s \end{array} \\ \begin{bmatrix} 0 & & c_1 & & c_2 & & c_s \\ \mu/(\lambda+\mu) & & 0 & & c_1\lambda/(\lambda+\mu) & & (c_{s-1}\lambda)/(\lambda+\mu) \\ 0 & & 2\mu/(\lambda+2\mu) & & 0 & & (c_{s-2}\lambda)/(\lambda+2\mu) \\ \vdots & & & & & & \vdots \\ 0 & & & & & & 0 \end{bmatrix} \end{array} \tag{12}$$

Since the expected time per visit spent in state j is $1/(\lambda + j\mu)$, we obtain

THEOREM 3.

$$E(\bar{Q}) = \sum_{j=0}^{s} V_{sj}/(\lambda + j\mu). \tag{13}$$

Let L and $\bar{L}$ denote the number of customers who arrive during a queueing period, and a non-queue period, respectively. Using Wald's Theorem (Ross, Chapter 3), the following results can be obtained.

THEOREM 4.

$$E(\bar{L}) = \lambda E(\bar{Q}) \tag{14}$$

$$E(L) = \lambda E(Q). \tag{15}$$

The probability that an arrival to an empty queue has a positive delay, p_d, was used in (1). Since Poisson arrivals see time averages (see Stidham) p_d is the fraction of all customers arriving to an empty queue who must wait for service. Therefore

COROLLARY 5.

$$p_d = 1/E(\bar{L}). \tag{16}$$

6. THE DISTRIBUTION OF BUSY SERVERS

We will derive the distribution of busy servers during a non-queue period, $\{\bar{q}_i\}$, and the distribution of busy servers during a queueing period, $\{q_i\}$. Using the expression for p_q obtained from (7), (8) and (13), the distribution at an arbitrary epoch can be easily obtained.

THEOREM 5.

$$\bar{q}_i = V_{si}/(\lambda + i\mu)E(\bar{Q}), \qquad i = 0, 1, \cdots, s \tag{17}$$

where the matrix V is obtained from (11) *and $E(\bar{Q})$ from* (13).

Proof. Let $\bar{N}_B(t)$ be defined as in (10). Then

$$\bar{q}_i = E(\text{time } \bar{N}_B(t) = i \text{ in 1 queueing cycle})/E(\bar{Q})$$

and the result follows from the analysis of the previous section.

We now use (15), (16) and (17) to obtain $\{q_i\}$.

THEOREM 6.

$$q_i = [E(L) \sum_{k=s-i+1}^{s} c_k + \sum_{j=i}^{s} \bar{q}_j \sum_{k=s-i+1}^{2} c_k]/i\mu E(Q)p_d \tag{18}$$

where $E(Q)$ is given by (8).

Proof. Define

$$N_B(t) = \begin{cases} \text{number of busy servers at } t & \text{if } I_Q(t) = 1 \\ 0 \quad \text{if } I_k(t) = 0 \end{cases}$$

where $I_Q(t)$ is the indicator function defined in (15). Clearly,

$$q_i = E(\text{time } N_B(t) = i \text{ in 1 queueing period})/E(Q). \tag{19}$$

Let V_i be the number of visits to state i in one queueing period. Then

$$E(\text{time } N_B(t) = i \text{ during 1 period}) = E(V_i) \cdot 1/i\mu. \tag{20}$$

Let C_n be the nth customer of the queueing period and define

$$J_i^n = \begin{cases} 1 & \text{if } C_n \text{ sees } i \text{ servers busy while first in queue} \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$V_i = \sum_{n=1}^{L+1} J_i^n. \tag{21}$$

Note that the customer that initiates the queueing period arrives during a nonqueue period.

Consider a customer who arrives to a queue and requires S servers. By Proposition 1, he waits for S server completion times from the time he becomes first in queue, and the number of busy servers during these completion times is $s, s-1, \cdots s-S+1$. The probability that there are exactly i servers busy for some part of the time during which this customer is first in queue is the probability that $S \geq s - i + 1$. Now consider C_1. The number of busy servers while he is waiting will be K, $K-1, \cdots K-S_1+1$ where K is the number of servers busy when he arrives and S_1 is the number of servers he requires. The probability that there are i servers busy while C_1 is first in queue is the joint probability of $K \geq i$ and $S_1 \geq s - i + 1$. From (21):

$$E(V)_i = E(L) \sum_{k=s-i+1}^{s} c_k + \sum_{j=i}^{s} \bar{q}_j \sum_{k=s-i+1}^{s} c_k/p_d \tag{22}$$

and the result follows from (19), (20) and (22).

7. EXISTENCE OF A STEADY-STATE SOLUTION

In the previous sections we have derived steady-state results for a queueing system which allows a random number of servers per customer. Implicit in these derivations is the assumption that sufficient conditions exist for a steady-state probability distribution $\{\pi_i\}$ of the number of customers in the system. We will now find a number ρ such that when $\rho < 1$, such a steady-state solution exists.

THEOREM 7. *If* $\rho = \lambda/\mu \sum_{k=1}^{s} \sum_{j=0}^{k-1} c_k/(s-j) < 1$, *there exists a steady state probability distribution* $\{\pi_i\}$.

Proof. A limiting distribution exists if the expected length of a queueing cycle is finite. Since $V = (1 - T)^{-1}$ is a nonsingular matrix, (13) implies $E(\bar{Q}) < \infty$. From (8) we have

$$E(Q) = E(D)/[1 - \lambda E(B)]$$

where $E(D)$ is given by (A6) and is clearly finite. Therefore $E(Q)$ will be finite if $\lambda E(B) < 1$ and the result follows from (A5).

APPENDIX—DISTRIBUTIONS OF INTERSERVICE TIME AND INITIAL DELAY

Define $F_{ki}(t)$ as the probability that k or more servers become free in the interval $(t_0, t_0 + t)$ given that i servers are busy at t_0. The probability of a server becoming free within a time period of length t is clearly $1 - e^{-\mu t}$. The probability that k out of i servers will complete service by time t is binomially distributed; therefore

$$F_{ki}(t) = \sum_{j=k}^{i} \binom{i}{j} (1 - e^{-\mu t})^j (e^{-\mu t})^{i-j}, \qquad k \leq i. \tag{A1}$$

Since all servers are busy just after an interservice time commences,

$$B(t) = \textstyle\sum_{k=1}^{s} F_{ks}(t) c_k. \tag{A2}$$

To obtain $D(t)$, we must consider the number of busy servers found by the initiating customer as well as the number of servers he requires. Define $H(i, j)$ as the joint probability that an initiating customer finds i busy servers at his arrival epoch and needs j servers. From Bayes' Theorem,

$$H(i, j) = \begin{cases} \bar{q}_i c_j / p_d & i > s - j \\ 0 & i \leq s - j \end{cases} \tag{A3}$$

where $\{\bar{q}_i\}$, given by (14), is the stationary distribution of the number of busy servers during a non-queue period, and p_d, given by (15), is the probability of an arrival to an empty queue having a positive delay. Therefore

$$D(t) = \textstyle\sum_{i=1}^{s} \sum_{k=1}^{i} F_{ki}(t) \bar{q}_i c_{s-i+k} / p_d. \tag{A4}$$

We now proceed to get the expected values for B and D. When i servers are busy, the mean time for a server to become free is $1/i\mu$. Thus

$$E(B) = \textstyle\sum_{k=1}^{s} [1/(s\mu) + 1/((s-1)\mu) + \cdots + 1/((s-k+1)\mu)] c_k. \tag{A5}$$

Similarly,

$$E(D) = \textstyle\sum_{i=1}^{s} \sum_{k=1}^{i} [1/(i\mu) + 1/((i-1)\mu) + \cdots + 1/((i-k+1)\mu)] \bar{q}_i c_{s-i+k} / p_d. \tag{A6}$$

ACKNOWLEDGMENTS

I am very grateful to Daniel P. Heyman, Matthew J. Sobel and Ward Whitt for their valuable comments.

REFERENCES

GIMPELSON, L. A. 1965. Analysis of Mixtures of Wide-and-Narrow Band Traffic. *IEEE Trans. Commun. Technol.* **13,** 258–266.

GREEN, L. 1978. Queues which Allow a Random Number of Servers per Customer. Ph.D. dissertation, Yale University, 1978.

GREEN, L. 1980. Comparing Operating Characteristics in Queues in Which Customers Require a Random Number of Servers (to appear in *Management Science*).

KAUFMAN, J. S. 1977. Sizing a Message Store Subject to Blocking Criteria (unpublished manuscript).

KEMENY, J. G., AND J. C. SNELL. 1960. *Finite Markov Chains.* D. Van Nostrand, Princeton, N.J.

ROSS, S. M. 1970. *Applied Probability Models with Optimization Applications.* Holden-Day, San Francisco.

→ STIDHAM, S., JR. 1972. Regenerative Processes in the Theory of Queues, with Applications to the Alternating-Priority Queue. *Advan. Appl. Prob.* **4,** 542–577.

TAKÁCS, L. 1962. *Introduction to the Theory of Queues.* Oxford University Press, Oxford, England.

→ WELCH, P. D. 1964. On a Generalized $M/G/1$ Queueing Process in Which the First Customer of Each Busy Period Receives Exceptional Service. *Opns. Res.* **12,** 736–752.

WOLMAN, E. 1972. The Camp-On Problem for Multiple-Address Traffic. *Bell System Tech. J.* pp. 1363–1422.