MANAGEMENT SCIENCE
Vol. 27, No. 1, January 1981
Printed in U.S.A.

COMPARING OPERATING CHARACTERISTICS OF QUEUES IN WHICH CUSTOMERS REQUIRE A RANDOM NUMBER OF SERVERS*

LINDA GREEN†

We examine the relative effects of several service order disciplines on important operating characteristics of queues in which customers request a random number of servers. This class of queues is characterized by customers who cannot begin service until all required servers are available. We show that for many systems in this class, it is possible to define a new service order discipline which is more efficient than FIFO with respect to one or more measures such as expected waiting time, probability of delay, etc.
(QUEUES; MULTI-SERVER; CUSTOMERS SERVED BY SEVERAL SERVERS SIMULTANEOUSLY)

1. Introduction

There exist many queueing situations in which it is sometimes necessary to provide simultaneous service from several servers in order to perform the requested task. If servers are identical and the number required by each customer is constant, the system is equivalent to one which provides a single server per customer. Therefore, consider the class of queues which is characterized by customers who require simultaneous service from a random number of servers.

The most crucial attribute of these systems which provide a random number of servers per customer is that a customer cannot begin service until all required servers are available. This has two significant implications:

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

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

Queueing systems belonging to this class are found in a variety of contexts. In computer systems, buffers and other temporary storage devices are used for programs and data of varying dimensions. A loss system situation of this type was studied by Arthurs and Kaufman (see [4]). Communications systems provide many examples. Gimpelson [1] examined a system in which a single wide-band facility is used to carry traffic of different bandwidths, and Wolman [7] studied a problem in which data traffic is directed to two or more destinations (and cannot be transmitted until all required receivers are free). Emergency systems such as firefighting, police and rescue units, also exhibit this characteristic. The number of servers (people and/or equipment) that must be dispatched in order to be effective, varies with the type and severity of the situation. Other applications, some of which will be mentioned later, are prevalent. See Green [2] for a general discussion of this class of queues as well as results for various models.

*Accepted by Marcel F. Neuts, former Departmental Editor; received November 13, 1979. This paper has been with the author 1 month for 2 revisions.
†Columbia University.

0025-1909/81/2701/0065$01.25

Since it is possible in these systems to have idle servers when customers are in queue, it is of interest to consider alternative service order disciplines to FIFO that use some of these servers sooner. One might guess that using such a discipline would result in greater efficiency with respect to one or more measures such as expected waiting time, probability of delay and server utilization.

This paper explores this issue for three categories of models in this class of queueing systems. Under the assumption of Poisson arrivals and exponentially distributed server completion times, we confirm that in most systems considered, an alternative discipline performs "better" than FIFO with respect to one or more measures of efficiency.

2. Categorization of Models

Apart from the usual variations in arrival and service distributions, total number of servers, and waiting room capacities, these systems can be categorized by the degree of independence or dependence among servers. Although in all the systems under consideration, servers associated with the same customer begin service together, they do not necessarily end service together. Models in which individual server completion times are independent once work is begun will be referred to as models with *independent servers.* Situations which are best represented by this type of model include the previously mentioned emergency contexts. The major firefighting effort, for instance, cannot begin until all required units are present, but at various stages of control, individual units will free up and leave the scene of the fire. Another application is in 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 pool of jurors. Most of the impanelled jurors will be released one at a time after questioning by the judge and/or lawyers. Analytic results for these queues can be found in Green [3].

Those models in which servers free up together will be called *joint service* models. A simple example is a loading dock where the number of people needed to lift an item varies according to the size and weight. Many communications systems and computer systems also fall into this category. See Kim [5] for some numerical and approximation methods for these systems.

Another type of dependence between servers will also be considered. *Constant service rate* models are defined by the following characteristic: under the assumption that at least one server is busy, the expected time until the next server becomes free is independent of the number of servers who are busy. For example, consider a maintenance system in which component failures are viewed as "customers" which are "served" by a bank of spare parts. The service time for each spare is defined to be the time until that spare (or its equivalent) again becomes available. This is therefore equivalent to the time it takes to repair the failed item. So if repairs are performed by a single repair facility, the rate of repairs, and thus the expected time until the next spare part becomes available, is constant whenever the number of failed items is positive.

The following results assume that there exists a steady-state probability distribution $\{\pi_i\}$ for the number of customers in the system. A sufficient condition for the model with independent servers and the constant total service rate model is $\lambda E(B) < 1$ (see [2], [3]), where λ is the arrival rate and $E(B)$ is the expected interservice time defined in §3. For the joint service model, the sufficient condition must be computed numerically for each particular problem (see [5]).

3. Definitions

Before proceeding with the analysis, we define some random variables.

Since servers can be idle when a queue exists, the traditional concept of a busy period does not apply in these models. Therefore we define an analogous random variable, the *queue period*, generically denoted as Q, which is the period of time beginning when a customer arrives to an empty queue and must wait for service, and ending when the queue next becomes empty. Similarly, the *nonqueue* period, $\bar{Q}$, begins when the preceding queue period ends and ends when a queue next forms. The *queueing cycle* is the sum of $\bar{Q}$ and Q.

The following definitions are for those models in which servers free one at a time. Let $\{t_n, n = 1, 2, \ldots\}$ be the times when the customers in a queue period enter service and define $B_{n+1} = t_{n+1} - t_n$, $n \geqslant 1$. We call B_n the *interservice time* of the nth customer in the queue period. B_n is the time it takes for the nth customer of the queue period to enter service, measured from the time he becomes first in queue. Similarly, define the *initial delay* random variable, D, as the delay in entering service encountered by a customer who initiates a queue period. If a customer arrives to an empty queue at time t_0 and enters service at t_1, then $D = t_1 - t_0$. Note that the initial delay is precisely the ordinary delay random variable (time spent in the queue) for the first customer of the queue period. However, the interservice time is only one component of the delay of all other customers in the queue period.

4. Constant Total Service Rate Model

We consider a queueing system with s identical servers with completion times that are exponentially distributed with rate μ/i when i servers are busy. Thus the service rate of the entire system is held constant at μ (unless, of course, the system is empty). Customers arrive according to a Poisson process and request simultaneous service from i servers with probability c_i, $1 \leqslant i \leqslant s$. The numbers of servers requested by successive customers are independent. Without loss of generality, we assume $c_0 = 0$.

We define SNOS (Smallest Number of Servers) to be the service order discipline under which the customer in queue needing the fewest number of servers goes into service first. That is, each time a server becomes free, the queue is scanned to see if there is a customer who can enter service because there are now enough available servers. In addition, an arriving customer who finds enough free servers goes into service immediately. In case of ties, the customer nearest to the head of the line enters service first.

The following proposition and corollary will be used extensively in obtaining subsequent results. Proofs for the FIFO system are identical to the ones for the model with independent servers which appear in Green [3]. Similar proofs can be constructed for the SNOS discipline.

PROPOSITION 1. *Under either the* FIFO *or* SNOS *service order discipline, all s servers are busy at every epoch at which a customer with positive delay enters service during the queue period.*

COROLLARY 1. *The sequence of queueing cycles forms a renewal process and the queue-length process is regenerative.*

Define Q as the length of a generic queue period and L as the number of customers who enter service during Q under FIFO, and Q' and L' as the corresponding random

variables under SNOS. We now prove

THEOREM 1. $\Pr(L = n) = \Pr(L' = n)$ *for* $n = 1, 2, \ldots$ *and* $\Pr(Q \leqslant x) = \Pr(Q' \leqslant x)$ *for all* $x > 0$.

PROOF. We will show that the amount of time that each arrival contributes to the queue period is the same under both disciplines.

Let C_i represent the ith arrival during the queue period. The queue period begins with the arrival of a customer C_0 to an empty queue, who needs more servers than are available. Let K_i be the number of servers required by C_i and let N_i be the number of idle servers at C_i's arrival epoch. Note that Proposition 1 implies that the nonqueue period for both FIFO and SNOS begins with all servers busy. Since the number of busy servers throughout the nonqueue period is affected only by server completion epochs and the arrival process, and is therefore independent of the service order discipline, the distribution of busy servers during the nonqueue period is independent of the service order discipline. In particular, N_0 has the same distribution under both disciplines. So C_0's contribution to the queue period (the length of the queue period if there are no arrivals before he enters service) is his delay D which is distributed as the sum of $K_0 - N_0$ server completion times in both systems. Note that the assumption of a constant total service rate implies that completion times are independent and identically distributed (i.i.d.). Assume there is another arrival C_1 during D. Under FIFO, C_1 joins the end of the line and therefore adds his interservice time B_1 to the length of the queue period. By Proposition 1, B_1 is distributed as the sum of K_1 completion times. Under SNOS, there are three possibilities:

CASE 1. $N_1 < K_1 < K_0$. C_1 becomes first in queue and enters service (if he's not "bumped") after $K_1 - N_1$ completion times, causing all servers to become busy. After an additional N_1 completion times, C_0 has the same remaining delay as he would have had without C_1's arrival. Therefore C_1's contribution to Q' has the same distribution as B_1.

CASE 2. $K_1 \geqslant K_0$. C_1 joins the end of the queue and adds K_1 completion times to Q'. This is true even if he is bumped out of the first position in queue by a subsequent arrival C_s. As illustrated in Case 1, the number of completions before C_1 is bumped can be considered part of C_s's contribution, and C_1's contribution is his ultimate waiting time as first in queue.

CASE 3. $K_1 \leqslant N_1 \leqslant K_0$. C_1 enters service immediately causing $N_1 + K_1$ servers to be busy. He adds K_1 completion times to C_0's wait in queue and thus, to Q'.

So in each case, C_1 adds K_1 server completion times to the time during which the first queue position is occupied. Since all subsequent arrivals to the queue period are faced with the same three possible situations, it is clear that C_i has a contribution to Q' which is distributed as B_i. Let $B_i = 0$ for all customers who arrive subsequent to the queue period. Then from above, $Q = Q' = D + \sum_{i=1}^{\infty} B_i$. Let L_0 and L_i be the number of arrivals during D and B_i, respectively. Since the arrival process is unaffected by service discipline, $L = L' = \sum_{i=0}^{\infty} L_i + 1$ thus completing the proof. Q.E.D.

Now let p_q and p'_q be the stationary probabilities of having a queue under FIFO and SNOS, respectively. Then

COROLLARY 2. $p_q = p'_q$.

PROOF. From Theorem 1, $E(Q) = E(Q')$ and from Proposition 1 and the fact that the arrival process is independent of the service order discipline, $E(\bar{Q}) = E(\bar{Q}')$.

Using Corollary 1 we can write (see e.g. [6, Chapter 5])

$$p_q = E(Q)/\left[E(\bar{Q}) + E(Q)\right], \qquad p_q' = E(Q')/\left[E(\bar{Q}') + E(Q')\right]$$

and the result follows. Q.E.D.

COROLLARY 3. Pr(*positive delay under* SNOS) $\leqslant$ Pr(*positive delay under* FIFO).

PROOF. Let N_B be the number of busy servers during a FIFO queue period, and $\bar{N}_B$ be the number of busy servers during a FIFO nonqueue period. Let N_B' and $\bar{N}_B'$ be the same measures in the SNOS system.

$$\Pr \text{ (positive delay under FIFO)} = p_q + (1 - p_q) \sum_{i=1}^{s} c_i \Pr(\bar{N}_B > s - i)$$

and

$$\Pr \text{ (positive delay under SNOS)} = p_q' \sum_{i=1}^{s} c_i \Pr(N_B' > s - i) + (1 - p_q') \sum_{i=1}^{s} c_i \Pr(\bar{N}_B' > s - i)$$

since a customer who arrives during a SNOS queue period will enter service immediately if there is a sufficient number of idle servers. By Proposition 1, $\bar{N}_B$ has the same distribution as $\bar{N}_B'$ and from Corollary 2, $p_q = p_q'$. Therefore the result is obtained. Q.E.D.

Now define W and W' as the steady-state waiting time in queue for the FIFO and SNOS systems, respectively. We are now prepared to prove

THEOREM 2. $E(W') \leqslant E(W)$.

PROOF. The two systems are clearly the same when there is no queue. Suppose there are two customers in queue. Customer C_1 needs i servers and C_2 needs j servers. If $i \leqslant j$, there is the same order of service under both disciplines. So suppose $i > j$. Under both FIFO and SNOS, both customers must wait until at least j servers are free. Call the time at which this first occurs T. Under FIFO, C_1 waits another $i - j$ completion times until entering service and C_2 waits $i - j + j = i$ completion times, measured from T. Under SNOS, C_2 goes into service at T and C_1 waits $j + i - j = i$ completion times from T. So under SNOS, C_1 waits j completion times more than in the FIFO system and C_2 saves i completion times of waiting. So the net saving in total waiting time using SNOS is $i - j$ completion times. In general, when the $m + 1$st customer in queue needs j servers and the first m customers need $K_1, K_2, \ldots, K_m$ servers, and $j < K_1, K_2, \ldots, K_m$, then the total waiting time saved by SNOS is $\sum_{n=1}^{m} K_n - mj > 0$ completion times.

When there is only one person in queue needing i servers, and a second customer arrives needing j servers and the number of free servers at his arrival is k, $j \leqslant k < i$, then in the FIFO system, customer 2 waits for $i - k + j$ completions. Under SNOS, customer 2 does not wait at all and customer 1 waits an additional j completion times. So again there is a net savings in total waiting time using SNOS equal to $i - k > 0$ completion times. In general, if this customer arrives to find m customers who need $K_1, K_2, \ldots, K_m$ servers with $j \leqslant k < K_1, K_2, \ldots, K_m$, then the net savings in waiting

time under SNOS is $\sum_{n=1}^{m} K_n - mk > 0$ completion times. Since $p_q = p'_q$ and $\Pr(L = n) = \Pr(L' = n)$, $n = 1, 2, \ldots$, the theorem is proved. Q.E.D.

THEOREM 3. *Let $D = \{$service order disciplines for which the expected length of the queue period is $E(Q)$ and the expected number of customers entering service during a queue period is $E(L)\}$. Then no other discipline in set D results in a smaller expected waiting time than* SNOS.

PROOF. From Theorem 1, SNOS and FIFO are members of D and it can be shown that LIFO is also in this set. Assume $d_0 \in D$ minimizes expected waiting time in queue and d_0 is not SNOS. Consider an arbitrary queue period under d_0. If using SNOS would result in the same service order of customers, the result is trivial. Therefore, assume there is a time T at which customer C_f, who requires the fewest servers K_f, would enter service under SNOS but is not the next to start service under d_0. Suppose there are m customers who precede C_f into service under d_0 at times $x_1, \ldots, x_m$, respectively, and who require $K_1, \ldots, K_m$ servers. If T is not C_f's arrival epoch, then by the same arguments as in the proof of Theorem 1, under d_0, each of these customers will add his interservice time to C_f's waiting time in queue. So C_f enters service at $x_{m+1} = T + \sum_{i=1}^{m} B_i$ where B_i is distributed as the sum of K_i completion times. However, if C_f starts service at T, the m customers can enter service at (or possibly before for a customer not in queue at T) $x_1 + B_f, x_2 + B_f, \ldots, x_m + B_f$. Since arrival epochs are unaffected by service order discipline, the total waiting time of these $m + 1$ customers will be reduced by at least $\sum_{i=1}^{m} B_i - B_f^{(m)}$ where $B_f^{(m)}$ is the m-fold convolution of B_f. Since $K_f \leqslant K_i$, $i = 1, \ldots, m$, the total expected wating time for the queue period will be reduced by at least $(\sum_{i=1}^{m} K_i - mK_f)/\mu > 0$. If C_f arrives at T and sees $k \geqslant K_f$ idle servers, by the second half of the proof of Theorem 2, the total expected waiting time will decrease by $(\sum_{i=1}^{m} K_i - mk)/\mu$ (or more if any of the m customers arrives after T) if C_f enters service at T. So in both cases, the expected waiting time using d_0 can be reduced by letting C_f enter service at T. Therefore d_0 does not minimize expected waiting time for D, thus proving the result. Q.E.D.

Since our investigation of SNOS was first prompted by the question of server utilization, we will now look at how this factor differs between the two disciplines. Let N be the number of busy servers in steady-state in the FIFO system and N' be the same measure under SNOS.

THEOREM 4. *$N \geqslant^{\text{st}} N'$ where $\geqslant^{\text{st}}$ denotes stochastic order.*

PROOF. Let $N(t)$ be the number of busy servers at time t.

$$\Pr(N \geqslant n) = \lim_{t\to\infty} \Pr(N(t) \geqslant n)$$

$$= \lim_{t\to\infty} \Pr\big(N(t) \geqslant n \,|\, t \in \bar{Q}\big)(1 - p_q) + \lim_{t\to\infty} \Pr(N(t) \geqslant n \,|\, t \in Q)p_q.$$

Since the distribution of busy servers during the nonqueue period is identical under FIFO and SNOS (see proof of Theorem 1),

$$\lim_{t\to\infty} \Pr\big(N(t) \geqslant n \,|\, t \in \bar{Q}\big) = \lim_{t\to\infty} \Pr\big(N'(t) \geqslant n \,|\, t \in \bar{Q}'\big),$$

and since $p_q = p'_q$ from Corollary 2, it is sufficient to prove that

$$\lim_{t\to\infty} \Pr(N(t) \geqslant n \,|\, t \in Q) \geqslant \lim_{t\to\infty} \Pr(N'(t) \geqslant n \,|\, t \in Q'). \tag{1}$$

Let C_i be the ith arrival during the queue period. As shown in the proof of Theorem 1, the queue period under both disciplines has a decomposition given by

$$Q = \sum_{i=0}^{L} Q_i = \sum_{i=0}^{L'} Q_i' = Q' \tag{2}$$

where Q_i is C_i's contribution to the length of the queue period and is distributed as D for $i = 0$ and as B for $i \geqslant 1$. Recall that L and L' have the same distribution. We will show that

$$\lim_{t\to\infty} P(N(t) \geqslant n \mid t \in Q_i) \geqslant \lim_{t\to\infty} P(N'(t) \geqslant n \mid t \in Q_i'), \qquad i = 0, 1, \ldots . \tag{3}$$

Under FIFO, Q_i is exactly the interservice time of C_i (initial delay if $i = 0$). Let K_i be the number of servers required by C_i, and let N_i and N_i' be the number of idle servers at C_i's arrival epoch under FIFO and SNOS, respectively. Q_0 will consist of exactly $K_0 - N_0$ server completion times during which $s - N_0, s - N_0 - 1, \ldots, s - K_0 + 1$ servers will be busy, respectively. For C_i, $i \geqslant 1$, Q_i will consist of K_i completion times during which $s, s - 1, \ldots, s - K_i + 1$ servers will be busy. Under SNOS, C_0 will find N_0' (distributed as N_0) idle servers at his arrival epoch. As shown in Case 1 of Theorem 1's proof, Q_0' will consist of $K_0 - N_0$ server completion times during which $s - N_0, s - N_0 - 1, \ldots, s - K_0 + 1$ servers will be busy, regardless of whether or not C_0 is "bumped" out of first position by a subsequent arrival. For C_i, $i \geqslant 1$, there are three possibilities under SNOS:

CASE 1. C_i becomes first in queue upon his arrival and enters service next. Then as shown in Case 1 of Theorem 1, C_i can be considered as "inheriting" the N_i' completion times from the customer who he bumped out of first position, and therefore Q_i' consists of K_i completion times with $s, s - 1, \ldots, s - K_i + 1$ servers busy, respectively.

CASE 2. C_i has a positive waiting time until he becomes first in queue (for the last time before entering service). As shown in Case 2 of Theorem 1, Q_i' is C_i's ultimate waiting time in the first queue position and therefore again consists of K_i completion times during which $s, s - 1, \ldots, s - K_i + 1$ servers are busy.

CASE 3. $N_i' \geqslant K_i$ and so C_i enters service immediately upon his arrival. As shown in Case 3 of Theorem 1, Q_i' consists of the K_i completion times associated with the servers he occupies. However, there is a positive probability that $j = N_i' - K_i > 0$ thus causing $s - j, s - j - 1, \ldots, s - j - K_i + 1$ to be busy during Q_i', instead of the $s, s - 1, \ldots, s - K_i + 1$ that would be busy during Q_i under FIFO.

So for any $i \geqslant 0$, C_i will belong to one of the three cases above and for each it is clear that (3) will hold. (1) then follows from (2) and (3), thus proving the theorem. Q.E.D.

5. Model with Independent Servers

We now consider the system characterized by identical and independent servers with completion times that are exponentially distributed with mean $1/\mu$. As before, arrivals are Poisson and customers request service from i servers with probability c_i, $1 \leqslant i \leqslant s$.

Since this system is similar to the one with constant total service rate in that both have the characteristic that servers free one at a time, it seems likely that SNOS would again minimize expected waiting time. However, in the general case of s servers, it is not clear whether or not this is the case. This ambiguity is due to the following result.

Let p_q be the steady-state probability that a queue exists in the s-server FIFO system and p_q' be the corresponding probability in the s-server SNOS system.

THEOREM 5. $p_q \leqslant p_q'$.

PROOF. From Corollary 1,

$$p_q = E(Q)/\left[E(\bar{Q}) + E(Q)\right], \qquad p_q' = E(Q')/\left[E(\bar{Q}') + E(Q')\right], \tag{4}$$

where Q, Q', $\bar{Q}$ and $\bar{Q}'$ are the same random variables which were defined in the last section, but for the systems with independent servers. As mentioned in the previous section, Proposition 1 is also true in the case of independent servers and therefore $E(\bar{Q}) = E(\bar{Q}')$. We will show that $E(Q) \leqslant E(Q')$ which from (4) is sufficient to prove the theorem.

Let C_i represent the ith arrival during the queue period, K_i be the number of servers required by C_i, and N_i be the number of idle servers at C_i's arrival epoch. As in the proof of Theorem 1, C_0's contribution to the queue period is distributed as the initial delay random variable D under both disciplines. Also, as in Theorem 1, under FIFO each subsequent C_i adds his interservice time B_i to Q. Under SNOS however, there is a positive probability for each C_i that $N_i > K_i$, resulting in his entering service immediately and causing fewer than s servers to be busy. Therefore he adds an expected time to Q' of

$$\frac{1}{\mu}\left[\frac{1}{s - N_i + K_i} + \frac{1}{s - N_i + K_i - 1} + \cdots + \frac{1}{s - N_i + 1}\right]$$

$$> \frac{1}{\mu}\left[\frac{1}{s} + \frac{1}{s-1} + \cdots + \frac{1}{s - K_i + 1}\right] = E(B_i). \tag{5}$$

If C_i has a positive delay, he causes all servers to be busy when he enters service and so contributes an expected time of $E(B_i)$. Therefore from (5)

$$E(Q) \leqslant E(Q')$$

thus proving the theorem. Q.E.D.

If the length of the queue period were the same under both disciplines, the expected waiting time would be smaller under SNOS than under FIFO, as it is in the constant total service rate case. However, the greater probability in the SNOS system of an arrival seeing a queue may cause significantly greater waiting times for some customers (those requiring a large number of servers) who would be arriving during a nonqueue period under FIFO but who encounter a queue under SNOS. We can eliminate this problem by modifying the SNOS discipline as follows:

Let SNOS* be the service order discipline which, at a server-freeing or arrival epoch, selects for service the customer, if any, that causes all servers to be busy. This discipline is identical to SNOS except when an arrival to a queue needs fewer than the number of servers that are idle. In this case, the customer who next enters service under SNOS* is the one who would have been next exclusive of the new arrival.

SNOS* eliminates the arrivals to a queue who enter service immediately and cause fewer than s servers to be busy. Since it is this possibility that increases the expected length of the queue period under SNOS relative to FIFO, we get the same results with SNOS* for the model with independent servers as we did with SNOS for the constant

service rate model, with one exception. The proofs of the following are almost identical to those in the previous section and are therefore omitted.

Let the non-primed letters represent the same measures as before for the FIFO model with independent servers, and the primed letters be the corresponding measures under SNOS*.

THEOREM 6. $\Pr(L = n) = \Pr(L' = n)$, $n = 1, 2, \ldots$ *and* $\Pr(Q \leqslant x) = \Pr(Q' \leqslant x)$, $x > 0$.

COROLLARY 4. $p_q = p'_q$.

COROLLARY 5. $\Pr($*positive delay under* SNOS*$) \leqslant \Pr($*positive delay under* FIFO$)$.

THEOREM 7. $E(W') \leqslant E(W)$.

THEOREM 8. *Let* $D = \{$*service order disciplines for which the expected length of the queue period is* $E(Q)$ *and the expected number of customers entering service during a queue period is* $E(L)\}$. *Then no other discipline in set* D *results in a smaller expected waiting time than* SNOS*.

Since all servers become busy whenever a customer enters service during a queue period under SNOS*, the next theorem follows from Theorem 6.

THEOREM 9. $\Pr(N = n) = \Pr(N' = n)$, $n = 0, 1, 2, \ldots$

6. Joint Service Model

Recall that the distinguishing feature of the joint service model is the assumption that servers who work on the same customer free up simultaneously. Assume that regardless of the number of servers required, all customers have an exponential service time with mean $1/\mu$. From this assumption, it is clear that the instantaneous customer departure rate is proportional to the number of customers in service. Since the SNOS discipline usually results in more customers in service earlier in the queue period, it again appears as though SNOS would be more efficient than FIFO. In general, this is not the case. Consider the following example:

EXAMPLE 1. Assume a system with 7 servers and $c_3 > 0, c_4 > 0, c_i = 0,\ i \neq 3, 4$, $c_3 + c_4 = 1$. Using the SNOS discipline in this case results in 3-server customers being served first and therefore, the accumulation of 4-server customers at the end of the queue. Since only one 4-server customer can be in service at a time, this will clearly result in longer expected queue periods and waiting times than if they are interspersed with the 3-server customers. Since servers don't free up one at a time as in the other systems, allowing a 3-server customer to precede a 4-server customer into service won't always result in having a customer enter service earlier. In fact, in this system, the expected length of the queue period will be shorter than with FIFO or SNOS if when 4 servers are available, the next 4-server customer in queue enters service, and if exactly 3 servers are free, the next 3-server customer is selected.

This leads to consideration of disciplines which will use more servers sooner. The obvious candidate is one that scans the queue at every service completion and arrival epoch for a set of customers which by next entering service, will maximize the number of busy servers. However, this doesn't necessarily produce a discipline with smaller expected waiting times than FIFO, as illustrated by the next example:

EXAMPLE 2. Assume a system with 4 servers and $c_1 > 0$, $c_2 > 0$, $c_3 = 0$, $c_4 > 0$, $c_1 + c_2 + c_4 = 1$. Suppose at a service completion epoch, 4 servers become idle and the customers in queue are, in order of arrival, a 1-server, a 2-server, and a 4-server customer. Then by choosing the 4-server customer as next to enter service, his expected wait in queue is reduced by $1/2\mu + 1/\mu = 3/2\mu$ over FIFO, while each of the other 2 customers will have his expected wait increased by $1/\mu$. Therefore there will be a net increase in total expected waiting time of $1/2\mu$ in this case and in general, it is not likely that the overall expected waiting time in steady state will be better than under FIFO.

Using a discipline which selects customers so as to maximize the number of busy servers may not always result in an improved expected waiting time, but it does appear to result in a decrease in the expected length of the queue period. More specifically, consider the following discipline. Define MXMN (maximize servers, minimize customers) to be the discipline which at every service completion and arrival epoch selects the set of customers to next enter service as follows: first identify those sets of customers which would maximize the number of busy servers and among those, select any set which minimizes the number of customers in service. Note that the secondary criterion will favor customers who require more servers. This results in the "smaller" customers accumulating at the end of the queue where they have more opportunity to enter service with other smaller customers. This becomes clearer in the 2 server system where MXMN results in first serving the 2-server customers until there are none left in queue, at which time the 1-server customers will all be served until there are none of them left, etc. In this system, MXMN can be shown to result in a shorter expected queue period than under FIFO. It is not clear, however, even in this small system, whether or not it is better with respect to expected waiting time.[1]

[1] I am very grateful to Daniel P. Heyman for his valuable comments. I also thank J. G. Shantikumar for his suggestions regarding §5.

References

1. GIMPELSON, L. A., "Analysis of Mixtures of Wide-and-Narrow Band Traffic," *IEEE Trans. Communication Technology*, Vol. 13 (1965), pp. 258–266.
2. GREEN, L., "Queues Which Allow a Random Number of Servers Per Customer," Ph.D. Dissertation, Yale University, 1978.
3. ———, "A Queueing System in Which Customers Require a Random Number of Servers," *Operations Res.*, Vol. 28 (1980).
4. KAUFMAN, J. S., "Sizing a Message Store Subject to Blocking Criteria," unpublished manuscript, 1977.
5. KIM, S., "$M/M/s$ Queueing System Where Customers Demand Multiple Server Use," Ph.D. Dissertation, Southern Methodist University, 1979.
6. ROSS, S. M., *Applied Probability Models with Optimization Applications*, Holden-Day, San Francisco, Calif., 1970.
7. WOLMAN, E., "The Camp-On Problem for Multiple-Address Traffic," *Bell Systems Tech. J.*, (1972), pp. 1363–1422.