The Amount of Overtaking in a Network of Queues

Ward Whitt
*AT&T Bell Laboratories,
Holmdel, New Jersey 07733*

Customer B *overtakes* customer A in a queueing system if A arrives before B but B departs first. To understand the phenomenon of overtaking in a network of queues and its impact on sojourn times, it is important to develop concepts of the amount of overtaking. In this paper, two stochastic measures of the amount overtaking are defined: the random number of customers overtaken by an arbitrary customer and the random permutation of the original order of n customers upon departure. Describing these measures, however, often presents rather difficult combinatorial and probabilistic problems. In this paper, we make several conjectures about the amount of overtaking in open Jackson networks and its impact on sojourn times. Then we describe the amount of overtaking in several smaller networks, for example, a single node with the first-come first-served discipline and instantaneous Bernoulli feedback.

1. INTRODUCTION AND SUMMARY

We say that customer B *overtakes* customer A in a queueing system if A arrives before B but B departs first. The queueing systems we have in mind are networks of queues or associated subnetworks. Overtaking is of interest, for example, in packet-switched communication networks because real time is required to reassemble packets into messages when the order of the packets can be altered in transmission. Overtaking is also of interest in networks of queues because of its impact on sojourn times (the time spent at each node in the network). In an open Jackson network of M/M/1 queues with one customer type, if a customer traveling through the network along some path cannot be overtaken, then given that the customer follows this path the successive sojourn times at the nodes along the way are independent; see Reich [17, 18], Walrand and Varaiya [21], and Melamed [16]. On the other hand when overtaking is possible, the sojourn times are in general dependent; see Burke [3] and Simon and Foley [19].

In a single-server queue, overtaking is intimately connected to the queue discipline. There is no overtaking with the FIFO (first-in first-out) discipline and maximum overtaking with the LIFO (last-in first-out) discipline (within the class of work-conserving disciplines that do not allow the server to be idle when a customer is present). Other intermediate disciplines such as ROS (random order of service) differ to a large extent by the degree of overtaking; see Kingman [13, 14] and Vasicek [20].

NETWORKS, Vol. 14 (1984) 411–426

CCC 0028-3045/84/030411-16$04.00

We believe that overtaking is an important property of queueing systems that deserves systematic study. The purpose of this paper is to begin such a study. The goal is to develop useful quantitative measures of the *amount of overtaking.* We want to know how model structure affects overtaking and how, in turn, overtaking affects the standard measures of congestion, in particular, sojourn times. We believe that sojourn times often become more variable as overtaking increases. In this paper we indicate ways to make this statement (the conditions and the conclusion) more precise.

The principal queueing systems we consider are open Jackson [8, 9] networks. (See Disney [4] and Kelly [11] for background on networks of queues.) However, we have few results at this level of generality. We make several conjectures about Jackson networks, which should hold more generally, and establish results for various special cases.

The basic *open Jackson network* has one type of customer, unlimited waiting space, customers served in order of their arrival at each node at rates that may depend on the number of customers at the node (thus covering the standard multiserver queue as a special case), arrivals from outside at a rate that may depend on the total number of customers in the network, and a probability distribution on the entering node independent of the network history; see pp. 456–464 of Heyman and Sobel [7]. In a Jackson network there is Markovian routing: a customer departing node i goes next to node j with probability q_{ij}, independently of the history of the network and past routing. In a Jackson network the vector describing the number of customers at each node is a continuous-time Markov chain.

Our first conjecture is about sojourn times in Jackson networks for which overtaking is possible.

Conjecture 1.1. In an open Jackson network, the sojourn times of a customer at the different nodes on his route through the network are positively dependent, for example, *associated* (see Barlow and Proschan [1]).

When random variables are associated, the correlations are nonnegative, but also the correlations between all nondecreasing functions of the random variables are nonnegative. All experimental results known to us are consistent with Conjecture 1.1. For example, Kiessler [12] reported that he found positive correlations by simulation in the Simon-Foley network (Section 2). However, the correlations are so small, they are often hard to detect. (Ralph Disney and Benjamin Melamed indicate that they have also done other work with all results being consistent with Conjecture 1.1.)

Before proceeding, we point out that in multitype Jackson networks the sojourn times can be dependent without the *direct overtaking* we are considering. To obtain independence, it is necessary to rule out *indirect overtaking* as well, in which the influence of later customers can overtake the designated customer; see Walrand and Varaiya [21] and Melamed [16]. For example, a customer of type 1 can be overtaken indirectly on the path (1, 2, 3) if a customer of type 2 can go from node 1 to node 4 and a customer of type 3 can go from node 4 to node 3. We do not consider such indirect overtaking here, but is obviously worth studying.

The next two conjectures express the idea that the sojourn times should become more variable as the overtaking increases. They involve an undefined measure of overtaking, but the idea should be clear.

Conjecture 1.2. In an open Jackson network, the correlations between sojourn times increase as the overtaking increases.

Conjecture 1.3. In an open Jackson network, the total sojourn time distribution becomes more variable, for example, increases in the ordering determined by the expected value of all nondecreasing convex real-valued functions (see Whitt [22]), when the overtaking increases.

In this paper we consider only a single customer type. Moreover, we consider overtaking for customers that enter and leave the network at common designated nodes. This is not as much of a restriction as it might appear because we can add two extra single-server FIFO nodes to the network and let all external arrivals enter via one of these nodes and let all departures from the network be routed through the other node. These extra nodes do not change the overtaking. If all external arrival processes were originally independent Poisson processes, then this is maintained in the modified network by having a single Poisson arrival process to an exponential server with Markovian routing. However, we actually consider more general arrival processes, such as the general stationary point processes of Franken et al. [6]. We also assume Markovian routing throughout, that is, that each departure from node i is routed to node j with probability q_{ij} independently of the history of the network. For general arrival processes, we cannot add an extra node with Markovian routing at the front end without loss of generality.

We consider two measures of overtaking. The first counts the number of customers overtaken, without regard to their identity. The second focuses on how the order of particular customers is altered, for example, whether or not the order of two successive arrivals is reversed upon departure. We assume throughout that the system is in equilibrium, that is, that we have appropriate stationary processes; see [6].

The first measure of overtaking can be defined in two related ways: Let N be the number of customers that an arbitrary customer overtakes (active or optimistic view) and let $\bar{N}$ be the number of customers that overtake an arbitrary customer (passive or pessimistic view). It appears that N is often much easier to analyze than $\bar{N}$, and we focus on it in this paper. However, if we are interested in $\bar{N}$, it is important to observe that in the reversed network (with time reversed) the pair $(N, \bar{N})$ is just $(\bar{N}, N)$ for the original network. Hence, if the network is *reversible* (the finite-dimensional distributions of the reversed processes are the same as the original processes, see p. 5 of Kelly [11]), then $\bar{N}$ and N have the same distribution. Even if $\bar{N}$ and N do not have the same distribution, under considerable generality we have $E\bar{N} = EN$. The idea of course is that every overtaking event is counted once by the overtaking customer and once by the overtaken customer, but the timing is not identical because the overtaken customer arrives first.

We now show that typically $EN = E\bar{N}$. With stationary versions, it is possible to define a stationary process $\{X(i, j), -\infty \leqslant i, j \leqslant +\infty\}$ where $X(i, j)$ is 1 if customer i overtakes customer j, $i < j$, and 0 otherwise. Then let

$$N_{in} = \sum_{j=i+n}^{\infty} X(i,j), \quad \bar{N}_{jn} = \sum_{i=-\infty}^{i=j-n} X(i,j), \tag{1.1}$$

$N_i = N_{i1}$, and $\bar{N}_j = \bar{N}_{j1}$. Then N_i and $\bar{N}_j$ are distributed as N and $\bar{N}$ for each i, j. If

$\{N_i\}$ and $\{\bar{N}_j\}$ are metrically transitive as well as stationary (see [6]), then

$$EN = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} N_i \quad \text{and} \quad E\bar{N} = \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \bar{N}_j. \tag{1.2}$$

We obviously have $EN = E\bar{N}$ if the tails of the series N_i and $\bar{N}_j$ in (1.1) are appropriately negligible, that is, if $N_{in} \to 0$ and $\bar{N}_{jn} \to 0$ as $n \to \infty$ uniformly in i and j.

The second measure of overtaking focuses on changes in order. The network of queues maps the original order $(1, 2, \ldots, n)$ of any n successive customers into one of the $n!$ permutations of the vector $(1, 2, \ldots, n)$. The second measure of overtaking is the probability distribution (one for each n) on this space Π_n of permutations π, representing the distributions of the order of these customers upon departure. We can compare to such probability distributions, say P_1 and P_2, using stochastic order after defining a partial order on the space of permutations; see Kamae et al. [10]. We define a partial order on the space of permutations by saying that $\pi_1 \leqslant \pi_2$ if π_2 can be obtained from π_1 by successive switches of ordered adjacent elements, that is, by switching (i, j) into (j, i) if $i > j$. See p. 159 of Marshall and Olkin [15] for a discussion of this and related orderings of permutations. A diagram of the partial order in the case $n = 3$ appears in Figure 1. The degree of overtaking can be conveniently summarized by the number d of pairwise switches required to bring the permutation to the original order $(1, 2, \ldots, n)$. From Figure 1, we see that $d(3, 1, 2) = 2 > 1 = d(2, 1, 3)$, but $(3, 1, 2)$ and $(2, 1, 3)$ are not comparable in the partial order. We say that probability measure P_1 is stochastically less than or equal to probability measure P_2 on Π_n, and write $P_1 \leqslant_{st} P_2$, if $E(f, P_1) \leqslant E(f, P_2)$ for all nondecreasing real-valued functions f on $(\Pi_n, \leqslant)$, where $E(f, P)$ is the expected value of f with respect to P.

We say that a network is *order preserving* if there can be no overtaking. We say that

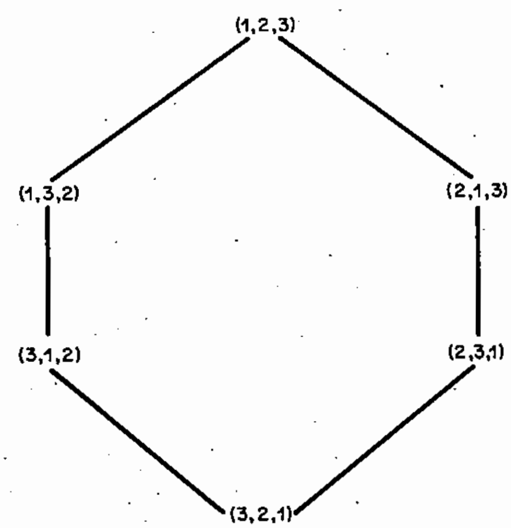

FIG. 1. The diagram of the partial order on the space Π_n of permutations: the case $n = 3$.

a network is *stochastically order preserving* if the order is more likely to be preserved than reversed or, more specifically, if the probability distribution on the departure permutation is stochastically greater or equal to the uniform distribution. The LIFO discipline yields simple examples of networks that are not stochastically order preserving. However, we make the following conjecture.

Conjecture 1.4. All open Jackson networks are stochastically order preserving.

There are certain obvious deficiencies in considering only the permutation of n successive arrivals. We do not account for other arrivals, so we cannot consider the effect of two networks in succession operating on a given arrival sequence, that is, where the departures from the first network are the arrivals to the second. Related measures of overtaking might be defined by letting an arbitrary arrival have the fixed label 0 and then considering other customers' position relative to this special customer. The network permutes the sequence $(\ldots, -n, -(n-1), \ldots, -1, 0, 1, 2, \ldots)$ into the order of departures relative to the designated customer. If the output of one network is the input to another network, then the output of the second is just the composition of the two network permutations. However, the distribution of the departure process is no doubt dependent on the permutation, making further analysis difficult.

In the remainder of this paper we focus on a special case of permutations: the simple switch. Let $\gamma(k)$ be the probability that an arbitrary customer does not overtake the kth previous customer. As a special case of Conjecture 1.4, we have

Conjecture 1.5. For all open Jackson networks and all $k \geqslant 1$,

$$\gamma(k) \geqslant 1/2.$$

Another natural conjecture is

Conjecture 1.6. For all open Jackson networks, $\gamma(k)$ is nondecreasing in k.

Since the system is assumed to be in equilibrium, the probability that an arbitrary customer has departed before k more customers arrive approaches 1 as $k \to \infty$. Hence, $\gamma(k) \to 1$ as $k \to \infty$. Examples 2.1 and 4.1 show that, for any k, $\gamma(k)$ can be arbitrarily close to $\frac{1}{2}$ in a large network of single-server FIFO nodes or in a single multiserver node.

Our definition of overtaking allows only a single comparison: We compare the order of customers upon arrival and departure. To understand complex networks, we can of course apply this single-comparison definition to subnetworks as well as the entire network, but more detailed concepts of overtaking may also be useful. If we can order the nodes and the waiting positions at each node, we can also define the number of times customers A and B switch order before departure. Instead of $\gamma(k)$, we could work with $\bar{\gamma}(k)$, the probability that an arbitrary customer *never* overtakes the kth previous customer during his stay in the network. Obviously the lower bound of $\frac{1}{2}$ in Conjecture 1.5 does not apply to $\bar{\gamma}(k)$. A minor modification of Example 2.1 shows that $\bar{\gamma}(k)$ can be arbitrarily close to 0.

In the rest of the paper, we study the amount of overtaking in several small networks. In Section 2 we consider the simple three-node network used by Simon and Foley [19] to show that sojourn times at different nodes in an acyclic network can be dependent. (See Fig. 2.) In Section 3 we consider a symmetric generalization in which departures from the first node are split and recombined after going through one of two or more intermediate nodes. (See Figs. 4 and 5.) In Section 4, we consider the standard multiserver queue with the first-come first-served discipline. Finally, in Section 5, we consider a single FIFO server with instantaneous Bernoulli feedback. (See Fig. 6.)

We reiterate that this study is a bare beginning. The conjectures remain unresolved and, even for the small network examples, useful relationships between the amount of overtaking and sojourn-time distributions remain to be determined.

2. THE SIMON-FOLEY NETWORK

In this section we investigate the simple three-node network used by Simon and Foley [19] and Walrand and Varaiya [21, Fig. 2] to show that sojourn times at the different nodes in acyclic networks are not all independent when there can be overtaking. See Figure 2. This network has three single-server nodes, each with the FIFO discipline and unlimited waiting space. Let all external arrivals come to node 1. Let successive departures be routed from node 1 independently (of the history of the network and past routing) to node 2 with probability p and to node 3 with probability $1 - p$. Let all departures from node 2 go to node 3 and let all departures from node 3 leave the system.

For the moment, the arrival process at node 1 and the three service processes can be general stationary point processes as in Franken et al. [6]. We assume that equilibrium exists and that the network is in equilibrium, that is, we have a stationary version of the vector queue-length process.

Let Q_2 be the queue-length (number in system) at node 2 seen by a departure from node 1. (This need not agree with the queue-length at node 2 seen by an arrival to node 1.)

For an arbitrary customer to overtake other customers, it is necessary and sufficient that two conditions be satisfied: (i) the customer must be routed from node 1 to node 3, and (ii) upon departure from node 1, node 2 must not be empty. The probability a customer overtakes exactly k customers is just $(1 - p) P(Q_2 = k)$.

In this general framework it is possible to bound the probability of a switch, that is, that an arbitrary customer will overtake any specific previous arrival. As before, let

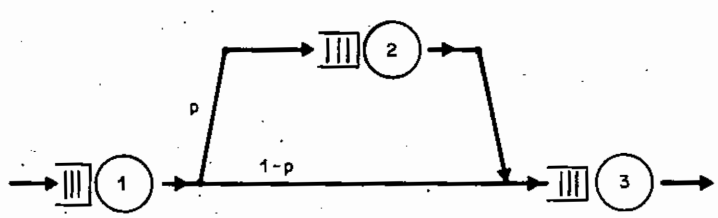

FIG. 2. The Simon-Foley acyclic network: Customers going directly from node 1 to node 3 overtake customers at node 2.

$\gamma(k)$ be the probability that the customer does not overtake the kth previous arrival. For overtaking, the arbitrary customer must go to 3 and the previous customer must go to 2, for which the probability is $p(1-p) \leq \frac{1}{4}$. Given this, the previous arrival may depart before the designated customer arrives. This gives $1-\gamma(k) \leq \frac{1}{4}$ for all k. With iid exponential service times, independent of the arrival process, $\gamma(k) > \frac{3}{4}$. For this model, it is easy to show that $\gamma(k)$ is nondecreasing in k. As k increases, the previous arrival is more likely to have departed before the designated customer arrives.

For the Simon-Foley network, $\gamma(k) \geq \frac{3}{4}$. However, we can use the Simon-Foley network as a component in a larger network to make $\gamma(k)$ arbitrarily close to $\frac{1}{2}$ for any k.

Example 2.1. The following open Jackson network of a single-server FIFO queues is a sequence of Simon-Foley networks. Let there be $2n+1$ nodes. From nodes $2k+1$, $0 \leq k \leq n-1$, let the probability be $\frac{1}{2}$ of going to node $2k+2$ and $\frac{1}{2}$ of going to node $2k+3$. Let all departures from node $2k$, $1 \leq k \leq n$, go to node $2k+1$. Let all departures from node $2n+1$ leave the system. (See Fig. 3.) Let all arrivals enter node 1 in a Poisson process at rate 1. Let the service rates be $\mu = 2$ at nodes $2k+1$, so that the traffic intensity is $\frac{1}{2}$ at each of these nodes. Let the service rates at nodes $2k$ be just slightly larger than $\frac{1}{2}$, so that the traffic intensities are only slightly less than 1. In particular, let the service rate be $[2(1-\epsilon^{-(n-k)})]^{-1}$ at node $2k$, so that the traffic intensity at node $2k$ is $1-\epsilon^{-(n-k)}$. By heavy traffic theory, this guarantees that the equilibrium queue length at node $2k$ is much larger than the equilibrium queue length at node $2(k+1)$ for $1 \leq k \leq n-1$. For any k, for sufficiently small ϵ, the probability is therefore approximately $\frac{1}{4}$ that an arbitrary customer overtakes the kth previous customer by node 3. This occurs if the kth previous customer is routed from 1 to 2 while the arbitrary customer is routed from 1 to 3. Since the queue-length at 2 is much larger than at any other node, the probability of reovertaking later is negligible. The probability that both customers are routed the same way from node 1 is $\frac{1}{2}$. The experiment is then repeated at node 3, etc. Hence, $1-\gamma(k)$ is approximately

$$1-\gamma(k) \approx \tfrac{1}{4} + (\tfrac{1}{2})\tfrac{1}{4} + \cdots + (\tfrac{1}{2})^{n-1}(\tfrac{1}{4}) = \tfrac{1}{2}(1-(\tfrac{1}{2})^n). \tag{2.1}$$

In other words, for ϵ sufficiently small and n sufficiently large, $\gamma(k)$ is arbitrarily close to $\frac{1}{2}$.

If, instead, we let the traffic intensities at nodes $2k$ be increasing by making the service rate by $[2(1-\epsilon^k)]^{-1}$ at node $2k$ (which is equivalent to reversing time in the original network), then the probability of reovertaking is significant at each opportunity and $\bar{\gamma}(k)$, the probability of never overtaking the kth previous customer (introduced toward the end of Section 1), can be made arbitrarily close to 0. ■

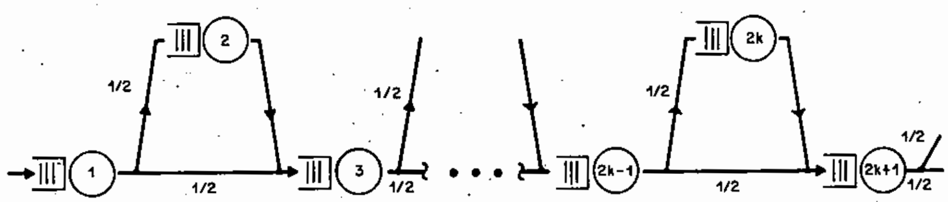

FIG. 3. A sequence of Simon-Foley networks.

We now return to the three-node Simon-Foley network (Fig. 2) and consider the special case of a Jackson network: Let the arrival process be Poisson at rate λ and let the service times be iid at each node with rate μ_i. To have stability, we assume $\rho_1 \equiv \lambda/\mu_1 < 1, \rho_2 \equiv \lambda p/\mu_2 < 1$, and $\rho_3 \equiv \lambda(1-p)/\mu_3 < 1$.

Since the departure process from node 1 is Poisson, see Burke [2], and since Poisson departures see time averages, see Wolff [23], Q_2 has the geometric distribution with parameter ρ_2. The expected number of customers overtaken by an arbitrary customer is thus

$$EN = (1-p)EQ_2 = (1-p)\rho_2/(1-\rho_2) = p(1-p)x/(1-px), \tag{2.2}$$

where $x = \lambda/\mu_2$. Thus, EN depends on just two parameters p and x, where p and x must satisfy $x < 1/p$ for stability. The following properties of $EN(x, p)$ are easily established.

Theorem 2.1. $EN(x, p)$ is increasing in x and a concave function of p with minima at $p = 0$ and 1 and a maximum at

$$p_{\max} = (1-\sqrt{1-x})/x, \quad 0<x<1. \tag{2.3}$$

When $0 < x < 1$, the possible values of EN range from 0 to p, with

$$\max_p EN = \frac{2(1-\sqrt{1-x})}{x} - 1 = 2p_{\max} - 1, \quad 0<x<1. \tag{2.4}$$

For $x > 1$, EN can be made arbitrarily large by letting $p \uparrow x^{-1}$. For $x = 1, EN \to 1$ as $p \to 1$.

A next step for this example, which we have not completed, is to examine the joint distribution of the sojourn times at nodes 1 and 3. In support of Conjectures 1.1 and 1.2, we would like to show that the correlation between the sojourn times is nonnegative, increasing in x and concave in p with a maximum at $p_{\max}$ in (2.3).

3. SPLITTING AND RECOMBINING

Consider the network in Figure 4 with four single-server nodes, each with the FIFO discipline and unlimited waiting space. Successive departures from node 1 are routed independently (of the history of the network and past routing) to node 2 with probability p and to node 3 with probability $1 - p$. Departures from nodes 2 and 3 proceed to node 4 and from there out of the system. Thus, customers going to node 2 (3) can overtake customers going to node 3 (2). However, unlike the Simon-Foley network, a customer who goes to node 3 will typically not pass some of the customers at node 2. Some of the customers at node 2 that the designated customer sees upon his arrival at node 3 may still depart from node 2 and arrive at node 4 before he does. Moreover, some of the customers that this customer sees at node 2 when he departs from node 3 may have arrived in the network after him. Hence, the distribution of the number N of customers overtaken is more complicated here.

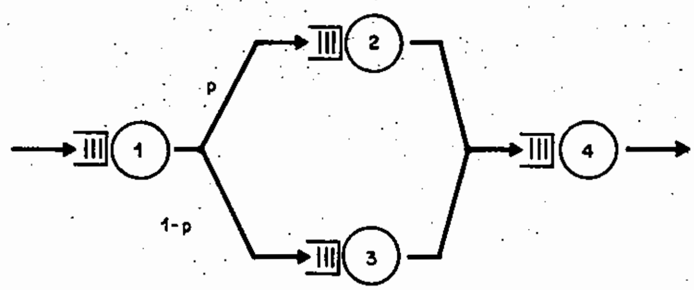

FIG. 4. Splitting and recombining: customers going to node 2 can overtake customers going to node 3 and vice versa.

We remark that this four-node network reduces to the Simon-Foley network when the service times at node 3 are all 0. Also nodes 2 and 3 together can be thought of as a single two-server node in which customers are randomly assigned to one of the servers upon arrival.

It is useful to consider a new random quantity, C, related to the number N, defined as the number of customers seen upon departure from node 1 at node 2 if he goes to node 3 or at node 3 if he goes to node 2. Obviously, C is the maximum number of customers that could be overtaken, so that

$$N \leqslant C. \tag{3.1}$$

Begin by letting the arrival and service processes be general stationary point processes and consider the network in equilibrium. Let Q_j be the queue length at node j seen by a departure from node 1. The probability distribution and mean of C are

$$P(C=k) = pP(Q_3=k) + (1-p)\,P(Q_2=k), \qquad k \geqslant 0,$$

and

$$EC = pEQ_3 + (1-p)\,EQ_2.$$

As in Section 2, in this general setting it is easy to bound the probability of a switch of order for any two customers: one must go to node 2 and the other must go to node 3, so that the probability a customer passes the kth previous arrival is bounded as: $1-\gamma(k) \leqslant 2p(1-p) \leqslant \frac{1}{2}$.

Suppose that the sequences of service times at nodes 2 and 3 are each iid and independent of the departure process from node 1. Then it is not difficult to obtain an expression for EN as a function of p, the distribution of (Q_2, Q_3), and the service rates μ_2 and μ_3:

$$EN(\mu_1, \mu_2, p) = \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} P(Q_2=j, Q_3=k)\left[p\sum_{m=0}^{k-1}(k-m)\,\alpha(j,m,q) + (1-p)\sum_{m=0}^{j-1}(j-m)\,\alpha(k,m,1-q)\right] \tag{3.2}$$

where $q = \mu_2/(\mu_2 + \mu_3)$ and $\alpha(j, m, q)$ is the probability in a sequence of Bernoulli trials with probability q of success that exactly m failures precede the jth success, which of course has the negative binomial distribution

$$\alpha(j, m, q) = \binom{m + j - 1}{m} q^j (1 - q)^m; \tag{3.3}$$

see Feller [5, p. 165]. Unfortunately, however, (3.2) is a bit cumbersome.

Now consider the case of a Jackson network. Since Q_j has a geometric distribution,

$$EC = \frac{p(1 - p)x_3}{1 - (1 - p)x_3} + \frac{p(1 - p)x_2}{1 - px_2}, \tag{3.4}$$

where $x_j = \lambda/\mu_j$. Suppose in addition that $x_2 = x_3 = x$, so that EC in (3.4) depends on the two parameters x and p. For stability, $1 - x^{-1} < p < x^{-1}$, which is binding for $x \geqslant 1$. Solutions are possible if and only if $0 < x < 2$.

One might conjecture that in the symmetric case with $x_2 = x_3 = x$ that EC is always maximized by $p = \frac{1}{2}$, but this is not the case.

Theorem 3.1. In the symmetric case, with $x_2 = x_3 = x$,

$$\sup_p EC(x, p) = \begin{cases} \infty, & 1 < x < 2 \\ \dfrac{x/2}{1 - (x/2)}, & 0 < x \leqslant 1, \end{cases} \tag{3.5}$$

with the extreme approached as $p \uparrow x^{-1}$ or $p \downarrow 1 - x^{-1}$ for $1 < x < 2$; the extreme attained for all p, $0 < p < 1$, for $x = 1$; and the extreme attained uniquely by $p = \frac{1}{2}$ for $0 < x < 1$.

Proof. For $x > 1$, let p approach the indicated extreme values. For $x < 1$, differentiate (3.4) with $x_2 = x_3 = x$ and observe that the polynomial must have zeroes less than 0 and greater than 1 because $EC(x, p) = 0$ for $p = 0$ and 1, so that there is at most one zero for $0 < p < 1$, this zero occurring at $p = \frac{1}{2}$. ∎

We can apply Theorem 3.1 to deduce results for EN in the symmetric case. First we establish the following lower bound for EN.

Theorem 3.2. In the symmetric case,

$$EN(x, p) \geqslant p(1 - p)x \left(\frac{x(1 - p)}{1 - x(1 - p)} - 2 \right). \tag{3.6}$$

Proof. The arbitrary customer goes to node 2 and finds it empty with probability $p \cdot P(Q_2 = 0) = p(1 - p)x$. If node 3 had infinitely many customers, the number of departures from node 3 before the arbitrary customer at node 2 departs would be geometrically distributed with parameter $\frac{1}{2}$ and thus mean 2. Hence, we have (3.6). ∎

Theorem 3.3. In the symmetric case,

$$\sup_p EN(x,p) = \begin{cases} \infty, & 1 < x < 2 \\ 1, & x = 1, \end{cases} \tag{3.7}$$

and

$$\sup_p EN(x,p) < \frac{x/2}{1-(x/2)}, \quad 0 < x < 1, \tag{3.8}$$

with the supremum approached as $p \uparrow x^{-1}$ for $x \geq 1$.

Proof. For $x > 1$, apply Theorem 3.1: use the fact that $EC(x, p) \to \infty$ as $p \uparrow x^{-1}$. For $x \leq 1$, combine (3.1) and (3.5) to obtain the upper bound

$$\sup_p EN(x,p) \leq \frac{x/2}{1-(x/2)}. \tag{3.9}$$

It is not difficult to show that $EN(x, p)$ is strictly less than this upper bound for $x < 1$ because there is positive probability that one of the C customers will not actually be overtaken. For $x = 1$, apply Theorem 3.2 with (3.9). With $x = 1$, the lower bound is $(1-p)^2 - 2p$, which approaches 1 as $p \to 0$. ∎

Remark. From (3.8) and (2.3), we conclude that in the Jackson case $\sup_p EN(x, p)$ is strictly larger for the Simon-Foley network in Section 2 than for the four-node network in Figure 4 for x just less than 1. The two agree for $x \geq 1$. We *conjecture* that the strict ordering holds for all x, $0 < x < 1$.

It is also possible to split the departures from node 1 into more than 2 parts and recombine, as shown in Figure 5. If the maximum probability p_j of going from node 1 to node j, $2 \leq j \leq n+1$, converges to 0 as $n \to \infty$, then in the limit the nodes $\{2, 3, \ldots, n+1\}$ behave like an infinite-server system. With iid exponential service times, with probability very close to 1, the arbitrary customer finds a free server and then passes each other customer with probability $\frac{1}{2}$. Hence we have

Theorem 3.4. If $n \to \infty$ and $\max_{2 \leq j \leq n+1} p_j \to 0$, then in the symmetric Jackson case

$$\lim_{n\to\infty} EN(x, p_2, \ldots, p_n) = x/2, \quad x > 0. \tag{3.10}$$

Proof. The expected number of customers in the system when the designated customer arrives is EQ for an $M/M/\infty$ system, which is x. As noted above, in the limit the designated customer arrives at an empty node. He then passes each other customer with probability $\frac{1}{2}$. ∎

Remark. Equation (3.10) is strictly less than the supremum of EN over $p_1, \ldots, p_n$ for $x \geq 1$ and any n with $n \geq 2$ because it is strictly less than (3.7). This shows that $p_j = n^{-1}$ does not cause the most overtaking in the symmetric case.

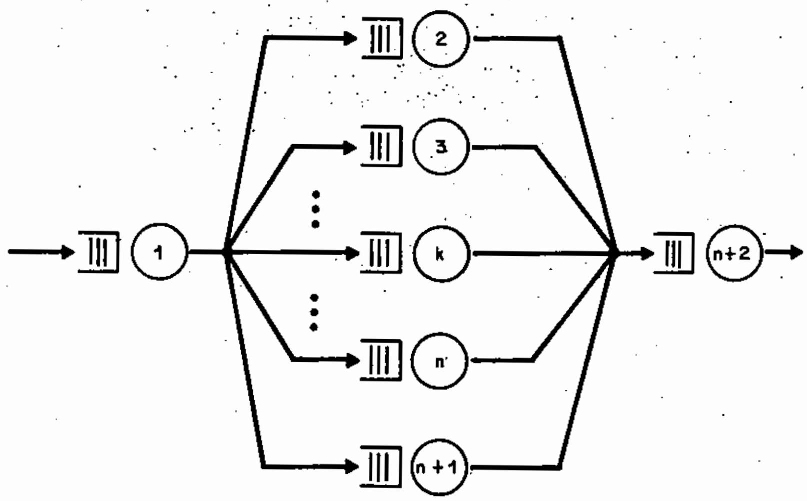

FIG. 5. Splitting into n parts and recombining.

4. A MULTISERVER QUEUE

Consider the standard s-server queue with unlimited waiting room and the first-come first-served discipline. Let the arrival process be a general stationary process but let the service times be iid and independent of the arrival process. If the service times are exponentially distributed, then

$$EN = \frac{P(Q \geq s-1)(s-1)}{2} + \sum_{k=1}^{s-2} \frac{P(Q=k)\,k}{2}, \tag{4.1}$$

where Q is the equilibrium number in system seen by an arrival. As $s \to \infty$, this approaches $EQ/2$. If the service-time distribution has increasing (decreasing) failure rate, then (4.1) is an (upper) lower bound.

In the $M/G/s$ case, Q has the time-stationary distribution. In this case the expected number of busy servers seen by an arrival is

$$EB = \lambda/\mu \tag{4.2}$$

{see (4.2.3) of [6]}, so that

$$EN = \frac{EB}{2} - \frac{P(Q \geq s)}{2}. \tag{4.3}$$

Hence, we have established:

Theorem 4.1. For the $M/G/s$ queue,

$$\sup_{s \geq 1} EN(s) = EN(\infty) = \frac{EB}{2} = \frac{\lambda}{2\mu}. \tag{4.4}$$

For $\lambda/\mu \geqslant 1$, (4.4) is strictly less than EN for the networks in Sections 2 and 3. We *conjecture* that this ordering holds for all $x = \lambda/\mu$.

If instead of a Poisson arrival process, we have a renewal arrival process with a NBUE (new better than used in expectation) or NWUE renewal interval distribution, then there is a stochastic order relation between Q as seen by the arrival and the time-stationary number in system; see p. 114 of Franken et al. [6]. Hence, (4.1) holds as an inequality with the time-stationary distribution and (4.4) is valid as an inequality.

In the setting of (4.1), that is, with exponential service times, it is easy to find bounds on $\gamma(k)$, the probability that an arbitrary customer does not overtake the kth previous arrival. We give bounds for $k = 1$.

Theorem 4.2. With exponential service times,

$$P(Q \geqslant s)\left(\frac{s-1}{2s}\right) + \sum_{j=1}^{s-1} P(Q=j)\left(\frac{j}{2s}\right) \leqslant 1 - \gamma(1)$$

$$\leqslant P(Q \geqslant s)\left(\frac{s-1}{2s}\right) + P(Q \leqslant s-1)\frac{1}{2} \leqslant \frac{1}{2}. \quad (4.5)$$

Proof. If $Q \geqslant s$, then the previous arrival could not have departed yet. Given $Q \geqslant s$, the previous arrival will still be in the system when the designated customer begins service with probability $(s-1)/s$. Once both customers are in service, overtaking occurs with probability $\frac{1}{2}$. The lower bound is attained by assuming the maximum number of departures since the previous customer arrived; the upper bound is attained assuming none: If $(Q \leqslant s-1)$, the previous customer might still be there. For the lower bound, if $Q = j$, the previous customer is still in service with probability at least

$$\frac{j}{s} = \left(\frac{j}{j+1}\right)\left(\frac{j+1}{j+2}\right)\cdots\left(\frac{s-1}{s}\right). \quad \blacksquare$$

We now give an example to show that $\gamma(k)$ can be arbitrarily close to $\frac{1}{2}$ in a multi-server node.

Example 4.1. Consider the standard $M/M/s$ queue with fixed service rate μ. Let $s \to \infty$ and let the arrival rate $\lambda \to \infty$ so that the traffic intensity $\rho \equiv \lambda/s\mu$ is fixed. For large λ and s, with probability near one, both customers will find free servers and thus enter service immediately upon arrival. Since $\lambda \gg \mu$, with probability near one the later customer will arrive before the earlier one departs. Conditional on both being in service, the probability of overtaking is obviously $\frac{1}{2}$.

5. A SINGLE SERVER WITH INSTANTANEOUS BERNOULLI FEEDBACK

Consider a single-server queue with unlimited waiting room and the FIFO discipline. (See Fig. 6.) Let each departure be fed back to the end of the queue for another service with probability p, independent of the history of the system. As before, assume that equilibrium exists for the system, that is, there are time-stationary and customer-stationary versions of the queue length process; see [6].

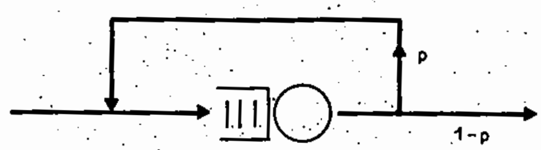

FIG. 6. Single server with instantaneous Bernoulli feedback.

First, note that to calculate the distribution of N, the number of customers overtaken by an arbitrary customer, it suffices to consider the customers in the system upon the arrival of the arbitrary customer. A customer can only overtake other customers that arrived earlier and only customers that have not already departed. Next, note that overtaking refers only to the final order upon departures; with feedback, two customers may switch order several times in the queue before departure; this preliminary switching is not counted. Hence, $EN = p^*EQ$, where p^* is the probability that the arbitrary customer overtakes a specific customer he finds upon arrival and Q is the number of customers in the system seen by the arrival. (The events indicating that different customers will be overtaken are in general dependent, but the dependency does not affect the expected value.) An arrival overtakes any customer it finds in the queue upon arrival with probability

$$\frac{p}{1+p} = \frac{p(1-p)}{1-p^2} = \sum_{k=0}^{\infty} (1-p)\, p^{2k+1}, \tag{5.1}$$

independently of the arrival and service processes. Thus

$$EN = (EQ)\, p/(1+p). \tag{5.2}$$

Note that $E\bar{N}$ is harder to compute directly. However, by the argument outlined in (1.1) and (1.2), it is possible to show that $E\bar{N} = EN$. It is also easy to compute $E\bar{N}$ and verify that $E\bar{N} = EN$ in the case of Poisson arrivals, using the following argument, due to D. R. Smith (personal communication). Each customer arriving during a test customer's stay has probability $p/(1+p)$ of overtaking it. The expected number arriving during the stay is just the arrival rate times the expected length of stay or EQ, by Little's law; p. 399 of [7].

In an $M/M/1$ queue with Bernoulli feedback, having parameters λ (arrival rate), μ (service rate), and p,

$$EQ = \rho/(1-\rho), \tag{5.3}$$

where

$$\rho = \lambda/(\lambda + \mu(1-p)), \tag{5.4}$$

which is increasing in λ and p and decreasing in μ. Hence, so is EN.

Back in the general setting, the probability $\gamma(k)$ that an arbitrary customer does not overtake the kth previous customer is bounded below by

$$\gamma(k) \geqslant 1/(1+p) \geqslant 1/2. \tag{5.5}$$

We would have $\gamma(k) = 1/(1 + p)$ if the kth preceding customer were certain not to have departed before the designated customer arrives. The probability of such a departure can be made arbitrarily small by making the traffic intensity of the queue large.

In the $M/M/1$ setting, it is not difficult to show that $\gamma(k)$ is a decreasing function of γ and p and an increasing function of μ.

This work was motivated by Alan Goldman's questions at one of Ralph Disney's lectures at Johns Hopkins University in June 1982. I am grateful to my colleague D. R. Smith and the referees for helpful comments.

References

[1] R. E. Barlow and F. Proschan, *Statistical Theory of Reliability and Life Testing.* Holt, Rinehart and Winston, New York (1975).

[2] P. J. Burke, The output of a queueing system. *Operations Research* 4 (1956) 699–704.

[3] P. J. Burke, The dependence of sojourn times in tandem *M/M/s* queues. *Operations Research* 17 (1969) 754–755.

[4] R. L. Disney, *Queueing Networks and Applications.* The Johns Hopkins Press, Baltimore, forthcoming.

[5] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. I, Third Edition. Wiley, New York (1968).

[6] P. Franken, D. König, U. Arndt, and V. Schmidt, *Queues and Point Processes.* Akademie-Verlag, Berlin (1981).

[7] D. P. Heyman and M. J. Sobel, *Stochastic Models in Operations Research*, Vol. I. McGraw-Hill, New York (1982).

[8] J. R. Jackson, Networks of waiting lines. *Operations Research* 5 (1957) 518–521.

[9] J. R. Jackson, Jobshop-like queueing systems. *Management Science* **10** (1963) 131–142.

[10] T. Kamae, U. Krengel, and G. L. O'Brien, Stochastic inequalities on partially ordered spaces. *Annals of Probability* 5 (1977) 899–912.

[11] F. P. Kelly, *Reversibility and Stochastic Networks.* Wiley, New York (1979).

[12] P. C. Kiessler, A simulation analysis of sojourn times in a Jackson network. Technical Report VTR 8016, Department of Industrial Engineering and Operations Research, Virginia Polytechnic Institute and State University (1980).

[13] J. F. C. Kingman, The effect of queue discipline on waiting time variance. *Proceedings of the Cambridge Philosophical Society* 58 (1962) 163–164.

[14] J. F. C. Kingman, Queue disciplines in heavy traffic. *Mathematics of Operations Research* 7 (1982) 262–271.

[15] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications.* Academic, New York (1979).

[16] B. Melamed, Sojourn times in queueing networks. *Mathematics of Operations Research* 7 (1982), 223–244.

[17] E. Reich, Waiting times when queues are in tandem. *Annals of Mathematical Statistics* 28 (1957) 768–773.

[18] E. Reich, Note on queues in tandem. *Annals of Mathematical Statistics* 34 (1963) 338–341.

[19] B. Simon and R. D. Foley, Some results on sojourn times in acyclic Jackson networks. *Management Science* 25 (1979) 1027–1034.

[20] O. A. Vasicek, An inequality for variance of waiting time under a general queueing discipline. *Operations Research* 25 (1977) 879–884.

[21] J. Walrand and P. Varaiya, Sojourn times and the overtaking condition in Jacksonian networks. *Advances in Applied Probability* 12 (1980) 1000–1018.
[22] W. Whitt, The effect of variability in the GI/G/s queue. *Journal of Applied Probability* 17 (1980) 1062–1071.
[23] R. W. Wolff, Poisson arrivals see time averages. *Operations Research* 30 (1982) 223–231.

Received February 21, 1983
Accepted December 8, 1983