COMMUN. STATIST.—STOCHASTIC MODELS, 8(4), 719-732 (1992)

COMPARISONS OF MULTI-SERVER QUEUES WITH FINITE WAITING ROOMS

Arthur W. Berger

AT&T Bell Laboratories
Holmdel, NJ 07733

Ward Whitt

AT&T Bell Laboratories
Murray Hill, NJ 07974-0636

Key words: multi-server queues, queues with finite waiting rooms, throughput, balking, stochastic comparisons, subsequence stochastic ordering.

ABSTRACT

In this paper we consider s-server queues with capacity c, $1 \le s \le c \le \infty$, the first-come first-served queue discipline and very general arrival and service processes. We show that the admission epochs and departure epochs decrease, so that the throughput increases, when any of the following changes occur: (1) the number of servers s increases, (2) the capacity c increases, (3) the external arrival counting process increases or (4) the service times decrease, provided that the service times are assigned in order of service initiation and that a subsequence ordering is used to compare arrival counting processes. The subsequence ordering for the arrival processes is very important for obtaining positive results with finite waiting rooms. The subsequence ordering holds between a superposition point process and its component point processes. The subsequence ordering can often be applied via its stochastic generalization, the stochastic subsequence ordering, which is implied by a failure rate ordering.

1. Introduction

In this paper we establish sample-path and stochastic comparisons for multi-server queues with finite waiting rooms. Our results extend previous work by Sonderman [4], [5] and Whitt [6]. Our extension is motivated by our study of the impact of a job buffer in a token-bank rate-control throttle in Berger and Whitt [1], and we apply our results there.

Let A/A/s/c denote an s-server queue with total capacity c, i.e., with an extra waiting room of size $c - s$, $1 \le s \le c \le \infty$, in which customers are served in order of their arrival by the first available server without defections after entering the system. If there is a finite waiting room and if the system is full when a customer arrives, then the customer leaves without receiving service or affecting future arrivals. The first A means that the arrival process is arbitrary, not necessarily renewal and not necessarily stationary. The second A means that the service times are also arbitrary. We insert a GI to indicate that the interarrival or service times are independent and identically distributed (i.i.d.). For the service times, GI also means that the service times are independent of the arrival processes. As usual, we use M and D as the special cases of GI with exponential and deterministic distributions.

A major concern is simultaneous events; i.e., there may be batch arrivals and arrivals at the same instant as departures. The treatment of simultaneous events is not critical to the comparison results, provided that we consistently treat the two systems being compared. To be specific, we stipulate that departures of customers in the system occur before new arrivals are considered. All arriving customers are labeled and thus ordered, including customers arriving in a batch. We treat (possibly admit and possibly immediately serve) the arriving customers one at a time in order of their index.

Typically customers depart faster when they arrive faster. Indeed, this is true in great generality for A/A/s/∞ queues; see Theorem 12(a) and (b) of Whitt [6].

However, for finite-capacity queues, positive comparisons are not so easy to make. For example, consider a D/D/1/1 model with service times 1. If interarrival times decrease from 1.0 to 0.9, then the interdeparture times *increase* from 1.0 to 1.8.

However, a positive comparison result does hold for M/GI/s/c models, as was established in Theorem 2 of Sonderman [4]. In Theorem 12(c) of Whitt [6] it was noted that this result extends to A/GI/s/c models using the failure-rate ordering $\leq_1$ defined on p. 206 of [6] for the arrival processes, by the same proof. Here we observe that in fact essentially the same proof applies again using the weaker subsequence stochastic ordering $\leq_2$ from [6] for the arrival processes, which we denote by $\subseteq_{st}$, following Budka and Yao [2] and Budka [3].

In particular, we say that one counting function $\{A^1(t) : t \geq 0\}$ (e.g., one sample path of a stochastic counting process) is less than or equal to another $\{A^2(t) : t \geq 0\}$ in the *subsequence ordering* and we write $A^1 \subseteq A^2$, if the arrival epochs of A^1 are a subsequence of the arrival epochs of A^2 and $A^1(t) - A^1(t-) \leq A^2(t) - A^2(t-)$. Thus, batch sizes in A^1 are smaller than the batch sizes at the same time in A^2. The subsequence ordering arises naturally when A^2 is the superposition of A^1 and another counting function.

The applications of the subsequence ordering are greater than may be apparent, because it may hold in a stochastic sense when it does not hold directly; see [6] for more discussion. For this purpose, we say that one counting process $\{A^1(t) : t \geq 0\}$ is less than or equal to another $\{A^2(t) : t \geq 0\}$ in the *subsequence stochastic ordering*, which we denote by $A^1 \subseteq_{st} A^2$, if it is possible to construct new processes $\{\tilde{A}^1(t) : t \geq 0\}$ and $\{\tilde{A}^2(t) : t \geq 0\}$ on a new probability space such that $\{\tilde{A}^i(t) : t \geq 0\}$ has the same probability law (as a stochastic process, i.e., finite-dimensional distributions) as $\{A^i(t) : t \geq 0\}$ for each i and $\tilde{A}_1 \subseteq \tilde{A}_2$. Note that $A^1 \subseteq A^2$ trivially implies that $A^1 \subseteq_{st} A^2$. However, it is significant that we can apply the ordering $A^1 \subseteq_{st} A^2$ above to make stochastic comparisons that do not stem from superpositions, because $\subseteq_{st}$ is

implied by the failure-rate ordering $\leq_1$ in [6]. Additional stochastic subsequence orderings are established by Budka and Yao [2].

Moreover, as is clear from Sonderman [4], the GI assumption for the service times is not needed if the successive service times are assigned to customers when they start service rather than upon external arrival. Otherwise, this assignment rule is critical; see Whitt [7].

In Section 2 we define the basic processes and state the main result. In Section 3 we give the proof. In Section 4 we state some consequences. The corollaries there show that schemes in which customers balk (leave after joining the queue but before starting service) or extra customers are rejected (not admitted even when there is space) reduce throughput in the strong sample-path sense. We conclude by stating stochastic comparisons that follow from the sample-path comparisons.

2. The Main Result

We start by indicating how to recursively define the successive admission and departure epochs in an A/A/s/c queue to facilitate inductive proofs. We focus on a single sample path, so that the analysis is deterministic. (See Corollary 4 in §4 for stochastic consequences.)

Let A_k be the epoch that the k^{th} arrival comes to try to enter the system; let B_k be the epoch that the k^{th} admitted customer is admitted; let I_k be the index j of the arrival epoch A_j corresponding to the k^{th} admitted customer; let S_k be the k^{th} service time, which we assume is assigned to the k^{th} customer to start service; let X_k be the departure epoch of the k^{th} admitted customer; and let D_k be the epoch of the k^{th} departure. (Note that we need not have $X_k = D_k$, because we allow multiple servers. Then overtaking in the service facility can occur.) Also, to account for batch arrivals, let N_k be the number of customers among the first k admitted customers that have been admitted at epoch B_k, i.e., $N_1 = 1$ and $N_{k+1} = 1$ if $B_{k+1} > B_k$, while $N_{k+1} = N_k + 1$ if $B_{k+1} = B_k$, $k \geq 1$.

We define the first c admission epochs and the first s departure epochs by setting

$$I_k = k \text{ and } B_k = A_k \text{ , } 1 \le k \le c \text{ , and } X_k = B_k + S_k \text{ , } 1 \le k \le s \quad (2.1)$$

We define the first departure epoch by

$$D_1 = \min\{X_1, \ldots, X_s\} \text{ .} \quad (2.2)$$

We then define I_{k+c}, B_{k+c}, X_{k+s} and D_k recursively for $k \ge 1$. (Of course, I_k and B_k are already defined for all k by (2.1) if $c = \infty$, and X_k is already defined for all k by (2.1) if $s = \infty$.) For this purpose, let $\min_k$ denote the k^{th} smallest number in a finite set of $j \ge k$ numbers. The definitions are:

$$I_{k+c} = \min\{j > I_{k+c-1} : A_j \ge D_k\} \text{ if } c < \infty \text{ , } k \ge 1 \text{ ,} \quad (2.3)$$

$$B_{k+c} = A_{I_{k+c}} \text{ if } c < \infty \text{ , } k \ge 1 \text{ ,} \quad (2.4)$$

$$X_{k+s} = S_{k+s} + \max\{D_k, B_{k+s}\} \text{ if } s < \infty \text{ , } k \ge 1 \text{ ,} \quad (2.5)$$

$$D_k = \min_k\{X_1, \ldots, X_{k+s-1}\} \text{ , } k \ge 1 \text{ .} \quad (2.6)$$

First, (2.3) stipulates that the $(k + c)^{\text{th}}$ admitted customer is the next arrival to occur after the $(k + c - 1)^{\text{st}}$ admitted customer and after or at the same time (since departures occur first) as the k^{th} departure epoch; this is the next customer that will find space in the system. (Formula (2.3) causes the difficulty in comparisons with finite rooms. Note that (2.3) does not appear when $c = \infty$.) Second, because we have the FIFO discipline in the waiting room, the $(k + s)^{\text{th}}$ admitted customer starts service at epoch $\max\{D_k, B_{k+s}\}$, and thus finishes at epoch X_{k+s} in (2.5). Finally, to determine the k^{th} ordered departure epoch it suffices to consider the k^{th} smallest of the departure epochs of the first $k + s - 1$ admitted customers as in (2.6), because subsequent admitted customers cannot begin service until after this departure epoch D_k.

Our main result is a direct sample-path comparison. In the following, we change the number of servers, the capacity and the service times as well as the

Theorem 1. *Consider two A/A/s/c queues with s and c indexed by subscripts 1 and 2, and the other model components indexed by superscripts 1 and 2. Let the service times be assigned to successive customers when service begins. Let the systems be initially empty. If* $s_1 \le s_2$, $c_1 \le c_2$, $A^1 \subseteq A^2$ *and* $S_k^1 \ge S_k^2$ *for all k, then* $B_k^1 \ge B_k^2$, $X_k^1 \ge X_k^2$ *and* $D_k^1 \ge D_k^2$ *for all k. Moreover, if* $B_k^1 = B_k^2$, *then* $N_k^1 \ge N_k^2$.

Remark 1. Note that we do not claim that $I_k^1 \ge I_k^2$ for all k or the reverse. Indeed, it is easy to see that these inequalities need not hold.

Remark 2. If $\overline{B}^i(t)$, $\overline{X}^i(t)$ and $\overline{D}^i(t)$ are the counting functions associated with the sequences B^i, X^i and D^i, respectively (counting the numbers in $[0,t]$), then Theorem 1 implies that $\overline{B}^1(t) \le \overline{B}^2(t)$, $\overline{X}^1(t) \le \overline{X}^2(t)$ and $\overline{D}^1(t) \le \overline{D}^2(t)$ for all t. However, we do *not* necessarily have the subsequence orderings $\overline{B}^1 \subseteq \overline{B}^2$, $\overline{X}^1 \le \overline{X}^2$ or $\overline{D}^1 \subseteq \overline{D}^2$.

Remark 3. Since customer $k + s$ starts service at $\max\{D_k, B_{k+s}\}$ for $k \ge 1$, the conclusion of Theorem 1 implies an ordering for these epochs as well.

3. Proof of Theorem 1

We apply mathematical induction. We first show that $N_k^1 \ge N_k^2$ whenever $B_k^1 = B_k^2$, assuming that we have established that $B_j^1 \ge B_j^2$ for all $j \le k$. Suppose that $B_k^1 = B_k^2 = B_{k-j}^2$ for some j with $j \ge 1$. Since $B_k^1 \ge B_{k-j}^1 \ge B_{k-j}^2 = B_k^2$, we have $B_k^1 = B_{k-j}^1$, so that indeed $N_k^1 \ge N_k^2$.

Since $A^1 \subseteq A^2$ by assumption, $A_k^1 \ge A_k^2$ for all k, so that

$$B_k^1 \ge A_k^1 \ge A_k^2 = B_k^2 , \quad 1 \le k \le c_2 , \tag{3.1}$$

by (2.1). Next, from (2.1) and (2.5),

$$X_k^1 \ge S_k^1 + B_k^1 \ge S_k^2 + B_k^2 = X_k^2 , \; 1 \le k \le s_2 . \tag{3.2}$$

Now, as our induction assumption, we assume that

$$B_k^1 \ge B_k^2 \text{ for all } k \le n + c_2 \tag{3.3}$$

and

$$X_k^1 \geq X_k^2 \text{ for all } k \leq n + s_2 \,, \tag{3.4}$$

which we have established for $n = 0$. Then

$$\begin{aligned} D_{n+1}^1 &= \min_{n+1}\{X_1^1, \ldots, X_{n+s_1}^1\} \geq \min_{n+1}\{X_1^2, \ldots, X_{n+s_1}^2\} \\ &\geq \min_{n+1}\{X_1^2, \ldots, X_{n+s_2}^2\} = D_{n+1}^2 \,. \end{aligned} \tag{3.5}$$

Moreover, for the principal case in which $c_2 < \infty$, by (2.3),

$$I_{n+1+c_2}^1 = I_{(n+1+c_2-c_1)+c_1}^1 = \min\{j > I_{n+c_2}^1 : A_j^1 \geq D_{n+1+c_2-c_1}^1\} \tag{3.6}$$

and

$$I_{n+1+c_2}^2 = \min\{j > I_{n+c_2}^2 : A_j^2 \geq D_{n+1}^2\} \,. \tag{3.7}$$

However, by (3.3) and (3.5),

$$B_{n+c_2}^1 \geq B_{n+c_2}^2 \text{ and } D_{n+1+c_2-c_1}^1 \geq D_{n+1}^1 \geq D_{n+1}^2 \,. \tag{3.8}$$

Moreover, if $B_{n+c_2}^1 = B_{n+c_2}^2$, then $N_{n+c_2}^1 \geq N_{n+c_2}^2$. Consequently, we can show that the $(n + c_2 + 1)^{st}$ admission occurs first in system 2, i.e.,

$$B_{n+1+c_2}^1 = A_{I_{n+1+c_2}^1}^1 \geq A_{I_{n+1+c_2}^2}^2 = B_{n+1+c_2}^2 \,. \tag{3.9}$$

To see that (3.9) indeed holds, we start with $I_{n+c_2+1}^1$, which is defined by the minimization in (3.6), and the corresponding admission epoch $A_{I_{n+c_2+1}^1}$. From the subsequence ordering $A^1 \subseteq A^2$, a customer also arrives to system 2 at this epoch. Moreover, if there is a batch arriving at this epoch in system 1, there is at least as large a batch arriving at that epoch in system 2. Suppose that customer number $I_{n+c_2+1}^1$ in system 1 is the k^{th} customer in the batch to arrive at epoch $B_{n+c_2+1}^1$; then let j' be the index of the k^{th} customer in the batch to arrive at that epoch in system 2. (There is such a customer in system 2 by the argument above.) It suffices to show that $j' \geq I_{n+c_2+1}^2$ in order to verify (3.9), because then

$$A_{I_{n+c_2+1}^1}^1 = A_{j'}^2 \geq A_{I_{n+c_2+1}^2}^2 \,. \tag{3.10}$$

We next apply (3.6) and (3.8) to deduce that

$$A^2_{j'} = A^1_{I^1_{n+c_1+1}} \geq D^1_{n+1+c_2-c_1} \geq D^2_{n+1} \tag{3.11}$$

and

$$A^2_{j'} = A^1_{I^1_{n+c_1+1}} = B^1_{n+c_2+1} \geq B^1_{n+c_2} \geq B^2_{n+c_2} \ . \tag{3.12}$$

In order to show that $j' \geq I^2_{n+c_2+1}$, by (3.7) and (3.11), it suffices to show that

$$j' > I^2_{n+c_2} \ . \tag{3.13}$$

We now proceed to verify (3.13).

First, if $A^2_{j'} > B^2_{n+c_2}$ (as occurs if $B^1_{n+c_2} > B^2_{n+c_2}$ by (3.12)), then $j' > I^2_{n+c_2}$, so that (3.13) holds. Hence, it suffices to consider the case in which $A^2_{j'} = B^2_{n+c_2}$. If $A^2_{j'} = B^2_{n+c_2}$, then, by (3.12), $B^1_{n+c_2+1} = B^1_{n+c_2}$, $I^1_{n+c_2+1} = I^1_{n+c_2} + 1$ and $N^1_{n+c_2+1} \geq 2$. Since $A^1 \subseteq A^2$, there is also a batch arrival to system 2 at epoch $B^1_{n+c_2+1}$ and the batch size is greater than or equal to $N^1_{n+c_2+1}$. Moreover, from the definition of j', customer j' is the $(N^1_{n+c_2+1})^{st}$ customer in this batch to arrive to system 2. Consider the $(j'-1)^{st}$ arrival to system 2. This customer is also part of the batch to arrive at epoch $B^1_{n+c_2+1}$ in system 2 since $N^1_{n+c_2+1} \geq 2$. In particular, customer $j'-1$ is the $(N^1_{n+c_2+1}-1)^{st}$ or, equivalently, the $(N^1_{n+c_2})^{th}$ customer in this batch. Now, since $A^2_{j'} = B^2_{n+c_2}$ by assumption above, $B^1_{n+c_2} = B^2_{n+c_2}$ by (3.12). Moreover, since $B^1_{n+c_2} = B^2_{n+c_2}$, $N^1_{n+c_2} \geq N^2_{n+c_2}$, as was shown at the beginning of the proof. Thus $j' - 1 \geq I^2_{n+c_2}$, which implies that $j' > I^2_{n+c_2}$, which is (3.13).

Finally, having established (3.9), we show that (3.4) holds for $n+1$ in the principal case in which $s_2 < \infty$. By (2.5),

$$X^1_{n+s_2+1} = S^1_{n+s_2+1} + \max\{D^1_{n+s_2+1-s_1}, B^1_{n+s_2+1}\}$$

$$\geq S^1_{n+s_2+1} + \max\{D^1_{n+1}, B^1_{n+s_2+1}\}$$

$$\geq S^2_{n+s_2+1} + \max\{D^2_{n+1}, B^2_{n+s_2+1}\} = X^2_{n+s_2+1} . \quad (3.14)$$

This completes the proof.

4. Consequences

Theorem 1 has useful implications for modifications of the A/A/s/c model in which some customers that could be admitted are rejected. For example, Theorem 1 directly covers the case of two classes of arrivals, with class 1 being admitted whenever there is any empty space in the system with capacity c, and class 2 being admitted only when there are at least r empty spaces (e.g., trunk reservation). More generally, Theorem 1 implies that any such scheme decreases the total throughput. Of course, such rejection policies may nevertheless be useful to discriminate between the classes if one class is more valuable than another.

Corollary 1. *Consider an A/A/s/c queue in which service times are assigned to successive customers when service begins; let it be indexed by the superscript 2. Consider a modified system indexed by a superscript 1 in which some arrivals that could be admitted are rejected. Then* $B^1_k \geq B^2_k$, $X^1_k \geq X^2_k$ *and* $D^1_k \geq D^2_k$ *for all* k.

Proof. The original sequence of potential arrival epochs is A^2. For each realization, we can regard the rejection scheme as an elimination of some terms from the sequence A^2, so that $A^1 \subseteq A^2$. Hence, we can apply Theorem 1. ■

Another application of Theorem 1 is to show that customers balking always reduces throughput. By balking, we mean that a customer in queue leaves before beginning service and before being assigned a service time. The balking introduces a new simultaneous event in addition to arrivals and departures. Again, the ordering of events at one epoch is not critical for the results, provided we are consistent, but to be specific we assume that balking occurs after all arrivals and departures at the same epoch. Moreover, multiple customers balking at the same epoch are treated in order of their arrival. Let C^i_k and Z^i_k be the epochs that the k^{th} customer to start service in system i is admitted to the system and finishes service, respectively.

In the spirit of (2.1)–(2.6), we first indicate how to construct the sequence $\{(C_k,Z_k) : k \geq 1\}$ when there is balking. As in the proof of Theorem 1, such a construction facilitates the comparison proof. By a *balking sequence*, we mean a sequence $\{(T_k,J_k) : k \geq 1\}$ with T_k being the time of the k^{th} balk and J_k be the external arrival index j of the balking customer. We assume that $\{T_k\}$ is nondecreasing. We let $\{(T_k,J_k)\}$ be defined in terms of the basic model data $\{(A_k,S_k)\}$ in an unspecified way. Of course, the J_k^{th} arrival must be in the system without having started service at time T_k; otherwise no balk takes place.

It is significant that C_k and Z_k record information only for customers that have started service. Hence, these variables are not changed if we reject customers immediately upon arrival who will eventually balk. However, we must be careful not to admit extra customers who would not be admitted.

Given (T_1,J_1), we construct a new arrival process $\{A_k^{(1)}\}$ in which customer J_1 never arrives and all arrivals from this time, A_{J_1}, until T_1 that were blocked in the original system are deleted as well. (For example, if the most recent admission before time T_1 filled the queue, then to properly account for this balk, we need to delete all arrivals after the most recent admission until time T_1. Since balking at time t takes place after all arrivals and departures at time t, it is indeed appropriate to delete all arrivals up to time T_1 after the most recent admission.) Next, using the revised basic model data $\{(A_k^{(1)},S_k) : k \geq 1\}$, with the original service times, we define a new arrival sequence $\{A_k^{(2)}\}$ to represent the second balk (T_2,J_2). We continue in this way, defining $\{A_k^{(m)}\}$ in terms of $\{(A_k^{(m-1)},S_k)\}$ in order to represent (T_m,J_m), and so on.

The desired sequence $\{(C_k,Z_k)\}$ is the sequence $\{(B_k,X_k)\}$ associated with all these rejections. In particular, consider the sequence $\{(A_k^{(m)},S_k,B_k^{(m)},X_k^{(m)},D_k^{(m)})\}$ obtained after making the adjustments for $T_1,\ldots,T_m$. Then $C_j = B_j^{(m)}$ and $Z_j = X_j^{(m)}$ for all customers j that start service before time T_m; i.e., for all $j > s$ such that $\max\{D_{j-s}^{(m)}, B_j^{(m)}\} < T_m$ (see Remark 3 above), and for $j \leq s$, $C_j = B_j = B_j^{(m)}$ and $Z_j = X_j = X_j^{(m)}$ since the first s arrivals enter service immediately, and thus never balk.

Since C_k and Z_k associated with balking have been constructed from finitely many rejections, we can apply Corollary 1 to prove the following result.

Corollary 2. *Consider an A/A/s/c model with $s < c$ in which service times are assigned when service begins; let it be indexed by the superscript 2. Consider a modified system indexed by superscript 1 in which some customers in queue balk (leave after joining the queue but before beginning service and before being assigned a service time). Then $C_k^1 \geq C_k^2$, $Z_k^1 \geq Z_k^2$ and $D_k^1 \geq D_k^2$ for all k.*

Remark 4. Note that we do not claim that $B_k^1 \geq B_k^2$ or $X_k^1 \geq X_k^2$ in Corollary 2, and indeed these inequalities need not hold. (If many balk, then it will be easier to admit more.)

In our motivating application [1] we actually want to make comparisons when balking occurs simultaneously with other changes. We apply Corollary 2, Theorem 1 without balking, and transitivity to obtain the following generalization.

Corollary 3. *Consider two A/A/s/c queues satisfying the conditions of Theorem 1. Suppose that in addition some customers balk in system 1. Then $C_k^1 \geq C_k^2$, $Z_k^1 \geq Z_k^2$ and $D_k^1 \geq D_k^2$ for all k.*

Proof. Let system 3 denote system 1 without the balking. By Corollary 2 and Theorem 1,

$$C_k^1 \geq C_k^3 = B_k^3 \geq B_k^2 = C_k^2 \quad \text{for } \textit{all } k\,.$$

The other processes are treated similarly. ■

Remark 5. Remark 3 applies to Corollaries 2 and 3, but now the epoch that the $(k+s)^{\text{th}}$ customer to start service starts service is $\max\{D_k, C_{k+s}\}$.

The counterexamples to throughput orderings in Whitt [7] occur because the service times are not assigned when service begins. However, in many systems the service times are naturally associated with the arriving customers, rather than being assigned at the beginning of service. Then we can still obtain *stochastic* comparison results if we assume that the service times are i.i.d. and independent of the arrival process; see [6]. The desired conclusion is stochastic order for stochastic sequences. We say that one sequence $\{Y_k^1 : k \geq 1\}$ is *stochastically*

less than or equal to another $\{Y_k^2 : k \geq 1\}$, and write $Y^1 \leq_{st} Y^2$, if it is possible to construct new sequences $\{\tilde{Y}_k^1 : k \geq 1\}$ and $\{\tilde{Y}_k^2 : k \geq 1\}$ on a common probability space, so that $\{\tilde{Y}_k^i : k \geq 1\}$ has the same distribution as $\{Y_k^i : k \geq 1\}$ for each i and $\tilde{Y}_k^1 \leq \tilde{Y}_k^2$ for all k. The ordering $Y^1 \leq_{st} Y^2$ is known to be equivalent to $Ef(Y^1) \leq Ef(Y^2)$ for each nondecreasing bounded real-valued function f on the space of sequences R^∞ with the usual ordering; see [6] and references there. Let $Y_1^1 \leq_{st} Y_1^2$ denote stochastic order for individual random variables.

The following stochastic comparison extends Theorem 12(c) of [6], which in turn extends results in [4] and [5]. When we allow balking (which is new), we assume that the sequence $\{(C_k,Z_k) : k \geq 1\}$ is a measurable function of the basic model data $\{(A_k,S_k) : k \geq 1\}$, which for practical purposes is without loss of generality, but we do not specify any specific balking rule. Given the construction of $\{(C_k,Z_k)\}$ before Corollary 2, it suffices for the balking sequence $\{(T_k,J_k) : k \geq 1\}$ to be a measurable function of the basic model data $\{(A_k,S_k)\}$. In actual stochastic applications, we would have the event $\{T_k \leq t\}$ depend on the $\{(A_k,S_k)\}$ through the history of the system over the time interval $[0,t]$, but we do not need to require this.

Corollary 4. *(a) Consider two A/A/s/c models in which the service times are assigned when beginning service and the service times are independent of the arrival process. If balking is allowed in system 1 but not in system 2, if* $s_1 \leq s_2$, $c_1 \leq c_2$, $A^1 \subseteq_{st} A^2$ *and* $S^1 \geq_{st} S^2$, *then* $C^1 \geq_{st} C^2$, $Z^1 \geq_{st} Z^2$ *and* $D^1 \geq_{st} D^2$.

(b) Consider two A/GI/s/c models in which the service times are assigned to customers in any order (that does not depend on the service times themselves). If balking is allowed in system 1 but not in system 2, if $s_1 \leq s_2$, $c_1 \leq c_2$, $A^1 \subseteq_{st} A^2$ *and* $S_1^1 \geq_{st} S_1^2$, *then* $C^1 \geq_{st} C^2$, $Z^1 \geq_{st} Z^2$ *and* $D^1 \geq_{st} D^2$.

Proof. (a) From $A^1 \subseteq_{st} A^2$, we can construct the special arrival processes $\tilde{A}^i$ with $\tilde{A}^i$ equal in distribution to A^i for each i and $\tilde{A}^1 \subseteq \tilde{A}^2$. Similarly, since the service times are independent of the arrival process and $S^1 \geq_{st} S^2$, we can construct the special service times $\tilde{S}_k^i$ with $\{\tilde{S}_k^i : k \geq 1\}$ equal in distribution to

$\{S_k^i : k \geq 1\}$ for each i and $\tilde{S}_k^1 \geq \tilde{S}_k^2$ for all k. Moreover, we can let $\{\tilde{S}_k^i : k \geq 1\}$ be independent of $\{\tilde{A}^i(t) : t \geq 0\}$ for each i. Then Corollary 3 implies that $\tilde{C}_k^1 \geq \tilde{C}_k^2$, $\tilde{Z}_k^1 \geq \tilde{Z}_k^2$ and $\tilde{D}_k^1 \geq \tilde{D}_k^2$ for all k. Since $(\tilde{A}^i, \tilde{S}^i)$ has the same distribution as (A^i, S^i) for each i, $f(\tilde{A}^i, \tilde{S}^i)$ has the same distribution as $f(A^i, S^i)$ for every measurable function f, even with a very general range. Consequently, $\{(\tilde{C}_k^i, \tilde{Z}_k^i, \tilde{D}_k^i) : k \geq 1\}$ has the same distribution as $\{(C_k^i, Z_k^i, D_k^i) : k \geq 1\}$ for each i, and we obtain the desired conclusion. (b) Since the service times are i.i.d. and independent of the arrival process, we can construct the special sequences $\{\tilde{S}_k^i : k \geq 1\}$ such that $\{\tilde{S}_k^i : k \geq 1\}$ is distributed the same as $\{S_k^i : k \geq 1\}$ for each i, $\tilde{S}_k^i$ is assigned to the k^{th} job to start service for each i and $\tilde{S}_k^1 \geq \tilde{S}_k^2$ for all k. We use the i.i.d assumption to assign the corresponding service times in order of service initiation. The rest of the proof is as in part (a). ■

See Section 7 of Berger and Whitt [1] for an interesting application of Theorem 1 and Corollary 3.

REFERENCES

[1] A. W. Berger and W. Whitt, The impact of a job buffer in a token-bank rate-control throttle, *Stochastic Models,* this issue.

[2] K. C. Budka and D. D. Yao, Monotonicity and convexity properties of rate control throttles, Department of Industrial Engineering and Operations Research, Columbia University, 1990. (Abbreviated version in *Proceedings of 29th IEEE conference on Decision and Control,* (1990) 883-884.)

[3] K. C. Budka, Stochastic monotonicity and concavity properties of rate-based flow control mechanisms, *IEEE Trans. Aut. Control,* to appear.

[4] D. Sonderman, Comparing multi-server queues with finite waiting rooms, I: same number of servers, *Adv. Appl. Prob.*, **11** (1979) 439-447.

[5] D. Sonderman, Comparing multi-server queues with finite waiting rooms, II: different number of servers, *Adv. Appl. Prob.*, **11** (1979) 448-455.

[6] W. Whitt, Comparing counting processes and queues, *Adv. Appl. Prob.*, **13** (1981) 207-220.

[7] W. Whitt, Counterexamples for comparisons of queues with finite waiting rooms, *Queueing Systems*, **10** (1992) 271-278.

Received: 10/3/91
Revised: 1/30/1992
Accepted: 3/24/1992

Recommended by Brad Makrucki, Editor