The Annals of Applied Probability
1992, Vol. 2, No. 2, 460–480

PERFORMANCE BOUNDS FOR SCHEDULING QUEUEING NETWORKS

BY JIHONG OU AND LAWRENCE M. WEIN

Operations Research Center, M.I.T., and Sloan School of Management, M.I.T.

The goal of this paper is to assess the improvement in performance that might be achieved by optimally scheduling a multiclass open queueing network. A stochastic process is defined whose steady-state mean value is less than or equal to the mean number of customers in a queueing network under any arbitrary scheduling policy. Thus, this process offers a lower bound on performance when the objective of the queueing network scheduling problem is to minimize the mean number of customers in the network. Since this bound is easily obtained from a computer simulation model of a queueing network, its main use is to aid job-shop schedulers in determining how much further improvement (relative to their proposed policies) might be achievable from scheduling. Through computational examples, we identify some factors that affect the tightness of the bound.

When viewed from a dynamic and stochastic standpoint, the job-shop scheduling problem is often modeled as a scheduling problem for a multiclass network of queues. Despite the recent development of effective heuristics for scheduling queueing networks in heavy traffic [see, e.g., Harrison (1988), Harrison and Wein (1990), Laws and Louth (1990) and Wein (1990a)], the exact problem remains mathematically intractable and the primary mode of analysis by scheduling researchers [see, for example, Panwalkar and Iskander (1977)] and practitioners is computer simulation. In these studies, a detailed computer simulation model of the queueing network (or job-shop) is developed, different job-shop scheduling heuristics are tested and the resulting performance measures are usually compared to a straw policy (such as the first-come first-served rule) in order to identify effective scheduling policies. One problem with this approach is that the scheduling analyst is unable to determine the proximity to optimality of the proposed scheduling policies.

In this paper, we derive a bound on the achievable performance of an optimal scheduling policy in a general open queueing network. In particular, a stochastic process is defined whose steady-state mean is less than or equal to the steady-state mean number of customers in the network under any possible scheduling policy. Moreover, this stochastic process is easily obtained from a computer simulation model of the queueing network, and thus offers a lower bound on performance when the objective of the job-shop scheduling problem is to minimize the mean work-in-process inventory on the shop floor (or the

Received October 1990; revised April 1991.

AMS 1980 *subject classifications*. 60K25, 90B15, 90B22.

Key words and phrases. Networks of queues, sample path analysis.

mean sojourn time, by Little's formula). This bound is useful in helping job-shop scheduling analysts determine the effectiveness of their policies.

The queueing network under study consists of a finite number of single-server stations and is populated by a variety of different types of customers, where each type has its own arrival stream and its own arbitrary deterministic route through the stations. Three bounds are derived in this paper, each under a different set of distributional assumptions. First, we assume that each stage of each type's route has a different exponential service time distribution and allow the arrival processes to be arbitrary. A *pathwise* lower bound is derived in Section 1 for this network; this bound is a stochastic process that is less than or equal to the number of customers in the network under any scheduling policy for all times t with probability 1. This simple bound ignores the interference at a station due to customers of different types, as well as the interference due to customers at different stages of a given type's route and is primarily used as a basis for comparison.

In Section 2 we maintain arbitrary arrival processes but assume that the processing times for all operations performed at a given station are independent and identically distributed (iid) exponential random variables. A pathwise bound is derived in a two-step procedure; first, we use linear programming to derive a lower bound on the total number of customers in the system at time t in terms of a vector whose ith component is the number of customers present in the network at time t that require at least one more service from station i before exiting. Then, a pathwise lower bound on this vector process is derived by constructing a pathwise upper bound for the cumulative departure process of exiting customers at each service station under an arbitrary scheduling policy.

In Section 3 we allow each stage of each type's route to have a different exponential service time distribution. However, the arrival streams for the various customer types are now restricted to be independent Poisson processes. For this network, we are only able to obtain a lower bound on steady-state, rather than pathwise, performance; that is, we define a stochastic process whose steady-state mean value is less than or equal to the steady-state mean number of customers in the network under any scheduling policy. A similar two-step procedure is used here, but steady-state mean value bounds, not pathwise bounds, are derived in each step.

All the bounds derived in this paper are valid for any nonpreemptive scheduling policy that is *nonanticipating* with respect to the service times of the various operations; that is, although the service time distribution of each operation is known by the scheduler, the actual service times do not become known until they are realized. The scheduler is also allowed to observe the vector queue length process at each point in time, and to observe each customer's deterministic route at the moment of their arrival.

In Section 4, we perform a simulation experiment on three two-station networks and a three-station network under a variety of load conditions. Three stochastic processes are simulated for each example: the total number of customers in the network under the first-come first-served (FCFS) policy, the total number of customers in the network under a proposed scheduling policy (which is derived by various analytic and ad hoc methods), and the stochastic process (which leads to the bound) derived in Section 2 or 3 (depending on the particular network).

The numerical results are moderately encouraging, with the time average value of the bound averaging (over the 32 simulated scenarios) 78.0% of the mean number of customers in the network under the proposed policy. Since the pathwise bound derived in Section 2 is more effective than the steady-state bound derived in Section 3, the bounds tend to be more effective for networks in which service rates depend on the station, rather than the customer class. Also, the bounds tend to become less effective as the amount of feedback in the routes increases. For all four examples, the bounds were tightest when the load on the network was very heavy and imbalanced across the stations. However, for three of the four examples, the bounds performed worst when the load on the network was heavy and balanced across the stations. For these same examples, the proposed policies offered a significant improvement in performance over FCFS when the load was heavy and imbalanced, and the lower bounds showed that most of the possible improvement from scheduling (relative to FCFS) had been obtained by these proposed policies. Although we did not test the bound on any network with a large number of stations, we suspect that the efficiency of the bound will deteriorate as the number of stations increases. We hope the slackness in these bounds will motivate others to further study this problem area.

Although some of the ideas employed here have been used by Harrison and Wein (1989) and Laws and Louth (1990) to derive pathwise bounds for particular scheduling problems, this paper appears to contain the first attempt to offer a systematic procedure to develop performance bounds for general multiclass queueing networks operating under arbitrary scheduling policies. Readers are also referred to Weiss (1990), who derives worst-case bounds for Smith's rule (that is, the weighted shortest expected processing time rule) for parallel machines serving a fixed set of jobs with stochastic processing times.

1. A naive pathwise bound for each customer type.

The networks considered in this paper have I single-server stations and are visited by a variety of different customer types, each with their own arbitrary deterministic *route* (that is, sequence of stations to be visited) through the system. As in Kelly (1979) and Harrison (1988), we define a different *class* of customer for each stage of each customer type's route. In this section, we describe an obvious pathwise bound that can be obtained by assuming that customer classes do not compete with each other for the network's service resources and by analyzing each customer type in isolation. In particular, if a given customer type visits a sequence of stations, for example, $1 \to 2 \to 3 \to 2 \to 4$, then we create a modified tandem queueing system for which there is a separate station for each visit; in the example, the tandem system would include five stations, with station 2 repeated. A different tandem system is created for each customer type and thus the modified systems ignore the interference at a station due to customers of different types, as well as the interference due to different classes arising from the same type. Clearly, the times at which customers of a given class arrive at their station in the original system is pathwise lower bounded by the arrival times in the modified systems of tandem queues; to avoid unnecessary notation in this section, the proof of this observation is deferred until Proposition 2 in the next section.

In order to write down this bound precisely, we need to specify the stochastic processes underlying the queueing network. Customers of class $k = 1, \dots, K$ arrive according to independent arbitrary arrival processes $\{N_k(t),\ t \geq 0\}$, where N_k is assumed to be nondecreasing, RCLL (that is, its sample paths are right continuous and have left limits with probability 1) and satisfy $N_k(0) = 0$; thus, $N_k(t) = 0$ for all $t \geq 0$ for any class that does not correspond to the first stage along some customer type's route. For $k = 1, \dots, K$, let $\{S_k(t),\ t \geq 0\}$ be a Poisson process with parameter μ_k, which is the service rate for customers of class k and suppose $S_k(0) = 0$. We interpret $S_k(t)$ as the number of service completions up to time t if a server was continuously serving class k customers during $[0, t]$.

Suppose a certain customer type has n stages on its route and let class k correspond to the kth stage of this route, for $k = 1, \dots, n$. Then consider an n-station FCFS tandem queueing system with arrival process $\{N_1(t),\ t \geq 0\}$ to the first station and let station $k = 1, \dots, n$ have a service time distribution characterized by the potential service process $\{S_k(t),\ t \geq 0\}$. As in Harrison and Wein (1989), we assume the tandem queueing system is run according to the following modified service mechanism that was introduced by Borovkov (1965). The potential service processes S_k, $k = 1, \dots, n$, are always turned on, and whenever a potential service completion occurs in S_k, then a customer is allowed to depart station k if at least one customer is present at this station. If a customer arrives at station k at time t to an idle server, then its service time is the residual portion of the potential service time that is in progress at time t; thus, the service time is still exponential with parameter μ_k. The pathwise performance bound is relative to the probability space introduced by the processes $\{N_k(t),\ t \geq 0\}$ and $\{S_k(t),\ t \geq 0\}$ and Borovkov's service mechanism.

Let $\{A_k^*(t),\ t \geq 0\}$ denote the arrival process of class k customers to station k of the tandem queueing system, for $k = 1, \dots, n$. These processes can be defined sequentially starting with $k = 1$. In particular, for $t \geq 0$, we have

$$A_1^*(t) = N_1(t) \tag{1}$$

and

$$A_k^*(t) = S_{k-1}(t) + \inf_{0 \leq s \leq t} \{A_{k-1}^*(s) - S_{k-1}(s)\} \quad \text{for } k = 2, \dots, n. \tag{2}$$

Notice that A_k^* in (2) is the cumulative departure process from station $k - 1$, which is expressed as the *potential* number of departures from station $k - 1$ minus the *lost* number of departures due to an empty queue; readers are referred to Chapter 2 of Harrison (1985) for a full development of this approach. It follows that the number of customers in the tandem queueing system at time t is

$$(3) \qquad N_1(t) - S_n(t) - \inf_{0 \le s \le t} \{A_n^*(s) - S_n(s)\} \quad \text{for } t \ge 0,$$

which is a lower bound on the total number of customers of this type in the actual queueing network at time t for all $t \ge 0$. If we index the customer types in the network by $j = 1, \dots, J$, and let $Z_j(t)$ be the number of customers in the jth tandem system at time t [as calculated from (3)], then $\{\Sigma_{j=1}^J Z_j(t), t \ge 0\}$ is a pathwise lower bound on the total number of customers in the original queueing network under any scheduling policy.

The main advantage of this bound over the bounds that will be derived in the next two sections is that each customer class contributes to the bound at each point in time, since we are summing over the number of customers of each class in a set of tandem queueing systems. However, the bound derived in this section ignores all of the queueing effects between the various classes at a station and hence this bound will not be useful unless the original network has low traffic intensity, or the majority of the offered load at each station is due to one customer class.

2. A pathwise bound. In this section, the arrival processes are allowed to be arbitrary, but the service times at each station are required to be iid exponential random variables. We also assume that Borovkov's service mechanism dictates the timing of customer departures from each station, although the particular exiting customer depends on the scheduling policy used at each station. Let $Q_k(t)$ be the number of class k customers in the network at time t, and let $Q = (Q_k)$ be the vector queue length process. The goal of this section is to derive a lower bound for $\Sigma_{k=1}^K Q_k(t)$ under any scheduling policy for all times t with probability 1.

Let $s(k)$ denote the particular station that serves class k customers for $k = 1, \dots, K$, and define the $I \times K$ matrix $M = (M_{ik})$, where $M_{ik} = 1$ if customers of class k require at least one more service from station i before exiting, and let $M_{ik} = 0$ otherwise; this definition implies that $M_{s(k)k} = 1$ for $k = 1, \dots, K$. Define the I-dimensional process $W = (W_i)$ by

$$(4) \qquad W(t) = MQ(t) \quad \text{for all } t \ge 0,$$

so that $W_i(t)$ is the number of customers present in the network at time t that require at least one more service from station i before exiting.

The derivation of the pathwise lower bound is a two-step procedure. First, a pathwise lower bound $W^*(t)$ is found for $W(t)$, meaning that

$$(5) \qquad W_i^*(t) \le W_i(t) \quad \text{for } i = 1, \dots, I \text{ and } t \ge 0$$

for all scheduling policies. (We will construct such a bound shortly.) Then, by (4) and (5), a lower bound on the number of customers in the network at time t under any scheduling policy can be obtained by solving the following linear program parametrically for all nonnegative values of $W^*(t)$:

$$\min_{Q(t)} \sum_{k=1}^{K} Q_k(t) \tag{6}$$

subject to

$$\sum_{k=1}^{K} M_{ik}Q_k(t) \geq W_i^*(t) \quad \text{for } i = 1, \dots, I, \tag{7}$$

$$Q_k(t) \geq 0 \quad \text{for } k = 1, \dots, K. \tag{8}$$

If we let $f(W_1^*(t), \dots, W_I^*(t))$ denote the optimal objective function value of this linear program, then for any scheduling policy,

$$f(W_1^*(t), \dots, W_I^*(t)) \leq \sum_{k=1}^{K} Q_k(t) \quad \text{for } t \geq 0. \tag{9}$$

In (9), the right side depends on the scheduling policy and the left side is independent of the scheduling policy and is a pathwise performance bound for the network. The function f in (9) is monotone nondecreasing and, as noted in the following proposition, has a very simple form in a special, but not uncommon, case.

PROPOSITION 1. *The inequality*

$$f(W_1^*(t), \dots, W_I^*(t)) \geq \max_{1 \leq i \leq I} W_i^*(t) \quad \textit{for } t \geq 0 \tag{10}$$

always holds, and is satisfied with equality if there is a customer type who visits every station in the network.

PROOF. Since M_{ik} takes on the value of 0 or 1 for all $i = 1, \dots, I$ and $k = 1, \dots, K$,

$$\sum_{k=1}^{K} Q_k(t) \geq \sum_{k=1}^{K} M_{ik}Q_k(t) \tag{11}$$

and thus for any feasible solution $(Q_1(t), \dots, Q_K(t))$ to (6)–(8), we have

$$\sum_{k=1}^{K} Q_k(t) \geq W_i^*(t) \quad \text{for } i = 1, \dots, I, \tag{12}$$

which implies (10). If some customer type visits every station in the network, then the class, say j, corresponding to the first stage of this type's route, satisfies $M_{ij} = 1$ for $i = 1, \dots, I$. Then $Q_j(t) = \max_{1 \leq i \leq I} W_i^*(t)$ and $Q_k(t) = 0$ for $k \neq j$ is a feasible solution to (6)–(8) that achieves the lower bound in (10). □

In summary, a pathwise performance bound (9) [and a weaker bound, (10)] have been derived in terms of a hypothetical vector process W^* that satisfies

(5). The remainder of this section is devoted to finding the pathwise lower bound W^*. The key to constructing W^* is to derive an upper bound on the cumulative departure process from each station (that is, customers visiting this station for the last time) under an arbitrary scheduling policy. In order to derive this bound, we find it useful to consider a modified network where each customer, upon arrival to the system, immediately splits into a number of different customers, one for each of the different stations that are visited by the original customer. Each customer in the modified network is served exclusively at one station. In particular, if a certain customer type in the original feedback network visits a certain station l times on its route, then the customer created for that station in the modified network will immediately (that is, without any delays) feedback $l - 1$ times after the first visit to that station; thus we are ignoring the time it would actually take to feedback to that station after visiting other stations along its route. Under this construction, each station in the modified network will behave as a multiclass queue with feedback.

If a customer arrives to the original queueing network at time t, then the corresponding customers (one for each station on the original customer's route) in the modified network will not necessarily arrive at their respective stations at time t; instead, we will delay the arrivals in the modified network in order to obtain a tighter bound. This delay is constructed as in Section 1; namely, the time at which a given customer first arrives at any particular station along its route can be given a pathwise lower bound by tracing the path of this customer along its route through the stations while pretending these stations form a tandem queueing system (repeating stations visited more than once) and ignoring customers of other types.

By appropriate scheduling and insertion of idle time, it is fairly obvious that one can obtain a sample path of the departure process for the feedback queue of the modified network that is identical to any achievable departure process in the corresponding queue of the original system. Therefore, if the scheduler of the modified system inserts no idle time and always serves a customer having the least number of feedbacks remaining at that station, then the cumulative departure process from each station in the modified system is pathwise larger than in the original system.

As in Section 1, we assume that customers of class $k = 1, \dots, K$ arrive according to independent arbitrary arrival processes $\{N_k(t), t \geq 0\}$. For station $i = 1, \dots, I$, let the potential service process $\{S_i(t), t \geq 0\}$ be a Poisson process with parameter μ_i, which is the service rate for station i. Although these two vector processes are the only primitive stochastic processes of the network, we will need to define several more processes to precisely define the bound W^*. Let $A_k(t)$ be the number of arrivals of class k customers to station $s(k)$ (which is the station that serves them in the original network) up to time t in the original network under any arbitrary network scheduling policy. Let $D_i(t)$ be the number of service completions by server i in $[0, t]$ that constitute the last visit by a customer to station i under any arbitrary scheduling policy. Since $N_k(t) = 0$, $t \geq 0$, for all classes that do not correspond to the first stage along some customer type's route, it follows that

$$W_i(t) = \sum_{k=1}^{K} M_{ik} N_k(t) - D_i(t). \tag{13}$$

As mentioned earlier, the key to obtaining a pathwise lower bound W_i^* of W_i is to find a pathwise upper bound of $\{D_i(t), t \geq 0\}$.

We will also define a K-dimensional vector $A^* = (A_k^*)$ of *delayed arrival processes* in a similar manner to definitions (1)–(2), where $A_k^*(t)$ is the number of class k arrivals in $[0, t]$ to station $s(k)$ in the modified network. However, we essentially ignore A_k^* if class k does not correspond to the first visit to a station by a customer type. Thus, let $I(k) = i$ if class k corresponds to the first visit to station i by the corresponding customer type, and let $I(k) = 0$ if class k is not the first visit to a station by some customer type. Then for classes $\{k\colon I(k) > 0\}$, the process A_k^* will be constructed so that $A_k^*(t)$ is greater than or equal to $A_k(t)$ for $t \geq 0$. For ease of notation, we assume without loss of generality that the classes are ordered so that consecutive stages of each customer type's route are also consecutively numbered classes. If class k corresponds to the first stage along some customer type's route, then

$$A_k^*(t) = N_k(t) \quad \text{for } t \geq 0, \tag{14}$$

and otherwise, let

$$A_k^*(t) = S_{s(k-1)}(t) + \inf_{0 \leq s \leq t} \{A_{k-1}^*(s) - S_{s(k-1)}(s)\} \quad \text{for } t \geq 0. \tag{15}$$

Notice that A_k^* is nondecreasing and RCLL for $k = 1, \dots, K$.

Proposition 2. *For all $t \geq 0$,*

$$A_k(t) \leq A_k^*(t) \text{ for all scheduling policies and all classes } \{k\colon I(k) > 0\}. \tag{16}$$

Proof. As explained in Section 1, if $\{k\colon I(k) > 0\}$, then $\{A_k^*(t), t \geq 0\}$ represents the departure process from a tandem queueing system (not to be confused with the original or modified queueing network) consisting of $n - 1$ single-server exponential stations, where customers arrive to the system according to the process $\{A_{s(k-n+1)}^*(t), t \geq 0\}$ (which equals $\{N_{s(k-n+1)}(t), t \geq 0\}$, since class $k - n + 1$ is the first stage along this customer type's route) and the service rate at station $i = 1, \dots, n - 1$ of the tandem system is $\mu_{s(k-n+i)}$. Thus, A_k^* represents the arrival process of class k customers to station $s(k)$ in the original network if they received preemptive priority at each previous stage of their route. Since each customer class in the original queueing system may be competing with other classes at their respective stations (and perhaps even with customers of its own type which have fed back to the same station), $A_k^*(t)$ is an upper bound on the number of class k arrivals in $[0, t]$ to station $s(k)$ in the original network under any scheduling policy, for all $t \geq 0$, and thus (16) holds. □

The arrival process to station i in the modified network is $\{\Sigma_{\{k:\, I(k)=i\}} A_k^*(t),\ t \geq 0\}$, which is a superposition of the delayed arrival processes for the various classes that visit this station for the first time. The potential service process for station i in the modified network is the same as in the original network, namely $\{S_i(t),\ t \geq 0\}$. Define $\{F_i(t), t \geq 0\}$, $i = 1, \dots, I$, to be the cumulative departure process of *exiting* customers (that is, customers visiting station i for the last time) from station i (which is a multiclass feedback queue) in the modified network under the shortest expected remaining processing time (SERPT) policy; this policy gives nonpreemptive priority to the customer class that requires the least expected remaining amount of work at station i before exiting. Since all service operations at station i of the modified network are iid, this policy awards priority to the class that has the least number of remaining feedbacks to station i on its route. Then define $W_i^*(t)$ for $i = 1, \dots, I$ and $t \geq 0$ by

$$W_i^*(t) = \sum_{k=1}^{K} M_{ik} N_k(t) - F_i(t), \tag{17}$$

which represents the number of customers arriving to the original queueing network in $[0, t]$ requiring at least one service at station i minus the number of customers departing (for the last time) station i of the modified queueing network in $[0, t]$ under the SERPT policy.

Proposition 3. *For all $t \geq 0$ and all scheduling policies, $W_i^*(t) \leq W_i(t)$, $i = 1, \dots, I$.*

Proof. We begin by proving the result for the special case where no feedback exists; that is, customers do not visit any station more than once on their route. In this case, station i of the modified network is a single-server queue with no feedback and

$$F_i(t) = S_i(t) + \inf_{0 \leq s \leq t} \left\{ \sum_{\{k:\, I(k)=i\}} A_k^*(s) - S_i(s) \right\} \quad \text{for } t \geq 0. \tag{18}$$

Although the departure process $D_i(t)$ in the original network depends on the scheduling policy employed, we have, for $i = 1, \dots, I$ and $t \geq 0$,

$$D_i(t) \leq S_i(t) + \inf_{0 \leq s \leq t} \left\{ \sum_{\{k:\, I(k)=i\}} A_k(s) - S_i(s) \right\} \tag{19}$$

$$\leq F_i(t) \quad \text{by (16) and (18)}. \tag{20}$$

Notice that the inequality in (19) is tight if the server at station i in the actual queueing network services customers whenever the queue is not empty. Our result follows by combining (13), (17) and (20).

Now let us consider the general feedback case. By (13) and (17), it suffices to show that $D_i(t) \leq F_i(t)$ for $t \geq 0$, $i = 1, \dots, I$, and all scheduling policies. Observe that if customers in the original feedback network, after visiting station i for the first time, skip subsequent stages of their route that are not at station i, then the same sequence of customer services at station i could be realized and hence the same departure process $\{D_i(t),\ t \geq 0\}$ could be observed, by possibly including inserted idle times. Moreover, if the actual arrival process of first time customers to station i $\{\Sigma_{\{k:\ I(k)=i\}} A_k(t),\ t \geq 0\}$, was replaced by our delayed arrival process $\{\Sigma_{\{k:\ I(k)=i\}} A_k^*(t),\ t \geq 0\}$, then, by Proposition 2, the same departure process of exiting customers could be realized, again by the possible insertion of idle times. Thus, any departure process of exiting customers that is feasible for station i of our original feedback queueing network is also feasible for the corresponding multiclass feedback queue in the modified network.

Therefore, a pathwise upper bound (for any scheduling policy) on the departure process of exiting customers for station i of the modified network will also be a pathwise upper bound on $D_i(t)$. For customer classes $\{k\colon I(k) = i\}$, let m_k denote the number of remaining visits to station i before exiting the network. The maximum number of service completions up to time t at station i of the modified network is

$$(21) \qquad S_i(t) + \inf_{0 \leq s \leq t} \left\{ \sum_{\{k:\ I(k)=i\}} m_k A_k^*(s) - S_i(s) \right\},$$

which is realized by any scheduling policy that always serves customers when the queue is not empty. Moreover, among this class of policies, the SERPT policy maximizes the departure process of exiting customers for all $t \geq 0$. Thus, $F_i(t) \geq D_i(t)$, for all scheduling policies and all times $t \geq 0$, which completes the proof. □

3. A steady-state bound. In this section, each customer class is allowed to have a different exponential service time distribution, but each customer type is constrained to have a Poisson arrival process; that is, N_k, $k = 1, \dots, K$, are now independent Poisson processes. We will use a similar procedure as in the last section (and will retain most of the notation), but will develop a steady-state, rather than pathwise, bound; thus, we will need to assume that the arrival rates, service rates and customer routes are such that the traffic intensity at each station in the network is less than 1. For $k = 1, \dots, K$, let q_k be defined by

$$(22) \qquad q_k = \lim_{T \to \infty} \frac{1}{T} E\left[\int_0^T Q_k(t)\, dt \right],$$

so that it represents the long run expected number of class k customers in the system under an arbitrary policy. Similarly, for $i = 1, \dots, I$, define

$$(23) \qquad w_i = \lim_{T \to \infty} \frac{1}{T} E\left[\int_0^T W_i(t)\, dt \right],$$

which is the long run expected number of customers in the network who require at least one more service at station i before exiting. Thus, it follows from (4) that

$$w_i = \sum_{k=1}^{K} M_{ik} q_k \quad \text{for } i = 1, \ldots, I \text{ and } k = 1, \ldots, K. \tag{24}$$

Suppose we can find an I-dimensional stochastic process W^* such that

$$w_i^* \triangleq \lim_{T \to \infty} \frac{1}{T} E\left[\int_0^T W_i^*(t)\, dt\right] \le w_i, \qquad i = 1, \ldots, I, \tag{25}$$

for all scheduling policies. Then a lower bound on the mean number of customers in the system in steady-state can be found by solving the following linear program parametrically in w^*:

$$\min_q \sum_{k=1}^{K} q_k \tag{26}$$

subject to

$$\sum_{k=1}^{K} M_{ik} q_k \ge w_i^* \quad \text{for } i = 1, \ldots, I, \tag{27}$$

$$q_k \ge 0 \quad \text{for } k = 1, \ldots, K. \tag{28}$$

Denoting the solution to the linear program by $f(w_1^*, \ldots, w_I^*)$, it follows that for any scheduling policy,

$$f(w_1^*, \ldots, w_I^*) \le \lim_{T \to \infty} \frac{1}{T} E\left[\int_0^T \sum_{k=1}^{K} Q_k(t)\, dt\right]. \tag{29}$$

By the convexity of f [for a proof of convexity, see, e.g., Proposition 4.1 in Wein (1990b)] and Jensen's inequality, it can be shown that

$$f(w_1^*, \ldots, w_I^*) \le \lim_{T \to \infty} \frac{1}{T} E\left[\int_0^T f(W_1^*(t), \ldots, W_I^*(t))\, dt\right] \tag{30}$$

and thus the steady-state bound is not as effective, in general, as the pathwise bound derived in Section 2. Our inability to find a pathwise bound W^* satisfying (5) for the network described in this section has led us to resort to the less effective steady-state bound. Also, observe that Proposition 1 still holds, with w_i^* in place of $W_i^*(t)$ for $i = 1, \ldots, I$.

In order to derive a stochastic process W^* satisfying (25), we again consider the modified queueing network described in Section 2; however, the network will be defined slightly differently, since each customer class can have its own exponential service time distribution. In particular, let S_k, $k = 1, \ldots, K$, be the Poisson process corresponding to the number of potential service completions in $[0, t]$ if class k customers were served continuously during that interval. The delayed arrival processes A_k^*, $k = 1, \ldots, K$, are defined exactly as in (14) and (15), except the service processes $S_{s(k)}$ in (15) are replaced here by S_k; that is, if class k corresponds to the first stage along some customer type's route, then let

$$A_k^*(t) = N_k(t) \quad \text{for } t \geq 0, \tag{31}$$

and otherwise, let

$$A_k^*(t) = S_{k-1}(t) + \inf_{0 \leq s \leq t} \{A_{k-1}^*(s) - S_{k-1}(s)\} \quad \text{for } t \geq 0. \tag{32}$$

The arrival process to station i of the modified network is $\sum_{\{k:\ I(k)=i\}} A_k^*$. Since $N = (N_k)$ are Poisson processes, it follows by the explanation of equation (15) in the proof of Proposition 2 and by Burke's theorem [Burke (1956)] that A_k^* is a Poisson process for all k. Since A_k^* are independent for all k such that $I(k) = i$, it follows that each station in the modified network behaves as a multiclass M/M/1 feedback queue. Furthermore, Proposition 2 holds true for this network, where, for all $\{k\colon I(k) > 0\}$, $\{A_k(t),\ t \geq 0\}$ is the arrival process of class k customers to station $s(k)$ in the original feedback network under any arbitrary scheduling policy. We also need to observe that the cumulative number of customers who have arrived at station i in the modified system is no more than the total number of customers who could have entered the original system and will eventually need service at station i. This observation clearly follows from the way the arrivals in the modified system were defined, and we state it without proof in the following proposition.

Proposition 4. *For all* $t \geq 0$,

$$\sum_{\{k:\ I(k)=i\}} A_k^*(t) \leq \sum_{k=1}^{K} M_{ik} N_k(t) \quad \textit{for } i = 1, \dots, I. \tag{33}$$

As in Section 2, we let $\{F_i(t),\ t \geq 0\}$, $i = 1, \dots, I$, be the cumulative departure process of exiting customers from station i (which is a multiclass M/M/1 feedback queue) in the modified network under the shortest expected remaining processing time (SERPT) policy, and define $W_i^*(t)$ for $i = 1, \dots, I$ and $t \geq 0$, by

$$W_i^*(t) = \sum_{k=1}^{K} M_{ik} N_k(t) - F_i(t). \tag{34}$$

Thus, for $i = 1, \dots, I$, the steady-state bound w_i^* is given by

$$w_i^* = \lim_{T \to \infty} \frac{1}{T} E\left[\int_0^T \left(\sum_{k=1}^{K} M_{ik} N_k(t) - F_i(t)\right) dt\right]. \tag{35}$$

Proposition 5. *For all scheduling policies,* $w_i^* \leq w_i$, $i = 1, \dots, I$.

Proof. Notice that (34) can also be expressed as

$$\begin{aligned} W_i^*(t) = {} & \left[\sum_{k=1}^{K} M_{ik} N_k(t) - \sum_{\{k:\ I(k)=i\}} A_k^*(t)\right] \\ & + \left[\sum_{\{k:\ I(k)=i\}} A_k^*(t) - F_i(t)\right], \end{aligned} \tag{36}$$

where the second bracketed term on the right side represents the number of customers in a multiclass M/M/1 feedback queue under the SERPT policy. For the original feedback queueing network, we again define $D_i(t)$ to be the number of service completions in $[0, t]$ that constitute the last visit by a customer to station i under any arbitrary scheduling policy. Then we have

$$W_i(t) = \sum_{k=1}^{K} M_{ik} N_k(t) - D_i(t) \tag{37}$$

$$= \sum_{k=1}^{K} M_{ik} N_k(t) - \sum_{\{k:\, I(k)=i\}} A_k^*(t) + \sum_{\{k:\, I(k)=i\}} A_k^*(t) - D_i(t). \tag{38}$$

If station i services m different customer types in the original network, then by (31)–(32),

$$\lim_{T\to\infty} \frac{1}{T} E\left[\int_0^T \sum_{k=1}^{K} M_{ik} N_k(t) - \sum_{\{k:\, I(k)=i\}} A_k^*(t)\, dt\right] \tag{39}$$

is the mean steady-state number of customers in a set of m different tandem queueing systems (readers are referred to the proof of Proposition 2 for the interpretation of A^*); this quantity is finite, since the traffic intensity at each station in the original queueing network is less than 1. Thus, by (23), (35)–(36) and (38), it suffices to show that, for all scheduling policies,

$$\begin{aligned} &\lim_{T\to\infty} \frac{1}{T} E\left[\int_0^T \sum_{\{k:\, I(k)=i\}} A_k^*(t) - F_i(t)\, dt\right] \\ &\quad \le \lim_{T\to\infty} \frac{1}{T} E\left[\int_0^T \sum_{\{k:\, I(k)=i\}} A_k^*(t) - D_i(t)\, dt\right], \end{aligned} \tag{40}$$

where the right side is dependent on the scheduling policy used in the original queueing network.

By the argument in the second to last paragraph in the proof of Proposition 3, any scheduling policy (and hence any corresponding departure process) that is feasible for station i of our original feedback queueing network is also feasible for the corresponding multiclass M/M/1 feedback queue in the modified network. Thus, inequality (40) follows by the fact that the SERPT policy minimizes the long run expected average number of customers in a multiclass M/M/1 feedback queue [see Klimov (1974) for a derivation of this classic result]. □

By (33) and (35)–(36), it follows that

$$w_i^* \ge \lim_{T\to\infty} \frac{1}{T} E\left[\int_0^T \sum_{\{k:\, I(k)=i\}} A_k^*(t) - F_i(t)\, dt\right], \tag{41}$$

which is the mean steady-state number of customers in a multiclass M/M/1 feedback queue (under the SERPT policy) that has the same traffic intensity as station i in the original queueing network. Thus, if the traffic intensity $\rho_i \geq 1$ for some station i in the original queueing network, then w_i^* will be infinite, as will our steady-state lower bound. Therefore, scheduling is unable to prevent an open queueing network from instability when $\max_{\{1 \leq i \leq I\}} \rho_i \geq 1$.

4. Examples. In this section, we test the bounds derived earlier on three two-station networks and one three-station network; the routing complexity for these examples ranges from a tandem network to a network with symmetric routing. For each network, we consider eight different scenarios, which consist of two levels of load balance (referred to as *balanced* and *imbalanced* in the tables and discussion to follow) crossed with four levels of load intensity (*light*, *medium*, *heavy* and *very heavy*). Let ρ_i be the traffic intensity at station i, which is the fraction of the time over the long run that server i is busy. For the balanced networks, the traffic intensity is the same at each station, and is 0.3, 0.6, 0.9 and 0.99 for the four respective load intensities. For the two-station imbalanced networks, the vector ρ of traffic intensities is $(0.3, 0.2)$, $(0.6, 0.4)$, $(0.9, 0.6)$ and $(0.99, 0.66)$ for the four load intensities and for the three-station imbalanced networks, the respective vectors are $(0.3, 0.2, 0.1)$, $(0.6, 0.4, 0.2)$, $(0.9, 0.6, 0.3)$ and $(0.99, 0.66, 0.33)$.

For each scenario of each network, we simulate and record the time average values of three stochastic processes: (i) the number of customers in the network under the FCFS policy, (ii) the number of customers in the network under a proposed policy (which is derived either from previous analysis or on a trial-and-error basis) and (iii) the lower bound (from either Section 2 or Section 3, depending on the particular network). The three stochastic processes are driven by the same customer arrival and service processes for a given scenario. The pathwise bound from Section 1 was also tested for each scenario, but it performed poorly, as expected. However, this bound is tighter than the Section 3 bound in the imbalanced light scenario for Example 3, although the two bounds are nearly identical in this case.

For each scenario, 20 independent runs were made, each consisting of 11,000 time units in Examples 1 and 2 and 91,000 time units for Examples 3 and 4. The observations in the first 1000 time units of each run were discarded to reduce the initialization effect. In Table 1 we provide the mean (and 95% confidence interval) over the 20 runs of the time average value of the three stochastic processes, where these means are abbreviated by FCFS, PROPOSED and BOUND, respectively.

Ideally, the effectiveness of our bounds should be measured by their proximity to the number of customers in the network under an optimal scheduling policy. Unfortunately, this is impossible to assess, since the optimal scheduling policy for each of these problems is unknown. Instead, Table 2 records the *efficiency* of the lower bound, which we define as

$$\text{lower bound efficiency} = \frac{\text{BOUND}}{\text{PROPOSED}} \times 100\%. \tag{42}$$

TABLE 1
Detailed results of simulation experiments

Example	Scenario	FCFS	Proposed	Bound
1	Balanced light	0.856(±0.010)	0.856(±0.008)	0.775(±0.010)
	Balanced medium	2.99(±0.058)	2.92(±0.051)	2.47(±0.046)
	Balanced heavy	17.9(±0.912)	15.5(±0.872)	13.2(±0.824)
	Balanced very heavy	114.(±24.8)	91.5(±23.3)	85.9(±23.0)
	Imbalanced light	0.677(±0.007)	0.677(±0.007)	0.621(±0.007)
	Imbalanced medium	2.16(±0.058)	2.14(±0.034)	1.85(±0.031)
	Imbalanced heavy	10.7(±0.565)	10.4(±0.548)	9.46(±0.544)
	Imbalanced very heavy	74.6(±24.6)	74.1(±24.8)	72.8(±24.6)
2	Balanced light	0.864(±0.014)	0.838(±0.015)	0.698(±0.010)
	Balanced medium	3.02(±0.058)	2.78(±0.053)	2.07(±0.030)
	Balanced heavy	18.5(±1.67)	13.6(±0.664)	7.99(±0.422)
	Balanced very heavy	97.2(±15.5)	56.2(±6.57)	34.7(±5.85)
	Imbalanced light	0.673(±0.011)	0.665(±0.010)	0.569(±0.008)
	Imbalanced medium	2.16(±0.042)	2.02(±0.036)	1.63(±0.032)
	Imbalanced heavy	10.4(±0.734)	7.53(±0.394)	6.03(±0.352)
	Imbalanced very heavy	56.0(±11.3)	28.8(±4.91)	26.2(±4.88)
3	Balanced light	0.951(±0.015)	0.900(±0.013)	0.734(±0.011)
	Balanced medium	3.76(±0.093)	3.06(±0.068)	2.18(±0.045)
	Balanced heavy	23.9(±1.98)	15.1(±1.14)	8.48(±0.611)
	Balanced very heavy	141.(±34.8)	102.(±25.5)	43.8(±12.9)
	Imbalanced light	0.742(±0.011)	0.709(±0.010)	0.560*(±0.006)
	Imbalanced medium	2.64(±0.060)	2.22(±0.044)	1.47(±0.031)
	Imbalanced heavy	13.2(±0.906)	8.67(±0.475)	6.94(±0.451)
	Imbalanced very heavy	81.9(±22.0)	45.9(±11.9)	43.5(±11.9)
4	Balanced light	1.37(±0.023)	1.35(±0.017)	1.17(±0.013)
	Balanced medium	5.15(±0.095)	4.68(±0.075)	3.32(±0.055)
	Balanced heavy	32.7(±2.89)	24.2(±1.79)	12.2(±0.702)
	Balanced very heavy	207.(±34.2)	151.(±27.1)	71.4(±22.0)
	Imbalanced light	0.860(±0.012)	0.840(±0.012)	0.717(±0.008)
	Imbalanced medium	2.79(±0.061)	2.51(±0.046)	1.78(±0.023)
	Imbalanced heavy	14.1(±1.40)	10.3(±0.903)	8.33(±0.890)
	Imbalanced very heavy	80.9(±23.1)	53.3(±14.8)	50.9(±14.8)

*Section 1 bound.

Since the main use of these bounds is to aid a job-shop scheduler in determining how much further improvement (relative to their proposed policy) might be achievable from scheduling, the efficiency seems like an appropriate measure for consideration. However, the gap between the proposed policy and the lower bound equals the gap between the proposed policy and an optimal policy plus the gap between an optimal policy and the lower bound and it is difficult to assess how much of the total gap is due to either portion; that is, some of our recorded gap may be due to our inability to specify a scheduling policy that is close to optimal.

Since scheduling analysts often compare their proposed policy to a straw policy, such as FCFS, Table 2 also includes the *effectiveness* of the proposed

TABLE 2
Summary results of simulation experiments

Example	**Scenario**	**Lower bound efficiency (%)**	**Proposed policy effectiveness (%)**
1	Balanced light	90.5	0.0
	Balanced medium	84.6	2.3
	Balanced heavy	85.2	13.4
	Balanced very heavy	93.9	19.7
	Imbalanced light	91.7	0.0
	Imbalanced medium	86.4	0.1
	Imbalanced heavy	91.0	2.8
	Imbalanced very heavy	98.2	0.1
2	Balanced light	83.3	2.8
	Balanced medium	74.5	7.9
	Balanced heavy	58.8	26.5
	Balanced very heavy	61.7	42.2
	Imbalanced light	85.6	2.7
	Imbalanced medium	80.7	3.7
	Imbalanced heavy	80.1	27.6
	Imbalanced very heavy	91.0	48.6
3	Balanced light	81.6	5.4
	Balanced medium	71.2	18.6
	Balanced heavy	56.2	36.8
	Balanced very heavy	42.9	27.7
	Imbalanced light	79.0*	4.4*
	Imbalanced medium	66.2	15.9
	Imbalanced heavy	80.0	34.3
	Imbalanced very heavy	94.8	44.0
4	Balanced light	86.7	1.5
	Balanced medium	70.9	9.1
	Balanced heavy	50.4	26.0
	Balanced very heavy	47.3	27.1
	Imbalanced light	85.4	2.4
	Imbalanced medium	70.9	10.0
	Imbalanced heavy	80.9	27.0
	Imbalanced very heavy	95.5	34.1

*Section 1 bound.

policy, which is defined as

$$(43) \quad \text{proposed policy effectiveness} = \frac{\text{FCFS} - \text{PROPOSED}}{\text{FCFS}} \times 100\%.$$

This quantity tells us the percentage reduction in the mean number of customers in the network achieved by the proposed policy relative to the FCFS policy.

We will describe the four example networks and their respective proposed policies before presenting the simulation results. The detailed network parameters for all 32 scenarios are listed in Table 3.

TABLE 3
Data for simulation experiments

	Load balance		**Load intensity**			
Example	**Balanced**	**Imbalanced**	**Light**	**Medium**	**Heavy**	**Very heavy**
1	$\mu_1 = 2.0$ $\mu_2 = 1.0$	$\mu_1 = 2.0$ $\mu_2 = 1.5$	$\lambda_A = 0.3$ $\lambda_B = 0.3$	$\lambda_A = 0.6$ $\lambda_B = 0.6$	$\lambda_A = 0.9$ $\lambda_B = 0.9$	$\lambda_A = 0.99$ $\lambda_B = 0.99$
2	$\mu_1 = 1.0$ $\mu_2 = 1.0$	$\mu_1 = 1.0$ $\mu_2 = 1.5$	$\lambda = 32/105$	$\lambda = 64/105$	$\lambda = 96/105$	$\lambda = 105.6/105$
3	$\mu_{A1} = 1/4$ $\mu_{A2} = 1.0$ $\mu_{B1} = 1/8$ $\mu_{B2} = 1/6$ $\mu_{B3} = 1/2$ $\mu_{B4} = 1/7$	$\mu_{A1} = 1/4$ $\mu_{A2} = 2/3$ $\mu_{B1} = 1/8$ $\mu_{B2} = 1/4$ $\mu_{B3} = 1/2$ $\mu_{B4} = 3/14$	$\lambda_A = 3/140$ $\lambda_B = 3/140$	$\lambda_A = 6/140$ $\lambda_B = 6/140$	$\lambda_A = 9/140$ $\lambda_B = 9/140$	$\lambda_A = 99/140$ $\lambda_B = 99/140$
4	$\mu_{A1} = 1/2$ $\mu_{A2} = 1/4$ $\mu_{A3} = 1/6$ $\mu_{B1} = 1/7$ $\mu_{B2} = 1/5$ $\mu_{B3} = 1/3$	$\mu_{A1} = 1/2$ $\mu_{A2} = 1/2$ $\mu_{A3} = 1.0$ $\mu_{B1} = 1/7$ $\mu_{B2} = 1/4$ $\mu_{B3} = 1/2$	$\lambda_A = 1/30$ $\lambda_B = 1/30$	$\lambda_A = 2/30$ $\lambda_B = 2/30$	$\lambda_A = 0.1$ $\lambda_B = 0.1$	$\lambda_A = 0.11$ $\lambda_A = 0.11$

EXAMPLE 1. This simple network appears in Figure 1, where type A customers visit station 1 and then exit and type B customers visit station 1, proceed to station 2 and then exit. Type A and type B customers arrive at station 1 according to independent Poisson processes with rates λ_A and λ_B, respectively. The exponential service rates μ_1 and μ_2 are associated with the two servers, not the three classes and thus the pathwise bound derived in Section 2 is valid for this network.

The only real scheduling decision in this problem is to dynamically decide which customer type to serve at station 1. Harrison and Wein (1989) studied this scheduling problem by analyzing an approximating Brownian control problem under balanced heavy loading conditions and proposed the following

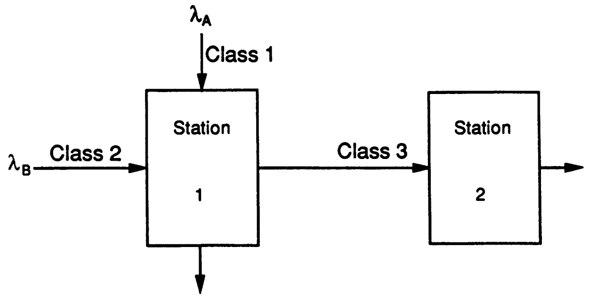

FIG. 1. *The network for example* 1.

scheduling policy, which is our proposed policy for this example: higher priority is awarded to type A customers at station 1, unless there are c or fewer customers in queue and in service at station 2. In the latter case, priority is given to type B customers in order to avoid idleness at station 2. The most effective value of the parameter c was chosen via computer simulation.

EXAMPLE 2. This example, which appears in Figure 2, is a simplified two-station version of the nine-station symmetric job-shop studied in Chapter 11 of Conway, Maxwell and Miller (1967). Customers arrive according to an independent Poisson process at rate λ to each station. When customers complete service at a station, they visit the other station with probability $\frac{1}{2}$ and exit the network with probability $\frac{1}{2}$, independent of all previous history. As in Conway, Maxwell and Miller (1967), a customer's entire route is chosen at the time of its arrival to the network and is made known to the scheduler. For ease in developing the simulation model, we did not allow a customer to have more than six operations on its route; hence there are 12 possible routes through the network. Since we assume that the exponential service rates are the same for each service operation performed at a given station, only 12 customer classes are required and the pathwise lower bound derived in Section 2 is employed.

Our proposed policy for this example is a dynamic policy developed by Wein and Ou (1991) using a Brownian approximation procedure. For $i = 1, 2$ and $k = 1, \dots, 12$, let A_{ik} be the expected remaining processing time for a class k customer at station i before that customer exits the network and define $\{V_i(t), t \geq 0\}$ by

$$V_i(t) = \sum_{k=1}^{12} A_{ik} Q_k(t) \quad \text{for } i = 1, 2, \tag{44}$$

where Q is the vector queue length process. Thus, $V_i(t)$ represents the total amount of work remaining in the network for station i at time t. When $V_1(t) > V_2(t)$, the proposed policy awards priority to classes with smaller values of A_{1k}, and if there is a tie among classes, then priority is given to larger values of A_{2k} at station 1 and smaller values of A_{2k} at station 2. Similarly, when $V_1(t) < V_2(t)$, priority is given to classes with smaller values of A_{2k} and

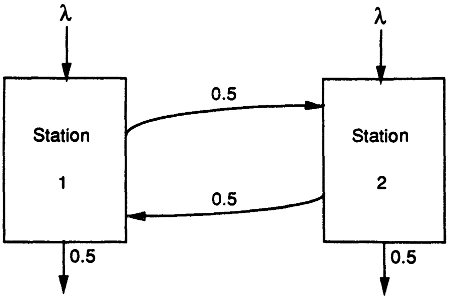

FIG. 2. *The network for example* 2.

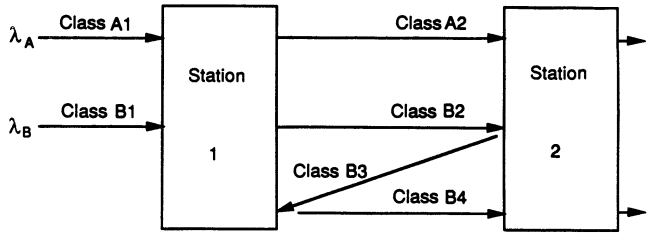

FIG. 3. *The network for example* 3.

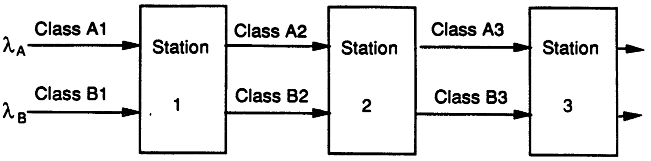

FIG. 4. *The network for example* 4.

when ties exist, priority is awarded to smaller values of A_{1k} at station 1 and larger values of A_{1k} at station 2.

EXAMPLE 3. This two-station example, which appears in Figure 3, not only allows customer feedback, but also allows each customer class to have its own exponential service rate; thus, the steady-state bound derived in Section 3 is required. There are two customer types, A and B, with two and four stages on their respective routes, and the six customer classes will be referred to by their type-stage pair. Although effective scheduling policies have been developed under balanced heavy loading conditions for two-station closed [that is, constant population size; see Harrison and Wein (1990)] networks and two-station networks with controllable inputs [see Wein (1990a)], the general two-station open network problem has not been successfully analyzed. We tested several static and dynamic scheduling policies by computer simulation and found that the simple shortest expected remaining processing time (SERPT) rule, which gives priority to customers who are closest to exiting the network, was most effective. Thus, our proposed policy is the SERPT policy.

EXAMPLE 4. Our last example is the three-station tandem queueing system pictured in Figure 4. The steady-state bound derived in Section 3 is required for this example, since each customer class has a different service rate. After testing several static and dynamic policies in trial simulation runs, we have used the shortest expected processing time policy, which gives priority to the class whose upcoming operation has the shortest expected processing time, as the proposed policy.

The simulation results for the four examples are summarized in Table 2, which displays the efficiency of the lower bound and the effectiveness of the proposed policy, which are defined in equations (42) and (43), respectively. The average efficiency of the lower bounds over the 32 scenarios is 78.0%. The bounds are most efficient for Example 1, where the efficiency averages 90.2% over the eight scenarios, and the average efficiency for Examples 2, 3 and 4 is 77.0%, 71.5% and 73.5%, respectively.

The large amount of feedback present in Example 2 is probably the main reason why the pathwise bound is less effective in Example 2 than Example 1. However, it is possible that the proposed policy is closer to optimality in Example 1 than in Example 2, which would also contribute to the discrepancy. Similarly, the simple structure of the three-station network in Example 4 probably allows its steady-state bound to be more efficient than the steady-state bound of the two-station feedback network in Example 3. The lower efficiencies in Examples 3 and 4 relative to Examples 1 and 2 may be partially due to the fact that, as explained in (30), the steady-state bound derived in Section 3 is not as efficient as the pathwise bound derived in Section 2.

The bound efficiencies in all four examples exhibit similar dependencies with respect to the magnitude and balance of the network's load. In particular, the bounds are generally more effective for imbalanced networks: The average efficiency for the 16 balanced and imbalanced scenarios is 71.2% and 84.8%, respectively. For the balanced networks, the bound efficiencies in Examples 2, 3 and 4 deteriorate as the load becomes heavier, although the efficiency increases slightly at very heavy loads for Example 2. For the balanced network in Example 1 and all four imbalanced networks, the bounds were least efficient under the medium load and most efficient under very heavy loading.

Most of this behavior is not difficult to explain. When the load on the system is light, there is little congestion in the network and one would not expect a large difference between our bounds and the proposed policy (or the FCFS policy). Moreover, since $f(x_1, x_2) = x_1 \vee x_2$ for our four examples by Proposition 1, it is clear why the bounds are most effective when the load on the network is very heavy and imbalanced; in this case, most of the congestion occurs at one station in the network and this congestion is captured by the function f. However, a smaller portion of the total congestion is at one station when the network becomes more balanced or the network becomes more lightly loaded; thus, the bounds become less effective in these cases. The function f also implies that our bounds will deteriorate as the number of stations in the network increases; in particular, it would appear that the bound would perform poorly in a well-balanced network with many stations. However, the bound may still be useful in a network with many stations if the network is heavily loaded and possesses a decisive bottleneck station.

Example 1 is the only network that possesses an efficient bound under balanced heavy loading conditions. When the pathwise bound derived in Section 2 is applied to the version of this network considered in Harrison and Wein (1989), it reduces to the bound denoted by $w_1^{\mathrm{LRPT}}(t) \vee w_2^{\mathrm{LRPT}}(t)$ in Proposition 2 of that paper. Harrison and Wein show that a pathwise bound that is smaller pathwise than $w_1^{\mathrm{LRPT}}(t) \vee w_2^{\mathrm{LRPT}}(t)$ weakly converges (under the standard heavy traffic scaling) to the optimal objective function value of a Brownian control problem that approximates this scheduling problem under heavy traffic conditions. Thus, it is not surprising that our bound performs well when the load on this network is very high.

As expected, the effectiveness of the proposed policy generally increases with the system's load for the various networks, although the effectiveness dips at very heavy loads for the imbalanced network of Station 1 and the balanced network of Station 3. The proposed policy achieves the smallest improvement relative to FCFS in Example 1. However, we can infer from the relatively tight bounds that this lack of effectiveness is not due to our inability to find an effective policy, but is intrinsic to the scheduling problem. When Examples 2, 3 and 4 are subject to an imbalanced and very heavy load, the proposed scheduling policies achieve substantial improvements in performance relative to FCFS and the lower bounds imply that only slight improvements beyond this are possible.

Acknowledgments. We are grateful to Michael Harrison for suggesting this research topic to us and to two referees for their suggestions that have improved the paper's readability. This research is partially supported by a research grant from the Leaders for Manufacturing Program at MIT.

REFERENCES

BOROVKOV, A. A. (1965). Some limit theorems in the theory of mass service, II. *Theory Probab. Appl.* **10** 375–400.

BURKE, P. J. (1956). The output of queueing systems. *Oper. Res.* **6** 699–704.

CONWAY, R. W., MAXWELL, W. L. and MILLER, L. W. (1967). *Theory of Scheduling*. Addison-Wesley, Reading, Mass.

HARRISON, J. M. (1985). *Brownian Motion and Stochastic Flow Systems*. Wiley, New York.

HARRISON, J. M. (1988). Brownian models of queueing networks with heterogeneous customer populations. In *Stochastic Differential Systems, Stochastic Control Theory and Applications* (W. Fleming and P. L. Lions, eds.) **10** 147–186. Springer, New York.

HARRISON, J. M. and WEIN, L. M. (1989). Scheduling networks of queues: Heavy traffic analysis of a simple open network. *Queueing Systems Theory Appl.* **5** 265–280.

HARRISON, J. M. and WEIN, L. M. (1990). Scheduling networks of queues: Heavy traffic analysis of a two-station closed network. *Oper. Res.* **38** 1052–1064.

KELLY, F. P. (1979). *Reversibility and Stochastic Networks*. Wiley, New York.

KLIMOV, G. P. (1974). Time sharing service systems I. *Theory Probab. Appl.* **19** 532–551.

LAWS, C. N. and LOUTH, G. M. (1990). Dynamic scheduling of a four station network. *Probability in the Engineering and Information Sciences* **4** 131–156.

PANWALKAR, S. S. and ISKANDER, W. (1977). A survey of scheduling rules. *Oper. Res.* **25** 45–61.

WEIN, L. M. (1990a). Scheduling networks of queues: Heavy traffic analysis of a two-station network with controllable inputs. *Oper. Res.* **38** 1065–1078.

WEIN, L. M. (1990b). Scheduling networks of queues: Heavy traffic analysis of a multistation network with controllable inputs. *Oper. Res.* To appear.

WEIN, L. M. and OU, J. (1991). The impact of processing time knowledge on dynamic job-shop scheduling. *Management Sci.* **37** 1002–1014.

WEISS, G. (1990). Approximation results in parallel machines stochastic scheduling. *Annals of Operations Research* **26** 195–242.

OPERATIONS RESEARCH CENTER
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
CAMBRIDGE, MASSACHUSETTS 02139

ALFRED P. SLOAN SCHOOL OF MANAGEMENT
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
50 MEMORIAL DRIVE E53-343
CAMBRIDGE, MASSACHUSETTS 02139