Performance Bounds for Scheduling Queueing Networks

by

Jihong Ou and Lawrence M. Wein

OR 227-90 September 1990

Performance Bounds for Scheduling Queueing Networks

Jihong Ou

Operations Research Center, M.I.T.

and

Lawrence M. Wein

Sloan School of Management, M.I.T.

Abstract

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.

September 1990

Performance Bounds for Scheduling Queueing Networks

Jihong Ou

Operations Research Center, M.I.T.

and

Lawrence M. Wein

Sloan School of Management, M.I.T.

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, for example, Harrison 1988, Laws and Louth 1989, Harrison and Wein 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 has no way of knowing 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 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 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. We begin by assuming that the processing times for all operations performed at a given station are independent and identically distributed exponential random variables; however, the arrival processes are allowed 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 one. The 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 i^{th} 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 2, we generalize the network under consideration to 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.

In Section 3, we derive a simple bound that ignores all the congestion effects across classes; this bound is primarily used as a basis for comparison. 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 1 or 2 (depending on the particular network).

The numerical results are moderately encouraging, with the time average value of the bound equaling 78.0%, on average, of the mean number of customers in the network under the proposed policy. Since the pathwise bound derived in Section 1 is more effective than the steady-state bound derived in Section 2, 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 examples 2,3, and 4, 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 Laws and Louth (1989) and Harrison and Wein (1989) 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 (1988), 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 Pathwise Bound

The network considered in this section has I single-server stations and is 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. Customers of class $k = 1, ..., K$ require service at a particular station $s(k)$, and we 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.

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 under any scheduling policy for $\sum_{k=1}^{K} Q_k(t)$ for all times t. Define the $I-$dimensional process $W = (W_i)$ by

$$W(t) = MQ(t) \text{ for all } t \geq 0, \tag{1}$$

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

$$W_i^*(t) \leq W_i(t) \text{ for } i = 1, ..., I, \text{ and } t \geq 0, \tag{2}$$

for all scheduling policies. (We will construct such a bound shortly.) Then, by (1) and (2), 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 $W^*(t)$:

$$\min_{Q(t)} \sum_{k=1}^{K} Q_k(t) \tag{3}$$

$$\text{subject to } \sum_{k=1}^{K} M_{ik} Q_k(t) \geq W_i^*(t) \text{ for } i = 1, ..., I, \tag{4}$$

$$Q_k(t) \geq 0 \text{ for } k = 1, ..., K. \tag{5}$$

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

$$f(W_1^*(t), ..., W_I^*(t)) \leq \sum_{k=1}^{K} Q_k(t) \text{ for } t \geq 0. \tag{6}$$

In (6), 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 remainder of this section is devoted to finding a pathwise lower bound W^* satisfying (2), but first we observe that the function f in (6) has a very simple form in a special, but not uncommon, case.

Proposition 1. *If there is a customer type who visits every station in the network, then*

$$f(W_1^*(t), ..., W_I^*(t)) = \max_{1 \leq i \leq I} W_i^*(t) \text{ for } t \geq 0. \tag{7}$$

Proof. If we denote the dual variables by $\pi_i, i = 1, ..., I$, the dual linear program to (3)-(5) is

$$\max_{\pi} \sum_{i=1}^{I} \pi_i W_i^*(t) \tag{8}$$

$$\text{subject to } \sum_{i=1}^{I} \pi_i M_{ik} \leq 1 \text{ for } k = 1, ..., K, \tag{9}$$

$$\pi_i \geq 0 \text{ for } i = 1, ..., I. \tag{10}$$

If there is a customer type who visits every station in the network, then there is a constraint in (9) of the form $\sum_{i=1}^{I} \pi_i \leq 1$. Since M_{ik} takes on the value of zero or one for all $i = 1, ..., I$, and $k = 1, ..., K$, all other constraints in (9) are redundant, and the result follows. ∎

In summary, a pathwise performance bound (6) has been derived in terms of a hypothetical vector process W^* that satisfies (2). In order to construct W^*, we need to complete the specification of the queueing network. Customers of class $k = 1, ..., K$ 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 one), 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 $i = 1, ..., I$, let $\{S_i(t), t \geq 0\}$ be a Poisson process with parameter μ_i, which is the service rate for station i, and suppose $S_i(0) = 0$. Thus, we interpret $S_i(t)$ as the number of service completions at station i up to time t if the server was always busy during $[0, t]$. As in Harrison and Wein (1989), we assume the network is run according to the following modified service mechanism that was introduced by Borovkov (1965). The potential service processes $S_i, i = 1, ..., I$, are always turned on, and whenever a potential service completion occurs in S_i, then a customer is allowed to depart station i; the particular exiting customer depends on the scheduling policy used at station i, and a departure only occurs if at least one customer is present at station i. If a customer arrives to station i 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_i.

The key to constructing W^* is to derive an upper bound on the cumulative departure process from each station 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 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 neccessarily 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. In particular, we define a $K-$ dimensional vector $N^* = (N_k^*)$ of *delayed arrival processes*, but we essentially ignore N_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 $\{k : I(k) > 0\}$, $N_k^*(t)$ represents the number of class k customers who have arrived to station $s(k)$ (which is the station that serves them in the original network) of the modified network in $[0,t]$. For $\{k : I(k) > 0\}$, the process N_k^* will be constructed so that $N_k^*(t)$ will be greater than or equal to $A_k(t)$, which we define to be the number of arrivals of class k customers to station $s(k)$ up to time t in the original network under any arbitrary network scheduling policy. 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. The processes $N_k^*, k = 1, ..., K$, will

be defined sequentially starting with $k = 1$. In particular, if class k corresponds to the first stage along some customer type's route, then let

$$N_k^*(t) = N_k(t) \text{ for } t \geq 0, \tag{11}$$

and otherwise, let

$$N_k^*(t) = S_{s(k-1)}(t) + \inf_{0 \leq s \leq t} \{N_{k-1}^*(s) - S_{s(k-1)}(s)\} \text{ for } t \geq 0. \tag{12}$$

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

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

$$A_k(t) \leq N_k^*(t), \text{ for all scheduling policies and all classes } \{k : I(k) > 0\}. \tag{13}$$

Proof. In order to explain equation (12), suppose class k is the n^{th} stage along a customer type's route, where $n \geq 2$, and suppose $I(k) > 0$. Then $\{N_k^*(t), t \geq 0\}$ 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 $\{N_{s(k-n+1)}^*(t), t \geq 0\}$ (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 $i = 1, ..., n-1$ of the tandem system is $\mu_{s(k-n+i)}$. Notice that the departure process in (12) is expressed as the *potential* number of departures 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. Thus, N_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, $N_k^*(t)$ 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 (13) holds. ∎

The arrival process to station i in the modified network is $\{\sum_{\{k:I(k)=i\}} N_k^*(t), t \geq 0\}$, which is a superposition of the delayed arrival processes for the various classes that visit this station for the first time, and the potential service process for station i in the modified network is $\{S_i(t), t \geq 0\}$. Define $\{F_i(t), t \geq 0\}, i = 1, ..., I$, 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 are independent and identically distributed, this policy reduces to awarding priority to the class that has the least number of remaining stages of service on their route. Then define $W_i^*(t)$ for $i = 1, ..., I$, and $t \geq 0$, by

$$W_i^*(t) = \sum_{k=1}^{K} M_{ik} N_k(t) - F_i(t), \tag{14}$$

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, $W_i^*(t) \leq W_i(t), i = 1, ..., I$.*

Proof. For $\{k : I(k) > 0\}$, recall that $\{A_k(t), t \geq 0\}$ is the arrival process of class k customers to station $s(k)$ in the actual queueing network under an arbitrary scheduling policy. For the original queueing network, 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.

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

$$F_i(t) = S_i(t) + \inf_{0 \le s \le t} \left\{ \sum_{\{k:I(k)=i\}} N_k^*(s) - S_i(s) \right\}, \quad \text{for} \quad t \ge 0. \tag{15}$$

Although $D_i(t)$ depends on the scheduling policy employed, we have, for $i = 1, ..., I$, and $t \ge 0$,

$$D_i(t) \le S_i(t) + \inf_{0 \le s \le t} \left\{ \sum_{\{k:I(k)=i\}} A_k(s) - S_i(s) \right\} \tag{16}$$

$$\le F_i(t) \quad \text{by} \quad (13) \quad \text{and} \quad (15). \tag{17}$$

Notice that the inequality in (16) is tight if the server at station i in the actual queueing network services customers whenever the queue is not empty. For $i = 1, ..., I$, and $t \ge 0$, we have

$$W_i(t) = \sum_{k=1}^{K} M_{ik} N_k(t) - D_i(t) \tag{18}$$

$$\ge \sum_{k=1}^{K} M_{ik} N_k(t) - F_i(t) \quad \text{by} \quad (17), \tag{19}$$

$$= W_i^*(t) \quad \text{by} \quad (14). \tag{20}$$

Now let us consider the general feedback case. By (14) and (18), it suffices to show that $D_i(t) \le F_i(t)$ for $i = 1, ..., I$, and for all scheduling policies. To repeat, $\{D_i(t), t \ge 0\}$ is the departure process of exiting customers from station i in the original feedback queueing network under any arbitrary scheduling policy. Furthermore, for k such that $I(k) = i$, class k customers arrive to station i in this network according to $\{A_k(t), t \ge 0\}$, which in turn depends on the network scheduling policy. Observe that if customers in this 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 $\{D_i(t), t \ge 0\}$ could be observed. Moreover, if the actual arrival process of first time customers to station i, $\{\sum_{\{k:I(k)=i\}} A_k(t), t \ge 0\}$, was

replaced by our delayed arrival process $\{\sum_{\{k:I(k)=i\}} N_k^*(t), t \geq 0\}$, then, by (13), the same scheduling policy (and hence the same departure process of exiting customers), could be realized. Thus, 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 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 $\{k : I(k) = i\}$, 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

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

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. ∎

2. 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, $N_k, k = 1, ..., K$, 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 one. For $k = 1, ..., K$, let q_k be defined by

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

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

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

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 (1) that

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

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

$$w_i^* \triangleq \lim_{T\to\infty} \frac{1}{T} E[\int_0^T W_i^*(t)dt] \leq w_i, \quad i = 1, ..., I, \tag{25}$$

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^*:

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

$$\text{subject to} \quad \sum_{k=1}^{K} M_{ik} q_k \geq w_i^* \quad \text{for } i = 1, ..., I, \tag{27}$$

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

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

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

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

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

and thus the steady-state bound is not as effective, in general, as the pathwise bound derived in Section 1. Our inability to find a pathwise bound W^* satisfying (2) 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 $i = 1, ..., I$.

In order to derive a stochastic process W^* satisfying (25), we again consider the modified queueing network described in Section 1; however, the network will be defined slightly differently, since each customer class can have its own exponential service time distribution. In particular, let $S_k, k = 1, ..., K$, 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 $N_k^*, k = 1, ..., K$, are defined exactly as in (11) and (12), except the service processes $S_{s(k)}$ in (12) are replaced here by S_k; that is, if class k corresponds to the first stage along some customer type's route, then let

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

and otherwise, let

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

The arrival process to station i of the modified network is $\sum_{\{k:I(k)=i\}} N_k^*$. Since $N = (N_k)$ are Poisson processes, it follows by the explanation of equation (12) in the proof of Proposition 2 and by Burke's theorem (Burke 1956) that N_k^* is a Poisson process for all k. Since N_k^* are independent for all k such that $I(k) = i$, it follows that the arrival process to each station in the modified network is a superposition of independent Poisson arrival processes, which is itself Poisson; thus 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 such that $I(k) > 0$, $\{A_k(t), t \geq 0\}$ is the arrival process of class k customers to station $s(k)$ in the original feedback network under any arbitrary scheduling policy. We will also need the following result.

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

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

Proof. We begin by supposing that all customer classes $\{k : I(k) = i\}$ visit station i on the first stage of their route. Then

$$\sum_{\{k:I(k)=i\}} N_k^*(t) = \sum_{\{k:I(k)=i\}} N_k(t) \quad \text{by} \quad (31), \tag{34}$$

$$= \sum_{k=1}^{K} M_{ik} N_k(t), \tag{35}$$

since $N_k(t) = 0$ for $t \geq 0$ for all classes not on the first stage of some customer type's route. Notice that $\sum_{k=1}^{K} M_{ik} N_k(t)$ is independent of whether the classes $\{k : I(k) = i\}$ visit station i on the first stage of their route or on a later stage of their route. Now suppose some customer classes in the set $\{k : I(k) = i\}$ visit station i at a later stage of their route. If classes $k-1$ and k belong to the same customer type's route, then it is clear from (32) that $N_{k-1}^*(t) \geq N_k^*(t)$ for $t \geq 0$. Thus, $\sum_{\{k:I(k)=i\}} N_k^*(t)$ is less than or equal to the left side of equation (34), and our result follows. ∎

As in Section 1, we let $\{F_i(t), t \geq 0\}, i = 1, ..., I$, 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 $i = 1, ..., I$, and $t \geq 0$, by

$$W_i^*(t) = \sum_{k=1}^{K} M_{ik} N_k(t) - F_i(t). \tag{36}$$

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

$$w_i^* = \lim_{T \to \infty} \frac{1}{T} E\left[\int_0^T \left(\sum_{k=1}^{K} M_{ik} N_k(t) - F_i(t) \right) dt \right]. \tag{37}$$

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

Proof. Notice that (36) can also be expressed as

$$W_i^*(t) = \left[\sum_{k=1}^{K} M_{ik}N_k(t) - \sum_{\{k:I(k)=i\}} N_k^*(t)\right] + \left[\sum_{\{k:I(k)=i\}} N_k^*(t) - F_i(t)\right], \tag{38}$$

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

$$W_i(t) = \sum_{k=1}^{K} M_{ik}N_k(t) - D_i(t) \tag{39}$$

$$= \sum_{k=1}^{K} M_{ik}N_k(t) - \sum_{\{k:I(k)=i\}} N_k^*(t) + \sum_{\{k:I(k)=i\}} N_k^*(t) - D_i(t). \tag{40}$$

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

$$\lim_{T\to\infty} \frac{1}{T} E\left[\sum_{k=1}^{K} M_{ik}N_k(t) - \sum_{\{k:I(k)=i\}} N_k^*(t)\right] \tag{41}$$

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 N^*); this quantity is finite, since the traffic intensity at each station in the original queueing network is less than one. Thus, by (23), (37)-(38), and (40), it suffices to show that, for all scheduling policies,

$$\lim_{T\to\infty} \frac{1}{T} E\left[\int_0^T \sum_{\{k:I(k)=i\}} N_k^*(t) - F_i(t)dt\right] \leq \lim_{T\to\infty} \frac{1}{T} E\left[\int_0^T \sum_{\{k:I(k)=i\}} N_k^*(t) - D_i(t)dt\right], \tag{42}$$

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

By the argument in the paragraph under equation (20) 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 (42) follows by the fact that the SERPT policy minimizes the long run expected number of customers in a multiclass $M/M/1$ feedback queue (see Klimov 1974 for a derivation of this classic result). ∎

By (33) and (37)-(38), it follows that

$$w_i^* \geq \lim_{T\to\infty} \frac{1}{T} E\left[\int_0^T \sum_{\{k:I(k)=i\}} N_k^*(t) - F_i(t)dt\right], \tag{43}$$

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 $\max_{\{1\leq i\leq I\}} \rho_i \geq 1$.

3. A Pathwise Bound for Each Customer Type

In this section, we briefly 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 (rather than each station) in isolation. This bound allows a different exponential service time distribution for each class (which again is represented by the potential service processes $\{S_k(t), t \geq 0\}, k = 1, ..., K$), and arbitrary interarrival time distributions (denoted by $\{N_k(t), t \geq 0\}, k = 1, ..., K$).

Suppose a certain customer type has n stages on its route and the first stage on its route corresponds to class k. Then consider an $n-$station FCFS tandem queueing system with arrival process $\{N_k(t), t \geq 0\}$ to the first station, and station $i = 1, ..., n$ has service

time distribution characterized by the potential service process $\{S_{k+i-1}(t), t \geq 0\}$. If we define $N_k^*, k = 1, ..., K$, as in (31)-(32), then the number of customers in the tandem queueing system at time t is

$$N_k(t) - S_{k+n-1}(t) - \inf_{0 \leq s \leq t} \{N_{k+n-1}^*(s) - S_{k+n-1}(s)\} \text{ for } t \geq 0, \tag{44}$$

which is a lower bound on the total number of customers of this type in the actual queueing network at time t, for all $t \geq 0$. If we index the customer types in the network by $j = 1, ..., J$, and let $Z_j(t)$ be the number of customers in the j^{th} tandem system at time t, then $\{\sum_{j=1}^{J} Z_j(t), t \geq 0\}$ 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 previous bounds 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. In contrast, the function f appearing in (6) and (29) does not allow us to incorporate a contribution from each customer class at each point in time. 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 network has low traffic intensity, or the majority of the offered load at each station is due to one customer class.

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 six different scenarios, which consist of two levels of load balance (abbreviated by BALANCED and IMBALANCED) crossed with three levels of load intensity (LIGHT, MEDIUM, and 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 .3, .6, and .9 for the three respective load intensities. For the two-station IMBALANCED networks, the vector ρ of traffic intensities is (.3,.2), (.6,.4), and (.9,.6) for the three load intensities, and for the three-station imbalanced networks, the respective vectors are (.3,.2,.1), (.6,.4,.2), and (.9,.6,.3).

For each scenario of each network, we simulate and record the time average values of three stochastic processes: (1) the number of customers in the network under the FCFS policy, (2) the number of customers in the network under a PROPOSED policy (which is derived from either previous analysis or on a trial-and-error basis), and (3) the lower BOUND (from either Section 1 or Section 2, depending on the particular network). The pathwise bound from Section 3 was also tested for each scenario, but we only record the results for the one case where it was tighter than the other bound; in the majority of cases, this bound performed poorly, as expected.

For each scenario, 20 independent runs were made, each consisiting 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 the tables to follow, we provide the mean (and 95% confidence interval) over the 20 runs of the time average value of the three stochastic processes.

Ideally, the effectiveness of our bounds should be measured by their proximity to the number of customers under an optimal scheduling policy. Unfortunately, this is impossible to assess, since the optimal scheduling policy for each of these problems is unknown. Instead, we will record the ratio of the mean of the pathwise lower bound divided by the mean number of customers in the system under the PROPOSED policy. This ratio will be multiplied by 100% and referred to as the *efficiency* of the lower bound in the tables and discussion that follow. 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 BOUND equals

the gap between the PROPOSED policy and an optimal policy plus the gap between an optimal policy and the 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.

Example 1. This simple network is pictured 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 to 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 1 is valid for this network. The service rates are $\mu_1 = 2$ and $\mu_2 = 1$ in the three BALANCED scenarios, and are $\mu_1 = 2$ and $\mu_2 = 1.5$ in the IMBALANCED scenarios. The arrival rates λ_A and λ_B are both set equal to .3 for the LIGHT scenarios, .6 for the MEDIUM scenarios, and .9 for the HEAVY scenarios.

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 under the heavy traffic assumptions that $\mu_1 = 2$, $\mu_2 = 1$, $\lambda_A = 1$, and λ_B was close to 1 (for example, .9). Under these conditions, they analyzed a Brownian approximation (developed in Harrison 1988) to this problem, and proposed the following 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.

The results are recorded in Table I. The average efficiency over the six scenarios is 88.2%, and the bound appears to be slightly more efficient when the network is IMBALANCED. It is also interesting to note that the efficiency of the bound is lowest under the MEDIUM traffic load. When the load on the system is low, there is little congestion in

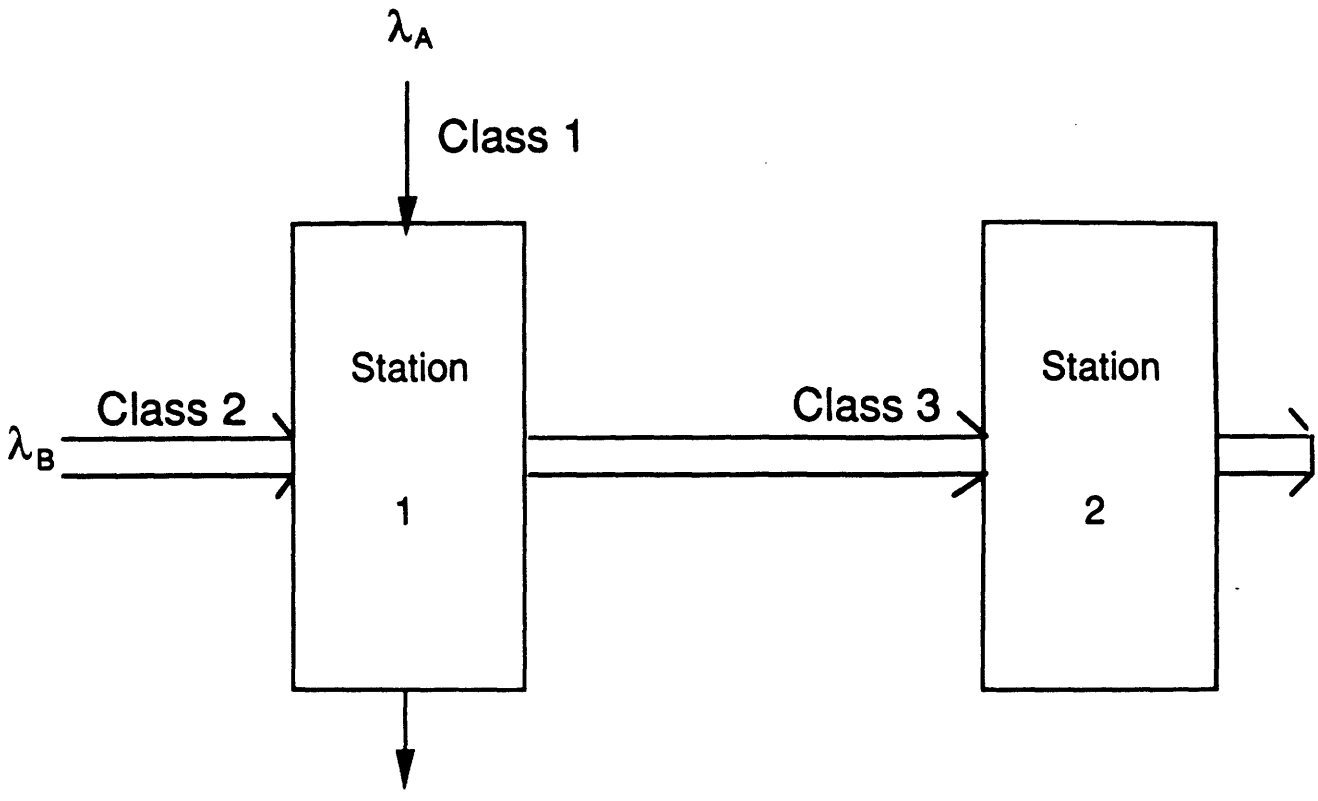

Figure 1. The network for example 1.

the network, and one would not expect a large difference between our pathwise bound and the proposed policy (or the FCFS policy). Moreover, when the pathwise bound derived in Section 1 is applied to the problem in Harrison and Wein (1989), it reduces to the bound denoted by $w_1^{LRPT}(t) \vee w_2^{LRPT}(t)$ in Proposition 2 of that paper. Harrison and Wein show that a pathwise bound that is smaller pathwise than $w_1^{LRPT}(t) \vee w_2^{LRPT}(t)$ 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 reasonable to presume that our bound should also perform well when the load on this network is very high. In order to test this hypothesis, we measured the efficiency of the lower bound when $\lambda_A = \lambda_B = .99$; the efficiency was 93.9% in the BALANCED case ($\rho_1 = \rho_2 = .99$), and 98.2% in the IMBALANCED case ($\rho_1 = .99, \rho_2 = .66$). Similarly, for the problem considered in Harrison and Wein (1989),

readers may refer to Table 2 of that paper to see that the efficiency of the pathwise bound is 89.2% when $\rho = (.95, .9)$ and is 95.5% when $\rho = (.995, .99)$.

SCENARIO	FCFS	PROPOSED	BOUND	EFFICIENCY
BALANCED				
LIGHT	.856 (±.010)	.856 (±.008)	.775 (±.010)	90.5%
MEDIUM	2.99 (±.058)	2.92 (±.051)	2.47 (±.046)	84.6%
HEAVY	17.9 (±.912)	15.5 (±.872)	13.2 (±.824)	85.2%
IMBALANCED				
LIGHT	.677 (±.007)	.677 (±.007)	.621 (±.007)	91.7%
MEDIUM	2.16 (±.058)	2.14 (±.034)	1.85 (±.031)	86.4%
HEAVY	10.7 (±.565)	10.4 (±.548)	9.46 (±.544)	91.0%

Table I. Simulation results for example 1.

Example 2. This example is a simplified two-station version of the nine-station symmetric job-shop studied in Chapter 11 of Conway et al. (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 one-half and exit the network with probability one-half, independent of all previous history. As in Conway et al. (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. For the BALANCED scenarios, the service rates are 1.0 at both stations, and for the IMBALANCED scenarios, the service rate is 1.0 at one station, and 1.5 at the other. The arrival rate λ is adjusted to achieve the desired loading levels for each of the six scenarios.

As in example 1, a Brownian approximation to this queueing network scheduling problem (under the BALANCED, HEAVY scenario) has already been addressed. In particular, Wein and Ou (1989) proposed the following dynamic scheduling policy, which is referred to as the PROPOSED policy in Table 2. For $i = 1, 2$ and $k = 1, ..., 12$, 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

$$V_i(t) = \sum_{k=1}^{12} A_{ik} Q_k(t) \quad \text{for } i = 1, 2, \tag{45}$$

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

The results for example 2 are displayed in Table II. Since all service operations at a given station have the same service rate, the lower bound is calculated using the pathwise bound in Section 1. The average efficiency over the six scenarios is only 77.2%, and thus the bound is not as efficient as it was in example 1. Once again, the pathwise bound appears to be more efficient in the IMBALANCED networks; in particular, the efficiency in the BALANCED networks deteriorates as the load becomes heavier, and is below 60% under a HEAVY load. However, the bound efficiency does improve under very heavy loads; the efficiency was 61.7% in the BALANCED network ($\rho_1 = \rho_2 = .99$) and 91.0% in the IMBALANCED network ($\rho_1 = .99, \rho_2 = .66$). Since $f(x_1, x_2) = x_1 \vee x_2$ by Proposition 1, it is clear why the bound is 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, our bound becomes 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.

SCENARIO	FCFS	PROPOSED	BOUND	EFFICIENCY
BALANCED				
LIGHT	.864 (±.014)	.838 (±.015)	.698 (±.010)	83.3%
MEDIUM	3.02 (±.058)	2.78 (±.053)	2.07 (±.030)	74.5%
HEAVY	18.2 (±1.27)	13.6 (±.709)	8.03 (±.422)	59.0%
IMBALANCED				
LIGHT	.673 (±.011)	.665 (±.010)	.569 (±.008)	85.6%
MEDIUM	2.16 (±.042)	2.02 (±.036)	1.63 (±.032)	80.7%
HEAVY	10.4 (±.734)	7.53 (±.394)	6.03 (±.352)	80.1%

Table II. Simulation results for example 2.

The large amount of feedback present in this example is probably the main reason why the 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.

Example 3. This two-station example, which is pictured in Figure 2, 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 2 is required. There are two customer types, A and B, with two and four stages on their respective routes. The six

customer classes will be indexed by $k = 1, ..., 6$, and referred to by their type-stage pair: $A1, A2, B1, B2, B3$, and $B4$. The mean service times (not rates) for the six scenarios are (8, 6, 2, 7, 4, 1) in the three balanced scenarios, and (8, 4, 2, 14/3, 4, 1.5) in the three IMBALANCED scenarios. The Poisson arrival rates λ_A and λ_B are 3/14 for the LIGHT scenarios, 6/14 for the MEDIUM scenarios, and 9/14 for the HEAVY scenarios.

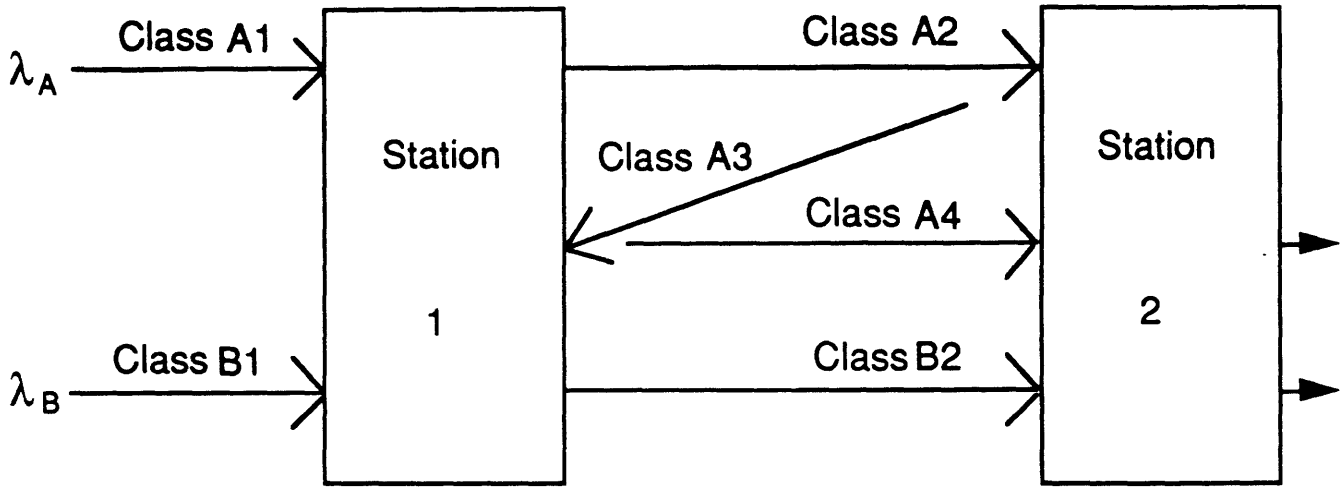

Figure 2. The network for example 3.

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 in Table III is the SERPT policy.

The results for example 3 are displayed in Table III. The average efficiency of the bounds over the six scenarios is 72.4%, which is lower than example 2; the lower efficiency may be partially due to the fact that, as explained in (30), the steady-state bound de-

rived in Section 2 is not as efficient as the pathwise bound derived in Section 1. Once again, the bound efficiency decreases with the load in the BALANCED networks, and the lowest efficiency was achieved under the BALANCED, HEAVY scenario. However, under LIGHT and MEDIUM loads, the efficiency was slightly higher under the BALANCED network than under the IMBALANCED network. As in example 2, there was a huge discrepancy in the bound efficiency under very heavy loads; the efficiency was only 49.4% in the BALANCED network ($\rho_1 = \rho_2 = .99$) and was 94.8% in the IMBALANCED network ($\rho_1 = .99, \rho_2 = .66$). Also, the simple bound derive in Section 3 was tighter than the Section 2 bound in the IMBALANCED, LIGHT scenario, although the two bounds were nearly identical in this case.

SCENARIO	FCFS	PROPOSED	BOUND	EFFICIENCY
BALANCED				
LIGHT	.951 (±.015)	.900 (±.013)	.734 (±.011)	81.6%
MEDIUM	3.76 (±.093)	3.06 (±.068)	2.18 (±.045)	71.2%
HEAVY	23.9 (±1.98)	15.1 (±1.14)	8.48 (±.611)	56.2%
IMBALANCED				
LIGHT	.742 (±.011)	.709 (±.010)	.560* (±.006)	79.0%
MEDIUM	2.64 (±.060)	2.22 (±.044)	1.47 (±.031)	66.2%
HEAVY	13.2 (±.906)	8.67 (±.475)	6.94 (±.451)	80.0%

*Section 3 bound

Table III. Simulation results for example 3.

Example 4. Our last example is the three-station tandem queueing system pictured in Figure 3. The six classes are indexed by $k = 1, ..., 6$ and ordered by ($A1, A2, A3, B1, B2, B3$). The service times for each class are exponential with mean (2, 4, 6, 7, 5, 3) in the three BALANCED cases, and (2, 2, 1, 7, 4, 2) in the IMBALANCED cases. The Posson arrival

rates λ_A and λ_B are both 1/3 for the LIGHT scenarios, 2/3 for the MEDIUM scenarios, and 1.0 for the HEAVY scenarios. The steady-state bound derived in Section 2 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 in Table 4.

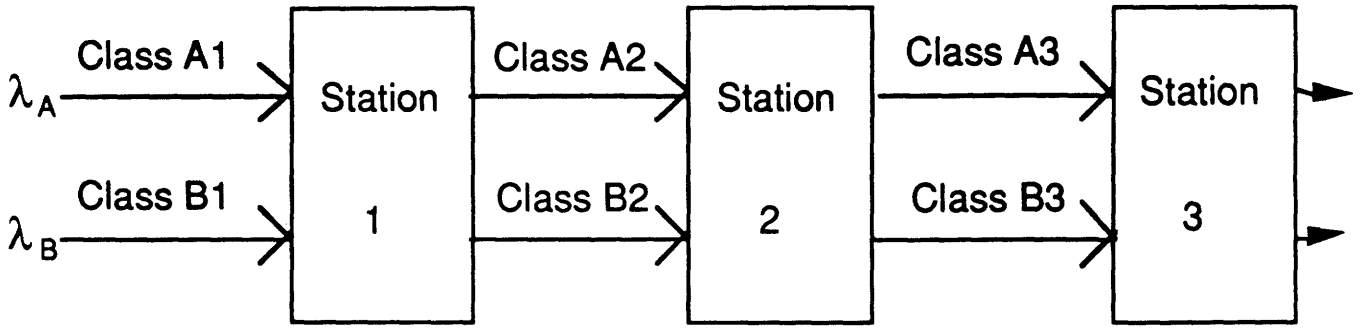

Figure 3. The network for example 4.

SCENARIO	FCFS	PROPOSED	BOUND	EFFICIENCY
BALANCED				
LIGHT	1.37 (±.023)	1.35 (±.017)	1.17 (±.013)	86.7%
MEDIUM	5.15 (±.095)	4.68 (±.075)	3.32 (±.055)	70.9%
HEAVY	32.7 (±2.89)	24.2 (±1.79)	12.2 (±.702)	50.4%
IMBALANCED				
LIGHT	.860 (±.012)	.840 (±.012)	.717 (±.008)	85.4%
MEDIUM	2.79 (±.061)	2.51 (±.046)	1.78 (±.023)	70.9%
HEAVY	14.1 (±1.40)	10.3 (±.903)	8.33 (±.890)	80.9%

Table IV. Simulation results for example 4.

The average bound efficiency over the six scenarios in Table 4 is 74.2%, which is

higher than the corresponding value in example 3; this higher efficiency is perhaps due to to the network's simple routing structure. The qualitative results are similar to the other examples, and only the BALANCED, HEAVY scenario results in a poor bound. The bound efficiencies were 57.1% when $\rho_1 = \rho_2 = \rho_3 = .99$ and 93.2% when $\rho = (.99, .66, .33)$.

Acknowledgements

We are grateful to J. Michael Harrison for suggesting this research topic to us. 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. *Theor. Probability Appl.* **10**, 375-400.

Burke, P. J. The Output of Queueing Systems. *Operations Research* **6**, 699-704.

Conway, R. W., W. L. Maxwell, and L. W. Miller. 1967. *Theory of Scheduling.* Addison-Wesley, Reading, Mass.

Harrison, J. M. 1985. *Brownian Motion and Stochastic Flow Systems.* John Wiley and Sons, New York.

Harrison, J. M. 1988. Brownian Models of Queueing Networks with Heterogeneous Customer Populations, in W. Fleming and P. L. Lions (eds.), *Stochastic Differential Systems, Stochastic Control Theory and Applications,* IMA Volume **10**, Springer-Verlag, New York, 147-186.

Harrison, J. M. and L. M. Wein. 1989. Scheduling Networks of Queues: Heavy Traffic Analysis of a Simple Open Network. *Queueing Systems* **5**, 265-280.

Harrison, J. M. and L. M. Wein. 1990. Scheduling Networks of Queues: Heavy Traffic Analysis of a Two-Station Closed Network. To appear in *Operations Research.*

Kelly, F. P. 1979. *Reversibility and Stochastic Networks*, John Wiley and Sons, New York.

Klimov, G. P. 1974. Time Sharing Service Systems I. *Th. Prob. Appl.* **19**,532-551.

Laws, C. N. and G. M. Louth. 1989. Dynamic Scheduling of a Four Station Network. To appear in *Probability in the Engineering and Information Sciences.*

Panwalkar, S. S. and W. Iskander. 1977. A Survey of Scheduling Rules. *Operations Research* **25**, 45-61.

Wein, L. M. 1990a. Scheduling Networks of Queues: Heavy Traffic Analysis of a Two-Station Network With Controllable Inputs. To appear in *Operations Research.*

Wein, L. M. 1990b. Scheduling Networks of Queues: Heavy Traffic Analysis of a

Multistation Network With Controllable Inputs. To appear in *Operations Research.*

Wein, L. M. and J. Ou. 1990. The Impact of Processing Time Knowledge on Dynamic Job-Shop Scheduling. Submitted for publication.

Weiss, G. 1988. Approximation Results in Parallel Machines Stochastic Scheduling, Industrial and Systems Engineering Department, Georgia Institute of Technology, Technical Report No. J-88-3.