Probability in the Engineering and Informational Sciences, 12, 1998, 329–350. Printed in the U.S.A.

BOUNDS ON THE MEAN DELAY IN MULTICLASS QUEUEING NETWORKS UNDER SHORTFALL-BASED PRIORITY RULES

SRIDHAR SESHADRI AND MICHAEL PINEDO
Department of Statistics and Operations Research & Operations Management Area
Leonard N. Stern School of Business
New York University
New York, New York 10012

A significant amount of recent research has been focused on the stability of multiclass open networks of queues (MONQs). It has been shown that these networks may be unstable under various queueing disciplines even when at each one of the nodes the arrival rate is less than the service rate. Clearly, in such cases the expected delay and the expected number of customers in the system are infinite. In this paper we propose a new class of scheduling rules that can be used in multiclass queueing networks. We refer to this class as the stable shortfall-based priority (SSBP) rules. This SSBP class itself belongs to a larger class of rules, which we refer to as the shortfall-based priority (SBP) rules. SBP is a generalization of the standard nonpreemptive priority rule in which customers of the same priority class are served first-come, first-served (FCFS). Rules from SBP can mimic FCFS as well as the so-called strict or head-of-the-line priority disciplines. We show that the use of any rule from the SSBP class ensures stability in a broad class of MONQs found in practice. We proceed with the construction of a sample path inequality for the work done by an SSBP server and show how this inequality can be used to derive upper bounds for the delay when service times are bounded. Bounds for the expected delay of each class of customers in an MONQ are then obtained under the assumptions that the external arrival processes have i.i.d. interarrival times, the routings are deterministic and the service times at each step of the route are bounded. In order to derive these bounds for the average delay in an MONQ we make use of some of the classical ideas of Kingman.

 0269-9648/98 $12.50

1. INTRODUCTION

Closed-form solutions for multiclass open networks of queues (MONQs) have been obtained only for very special cases, such as single-stage queueing systems (Wolff [37]) and reversible queueing networks (Kelly [23]). The same is true with regard to bounds on the delays in MONQs. While some bounds have been found for certain types of telecommunications networks (Cruz [11,12]), a number of examples in the literature show that MONQs can be unstable even when service times are deterministic, the arrival rate at each node is less than the service rate, and each node operates according to the first-come, first-served (FCFS) rule.

In this paper, we establish upper bounds on the expected delay in MONQs under the assumption that every node operates according to a certain type of priority rule. In order to obtain these results, we first select a priority rule that guarantees a stable system. To this end, we introduce a new class of rules which we call the stable shortfall-based priority (SSBP) rules. A rule from this class is a function of a set of parameters and its behavior can be controlled to a certain extent through proper selection of these parameters. The parameters represent the proportion of busy time the server allocates to each class of customer. We construct a sample path inequality for the work done in a multiclass single-stage, single-server system under a SSBP rule by comparing it with the work done in a specific single-class FCFS queue. This inequality is used to establish upper bounds for the expected delay of each class of customer in an MONQ under the assumption that the external arrival processes have i.i.d. inter-arrival times, the routings are deterministic, and the i.i.d. service times are bounded from above. In an intermediate step, we also obtain upper and lower bounds for the delay in a tandem queueing system without requiring the service times to be bounded. These delay bounds for MONQs are useful, as it has been shown in the literature that under a number of service disciplines the delays may be infinite; here, we show that, for this particular class of rules, there is a closed form expression for an upper bound on the delay.

The SSBP class of rules belongs to a larger class of rules that is of interest for several reasons. This larger class, called the shortfall-based priority (SBP) rules, contains the classic nonpreemptive priority rule introduced by Cobham [10] and analyzed by many researchers (see Kleinrock [25] and Buzacott and Shanthikumar [6]). According to this classic priority rule, customers of a higher priority class always have nonpreemptive priority over customers of a lower priority class. Customers within the same class are served according to FCFS. The class of SBP rules is of interest because an SBP rule enables the user to control the allocation of service time over the various priority classes. For example, consider a queue with two priority classes. Assume that 99% of the arrivals are of the higher priority class and 1% of the arrivals are of the lower priority class. The low priority customers always have to wait for the system to be free of high priority customers. Not only do the low priority customers have a large expected waiting time, there is also a large variance in their waiting time. An SBP rule may allocate the time of the server to the different priority classes in such a way that the mean as well as the variance of the waiting times of the lower priority customers are kept within limits.

2. RELATED LITERATURE

Bounds for the expected delay in queueing systems fall into two categories, namely bounds for single-stage systems and bounds for networks. The books by Buzacott and Shanthikumar [6] and Wolff [37] contain a number of bounds for single-stage systems. For results in the context of networks, we refer to the surveys by Harrison and Nguyen [22] and Dai [14]. In what follows, we give a brief overview of the results in the literature on MONQs.

There have been a number of negative results that indicate that MONQs need not be stable under certain scheduling rules, see, e.g., Bramson [2,3], Dai and Wang [15], Seidman [32], and Whitt [36]. In most of these papers, both the arrival and service processes are not stochastic, which makes the phenomenon of instability even more surprising. The recent work of Dai [14] links Harris recurrence to the stability of queueing networks by using fluid limits. Dai's definition of stability under a scheduling rule is based on the underlying Markov process of the network dynamics being positive Harris recurrent. The network is modeled by Dai using three assumptions: (i) the interarrival times as well as service times are mutually independent i.i.d. sequences, (ii) the interarrival times as well as service times have finite expectations, (iii) the interarrival times are unbounded and spread out. As an example, Dai shows that multiclass feed-forward networks are stable under these assumptions. We conjecture that, under the class of scheduling rules proposed in our paper, the feed-forward assumption can be dropped and the MONQ remain stable. In what follows, we prove this conjecture under the additional assumption that all the service times have an upper bound. Chen and Yao [9] recently showed that under the condition of "acyclic class transfer" (i.e., an MONQ in which customers can switch classes, but with no loops in class transfers) a simple priority rule can be identified under which MONQs are stable given that the maximum station utilization is less than unity. Chen and Yao also use the link with Harris recurrence to prove their result. In contrast to the approaches described above, we define a network to be stable under a given rule when the arrival rates of customers equal the corresponding departure rates (see Buzacott and Shanthikumar [6] and Stidham and El-Taha [34]), we develop a scheduling rule that leads to stability of the class of MONQs modeled, we use sample path analysis to provide delay bounds under certain assumptions, and we indicate how stable processor sharing rules can be used to control the MONQ. Our stability results hold when service times are bounded, and they continue to hold without this condition but with an additional restriction on the growth rate of the supremum of the service times, as indicated in Remark 4, Section 5.

Another direction of research was initiated by Bramson using fluid flow models (see [4] and [5]). In [4], he introduces a rule called head-of-the-line proportional processor sharing (HOLPPS). This rule works like processor sharing in the sense that at each instant the service given to a class is proportional to the number of customers present in that class. However, unlike processor sharing, the HOLPPS server provides the entire service allocated to a class to the first customer in that class. It is shown that the (fluid model of the) MONQ is stable under HOLPPS under

the usual traffic conditions. Bramson conjectures that the stability results extend to the processor sharing case. Our results bear out the conjecture for MONQs with bounded service times (see also Section 5) and deterministic routing. In [5], Bramson proves stability of a fluid model of an MONQ under the FIFO rule when the service times at a station do not depend on the class of the customer.

Cruz [11,12] shows how to compute the delay in the context of telecommunications networks. The ideas used by Cruz have been seminal to several other papers dealing with the estimation of network delays in the context of telecommunication; see, e.g., Chang [7], Cruz and Liu [13], Parekh [28], and Parekh and Gallagher [29,30]. In these papers the arrival processes are assumed to be deterministic. To the best of our knowledge, the ideas of Cruz have not been used for the construction of delay bounds for either tandem systems or MONQs when the arrival processes are of the renewal type. Parekh [28] introduced a scheduling rule called generalized processor sharing (GPS) and states that under certain conditions GPS is identical to the Fair Queueing rule proposed by Demers, Keshav, and Shenker [16]. Parekh used his GPS rule to compute upper bounds on the mean delay in networks with arbitrary topology assuming deterministic arrival processes (see [29] and [30]). Parekh characterizes the arrival processes by the maximum burst size and the long run average rate of traffic flow, as defined in Cruz [11]. A key aspect of the bounds given by Parekh (and also by Cruz [11,12]) is that arrival patterns can be found that achieve the worst case bound on the queue size and delay. As these results apply to deterministic networks, determining the worst case delay is equivalent to showing that the number of customers in the system over any time interval is uniformly bounded. This statement could serve as a definition of stability for such networks (see also the discussion in Section 5).

Subsequent work related to GPS provides tight bounds on the delay and on the number of packets (customers) in the system (see Georgadis et al. [18,19]). Zhang, Towsley, and Kurose [39] obtained bounds for the delay under the Generalized Processor Sharing rule using the so-called exponentially bounded burstiness (E.B.B.) model introduced by Yaron and Sidi [38]. The SBP class of rules are somewhat similar to those of GPS, but they also differ in at least two ways: SBP rules are nonanticipative, whereas GPS needs information about the service times of customers in the system, and SBP rules are less cumbersome to implement than GPS rules. Another stream of research on the stability of MONQs goes back to the work of Perkins and Kumar [31]. They examined the stability of MONQs with respect to several scheduling rules (see Kumar and Seidman [26] and Lu and Kumar [27] for examples). In these papers, stability is established using sample path arguments and recursive relations. The methods we use for obtaining the sample path inequality in Section 3 are similar. To the best of our knowledge, these ideas have not been used before to obtain bounds on the mean delay in MONQs.

Our proof techniques use some ideas that have appeared in the literature. Chang, Thomas, and Kiang [8] demonstrated the stability of open Jackson networks using the concept of scaled service times. We developed this idea of scaled service times independently, and use it in the Appendix to obtain a bound for the mean delay in a

tandem system of single-server queues. In Section 3, we define the shortfall-based scheduling rules. We then combine the bounds derived for a single-server queue in Section 3 and the delay bounds for tandem systems to derive a closed form expression for the delay in an MONQ.

3. BOUNDS FOR DELAY IN A SINGLE-SERVER QUEUE

In view of the negative results with regard to the stability of MONQs, it is of interest to define scheduling rules that enable us to obtain bounds on the expected delays in MONQs. In what follows we define a broad class of rules that are nonanticipative, nonpreemptive, and work conserving. We refer to this class of rules as the shortfall-based priority rules. We further restrict ourselves to a sub-class of SBP, called the stable shortfall-based priority rules, and compare the performance of a rule from this subclass to the performance of FCFS.

Consider a single-server queue with C classes of customers. The arrival processes are arbitrary. The service times sequence of each class is given by an i.i.d. sequence of random variables. The service times are assumed to be bounded both from above and below, i.e., away from zero. The following notation and terminology is used.

- $A_i(t)$ = number of customers of class i that arrive during time $[0,t]$.
- λ_i, $i = 1,2,\ldots,C$, are finite, strictly positive, constants. In the next section they will be interpreted as arrival rates. In this section, however, they are just fixed numbers subject to a constraint given below.
- $S_{i,n}$, $n = 1,2,\ldots$, is a sequence of i.i.d. service times, where $S_{i,n}$ is the service time of the nth customer of class i, $i = 1,2,\ldots,C$.
- $\rho_i = \lambda_i E(S_{i,1})$, $\rho_i > 0$, $i = 1,2,\ldots,C$, and $\sum_{i=1}^{C} \rho_i = \rho < 1$.
- $0 < L_i \leq S_{i,1} \leq U_i < \infty$, $i = 1,2,\ldots,C$.
- $U = \max_{i=1,\ldots,C} U_i$.
- θ_i, $i = 1,2,\ldots,C$ are constants used to specify a rule in SBP. The fraction θ_i is to be interpreted as the proportion of the busy time that must be dedicated to serving class i customers. A rule in SBP will be referred to as SBP(θ).
- $D_i(\sigma,t)$ = number of departures of class i customers from the queue during $[0,t]$ under scheduling rule σ.
- $N_i(\sigma,t) = A_i(t) - D_i(\sigma,t)$, is the number of class i customers in the system at time t, when scheduling rule σ is used.

In order to define the rule we need the following notation. Let

$$l_i(t) = \begin{cases} \sup\{s \geq 0 : N_i(\text{SBP}(\theta),s) = 0\} & \text{if } N_i(\text{SBP}(\theta),s) = 0 \text{ for some } s \in [0,t] \\ 0 & \text{otherwise} \end{cases}. \tag{1}$$

So the value of $l_i(t)$ is the last instant prior to time t under SBP(θ) with no class i customers present. At that point in time a customer of class i arrives and finds no-

body of class i in the system. This customer starts a busy period of class i customers that lasts until the current time t. Let

$$Y_i(t) = (t - l_i(t))\theta_i, \qquad i = 1,2,\ldots,C. \tag{2}$$

$Y_i(t)$ may be interpreted as the fraction of time that should have been devoted to class i customers during the time $[l_i(t), t]$ under the SBP(θ) rule. The Y_i and θ_i processes are quantities similar to those defined in Wein [35]. Let

$V_i(\sigma, a, b)$ = the work done during the time interval $[a, b]$ on class i customers, $i = 1,2,\ldots,C$, under scheduling rule σ.

$P(t) = \{i: N_i(\text{SBP}(\theta), t) > 0\}$, the set of customer classes present at time t under SBP(θ).

DEFINITION 1: *The SBP(θ) Rule. If a customer service is completed at time t, then the next service is initiated without any delay providing at least one customer is waiting in queue, and the next customer to be served will be from the (head of the line of) class*

$$\arg\max_{i \in P(t)} \{Y_i(t) - V_i(\text{SBP}(\theta), l_i(t), t)\}.$$

Customers within a class are treated on an FCFS basis and, once a customer service is started, it will not be interrupted until its completion.

The SBP(θ) rule is work conserving, i.e., the server is not idle when there is work to be done. The rule is also nonanticipative and nonpreemptive, i.e., the server does not interrupt the service of a customer (Wolff [37]). In other words, the SBP(θ) rule keeps track of the work that should have been done on each class of customer. This quantity is compared to the actual work done on that class and priority assigned to the class most behind with regard to shortfall. The class SBP is very broad. It can mimic FCFS by treating all classes as one, and it can imitate strict priority disciplines by assigning a high value to θ_1, a value one order of magnitude smaller to θ_2, etc. The choice of the values of θ_i can be used to control the amount of service provided to each class of customer, and the delays of different classes of customers can be traded off against one another.

In order to ensure stability, however, some restrictions have to be placed on the parameters. To this end, given that ρ_i is the utilization of the server by class i, define the class of stable shortfall-based priority rules as follows.

DEFINITION 2: *The SSBP Subclass of Rules. This is the collection of rules that belong to SBP, and for which $\sum_{i=1}^{C} \theta_i = 1$ and $\theta_i > \rho_i$ for all i. A rule from the subclass SSBP will be referred to as* SSBP(θ).

DEFINITION 3: *The Fair Processor Sharing (FPS) Rule. The* SSBP(θ) *rule with θ_i set equal to ρ_i/ρ for each class i, will be called the fair processor sharing rule. This rule is said to be "fair," because the server's capacity is allocated in proportion to server utilization by each class.*

It is important to see the necessity for a bookkeeping device that tracks the "shortfall"—it ensures fairness in a multiclass system. For example, under FCFS, the bursty arrival of one class of customer can make the server unavailable to other classes arriving after the burst. The departure process of each class of customer can also become extremely bursty under FCFS. As shown in Whitt [36], this burstiness can be aggravated to such an extent that the MONQ becomes unstable (also see Cruz [12]). The avoidance of bursty departures (see Theorem 1 below) is another appealing feature of SSBP. The reason for imposing the stability conditions that define the subclass SSBP will become obvious in Theorem 1 below.

The reader will observe that the SBP class of rules (developed independently by us) is closely connected to the head-of-the-line processor sharing rule (HOL-PS). Parekh used his scheduling rule, the generalized processor sharing (GPS) rule to mimic HOL-PS. GPS was designed to control the amount of service performed on each class under HOL-PS. The idea behind GPS is to assign a positive number φ_i to each class i. Denote the service provided to class j during $[t_1, t_2]$ as $S_j[t_1, t_2]$. Then, if class i is continuously backlogged during an interval $[t_1, t_2]$, GPS ensures that

$$\frac{S_i[t_1, t_2]}{S_j[t_1, t_2]} \geq \frac{\varphi_i}{\varphi_j}.$$

In order to implement GPS without having to share the processor, Parekh introduced a scheme called the packet-by-packet generalized processor sharing (PGPS) rule. PGPS tracks the order in which GPS would have scheduled the customers and implements the GPS schedule in a nonpreemptive manner. This implementation of PGPS has properties similar to SBP. PGPS needs information about the service of the customer before the service is completed, and therefore does not satisfy our nonanticipation criterion. (This requirement for PGPS is met in ATM networks where the packet length is known upon arrival.) Moreover, some simulation is required to track the order in which GPS would have scheduled customers. This computational burden is avoided under SBP. As a historical note, GPS is a generalization of the processor sharing and the FB scheduling algorithms discussed in Kleinrock [25] and the weighted fair queueing rule presented in Demers et al. [16].

In what follows, we analyze the performance of an SSBP(θ) rule with respect to system departures of class 1 customers only. For ease of notation define

$$H_i(t) = Y_i(t) - V_i(\text{SSBP}(\theta), l_i(t), t), \qquad i = 1,2,\ldots,C. \tag{3}$$

The $H_i(t)$ has to be regarded as the shortfall in the work done on class i customers at time t. By the definitions of l_i, Y_i, and H_i, $H_i(t)$ equals zero (regenerates) every time the SSBP(θ) system gets rid of all its class i customers, $i = 1,2,\ldots,C$.

To analyze the performance of the SSBP(θ) rule, we use another single-server queue. We shall refer to this second system as the FCFS system and to the original one as the SSBP(θ) system. Only class 1 customers arrive at the FCFS system, and their arrival process is given by $A_1(t)$. This arrival process is identical to the arrival process of class 1 customers in the SSBP(θ) system. The service times of the nth

customer in the FCFS system are set equal to $S_{1,n}/\theta_1$, $n = 1,2,\ldots$. This slows down the server of the FCFS system. Moreover, if the ρ_i were the load brought in by the various customer classes, the load on the server in the FCFS system would be $\lambda_1 E(S_{1,1})/\theta_1 = \rho_1/\theta_1 < 1$. We shall first establish a sample path inequality for the work done in the SSBP(θ) system.

LEMMA 1: *If*

$$\sum_{i=1}^{C} H_i(0)^+ \leq CU,$$

then

$$\sum_{i=1}^{C} H_i(t)^+ \leq CU, \quad \text{for } t \geq 0, \tag{4}$$

where A^+ denotes the positive part of A.

PROOF: Fix x such that $0 \leq x \leq S_{j,n}$. Let the nth customer of class j start its service at time t. In the interval $[t, t+x]$, V_i is nonincreasing for $i \neq j$. By Eq. (2), $Y_i(t+x) - Y_i(t)$ is less than or equal to $x\theta_i$. Therefore,

$$H_i(t+x) \leq H_i(t) + x\theta_i, \quad \text{for all } i \neq j. \tag{5}$$

As work is being done on class j during $[t, t+x]$,

$$H_j(t+x) = H_j(t) - x(1-\theta_j). \tag{6}$$

If $H_j(t) \geq x(1-\theta_j)$, we obtain from (5) and (6)

$$H_i(t+x)^+ \leq H_i(t)^+ + x\theta_i, \quad \text{for all } i \neq j, \tag{7}$$

$$H_j(t+x)^+ = H_j(t) - x(1-\theta_j) = H_j(t)^+ - x(1-\theta_j). \tag{8}$$

By using (7), (8), and the fact that $\sum_i \theta_i = 1$, we get

$$\sum_{i=1}^{C} H_i(t+x)^+ \leq \sum_{i=1}^{C} H_i(t)^+. \tag{9}$$

Otherwise, if $H_j(t) < x(1-\theta_j)$, then

$$x(1-\theta_j) > H_j \geq H_i, i \in P(t) \tag{10}$$

because the SSBP(θ) rule chose to serve the customer class with the largest value of $H_i(t)$, $i \in P(t)$. For customer classes not present at time t

$$\begin{aligned} l_i(t+x) > t, \quad \text{for all } i \notin P(t) &\Rightarrow Y_i(t+x) \leq x\theta_i, \quad \text{for all } i \notin P(t) \\ &\Rightarrow H_i(t+x) = Y_i(t+x) - V_i(\text{FPS}, l_i(t+x), t) \\ &\leq x\theta_i, \quad \text{for all } i \notin P(t). \end{aligned} \tag{11}$$

Using (5), (10), and (11) we obtain

$$\sum_{i\in P(t),\, i\neq j} H_i(t+x)^+ \leq \sum_{i\in P(t),\, i\neq j} (H_i(t)^+ + x\theta_i)$$

$$\leq (|P(t)| - 1)x(1-\theta_j) + \sum_{i\in P(t),\, i\neq j} x\theta_i, \tag{12}$$

$$\sum_{i\notin P(t)} H_i(t+x)^+ \leq \sum_{i\notin P(t)} x\theta_i, \tag{13}$$

where $|P(t)|$ is the cardinality of the set $P(t)$. By the assumption for this case,

$$H_j(t+x)^+ = 0. \tag{14}$$

Combining (12)–(14) we finally have

$$\sum_{i=1}^{C} H_i(t+x)^+ \leq |P(t) - 1|x(1-\theta_j) + \sum_{i\neq j} x\theta_i \leq CS_{j,n} \leq CU. \tag{15}$$

Inequalities (9) and (15) complete the proof of the lemma. The steps in the proof can be examined to conclude that the bound in (15) does not depend in any way on the stochastic nature of either the arrival processes or the service times. ■

In the next theorem we compare two single-server systems operating under different scheduling rules. The first system is a single-server queue subject to C arbitrary customer arrival processes $A_i(t)$, $i = 1,2,\ldots,C$, and a set of strictly positive numbers λ_i, $i = 1,2,\ldots,C$. Class i customers have i.i.d. service times $S_{i,n}$, $n = 1, 2,\ldots$, which are bounded from above by U and from below by L_i, $i = 1,2,\ldots,C$. The system operates under an SSBP(θ) scheduling rule. The second system is a single-server system that operates under the FCFS rule. It has only one class of customer arriving according to $A_1(t)$. The service time sequence in the FCFS system is given by $S_{1,n}/\theta_1$, $n = 1,2,\ldots$, with $\theta_1 > \rho_1$.

THEOREM 1:

$$D_1(\text{FCFS},t) - D_1(\text{SSBP}(\theta),t) \leq CU/L_1, \qquad t \geq 0.$$

PROOF: Note that the server is working more slowly in the FCFS system than in the SSBP(θ) system. Therefore, if the server in the FCFS system works for one unit of time, then the work done by that server will be counted as θ_1. It suffices to show that

$$(V_1(\text{FCFS},0,t) - V_1(\text{SSBP}(\theta),0,t)) \leq CU. \tag{16}$$

Proving inequality (16) is equivalent to showing that, with regard to class 1 customers, the work done in the FCFS system cannot exceed the work done in the SSBP(θ) system by more than CU. Using the definition of $H_1(t)$, we have

$$H_1(t) = \theta_1(t - l_1(t)) - V_1(\text{SSBP}(\theta), l_1(t), t)$$

$$\geq V_1(\text{FCFS}, l_1(t), t) - V_1(\text{SSBP}(\theta), l_1(t), t), \tag{17}$$

because the server in the FCFS system need not always be busy during $[l_1(t), t]$, and is slowed down by a factor of θ_1. Inequalities (4) and (17) imply that

$$(V_1(\text{FCFS}, l_1(t), t) - V_1(\text{SSBP}(\theta), l_1(t), t)) \leq CU. \tag{18}$$

By the definition of $l_1(t)$,

$$N_1(\text{SSBP}(\theta), l_1(t)) = 0. \tag{19}$$

As $A_1(t)$ is the same for the two systems, Eq. (19) implies that

$$V_1(\text{SSBP}(\theta), 0, l_1(t)) \geq V_1(\text{FCFS}, 0, l_1(t)). \tag{20}$$

Finally, using (18) and (20), we have

$$\begin{aligned}(V_1(\text{FCFS}, 0, t) &- V_1(\text{SSBP}(\theta), 0, t)) \\ &\leq (V_1(\text{FCFS}, 0, l_1(t)) - V_1(\text{SSBP}(\theta), 0, l_1(t)))^+ \\ &+ (V_1(\text{FCFS}, l_1(t), t) - V_1(\text{SSBP}(\theta), l_1(t), t))^+ \\ &\leq 0 + CU = CU. \end{aligned} \tag{21}$$

■

Remarks:

i. It is important to note that the bound in Theorem 1 does not depend on the stochastic nature of arrivals. The only common elements between the FCFS and the SSBP(θ) systems are the single common arrival process $A_1(t)$, the common values of $S_{1,n}, n = 1, 2, \ldots$, and the value of the parameter θ_1.

ii. The condition that $\theta_i > \rho_i$ is necessary for stability; otherwise, the decelerated single-server system becomes unstable. Further generalizations of the class SSBP are discussed in the conclusions.

We now relax the assumption that the arrival processes in the two systems are the same. We state a corollary of the theorem, which will be used in the next section.

COROLLARY 1: *If in Theorem* 1 *the arrival processes of class* 1 *customers are* $A_1^{\text{FCFS}}(t)$, *and* $A_1^{\text{SSBP}(\theta)}(t)$, *and if* $(A_1^{\text{FCFS}}(t) - A_1^{\text{SSBP}(\theta)}(t)) \leq Q, t \geq 0$, *then*

$$D_1(\text{FCFS}, t) - D_1(\text{SSBP}(\theta), t) \leq \frac{(Q + C)U}{L_1}, \qquad t \geq 0.$$

PROOF: Instead of using inequality (20) (as we did in Theorem 1), we use the assumption that $(A_1^{\text{FCFS}}(t) - A_1^{\text{SSBP}(\theta)}(t)) \leq Q, t \geq 0$, and the fact that there are no class 1 customers in the system at time $l_1(t)$ to conclude that

$$\begin{aligned} D_1(\text{SSBP}(\theta), l_1(t)) &= A_1^{\text{SSBP}(\theta)}(l_1(t)) - N_1(\text{SSBP}(\theta), l_1(t)) = A_1^{\text{SSBP}(\theta)}(l_1(t)) \\ &\geq A_1^{\text{FCFS}}(l_1(t)) - Q \geq D_1(\text{FCFS}, l_1(t)) - Q. \end{aligned}$$

This implies

$$V_1(\text{FCFS}, 0, l_1(t)) - V_1(\text{SSBP}(\theta), 0, l_1(t)) \leq QU. \tag{22}$$

As inequality (18) still holds, using (22) we obtain

$$D_1(\text{FCFS},t) - D_1(\text{SSBP}(\theta),t) \leq \frac{QU + (V_1(\text{FCFS}, l_1(t),t) - V_1(\text{SSBP}(\theta), l_1(t),t))^+}{L_1} \leq \frac{(Q+C)U}{L_1}. \quad (23)$$

■

4. MULTICLASS OPEN QUEUEING NETWORKS WITH BOUNDED SERVICE TIMES

In this section we use the results from the previous section to construct an upper bound for the average delay in the system for an MONQ with deterministic routing, uniformly bounded and i.i.d. service times at each step of every customer route, and i.i.d. interarrival times for each class of customers. For simplicity of exposition, we shall carry out the analysis for the fair processor sharing rule, i.e., the SSBP(θ) rule with θ_i set equal to ρ_i/ρ for each class i. The extension to the entire SSBP subclass is straightforward.

The notation is complex, but it will help the reader to know that only the route of class 1 customers will be analyzed in detail. There are C classes of customers arriving at d single-server stations in the MONQ. Each class has a deterministic route assigned to it. A route is a sequence of station numbers that the customers have to visit in the given order. On completion of the route the customer departs the network. The MONQ is defined as follows:

- $T_{i,n}$ = interarrival time between the nth and $(n+1)$st class i customer to arrive at the MONQ. $T_{i,n}, n = 1,2,\ldots$ is an i.i.d. sequence, with $E(T_{i,1}) = 1/\lambda_i$, and the squared coefficient of variation of $T_{i,1}$ is given by $c^2_{T_i}, i = 1,2,\ldots,C$.
- $r_i^{(1)}, r_i^{(2)}, \ldots, r_i^{(n_i)}$ with $r_i^{(s)} \in \{1,2,\ldots,d\}$, $1 \leq s \leq n_i$ denotes the route followed by a customer of class $i, i = 1,2,\ldots,C$. A class i customer arrives at station $r_i^{(1)}$; after being served there it goes to station $r_i^{(2)}$, and so on, until it exits the system after being served at station $r_i^{(n_i)}$ (n_i is said to be the number of steps in this route).
- $S^j_{i,n}$ is the service time at the jth step of the route of the nth customer of class i. It is assumed that $\{S^j_{i,n}, n = 1,2,\ldots\}$ is an i.i.d. sequence and independent of all else in the network as well as the external arrival processes of customers. We assume that all the service times are uniformly bounded from above by the constant U.

In order to apply the results of the previous section, we need to ensure that the service times are bounded from below (i.e., establish the value of L_1 in Theorem 1). To that end, we modify the service times of a customer from class i as follows:

$$\bar{S}^j_{i,n} = \begin{cases} S^j_{i,n} + (1 - \lambda_i E(S^j_{i,1}))/(2\lambda_i) & \text{if } S^j_{i,n} \leq (1 - \lambda_i E(S^j_{i,1}))/(2\lambda_i) \\ S^j_{i,n} & \text{otherwise} \end{cases}, \quad n = 1,2,\ldots; \quad j = 1,2,\ldots,n_i. \quad (24)$$

Note that the modified service time has a bar over the S. One way of implementing the new service times is by allowing the (corresponding) server to take that much longer to serve customers. In the rest of this section we shall use only these modified service times. Define:

- $c^2_{\bar{S}^j_i}$ = the squared coefficient of variation of $\bar{S}^j_{i,1}, j = 1,2,\ldots,n_i; i = 1,2,\ldots,C.$
- $c^2_{S^j_i}$ = the squared coefficient of variation of $S^j_{i,1}, j = 1,2,\ldots,n_i; i = 1,2,\ldots,C.$
- $\rho^j_{r_i^{(j)},i} = \lambda_i E(\bar{S}^j_{i,1})$ is the average load due to the jth processing step in the route of class i customers at station $r_i^{(j)}, j = 1,2,\ldots,n_i; i = 1,2,\ldots,C.$
- $\rho_m = \sum_{i=1}^{C} \sum_{j=1}^{n_i} I\{r_i^{(j)} = m\} \rho^j_{r_i^{(j)},i}$ is the total load on station $m, m = 1,2,\ldots,d.$ We assume that $\rho_m < 1$ for all m.
- $\rho_{\max} = \max_m \{\rho_m\}$.
- $A^{j,\sigma}_{i,n}$ = the arrival instant of the nth customer of class i to the jth step, i.e., to station $r_{i,n}^{(j)}$ when the scheduling rule σ is used at all stations, $j = 1,2,\ldots,n_i;$ $i = 1,2,\ldots,C; n = 1,2,\ldots.$
- $A_i^{j,\sigma}(t) = \sup\{n : A^{j,\sigma}_{i,n} \le t\}$ is the number of arrivals of class i customers during $[0,t]$ on the jth step of their route, when the rule σ is used at all stations.
- $D^{j,\sigma}_{i,n}$ = the departure epoch of the nth customer of class i at the jth step, i.e., from station $r_i^{(j)}$, when rule σ is used at all stations, $j = 1,2,\ldots,n_i;$ $i = 1,2,\ldots,C; n = 1,2,\ldots.$
- $D_i^{j,\sigma}(t) = \sup\{n : D^{j,\sigma}_{i,n} \le t\}$ is the total number of class i customers who have finished the processing of their jth step during $[0,t]$, when rule σ is used at all stations.

With these modified service times, the load due to class i customers at their jth processing step is given by

$$\rho^j_{r_i^{(j)},i} = \lambda_i E(\bar{S}^j_{i,1}) \le \lambda_i E(S^j_{i,n} + (1 - \lambda_i E(S^j_{i,1}))/(2\lambda_i)) = (1 + \lambda_i E(S^j_{i,1}))/2. \tag{25}$$

Denote the maximum load at any station in the network prior to the modification of service times by

$$\rho^O_{\max} = \max_{m=1,2,\ldots,d} \left[\sum_{i=1}^{C} \sum_{j=1}^{n_i} I\{r^{(j)} = m\} \lambda_i E(S^j_{i,1}) \right]. \tag{26}$$

Then

$$\begin{aligned} \rho_{\max} &= \left(1 + \max_{m=1,2,\ldots,d} \left[\sum_{i=1}^{C} \sum_{j=1}^{n_i} I\{r^{(j)} = m\} \lambda_i E(S^j_{i,1}) \right] \right) \Big/ 2 \\ &\le (1 + \rho^O_{\max})/2. \end{aligned} \tag{27}$$

We see from Eq. (27) that if the maximum load was less than one prior to the modification given in (24), then the modified network also has a maximum station load less than one. The special form in (27) is worth noting. It implies that

$$\frac{1}{(1-\rho_{\max})} \leq \frac{2}{(1-\rho^{o}_{\max})}.$$

To apply Theorem 1, we may now use the values of $(1-\rho^{j}_{r_i^{(j)},i})/(2\lambda_i)$ for the lower bounds on the corresponding sequence of service times. We compare two "systems," namely the route of a class 1 customer and a (single class) tandem system with n_1 stations that have single-servers. The MONQ operates under the FPS rule. The tandem system operates under the FCFS rule. The external arrival process at the tandem system is $A_1^{j,\mathrm{FPS}}(t)$. The reader will notice that this quantity does not depend on the scheduling rule, but the superscript underscores the fact that the external arrival processes of class 1 customers are the same in both systems. Define the quantities

$$\theta_j = \rho^{j}_{r_1^{(j)},1}/\rho_{r_1^{(j)}}, \qquad j = 1,2,\ldots,n_1. \tag{28}$$

The service time of the nth customer at the jth station of the tandem system is distinguished by a hat and is

$$\hat{S}^{j}_{1,n} = \bar{S}^{j}_{i,n}/\theta_j, \qquad j = 1,2,\ldots,n_i;\ n = 1,2,\ldots. \tag{29}$$

In (28) and (29), the reader recognizes the parameters used to define the FPS rule. This is not the only possible choice of θ_i parameters—the MONQ will be stable (under the usual traffic conditions) for any rule in SSBP. Let $A_1^{j,\mathrm{FCFS}}(t), D_1^{j,\mathrm{FCFS}}(t)$ denote the total number of arrivals at and departures from the jth station in the tandem system during the interval $[0,t]$.

In our next theorem we again compare two systems. The first system we consider is an MONQ, with d single-server stations, C customer classes, with each class of arrivals having i.i.d. interarrival times and arrival rates $\lambda_i > 0, i = 1,2,\ldots,C$, with deterministic routes $\{r_i^{(j)}, j = 1,2,\ldots,n_i\}$, i.i.d. sequences of service times

$$\left\{\bar{S}^{j}_{i,n} = S^{j}_{i,n} + I\{S^{j}_{i,n} \leq (1-\lambda_i E(S^{j}_{i,1}))/(2\lambda_i)\}\,\frac{(1-\lambda_i E(S^{j}_{i,1}))}{2\lambda_i}, n = 1,2,3,\ldots\right\}$$

for the jth processing step of the class i customers that are uniformly bounded from above by U, infinite buffer capacity at all stations and operating under the FPS scheduling rule. The second system we consider is a tandem system with n_1 single-server stations, that is only subject to arrivals of class 1 customers, with the service time of the nth customer at the jth station equal to

$$\bar{S}^j_{1,n}\left(\frac{\left(\sum_{i=1}^{C}\sum_{l=1}^{n_i} I\{r_i^{(l)} = r_1^{(j)}\}\lambda_i E(\bar{S}^l_{i,1})\right)}{\lambda_1 E(\bar{S}^j_{1,1})}\right), \qquad j = 1,2,\ldots,n_1;\ n = 1,2,\ldots.$$

THEOREM 2:

$$D_1^{n_1,\mathrm{FPS}}(t) \geq D_1^{n_1,\mathrm{FCFS}}(t) - Q, \quad \text{for all } t \geq 0 \tag{30}$$

where

$$Q = C\sum_{j=1}^{n_1}\left(\frac{2\lambda_1 U}{(1-\lambda_1 E(S^i_{1,1}))}\right)^j. \tag{31}$$

PROOF: The proof is obtained by one application of Theorem 1 and repeated applications of Corollary 1. Consider the first step in the processing of class 1 customers in the MONQ and consider the first station of the tandem system. For the purpose of comparing the departures of class 1 customers from their first processing at the first station in the two systems, we can verify that all the conditions of Theorem 1 are satisfied. In particular: (i) the arrival process of class 1 customers is the common external arrival process $A_1^{1,\mathrm{FPS}}(t)$; (ii) the λ_i's are equal to the arrival rates (in Theorem 1 we only needed these to be strictly positive numbers); (iii) because of (24), we may also set the value of the lower bound for the processing times of class 1 customers equal to $(1-\lambda_i E(S^1_{1,1}))/2\lambda_1$. We then obtain, using Theorem 1,

$$(D_1^{1,\mathrm{FCFS}}(t) - D_1^{1,\mathrm{FPS}}(t))^+ \leq C\left(\frac{2\lambda_1 U}{(1-\lambda_1 E(S^1_{1,1}))}\right), \quad \text{for all } t \geq 0. \tag{32}$$

For the remaining stations, we use induction on j and Corollary 1 to conclude that if

$$(D_1^{j,\mathrm{FCFS}}(t) - D_1^{j,\mathrm{FPS}}(t))^+ \leq C\sum_{l=1}^{j}\left(\frac{2\lambda_1 U}{(1-\lambda_1 E(S^i_{1,1}))}\right)^l, \quad \text{for all } t \geq 0$$

then

$$(D_1^{j+1,\mathrm{FCFS}}(t) - D_1^{j+1,\mathrm{FPS}}(t))^+ \leq C\sum_{l=1}^{j+1}\left(\frac{2\lambda_1 U}{(1-\lambda_1 E(S^i_{1,1}))}\right)^l, \quad \text{for all } t \geq 0. \quad \blacksquare$$

In the next theorem we assume that the conditions stated for Theorem 2 hold and that there exists either a limiting distribution or a time average for the total delay of class 1 customers in the MONQ. Let the average delay in queue for class 1 customers be $E(TD_1)$.

THEOREM 3:

$$E(TD_1) \le \frac{n_1\left(2n_1 + c_{\tau_1}^2 + 2\left(\sum_{j=1}^{n_1-1} c_{S_1^j}^2\right) + c_{S_1^{n_1}}^2\right)}{\lambda_1(1-\rho_{\max}^0)} + \frac{C}{\lambda_1}\left(\sum_{j=1}^{n_1}\left(\frac{2\lambda_1 U}{(1-\lambda_1 E(S_{1,1}^i))}\right)^j\right), \tag{33}$$

where $c_{\tau_1}^2$ is the squared coefficient of variation of the interarrival time and $c_{S_1^j}^2$, $j = 1,2,\ldots,n_1$ are the squared coefficient of variation of the S_1^j's, and

$$\rho_{\max}^0 = \max_{m=1,2,\ldots,d}\left\{\sum_{i=1}^{C}\sum_{j=1}^{n_i} I\{r^{(j)} = m\}\lambda_i E(S_{i,1}^j)\right\}.$$

PROOF: We first bound the mean delay in the tandem system. The maximum load in the tandem system is less than or equal to $\rho_{\max}$ because

$$\lambda_1 E(\bar{S}_{1,n}^j)\left(\frac{\left(\sum_{i=1}^{C}\sum_{l=1}^{n_i} I\{r_i^{(l)} = r_1^{(j)}\}\lambda_i E(\bar{S}_{i,1}^l)\right)}{\lambda_1 E(\bar{S}_{1,1}^j)}\right) = \sum_{i=1}^{C}\sum_{l=1}^{n_i} I\{r_i^{(l)} = r_1^{(j)}\}\lambda_i E(\bar{S}_{i,1}^l)$$

$$= \rho_{r_1^{(j)}} \le \rho_{\max}.$$

The assumptions stated for Theorem 2, and the fact that the maximum station load is less than 1 in the tandem system, allow us to use Theorem A1 given in the Appendix in order to bound the total mean delay in queue in the tandem system, denoted by $E(TD_1^{\text{tandem}})$ as follows:

$$E(TD_1^{\text{tandem}}) \le n_1\left(2n_1 + c_{\tau_1}^2 + 2\left(\sum_{j=1}^{n_1-1} c_{\bar{S}_1^j}^2\right) + c_{\bar{S}_1^{n_1}}^2\right)\Big/(2\lambda_1(1-\rho_{\max})), \tag{34}$$

where the $c_{\bar{S}_1^j}^2$'s stand for the squared coefficient of variations of the modified service times. But the squared coefficient of variations of the modified service times are smaller than those of the original service times due to the shift away from the origin (see (24)). Using this observation and (27), and accordingly modifying the bound in (34), provides the first expression on the right-hand side of the inequality in (33). The second expression on the right-hand side of (33) follows from the sample path-wise bound of Theorem 2 and the use of Little's law. ■

5. REMARKS

The following remarks are in order.

1. It is well known that the first-come, first-served (FCFS) discipline is the fairest from the perspective of an individual customer. When it comes to fairness with respect to a class of customers, however, the head-of-the-line processor sharing (HOL-PS) rule has been singled out by researchers as the rule that provides the most

equitable treatment between classes. Greenberg and Madras [20] state that this rule "provides an appealing paradigm for the fair sharing of a service." Demers et al. analyze fairness in an axiomatic setting and show that HOL-PS achieves max–min fair throughputs. The definition of max–min fairness can be found in Gafni and Bertsekas [17], and a discussion of the concept can be found in Chapter 6 of Bertsekas and Gallagher [1]. Under this definition, an allocation is considered to be fair when (i) no class receives more than what it has requested, (ii) no allocation scheme that satisfies condition (i) has a larger minimum allocation, and (iii) the conditions given in (i) and (ii) continue to hold when the minimum allocation is removed.

2. The FPS rule is intended to work in situations where the total demand of all classes does not exceed the capacity of any node. It is easy to show that the network is stable when this condition is met along with other standard conditions for the stability of a single-stage queueing system. Stability of a system in this discussion is defined to be the arrival rates being equal to the departure rates for each class of customers. This definition is equivalent to rate stability (see Lemma 5.58) in Stidham and El-Taha [34]. For example, Buzacott and Shanthikumar [6] give three conditions for the stability of a single-stage system: (a) the number of customers that arrive by time t, $A(t)$, has a uniform limit, $\lambda = \lim_{t\to\infty} A(t)/t$; (b) the service times, S_n, have a uniform limit $1/\mu = \lim_{k\to\infty} (\sum_{j=1}^{k} S_j)/k$; and (c) $\lambda < \mu$.

3. For stable systems, FPS automatically achieves max–min fair throughputs.

4. Let the longest route in the MONQ have $n^{\max}$ steps. If service times are either bounded or such that for each i, $\sup_{i,k} S_{i,k}$ grows slower than $k^{1/n^{\max}}$, then from equation (15), Theorems 1 and 2, and Corollary 1, the stability of the network follows from the stability of each stage of the slowed down FCFS system.

5. An open question is to compare the performance of the FPS system with that of a weighted fair queueing or generalized processor sharing rule. The FPS rule as defined in this section does not fully mimic HOL-PS. The main reason is that the definition of the work that should be done on each class is predicated on setting aside capacity for each class. When that class of customers is absent, however, HOL-PS dedicates the extra available capacity to serving other classes that might be present at that time. To make FPS more compatible with HOL-PS (or GPS), it is necessary to redefine the "share" of service, θ_i. One possible redefinition is to set, $\theta_i(t) = \theta_i/(\sum_{j\in P(t)} \theta_j)$. The shortfall in work can then be expressed as

$$H_i(t) = \int_{l_i(t)}^{t} \theta_i(t)\, dt - V_i(l_i(t), t).$$

The condition for stability becomes $\lim_{T\to\infty} \int_0^T \theta_i(t)\, dt/T > \rho_i$, $i = 1,2,\ldots,C$. It can be shown that Theorem 1 continues to hold with this definition. Moreover, if all classes of customers are almost always present, then SSBP rules perform similar to the corresponding GPS rules.

Acknowledgments

We thank Professor J.G. Shanthikumar, University of California at Berkeley, for suggesting the problem of determining bounds for the mean delay in MONQ, and the late Professor Laurence Baxter, New York

State University, for several suggestions regarding the exposition and presentation of the results. We also thank an anonymous referee for several helpful suggestions.

References

1. Bertsekas, D. & Gallagher, R.G. (1992). *Data networks.* Englewood Cliffs, NJ: Prentice Hall.
2. Bramson, M. (1994). Instability of FIFO queueing networks. *Annals of Applied Probability* 4: 414–431.
3. Bramson, M. (1994). Instability of FIFO queueing networks with quick service times. *Annals of Applied Probability* 4: 693–718.
4. Bramson, M. (1996). Convergence to equilibria for fluid models of head-of-the-line proportional processor sharing queueing networks. *Queueing Systems* 23: 1–26.
5. Bramson, M. (1996). Convergence to equilibria for fluid models of FIFO queueing networks. *Queueing Networks* 22: 5–45.
6. Buzacott, J.A. & Shanthikumar, J.G. (1993). *Stochastic models of manufacturing systems.* Englewood Cliffs, NJ: Prentice Hall.
7. Chang, C.-S. (1994). Stability, queue length, and delay of deterministic and stochastic queueing networks. *IEEE Transactions on Automatic Control* 39(5): 913–931.
8. Chang, C.S., Thomas, J.A., & Kiang, S.H. (1994). On the stability of open networks: A unified approach by stochastic dominance. *Queueing Systems* 15: 239–260.
9. Chen, H. & Yao, D.D. (1994). Stable priority disciplines for multiclass networks. In P. Glasserman, K. Sigman, & D.D. Yao (eds.), *Stochastic networks: Stability and rare events.* New York: Springer.
10. Cobham, A. (1954). Priority assignment in waiting line problems. *Operations Research* 2: 70–76.
11. Cruz, R.L. (1991). A calculus for network delay, Part I: Network elements in isolation. *IEEE Transactions on Information Theory* 37(1): 114–131.
12. Cruz, R.L. (1991). A calculus for network delay, Part II: Network analysis. *IEEE Transactions on Information Theory* 37(1): 132–141.
13. Cruz, R.L. & Liu, H. (1993). Single-server queues with loss: A formulation. *Proceedings 1993 CISS, Johns Hopkins Univ.*
14. Dai, J.G. (1994). On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Annals of Applied Probability* 5: 49–77.
15. Dai, J.G. & Wang, Y. (1993). Nonexistence of Brownian models of certain multiclass queueing networks. *Queueing Systems* 13: 41–46.
16. Demers, A., Keshav, S., & Shenker, S. (1990). Analysis and simulation of a fair queueing algorithm. *Internetworking: Research and Experience* 1.
17. Gafni, E. & Bertsekas, D. (1984). Dynamic control of session input rates in communication networks. *IEEE Transactions* 29(10): 1009–1016.
18. Georgadis, L., Guerin, R., & Parekh, A. Optimal multiplexing on a single link: Delay and buffer requirements. [Unpublished manuscript], Yorktown Heights, NY: IBM TJ Watson Research Center.
19. Georgadis, L., Guerin, R., Peris, V., & Sivarajan, K.N. Efficient network provisioning based on per node traffic shaping. [Unpublished manuscript], Yorktown Heights, NY: IBM TJ Watson Research Center.
20. Greenberg, A. & Madras, N. (1992). How fair is fair queueing? *Journal of the Association for Computing Machinery* 39(3): 568–598.
21. Harrison, J.M. (1973). The heavy traffic approximation for single-server queues in series. *Journal of Applied Probability* 10: 613–629.
22. Harrison, J.M. & Nguyen, V. (1993). Brownian models of multiclass queueing networks: Current status and open problems. *Queueing Systems: Theory and Applications* 13: 5–40.
23. Kelly, F.P. (1979). *Reversibility and stochastic networks.* New York: John Wiley.
24. Kingman, J.F.C. (1962). Some inequalities for the queue GI/G/1. *Biometrika* 49: 315–324.
25. Kleinrock, L. (1976). *Queueing systems.* Vol. 2: *Computer applications.* New York: John Wiley.

26. Kumar, P.R. & Seidman, T.I. (1990). Dynamic instabilities and stabilization methods in distributed real-time scheduling of manufacturing systems. *IEEE Transactions on Automatic Control* 35: 289–298.
27. Lu, S.H. & Kumar, P.R. (1991). Distributed scheduling based on due date and buffer priorities. *IEEE Transactions on Automatic Control* 36: 1406–1416.
28. Parekh, A.J. (1992). A generalized processor sharing approach to flow control in integrated service networks. Ph.D. thesis, MIT, Boston, MA.
29. Parekh, A.J. & Gallagher, R.G. (1993a). A generalized processor sharing approach to flow control in integrated service networks: The single node case. *IEEE/ACM Transactions on Networking* 1(1): 344–357.
30. Parekh, A.J. & Gallagher, R.G. (1993b). A generalized processor sharing approach to flow control in integrated service networks: The multiple node case. *Proceedings IEEE/INFOCOM '93: 521–530.*
31. Perkins, J.R. & Kumar, P.R. (1989). Stable distributed real-time scheduling of manufacturing/assembly/disassembly systems. *IEEE Transactions on Automatic Control* AC-34: 139–148.
32. Seidman, T.I. (1993). First come, first served is unstable! [Manuscript], University of Maryland, Baltimore County.
33. Shanthikumar, J.G. & Yao, D.D. (1989). Stochastic monotonicity in general queueing networks. *Journal of Applied Probability* 26: 413–417.
34. Stidham, S. & El-Taha, M. (1995). Sample-path techniques in queueing theory. In J.H. Dshalalow (ed.), *Advances in queueing: Theory, methods, and open problems.* Boca Raton, FL: CRC Press.
35. Wein, L.M. (1990). Optimal control of a two-station Brownian network. *Mathematics of Operations Research* 15(2): 215–242.
36. Whitt, W. (1993). Large fluctuations in a deterministic multiclass network of queues. *Management Science* 39: 1020–1028.
37. Wolff, R.W. (1989). *Stochastic modeling and the theory of queues.* Englewood Cliffs, NJ: Prentice Hall.
38. Yaron, O. & Sidi, M. (1993). Performance of stability of communication networks via robust exponential bounds. *IEEE/ACM Transactions on Networking* 1(3): 372–385.
39. Zhang, Z.L., Towsley, D., & Kurose, J. (1995). Statistical analysis of the generalized processor sharing discipline. *IEEE Journal on Selected Areas of Communications* 13: 1071–1080.

APPENDIX
Bounds on the Delay in Tandem Queues with Single Servers

In this appendix, we establish upper and lower bounds on the mean delay in a tandem queueing system. Consider a tandem queueing system in which there are K single-server stations and infinite waiting room at all stations. All customers are of the same class and arrive at station 1. After being served in a first-come, first-served (FCFS) order at station $i, i = 1,2,\ldots,K-1$, customers proceed to station $i+1$. After being served at station K, customers leave the system. Define the following quantities for the tandem system:

- Let S_n^k denote the service time of the nth customer served at station k. The service times at station k are i.i.d.; the service time distribution has a finite second moment and a squared coefficient of variation, denoted as c_k.
- Let $\{T_n, n = 1,2,3,\ldots\}$ denote the sequence of i.i.d. interarrival times of customers at the first station, where T_n is the time between the arrivals of the nth and the $(n+1)$st customer. Let $E(T_1) = 1/\lambda$. The interarrival time distribution has a finite second moment, and a squared coefficient of variation, given by c_a^2.

- Let $\rho_k = \lambda E(S_1^k)$, $k = 1,2,\ldots,K$, denote the utilization of station $k, k = 1,2,\ldots,K$. We assume that $\rho_k < 1$, for all k.
- Let W_n^k and D_n^k denote the time spent at the kth station and the delay in queue at the kth station as experienced by the nth customer. Set $W_0^k = D_0^k = 0$.
- Let

$$TS_n^k = \sum_{i=1}^{n} S_i^k;\ TS_0^k = 0, \qquad n = 1,2,\ldots;\ k = 1,2,\ldots,K,$$

$$TT_n = \sum_{i=1}^{n} T_i;\ TT_0 = 0, \qquad n = 1,2,\ldots;\ k = 1,2,\ldots,K,$$

$$B_n^k = TS_n^k - TT_n, \qquad n = 1,2,\ldots;\ k = 1,2,\ldots,K,$$

$$TW_n^k = \sum_{i=1}^{k} W_n^i;\ TW_0^k = 0, \qquad n = 1,2,\ldots;\ k = 1,2,\ldots,K,$$

$$TD_n^k = \sum_{i=1}^{k} D_n^i;\ TD_0^k = 0, \qquad n = 1,2,\ldots;\ k = 1,2,\ldots,K.$$

In these definitions, TS_n^k is the sum of the first n service times at station k, TT_n is the sum of the first n interarrival times to the first station, TW_n^k is the total time spent by the nth customer at the first k stations, and TD_n^k is the total delay experienced by the nth customer at the first k stations. We are interested in TD_n^K, the total mean delay in queue in the system. It is shown by Harrison [21] that

$$TW_n^k = \max_{0 \le i_0 \le i_1 \le \cdots \le i_k = n-1} ((B_{i_k}^k - B_{i_{k-1}}^k) + (B_{i_{k-1}}^{k-1} - B_{i_{k-2}}^{k-1}) + \cdots + (B_{i_1}^1 - B_{i_0}^1)) + S_n^k, \quad (1)$$

where the quantity B_n^k represents the position of a random walk with increments $S_n^k - T_n$. Furthermore, if the station loads are strictly less than one, then as $n \to \infty$, the random vector $(TW_n^1, TW_n^2, \ldots, TW_n^K)$ converges in distribution to a nondefective random vector (see Thm. 1, Harrison [21]).

From Eq. (1), a lower bound for the delay in queue in the system can be obtained. Consider the delay in queue in a modified system in which the service times at all stations except station k have been set to zero. Denote the delay in queue of the nth customer in this modified system by $\hat{D}_n^k$. We assume that the sequence of interarrival times, $T_n, n = 1,2,\ldots$, and the sequence of service times, $S_n^k, n = 1,2,\ldots$, are the same in the original as well as the modified system. Then using (1) it can be shown that $TD_n^m \ge \hat{D}_n^k$, for all $K \ge m \ge k$. Thus we have

$$E(TD_n^K) \ge \max_{1 \le k \le K} E(\hat{D}_n^k). \quad (2)$$

We now derive an upper bound on the total delay in the tandem system. The relation for the delay experienced by the $(n+1)$st customer at station k can be written as

$$D_{n+1}^k = \max_{i_{k-1} \le n} (TW_{i_{k-1}}^{k-1} - TW_{n+1}^{k-1} + B_n^k - B_{i_{k-1}}^k)$$

$$= (D_n^k + S_n^k - (TW_{n+1}^{k-1} + T_n - TW_n^{k-1}))^+, \quad (3)$$

where $(TW_{n+1}^{k-1} + T_n - TW_n^{k-1})$ is the interarrival time between the nth and the $(n+1)$st customers to the kth station (to see this, set time to be zero when the nth customer arrives, then the nth customer reaches station k at time TW_n^{k-1}, and the $(n+1)$st customer reaches station k at time $TW_{n+1}^{k-1} + T_n$). Rewritten in this form, (3) is the standard Lindley recursion for the delay in a single-server queue (e.g., see Wolff [37]).

The $(n+1)$st customer has to wait for a time at least equal to $(TW_n^{k-1} - T_n)$, i.e., the total time spent by the nth customer at the first $(k-1)$ stations less the interarrival time between the nth and the $(n+1)$st customer, before being served at the $(k-1)$st station. The service time at the $(k-1)$st station when added to $(TW_n^{k-1} - T_n)$ gives a lower bound for the flow time TW_{n+1}^{k-1}:

$$TW_{n+1}^{k-1} \geq TW_n^{k-1} - T_n + S_{n+1}^{k-1}. \tag{4}$$

We shall now construct a modified system to obtain an upper bound for the mean delay in the system. Denote all modified quantities in this new system by a bar over the corresponding quantities in the original system. Let

$$\rho_{\max} = \max_{1 \leq k \leq K} (E(S_1^k)/E(T_1)).$$

In the modified system, retain the sequence of interarrival times for the original system, but scale the service times at station i with constants $a_k, k = 1,2,\ldots,K$, such that

$$a_k E(S_1^k)/E(T_1) = \rho_{\max} + \frac{(K-k)}{K}(1 - \rho_{\max}). \tag{5}$$

In other words, the new service times are $\bar{S}_n^k = a_k S_n^k, n = 1,2,\ldots; k = 1,2,\ldots,K$. Note that all the a_k's are greater than or equal to one. We also see that if $\rho_{\max} < 1$, then

$$\rho_{\max} + \frac{(K-1)}{K}(1 - \rho_{\max}) = (K-1)/K + \rho_{\max}/K < 1. \tag{6}$$

Identities (5) and (6) show that even with the scaling of the service times, the largest station load in the modified system does not exceed unity if the largest station load without the scaling was less than one. This, in turn, implies that under the assumption $\rho_{\max} < 1$, there exists a limiting distribution for the delay in queue at each station in this modified system (see Harrison [21]). It can be shown, using a sample path construction as given in Shanthikumar and Yao [33], that the total mean delay in queue in the modified system is larger than that in the original system. While, the device of scaling the service times in a queueing system to demonstrate stability has been used previously by other authors, for example, see Chang et al. [8], our construction and the choice of scaling factors are new. We observe that by equation (5), for $k = 2,3,\ldots,K$,

$$\begin{aligned} E(\bar{S}_n^k - \bar{S}_{n+1}^{k-1})/E(T_1) &= a_k \rho_k - a_{k-1}\rho_{k-1} \\ &= \rho_{\max} + \frac{(K-k)}{K}(1 - \rho_{\max}) - \left(\rho_{\max} + \frac{(K-k+1)}{K}(1 - \rho_{\max})\right) \\ &= -(1 - \rho_{\max})/K, \end{aligned} \tag{7a}$$

and

$$E(\bar{S}_n^1 - T_{n+1})/E(T_1) = \rho_{\max} + \frac{(K-1)}{K}(1-\rho_{\max}) - 1 = -(1-\rho_{\max})/K. \tag{7b}$$

Using relations (3) and (4) for the modified system (and bars as noted before when referring to the modified system), we obtain that

$$\bar{D}_{n+1}^k = (\bar{D}_n^k + \overline{TW}_n^{k-1} - \overline{TW}_{n+1}^{k-1} + \bar{S}_n^k - T_n)^+ \le |\bar{D}_n^k + \bar{S}_n^k - \bar{S}_{n+1}^{k-1}|, \qquad k = 2,3,\ldots,K, \tag{8}$$

$$\bar{D}_{n+1}^1 = (\bar{D}_n^1 + \overline{TW}_n^0 - \overline{TW}_{n+1}^0 + \bar{S}_n^1 - T_n)^+ \le |\bar{D}_n^1 + \bar{S}_n^1 - T_{n+1}|, \tag{9}$$

$$\bar{D}_{n+1}^0 = 0. \tag{10}$$

The delay in queue at the kth station, $\bar{D}_n^k$, is independent of $(\bar{S}_n^k - \bar{S}_{n+1}^{k-1})$, for $k = 2,3,\ldots K$. Using this fact, squaring both sides of (8), dividing by $E(T_1)$, using the identities (7), and rearranging, we obtain

$$\frac{(E[(\bar{D}_{n+1}^k)^2] - E[(\bar{D}_n^k)^2])}{E[T_1]} + \frac{2E[\bar{D}_n^k](1-\rho_{\max})}{K} \le \frac{E|\bar{S}_n^k - \bar{S}_{n+1}^{k-1}|^2}{E[T_1]}, \qquad k = 2,3,\ldots,K. \tag{11}$$

Similarly, as $\bar{D}_n^1$ is independent of $(\bar{S}_n^1 - T_{n+1})$, we obtain

$$\frac{(E[(\bar{D}_{n+1}^1)^2] - E[(\bar{D}_n^1)^2])}{E[T_1]} + \frac{2E[\bar{D}_n^1](1-\rho_{\max})}{K} \le \frac{E|\bar{S}_n^1 - T_{n+1}|^2}{E[T_1]}. \tag{12}$$

THEOREM A1: *There exist upper and lower bounds, $U(\rho_{\max})$ and $L(\rho_{\max})$ for the mean number of customers in a tandem queueing system, $E[N_T]$, that are functions of the first and second moments of the service times and the interarrival time. The values of the upper and lower bounds can be chosen as*

$$U(\rho_{\max}) = \frac{K(2K + c_a^2 + 2(c_1^2 + c_2^2 + \cdots + c_{K-1}^2) + c_K^2)}{2(1-\rho_{\max})} + \sum_{k=1}^{K} \rho_k \ge E[N_T] \ge L(\rho_{\max})$$

$$= \frac{\lambda^2 E((s_1^{\psi} - t_1)^+)^2}{2(1-\rho_{\max})} + \sum_{k=1}^{K} \rho_k,$$

where $\psi = \arg\max_{k=1,2,\ldots,K} \rho_k$, and $1/\lambda$ is the mean interarrival time.

PROOF: Denote the stationary total mean delay in queue in the original system as $E(TD^K)$ and the stationary mean delay at the kth station in the original and the modified system as $E(D^k)$ and $E(\bar{D}^k)$. From inequality (2) and equation (22) of Wolff [37, p. 478], we obtain,

$$E(TD^K) \ge \lambda \max_k E((S_1^k - T_1)^+)^2/(2(1-\rho_k)) \ge \lambda E((S_1^{\psi} - T_1)^+)^2/(2(1-\rho_{\max})). \tag{13}$$

The lower bound for the total number of customers in the system now follows from adding the mean service times at all stations to the lower bound for delay in (13) and using Little's law,

$$E(N_T) = \lambda\left(E(TD^K) + \sum_{k=1}^{K} E(S_1^k)\right) \geq \lambda^2 E((S_1^{\psi} - T_1)^+)^2/(2(1-\rho_{\max})) + \sum_{k=1}^{K} \rho^k.$$

Combining inequalities (11) and (12), and using the fact that the delays in queue at each station converge to a stationary distribution, we obtain

$$E(D^k) \leq E(\bar{D}^k) \leq \lambda K\left(\frac{E(\bar{S}_1^k)^2 + E(\bar{S}_1^{k-1})^2}{2(1-\rho_{\max})}\right), \qquad k = 2,3,\ldots,K, \tag{14}$$

$$E(D^1) \leq E(\bar{D}^1) \leq \lambda K\left(\frac{E(\bar{S}_1^1)^2 + E(T_1^2)}{2(1-\rho_{\max})}\right). \tag{15}$$

From Eq. (5), $E[(\bar{S}_1^k)^2] = a_k^2 E[(S_1^k)^2]$

$$= \left(\left(\rho_{\max} + \frac{(K-k)}{K}(1-\rho_{\max})\right)^2 \frac{E(T_1)^2}{E(S_1^k)^2}\right) E[(S_1^k)^2] \leq \frac{1}{\lambda^2}(1+c_k^2). \tag{16}$$

Using (14), (15), and (16), we obtain

$$E(TD^K) = \sum_{k=1}^{K} E(D^k) \leq \frac{K}{\lambda}\left(\frac{2K + c_a^2 + 2(c_1^2 + c_2^2 + \cdots + c_{K-1}^2) + c_K^2}{2(1-\rho_{\max})}\right). \tag{17}$$

Using (17) and Little's law

$$E(TN) = \lambda\left(E(TD^K) + \sum_{k=1}^{K} E(S_1^k)\right)$$

$$\leq K\left(\frac{2K + c_a^2 + 2(c_1^2 + c_2^2 + \cdots + c_{K-1}^2) + c_K^2}{2(1-\rho_{\max})}\right) + \sum_{k=1}^{K} \rho_k. \qquad \blacksquare$$

Remark: The delay bounds are tight if the utilization of the stations is decreasing and the interarrival times have the DMRL (decreasing mean residual lifetime) property. In particular, if the service times are equal (or nonincreasing) and deterministic at all stations, then the difference between the upper and lower bounds for the average number of customers in the system is less than one.

These bounds are similar to Kingman's [24] bounds for the delay in the GI/G/1 queue.