Probability in the Engineering and Informational Sciences, **18**, 2004, 445–472. Printed in the U.S.A.

JOIN MINIMUM COST QUEUE FOR MULTICLASS CUSTOMERS

STABILITY AND PERFORMANCE BOUNDS

Rahul Tandra
Department of EECS
University of California
Berkeley, CA
E-mail: tandra@eecs.berkeley.edu

N. Hemachandra
IE and OR Interdisciplinary Programme
Indian Institute of Technology, Bombay
Powai Mumbai, 400 076 India
E-mail: nh@iitb.ac.in

D. Manjunath
Department of Electrical Engineering
Indian Institute of Technology, Bombay
Powai Mumbai, 400 076 India
E-mail: dmanju@ee.iitb.ac.in

We consider a system of K parallel queues providing different grades of service through each of the queues and serving a multiclass customer population. Service differentiation is achieved by specifying different join prices to the queues. Customers of class j define a cost function $\psi_{ij}(c_i, x_i)$ for taking service from queue i when the join price for queue i is c_i and congestion in queue i is x_i and join the queue that minimizes $\psi_{ij}(\cdot,\cdot)$. Such a queuing system will be called the “join minimum cost queue” (JMCQ) and is a generalization of the join shortest queue (JSQ) system. Non-work-conserving (called Paris Metro pricing system) and work-conserving (called the Tirupati system) versions of the JMCQ are analyzed when the cost to an arrival of joining a queue is a convex combination of the join price for that queue and the expected waiting time in that queue at the arrival epoch. Our main results are for a two-queue system.

0269-9648/04 $16.00

https://doi.org/10.1017/S0269964804184027 Published online by Cambridge University Press

We obtain stability conditions and performance bounds. To obtain the lower and upper performance bounds, we propose two quasi-birth–death (QBD) processes that are derived from the original systems by suitably truncating the state space. The state space truncation in the non-work-conserving JMCQ follows the method of van Houtum and colleagues. We then show that this method is not applicable to the work-conserving JMCQ and provide sample-path-based proofs to show that the number in each queue is bounded by the number in the corresponding queues of these QBD processes. These sample-path proof techniques might also be of independent interest. We then show that the performance measures like mean queue length and revenue rate of the system are also bounded by the corresponding quantities of these QBD processes. Numerical examples show that these bounds are fairly tight. Finally, we generalize some of these results to systems with more queues.

1. INTRODUCTION

We consider a system of K parallel queues providing different grades of service in each of the queues to a multiclass customer population with J classes. Service differentiation is achieved by different join prices for the queues and service rates in them. A join price c_i is prescribed for service from queue i. Customers also incur a congestion cost due to, say, delays. These two costs are reflected in the customers of class j defining a cost function $\psi_{ij}(c_i, x_i)$ for service in queue i when the join price is c_i and congestion is x_i. Obviously, $\psi_{ij}(c_i, x_i)$ should be increasing in c_i and x_i. Let $\mathbf{c} = [c_1, \ldots, c_K]^T$ be the price vector and $\mathbf{x}(t) = [x_1(t), \ldots, x_K(t)]^T$ be the queue length vector at time t. The queue system posts both $\mathbf{c}$ and $\mathbf{x}(t)$. A customer of class j arriving at time t calculates its cost for service from queue i for $i = 1, \ldots, K$ and joins that queue for which the cost is minimum. This queuing system will be called the "join minimum cost queue" (JMCQ). A customer class is determined by the set $\{\psi_{1j}(\cdot,\cdot), \ldots, \psi_{Kj}(\cdot,\cdot)\}$. Thus, the JMCQ is a generalization of the well-known join shortest queue (JSQ) system.

An important motivation for this problem is to price quality of service in the Internet through an access charge. Specifically, we target the multiclass service system as defined by the DiffServ model of the IETF of Blake et al. [2]. In DiffServ, the per hop behaviors are implemented by means of appropriate scheduling mechanisms and users select the service class that best fits its requirements of quality. The JMCQ system described above can be seen to be amenable to this service as follows. A set of K service classes is defined. The total link capacity μ is divided among the K queues such that queue i is serviced at rate μ_i, $\mu_i > 0$ with $\sum_{i=1}^{K} \mu_i = \mu$. The price for service and the congestion in queue i (c_i and x_i, respectively) are posted. The instantaneous queue length or the unfinished work in the queues are examples of congestion information. In the extreme, each packet can calculate its cost for service from each class and take service from the queue that minimizes the cost.

The applicability of a multiclass service system to pricing quality of service in the Internet has been recognized for quite some time now. For example, Odlyzko

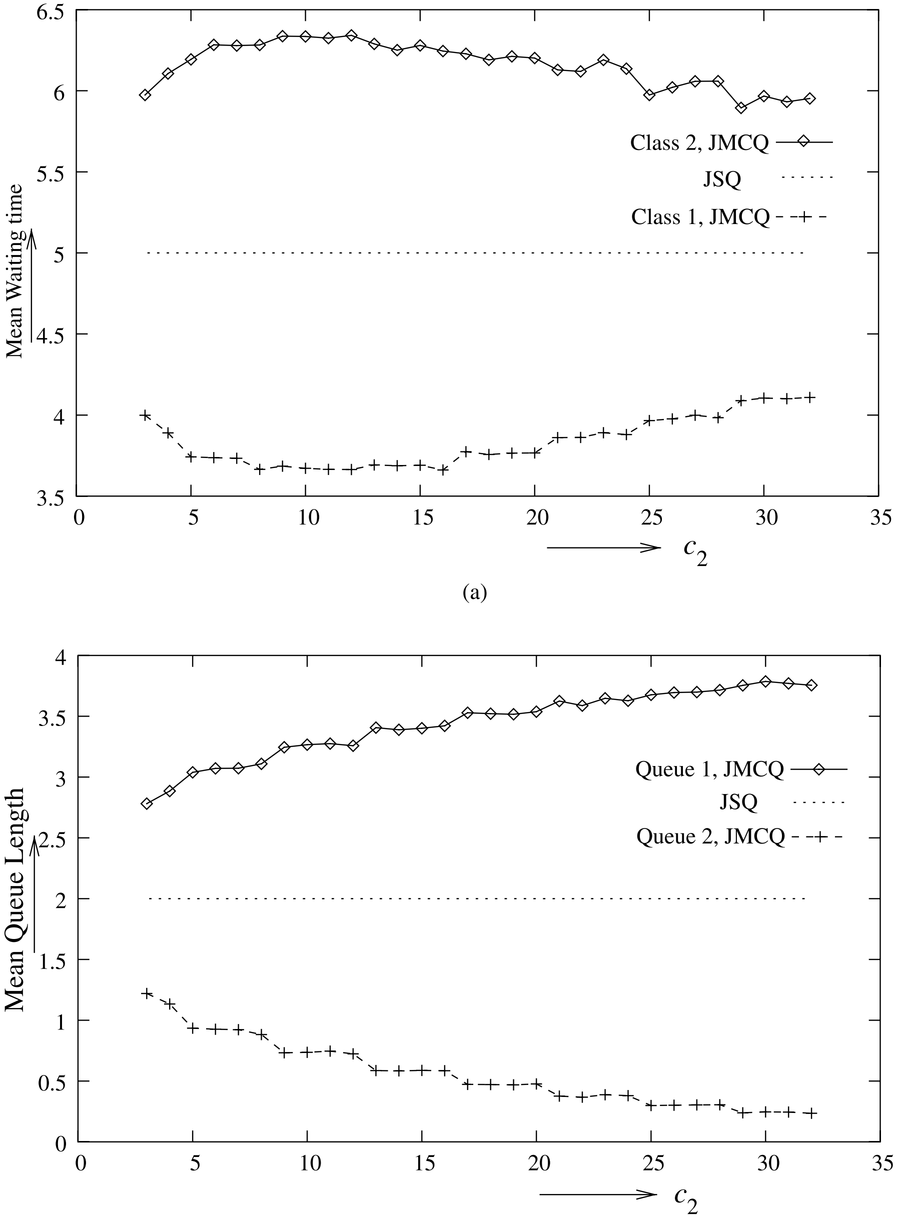
https://doi.org/10.1017/S0269964804184027 Published online by Cambridge University Press

[16] argues that the pricing scheme in the Paris Metro could be extended to pricing differential quality in the Internet. The network is partitioned into multiple logical networks with identical resources and the service in each partition is priced differently. If occupancy information in each partition is provided, prices would act as a control to provide differential service. This pricing model is called the Paris Metro pricing (PMP) scheme. Jain, Mullen, and Hausman [10] report an analysis to model the profitability of this pricing scheme. Gibbens, Mason, and Steinberg [9] describe more results. The PMP, as it is proposed, is a non-work-conserving scheme with the link capacity statically partitioned among the service grades. In Dube, Borker, and Manjunath [6], a work-conserving version of PMP called the *Tirupati scheme* is proposed. In this scheme, if there are no customers in a queue, the capacity allocated to it is distributed among the nonempty queues. This scheme takes inspiration from the queue management scheme in Tirupati, a major pilgrimage center in southern India, where it has been operating with remarkable efficiency for quite some time now. Dube et al. [6] analyzed the social optimality of the Tirupati pricing model and showed that the difference between the social cost of the optimally priced system and that of the Tirupati system is $C\epsilon$ for some constants C and ϵ. Another, more practical, contribution of Dube et al. [6] was dynamic pricing using a dynamic programming equation and a reinforcement-learning-based online pricing algorithm. A simple learning-scheme-based pricing mechanism to dynamically determine the join prices to provide a specified average grade of service from each queue is analyzed and described in Borkar and Manjunath [3]. A preliminary comparative analysis of the Tirupati and PMP queuing systems is presented in Manjunath, Goel, and Hemachandra [12], where it is shown numerically that the revenue rate is neither monotonic in nor a convex function of the prices. Refer to Falkner, Devetsikiotis, and Lambadaris [7] for a recent survey of Internet pricing.

Another application of the JMCQ is in pricing service at popular websites. There are a number of websites that now offer a faster service for a charge. The service offering is the same for the free and the priced version and the user pays for faster access.

We now show by way of a numerical example how pricing can selectively improve the grade of service of specific classes in a multiclass environment. Consider customers that use a convex combination of the join price and the expected waiting time as the cost [i.e., $\psi_i(c,x) = (1 - a_i)x + a_i c$]. Consider a work-conserving queue with two classes of customers with $a_1 = 0.8$ for a delay-sensitive class and $a_2 = 0.3$ for a price-sensitive class. Let the arrival rate and mean service time of both classes be the same. In a work-conserving JSQ system, both classes would get the same grade of service. To provide a better grade of service to the delay-sensitive class, let one of the queues prescribe a join price, say p. For an arrival rate of 0.4 for both classes of customers and a service rate of 0.5 in both the queues, Figure 1a shows the mean waiting time for each class and Figure 1b shows the mean queue lengths. Observe the significant decrease in mean delays for the delay-sensitive customers. See [19] for more detailed numerical results.

https://doi.org/10.1017/S0269964804184027 Published online by Cambridge University Press



(a)

(b)

FIGURE 1. Plots illustrating the differentiated service provided by JMCQ. (a) Mean waiting time of customers of different classes for the work-conserving JMCQ system compared with that of the JSQ system. (b) Mean queue lengths for the work-conserving JMCQ compared with the mean queue length in the JSQ system.

https://doi.org/10.1017/S0269964804184027 Published online by Cambridge University Press

In this article, we analyze the JMCQ applied to one link of a network or to a web service under a static pricing regime. After describing the model assumptions and notations in the next section, we first derive stability conditions for the queues under both the work-conserving Tirupati JMCQ and the non-work-conserving PMP JMCQ in Section 3. In addition to the usual performance measures of moments of the delay and queue length, the revenue rate for the queue system and the social cost are important measures. JMCQ is a generalization of JSQ and one can expect that it will be hard to obtain exact results. We focus on computable bounds for the performance measures.

There is considerable literature on JSQ. Boxma, Koole, and Liu [4] presented a recent survey of the results for JSQ. van Houtum, Zijm, Adan, and Wessels [20] gave a methodology for obtaining computable bounds for performance measures of JSQ that are related to mean rewards of the associated Markov chain. van Houtum, Adan, Wessels, and Zijm [21] considered a generalization of JSQ, where jobs of a class join the shortest of the queue that is capable of serving it. They proposed computable bounds for useful performance measures. For asymptotic results, Foley and McDonald [8] presented a recent example. In Section 4, we derive computable bounds for the performance measures mentioned above. For the non-work-conserving PMP model, we show how the methodology of [20] can be adopted for obtaining computable bounds for revenue rate and stationary mean number in each queue. We next observe that this methodology is not suited for the work-conserving Tirupati model. Our main result here is to show that the state space can be truncated in such a way as to form a quasi-birth–death (QBD) process in which the number in each component of the QBD processes gives bounds for the number in the queues of the JMCQ model. We show this by sample-path arguments and see that these bounds can be made fairly tight. We provide two numerical examples in Section 5 and conclude with discussions on generalization in Section 6.

2. MODEL DESCRIPTION

Without loss of generality, we let the queue join prices be ordered such that $c_1 \leq c_2 \leq \cdots \leq c_K$. Customers of class j arrive according to a Poisson process of rate λ_j and these arrival streams are independent. Denote $\sum_i^J \lambda_i = \lambda$, so that λ is the total customer arrival rate into the system. An arrival at time t selects the queue to join as described in the previous section. Ties in cost are awarded to the queue with the lowest join price.

The service requirements are assumed identical among the classes. This assumption is not unreasonable in modeling web service, in the urban transport system, in the Paris Metro system, or in the queue at Tirupati. It is especially not an unreasonable assumption in the context of Internet bandwidth, where a fixed access charge is the most prevalent charging mechanism.

The service times are independent and identically distributed (i.i.d.) exponential with unit mean. The total service capacity is μ and is partitioned among the queues as follows. The non-work-conserving system has static partitioning and queue i is served at rate μ_i, $\sum_{i=1}^{K} \mu_i = \mu$. The work-conserving system uses the generalized processor sharing model of Parekh and Gallager [17] with weight w_i for queue i; that is, at time t, queue i receives service at rate $\mu_i(t)$, where

$$\mu_i(t) = \frac{w_i \mu}{\sum_{j:\text{ Queue } j \text{ is nonempty at time } t} w_j}.$$

The cost of service from queue i for a class j customer, $\psi_{ij}(\cdot,\cdot)$, will be assumed to be a convex combination of the queue length and join price of the ith queue. (Because the service times are exponential, the expected waiting time from queue i is equal to the instantaneous queue length at the arrival epoch.) This is a simple and effective way of capturing price and delay sensitivities for different classes of customers. Thus, in the following, $\psi_{ij}(\cdot,\cdot)$ will be of the form

$$\psi_{ij}(c_i, x_i) = (1 - a_{ij})c_i + a_{ij} x_i, \qquad a_{ij} \in (0,1).$$

The a_{ij} will be called the delay sensitivity of class i with respect to queue j. We immediately simplify the model by letting $a_{ij} = a_i$ for $j = 1, \dots, J$; that is, the delay sensitivities are independent of the queue. Thus, the $\psi_{ij}(\cdot,\cdot)$ will be given by

$$\psi_{ij}(c_i, x_i) = (1 - a_j)c_i + a_j x_i.$$

This is a reasonable assumption when, for example, the service rate is the same in both queues.

In much of the rest of the article, we will consider a system with two queues, $K = 2$, and two customer classes, $J = 2$. Without loss of generality, we assume $a_1 > a_2$; that is, class 1 customers are more delay sensitive than the class 2 customers, who can be called price sensitive. We indicate generalizations to systems with $K > 2$ and $J > 2$ in Section 6.

We now introduce the concept of an attractor line which will help in a better understanding of the system. First, consider the system with only one class of customers, say class 1. Recall that we let $c_1 < c_2$. On the x_1–x_2 plane, an "attractor" line can be defined such that an arrival to the system when it is in a state on the left of this line will join queue 1. Similarly, an arrival to the system when it is in a state on the right of the attractor line will join queue 2; that is, arrivals tend to move the system toward the attractor line. For $\psi_j(\cdot)$ as above, the attractor line is defined by

$$x_2 = x_1 - \left(\frac{1 - a_1}{a_1}\right)(c_2 - c_1).$$

In a JMCQ system supporting multiclass traffic, an attractor line is defined for each class.

3. STABILITY ANALYSIS

We consider the stability of the queuing systems in terms of the ergodicity of the associated Markov chains.

https://doi.org/10.1017/S0269964804184027 Published online by Cambridge University Press

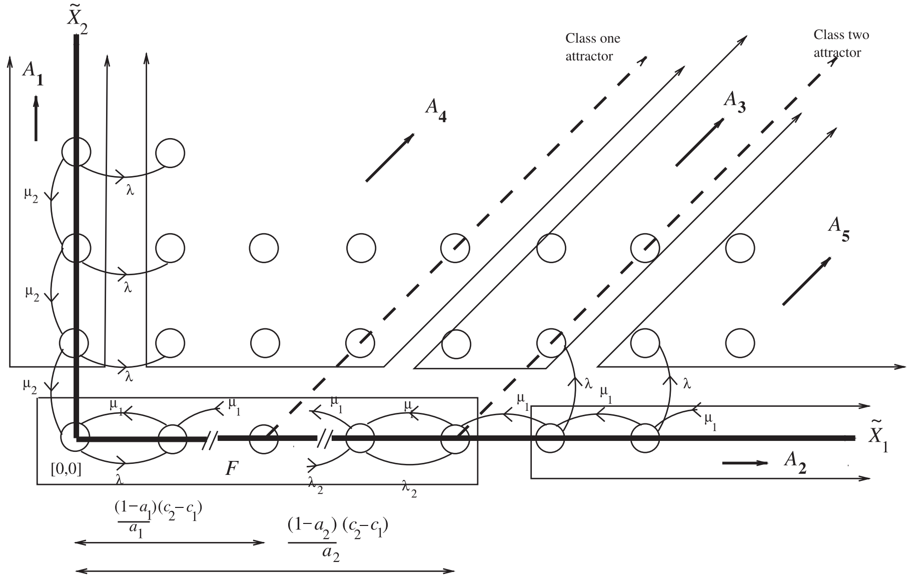


FIGURE 2. The transition rates for $\{\mathbf{x}(t)\}_{t\geq 0}$, the CTMC for the non-work-conserving queue system. Partitions $A_1,\ldots,A_5$ and F used in the stability analysis of Theorem 1 are also shown.

3.1. Non-Work-Conserving PMP System

Let $\mathbf{x}(t) = [x_1(t), x_2(t)]^T$ be the state of the system at time t, where $x_i(t)$ is the number of customers in queue i at time t in the non-work-conserving JMCQ. $\{\mathbf{x}(t)\}_{t\geq 0}$ evolves as an irreducible continuous-time Markov chain (CTMC) over the state space $\mathcal{Z}_+^2$. The transition rates for $\{\mathbf{x}(t)\}_{t\geq 0}$ are as shown in Figure 2. Let $\{\mathbf{x}_n\}_{n\geq 0}$, n an integer, be the jump chain (see, e.g., Asmussen [1] or Norris [16]) derived from $\{\mathbf{x}(t)\}_{t\geq 0}$ and let $\{p_{\mathbf{x}:\mathbf{x}'}\}$ be its transition probabilities.

LEMMA 1: *If $\lambda_1 + \lambda_2 > \mu_1 + \mu_2$, then both the jump chain $\{\mathbf{x}_n\}_{n\geq 0}$ and $\{\mathbf{x}(t)\}_{t\geq 0}$ are transient.*

PROOF: Define $h(\mathbf{x}) : \mathcal{Z}_+^2 \to \Re$ to be $h(\mathbf{x}) = \big((\mu_1 + \mu_2)/(\lambda_1 + \lambda_2)\big)^{(x_1+x_2)}$. When $\lambda_1 + \lambda_2 > \mu_1 + \mu_2$, we have $h(\mathbf{x})$ bounded, $h([0,0]) = 1 > h(\mathbf{x}), \mathbf{x} \neq [0,0]$. Also, for any state $\mathbf{x} \in \mathcal{Z}_+^2 \backslash \{[0,0]\}$,

$$\sum_{\mathbf{x}'} p_{\mathbf{x}:\mathbf{x}'} h(\mathbf{x}') \leq h(\mathbf{x}), \qquad \mathbf{x} \neq [0,0],$$

can be verified. Hence, from Theorem 2 of Mertens, Samuel-Cahn, and Zamir [13], it follows that the jump chain is transient. From Theorem 3.4.1 of Norris [15], the lemma now follows. ■

https://doi.org/10.1017/S0269964804184027 Published online by Cambridge University Press

THEOREM 1: $\{\mathbf{x}(t)\}_{t\geq 0}$ *is positive recurrent if* $\lambda_1 + \lambda_2 < \mu_1 + \mu_2$.

PROOF: For uniformizable CTMCs, Kingman [11] showed that Foster's criteria can be stated as follows. An irreducible CTMC is positive recurrent if and only if there exist nonnegative $y_{\mathbf{x}}$ such that

$$\sum_{\mathbf{x}\neq\mathbf{x}'} q_{\mathbf{x}\mathbf{x}'} y_{\mathbf{x}'} < \infty \quad \text{for all states } \mathbf{x}, \tag{1}$$

$$\sum_{\mathbf{x}\neq\mathbf{x}'} q_{\mathbf{x}\mathbf{x}'} (y_{\mathbf{x}} - y_{\mathbf{x}'}) \geq 1 \quad \text{for all but a finite number of states } \mathbf{x}. \tag{2}$$

As in Kingman [11], we use a quadratic Lyapunov function $y_{\mathbf{x}} := x_1^2 + x_2^2$. As with a typical queuing system, it can be easily verified that (1) holds when $\lambda_1 + \lambda_2 < \mu_1 + \mu_2$. The finite number of states referred to in (2) will include set F (see Fig. 2) and some more states identified below.

In region A_1 of the state space, (2) reduces to requiring $-\lambda + 2x_2\mu_2 - \mu_2 \geq 1$, which holds for all sufficiently large x_2. A similar relation is satisfied in A_2 for all but finitely many x_1. In region A_3, (2) simplifies to

$$2x_1(-\lambda_2 + \mu_1) + 2x_2(-\lambda_1 + \mu_2) - d \geq 1. \tag{3}$$

Let $\epsilon_1 := \mu_1 - \lambda_2$ and $\epsilon_2 := \mu_2 - \lambda_1$. If $\epsilon_i > 0, i = 1,2$, then (3) is true for all large x_1 and x_2. Suppose $\epsilon_1 > 0$ and $\epsilon_2 < 0$ such that $\epsilon_1 + \epsilon_2 = \mu - \lambda =: \delta > 0$. λ and μ are as defined earlier. Then, (3) reduces to

$$2\epsilon_1(x_1 - x_2) + 2x_2\delta - d \geq 1, \tag{4}$$

where $d := \lambda_1 = \lambda_2 + \mu_1 + \mu_2$. Since the attractor lines have positive x_1 intercepts, we have $x_1 > x_2$ and, hence, (4) is valid for all large x_1 and x_2. Finally, suppose $\epsilon_1 < 0$ and $\epsilon_2 > 0$. As in the previous case, (3) reduces to

$$2x_1\delta - 2\epsilon_2(x_1 - x_2) - d \geq 1. \tag{5}$$

For a given x_2, we have that $x_2 \leq x_1 \leq x_2 + k_2$, where k_2 is the intercept of the class 2 attractor line and, hence, (5) can be written as

$$2x_1\delta - 2\epsilon_2(x_1 - x_2) - d \geq 2x_1\delta - (2\epsilon_2 k_2 + d) \geq 1,$$

which is true for all large x_1.

In region A_4, (2) becomes $2x_1(\mu_1 - \lambda_1 - \lambda_2) + 2x_2\mu_2 - d \geq 1$. Now, let $\epsilon := \mu_1 - (\lambda_1 + \lambda_2)$ and $\epsilon + \mu_2 = (\mu_1 + \mu_2) - (\lambda_1 + \lambda_2) =: \delta > 0$. If $\epsilon \geq 0$, then (2) holds for all large x_1 and x_2. Suppose $\epsilon < 0$; then, (2) reduces to

$$2x_1\delta + 2\mu_2(x_2 - x_1) - d \geq 1. \tag{6}$$

For that part of A_4 with $x_2 \geq x_1$, (6) holds for large values of x_1 and x_2. If $x_1 > x_2$ in A_4, we have $x_1 - x_2 < k_1$, where k_1 is the x_1 intercept of the class 1 attractor line. In this case, (6) can be reduced to

$$2x_1\delta + 2\mu_2(x_2 - x_1) - d \geq 2x_1\delta - 2\mu_2 k_1 - d \geq 1,$$

which holds for all large values of x_1 and, hence, (2) is true in A_4 also.

https://doi.org/10.1017/S0269964804184027 Published online by Cambridge University Press

In region A_5, (2) reduces to $2x_1\mu_1 + 2x_2(\mu_2 - \lambda_1 - \lambda_2) - d \geq 1$. If $\mu_2 \geq \lambda_1 + \lambda_2$, we have the desired inequality for all but finitely many states of A_5. On the other hand, if $\mu_2 < \lambda_1 + \lambda_2$, let $\epsilon := \lambda_1 + \lambda_2 - \mu_2$ and $\mu_1 = \epsilon + \delta$ for some $\delta > 0$. In this case, (2) can be written as

$$2\mu_1(x_1 - x_2) + 2x_2\delta - d \geq 1$$

and (2) follows from the fact that $x_1 > x_2$. ■

3.2. Work-Conserving Tirupati System

Recall that in the work-conserving system the capacity of an empty queue is distributed among the nonempty queues. Let $\tilde{\mathbf{x}}(t) = [\tilde{x}_1(t), \tilde{x}_2(t)]^T$ be the state of the system at time t, where $\tilde{x}_i(t)$ is the number of customers in queue i at time t. Under the model assumptions, $\{\tilde{\mathbf{x}}(t)\}_{t\geq 0}$ evolves as an irreducible CTMC over the state space $\mathcal{Z}_+^2$, also denoted by $\mathcal{S}$. The transition rates for $\{\tilde{\mathbf{x}}(t)\}_{t\geq 0}$ is as shown in Figure 3.

Consider a process $\{\tilde{\mathbf{Y}}(t)\}_{t\geq 0}$ on $\{0,1,2,\ldots\}$ such that $\tilde{\mathbf{Y}}(t) = n$ iff $\tilde{x}_1(t) + \tilde{x}_2(t) = n$. $\tilde{\mathbf{Y}}(t)$ satisfies the conditions for Theorem 2.4 of Bremaud [5, Chap. 9] and is, hence, identical to an $M/M/1$ queue with arrival rate $\lambda_1 + \lambda_2$ and service rate $\mu_1 + \mu_2$. Thus, from the stability conditions of the $M/M/1$ queue, we can state the following theorem.

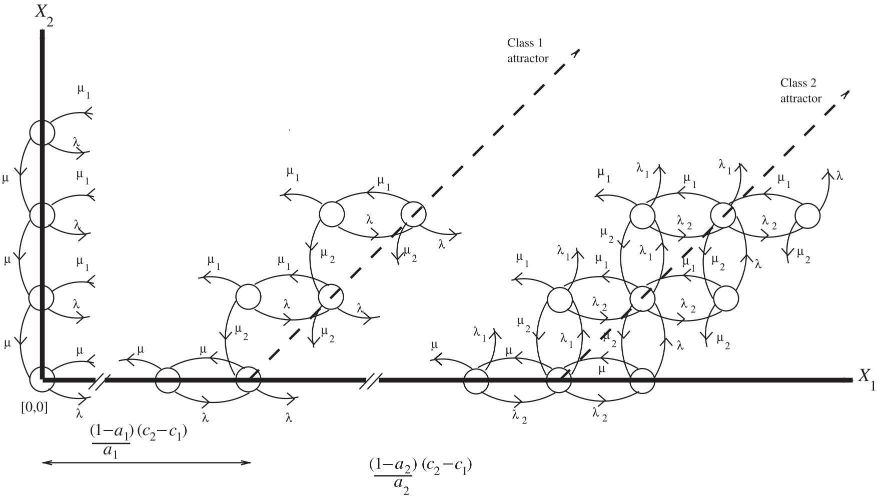


FIGURE 3. The transition rates for $\{\tilde{\mathbf{x}}(t)\}_{t\geq 0}$, the CTMC for the evolution of the work-conserving JMCQ. Observe that the departure rates for states on the axes are $\mu = \mu_1 + \mu_2$.

https://doi.org/10.1017/S0269964804184027 Published online by Cambridge University Press

THEOREM 2: $\{\tilde{\mathbf{x}}(t)\}_{t\geq 0}$ *is*

1. *transient iff* $\lambda_1 + \lambda_2 > \mu_1 + \mu_2$,
2. *positive recurrent iff* $\lambda_1 + \lambda_2 < \mu_1 + \mu_2$,
3. *null recurrent only iff* $\lambda_1 + \lambda_2 = \mu_1 + \mu_2$.

4. PERFORMANCE BOUNDS

We first define the performance measures of interest. Consider the non-work-conserving system. Let $n_j(t)$ be the number of class j arrivals in $(0,t)$ and $n(t) := n_1(t) + n_2(t)$. Let the kth arrival join queue δ_k, $\delta_k \in \{1,2\}$. The revenue rate, $\mathcal{R}$, of the system is defined as

$$\mathcal{R} = \lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{n(t)} [c_1 I(\delta_k = 1) + c_2 I(\delta_k = 2)].$$

Here, $I(\cdot)$ is the indicator function, with the usual meaning of taking a value one if the argument is true and zero otherwise.

An arriving customer of class j joining queue i when queue i has x_i customers incurs a cost of $\psi_j(c_i, x_i)$. The rate of social cost $\mathcal{D}_j$ for class j is defined below. Let k_j be the kth arrival of class j and let its arrival time be t_{k_j}. Then,

$$\mathcal{D}_j = \lim_{t\to\infty} \frac{1}{t} \sum_{k_j=0}^{n_j(t)} I(\delta_{k_j} = 1)\psi_j(c_1, x_1(t_{k_j})) + I(\delta_{k_j} = 2)\psi_j(c_2, x_2(t_{k_j})).$$

The social cost for the system $\mathcal{D}$ can be similarly defined. Also, $\bar{x}_i$ will denote the mean queue length in queue i and $\bar{x}$ the mean number in the system obtained as time averages. Similarly, we will denote the mean waiting time of a class j customer by $\overline{w}_j$, which is obtained as a customer average.

The corresponding measures for the work-conserving system will be $\tilde{\mathcal{R}}$, $\tilde{\mathcal{D}}_j$, $\tilde{\mathcal{D}}$, $\bar{\bar{x}}_i$, $\bar{\bar{x}}$, and $\overline{\overline{w}}_j$.

4.1. Non-Work-Conserving System

Consider the PMP system first. Define $\delta_{\mathbf{x}}^j$ to be the queue that an arriving class j customer to state $\mathbf{x}$ will join. For example, when the system is in state $\mathbf{x} = [x_1, x_2]$, $\delta_{\mathbf{x}}^1 = 1$ if $\psi_j(c_1, x_1) \leq \psi_j(c_2, x_2)$ and $\delta_{\mathbf{x}}^1 = 2$ otherwise. The transition rate matrix $\mathbf{Q}_{\mathbf{x}} = \{q_{\mathbf{x}:\mathbf{x}'}\}$ is easily determined. The previous section contains sufficient conditions for the ergodicity of this Markov chain, and in the following, we assume that they hold. Let $\underline{\pi} = \{\pi_{\mathbf{x}}\}$ be the stationary distribution of this Markov chain. The revenue rate $\mathcal{R}$ and the per class rate of social cost are obtained from the Law of Large Numbers (see Serfozo [18]) as follows:

$$\mathcal{R} = \sum_{\mathbf{x}} \pi_{\mathbf{x}} [\lambda_1 c_1 I(\delta_{\mathbf{x}}^1 = 1) + \lambda_1 c_2 I(\delta_{\mathbf{x}}^1 = 2) \\ + \lambda_2 c_1 I(\delta_{\mathbf{x}}^2 = 1) + \lambda_2 c_2 I(\delta_{\mathbf{x}}^2 = 2)], \tag{7}$$

$$\mathcal{D}_j = \sum_{\mathbf{x}} \pi_{\mathbf{x}} \lambda_j [I(\delta_{\mathbf{x}}^j = 1)((1 - a_j)c_1 + a_j x_1) \\ + I(\delta_{\mathbf{x}}^j = 2)((1 - a_j)c_2 + a_j x_2)]. \tag{8}$$

Closed-form expressions for $\mathcal{R}$, $\mathcal{D}_j$, and $\bar{x}_j$ are difficult to obtain and we look for approximate results in the form of computable bounds. First, consider the revenue rate, $\mathcal{R}$. We use the framework of van Houtum et al. [20] to obtain computable bounds by considering systems on a truncated state space.

Denote the original non-work-conserving JMCQ system defined over the state space $\mathcal{Z}_+^2$ by J and let $\{\mathbf{x}(t)\}_{t\geq 0}$ be the CTMC of this system. Further, let $\{\mathbf{x}_n\}_{n\geq 0}$ be the corresponding uniformized jump chain of the J system obtained from a uniformizing Poisson process of rate $d := \mu_1 + \mu_2 + \lambda_1 + \lambda_2$ (see Bremaud [5]). Since the steady state distribution is the same both in the CTMC and its corresponding uniformized chain, we work with $\{\mathbf{x}_n\}_{n\geq 0}$ in the rest of this section.

Let $\hat{\delta}_i$ be the indicator variable which captures the queue from which ith customer has departed and let $\hat{n}(t)$ be the number of departures from the system up to time t. Let $\hat{\mathcal{R}}$ be the revenue rate accrued if each customer is charged while departing from the system (instead of charging while entering the system). We claim that $\mathcal{R} = \hat{\mathcal{R}}$ almost surely. We first note that the following two limits exist almost surely:

$$\mathcal{R} = \lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{n(t)} [c_1 I(\delta_k = 1) + c_2 I(\delta_k = 2)],$$

$$\hat{\mathcal{R}} := \lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{\hat{n}(t)} [c_1 I(\hat{\delta}_k = 1) + c_2 I(\hat{\delta}_k = 2)].$$

Let $\{\tau_n\}_{n\geq 0} \uparrow \infty$ be the sequence of the end of busy periods of the stable system (i.e., return times to $\mathbf{0}$). Then, for the system that starts empty,

$$\sum_{k=0}^{n(\tau_n)} [c_1 I(\delta_k = 1) + c_2 I(\delta_k = 2)] = \sum_{k=0}^{\hat{n}(\tau_n)} [c_1 I(\hat{\delta}_k = 1) + c_2 I(\hat{\delta}_k = 2)]$$

for each n. Since $\mathcal{R}$ and $\hat{\mathcal{R}}$ exist, we can divide the above by τ_n and take limits along this subsequence, to have $\mathcal{R} = \hat{\mathcal{R}}$ almost surely. So, (7) can be written as, almost surely,

$$\mathcal{R} = \sum_{[x_1, x_2]} \pi_{[x_1, x_2]} [\mu_1 c_1 I_{(x_1 > 0)} + \mu_2 c_2 I_{(x_2 > 0)}]. \tag{9}$$

https://doi.org/10.1017/S0269964804184027 Published online by Cambridge University Press

We rewrite (9) as

$$\mathcal{R} = \sum_{\mathbf{x}} c(\mathbf{x})\pi(\mathbf{x}), \tag{10}$$

where $c(\mathbf{x})$ is the revenue rate in state $\mathbf{x}$ given by

$$c([x_1, x_2]) = \begin{cases} c_1\mu_1 + c_2\mu_2 & \text{if } x_1 > 0 \text{ and } x_2 > 0 \\ c_1\mu_1 & \text{if } x_1 > 0 \text{ and } x_2 = 0 \\ c_2\mu_2 & \text{if } x_2 > 0 \text{ and } x_1 = 0 \\ 0 & \text{if } x_1 = 0 \text{ and } x_2 = 0. \end{cases}$$

Following [20], we now establish precedences between states of $\{\mathbf{x}_n\}_{n\geq 0}$. These precedences are defined on the basis of the n-period revenue $v_n(\mathbf{x})$, which denote the expected revenue in the first $n \geq 0$ periods when starting from state $\mathbf{x}$. We say that the state $\mathbf{m} \in \mathcal{S}$ has precedence over state $\mathbf{n} \in \mathcal{S}$, if $\mathbf{m}$ and $\mathbf{n}$ satisfy the precedence relation

$$v_n(\mathbf{m}) \leq v_n(\mathbf{n}) \quad \text{for all } n = 0,1,2,\ldots. \tag{11}$$

Denote the unit vectors $[1,0]$ and $[0,1]$ by $\mathbf{e}_1$ and $\mathbf{e}_2$, respectively. We claim that state $\mathbf{m}$ has precedence over its neighboring states $\mathbf{m} + \mathbf{e}_1$ and $\mathbf{m} + \mathbf{e}_2$ for all $\mathbf{m} \in \mathcal{S}$. Also, note that the $\leq$ operation is transitive.

Let P be the set of all ordered pair of states $(\mathbf{m}, \mathbf{m} + \mathbf{e}_1)$ and $(\mathbf{m}, \mathbf{m} + \mathbf{e}_2)$, $\mathbf{m} \in \mathcal{S}$. We want to prove for all $n = 0,1,2,\ldots,$

$$v_n(\mathbf{m}) \leq v_n(\mathbf{n}), \quad \forall(\mathbf{m}, \mathbf{n}) \in P. \tag{12}$$

We use induction over n to prove (12). Taking $n = 1$ in (12) leads to

$$c(\mathbf{m}) \leq c(\mathbf{n}), \quad \forall(\mathbf{m}, \mathbf{n}) \in P. \tag{13}$$

We can easily verify that (13) holds. Assume that (12) holds for n and we prove it from $n + 1$. To establish (12) for $n + 1$, we have to show for each $(\mathbf{m}, \mathbf{n}) \in P$,

$$\begin{aligned} v_{n+1}(\mathbf{m}) &= c(\mathbf{m}) + \sum_i p(\mathbf{m}, \mathbf{i}) v_n(\mathbf{i}) \\ &\leq c(\mathbf{n}) + \sum_j p(\mathbf{n}, \mathbf{j}) v_n(\mathbf{j}) \\ &= v_{n+1}(\mathbf{n}), \end{aligned} \tag{14}$$

where $p(\mathbf{m}, \mathbf{n})$ denotes the corresponding transition probabilities of $\{\mathbf{x}_n\}_{n\geq 0}$. From (13), it suffices to show that

$$\sum_i p(\mathbf{m}, \mathbf{i}) v_n(\mathbf{i}) \leq \sum_j p(\mathbf{n}, \mathbf{j}) v_n(\mathbf{j}). \tag{15}$$

https://doi.org/10.1017/S0269964804184027 Published online by Cambridge University Press

In the non-work-conserving system, we can check (15) for all $(\mathbf{m},\mathbf{n}) \in P$. A convenient method of checking is by grouping terms corresponding to the same event (such as an arrival or departure). The attractor lines of the different customer classes divide the state space into many regions, and the transition probabilities depend on the region in which the state is located. Thus, we will have to verify (15) for the various regions. We illustrate this for the region A_3 in Figure 2. Let $\mathbf{n} = \mathbf{m} + \mathbf{e}_1$ in (15). The left-hand side of (15) is

$$\frac{\lambda_1}{d}\, v_n(\mathbf{m} + \mathbf{e_2}) + \frac{\lambda_2}{d}\, v_n(\mathbf{m} + \mathbf{e_1}) + \frac{\mu_1}{d}\, v_n(\mathbf{m} - \mathbf{e_1}) + \frac{\mu_2}{d}\, v_n(\mathbf{m} - \mathbf{e_2})$$

and the right-hand side is

$$\frac{\lambda_1}{d}\, v_n(\mathbf{m} + \mathbf{e_1} + \mathbf{e_2}) + \frac{\lambda_2}{d}\, v_n(\mathbf{m} + \mathbf{2e_1}) + \frac{\mu_1}{d}\, v_n(\mathbf{m}) + \frac{\mu_2}{d}\, v_n(\mathbf{m} + \mathbf{e_1} - \mathbf{e_2}).$$

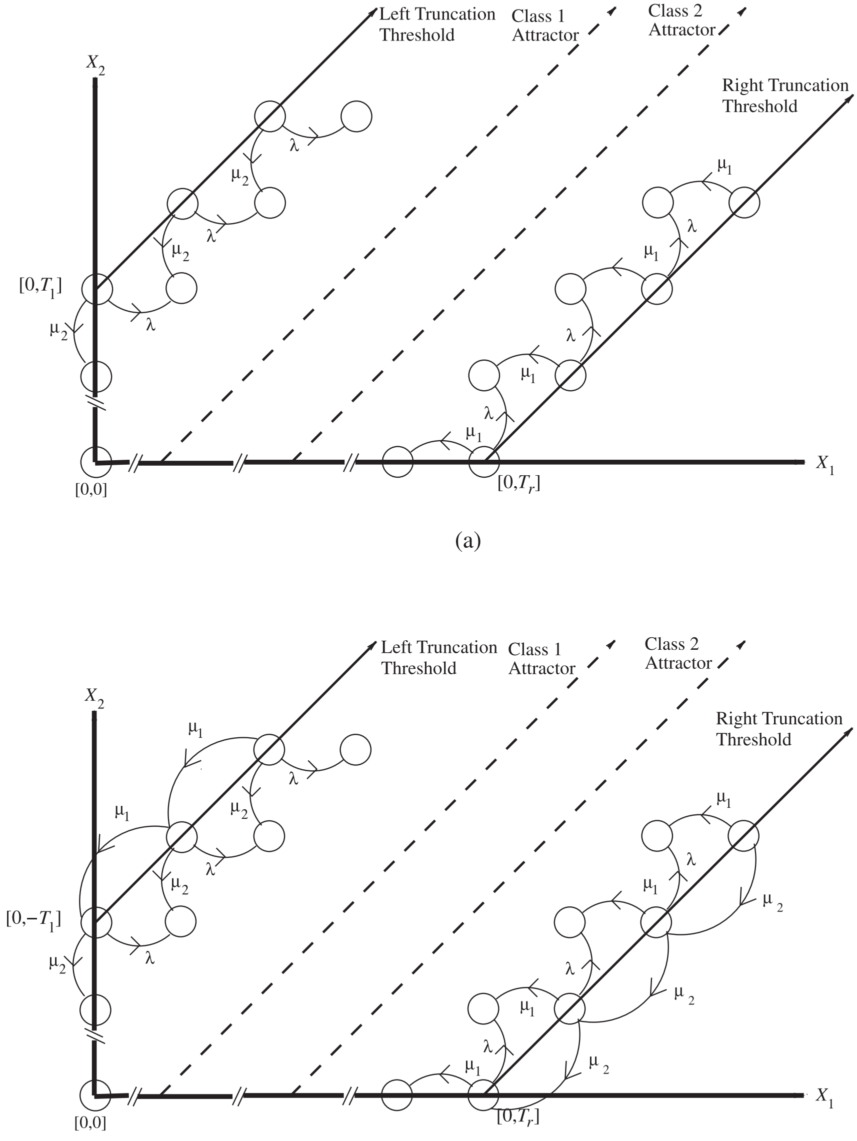
Therefore, proving (15) implies proving

$$\frac{\lambda_1}{d}\,[v_n(\mathbf{m} + \mathbf{e_2}) - v_n(\mathbf{m} + \mathbf{e_1} + \mathbf{e_2})] + \frac{\lambda_2}{d}\,[v_n(\mathbf{m} + \mathbf{e_1}) v_n(\mathbf{m} + \mathbf{2e_1})]$$
$$+ \frac{\mu_1}{d}\left[v_n(\mathbf{m} - \mathbf{e_1}) - \frac{\mu_1}{d}\, v_n(\mathbf{m})\right] + \frac{\mu_2}{d}\,[v_n(\mathbf{m} - \mathbf{e_2}) - v_n(\mathbf{m} + \mathbf{e_1} - \mathbf{e_2})] \le 0.$$

This is true because each term in the above inequality is nonnegative from the induction hypothesis. Similarly, we can verify (15) for the remaining regions.

Now, we propose two truncated systems and prove that the revenue rates in these systems act as bounds for the revenue rate in the J system. To motivate the truncation, we argue that $\pi_{\mathbf{x}}$ becomes very small for states $\mathbf{x}$ away from the attractor lines; hence, most of the probability mass of $\{\pi_{\mathbf{x}}\}$ is concentrated between the attractor lines and close to it and a state space truncated at some distance from the attractors will provide a good approximate solution. In the following, we show that if the truncation is done such that the transition from the states on the threshold lines are suitably modified, we can obtain bounds on the performance measures defined earlier, which can be made fairly tight. Specifically, we will consider the state space of the JMCQ truncated to contain only those states for which $T_l \le x_1 - x_2 \le T_r$, where T_l and T_r are integers with $T_l \le 0$ and $T_r \ge \big((1-a_2)/a_2\big)(c_2 - c_1) > 0$. T_l and T_r will be called the left and right truncation thresholds, respectively. Denote by $\mathcal{S}'$ the state space obtained after truncation. Also, let $\mathcal{S}'_{T_l}$ be the states on the left threshold line $(x_1 - x_2 = T_l)$, except the state $[0,|T_l|]$. Further, let $\mathcal{S}'_{T_r}$ be states on the right threshold line $(x_1 - x_2 = T_r)$, except the state $[T_r,0]$, and $\mathcal{S}'_I := \mathcal{S}' \backslash \mathcal{S}'_{T_r} \backslash \mathcal{S}'_{T_l}$. Figure 4 shows this truncation.

Denote the original work-conserving JMCQ system defined over the state space $\mathcal{Z}_+^2$ by J. We define systems $\mathsf{J}^{(u)}$ and $\mathsf{J}^{(l)}$ over $\mathcal{S}'$ as follows. For system $\mathsf{J}^{(u)}$, let $\mathbf{Q}^{(u)} = \{q^{(u)}_{\mathbf{x}:\mathbf{x}'}\}$ be its transition rate matrix obtained from $\mathbf{Q}$ as follows:



(a)

(b)

Figure 4. The truncated state space $\mathcal{S}'$ for the $\mathsf{J}^{(u)}$ and $\mathsf{J}^{(l)}$ systems. The modified transitions rates for states on $\mathcal{S}'_{T_r}$ and $\mathcal{S}'_{T_l}$ are shown. The transition rates for the states between the truncation lines are the same as that of J system shown in Figure 2. (a) Transition rates for the $\mathsf{J}^{(u)}$ system. For states in $\mathcal{S}'_{T_l}$ ($\mathcal{S}'_{T_r}$), transitions due to departures from queue 1 (2) are disallowed. (b) Transition rates for the $\mathsf{J}^{(l)}$ system. For states $\mathcal{S}'_{T_l}$ ($\mathcal{S}'_{T_r}$), there is a diagonal transition with a rate μ_1 (μ_2).

$$q^{(u)}_{\mathbf{x}:\mathbf{x}'} = \begin{cases} q_{\mathbf{x}:\mathbf{x}'} & \text{for } \mathbf{x},\mathbf{x}' \in \mathcal{S}' \text{ and } \mathbf{x} \neq \mathbf{x}' \\ -\sum_{\mathbf{x}',\mathbf{x}' \neq \mathbf{x}} q_{\mathbf{x}:\mathbf{x}'} & \text{for } \mathbf{x} = \mathbf{x}' \in \mathcal{S}'. \end{cases}$$

From the above, for $\mathbf{x} \in \mathcal{S}'$ an arrival will move $\mathsf{J}^{(u)}$ toward the attractor line of its class (the attractor lines of all the classes are included in the truncated state space) and will not cause the system to go out of $\mathcal{S}'$. For $\mathbf{x} \in \mathcal{S}'_{T_l}$, a departure from queue 1 is disallowed, whereas for $\mathbf{x} \in \mathcal{S}'_{T_r}$ a departure from queue 2 is disallowed to keep the system in $\mathcal{S}'$. Other transition rates are the same as that in J.

The transition rate matrix $\mathbf{Q}^{(l)} = \{q^{(l)}_{\mathbf{x}:\mathbf{x}'}\}$ for system $\mathsf{J}^{(l)}$ is obtained from $\mathbf{Q}$ as follows. As with $\mathsf{J}^{(u)}$, the transition rates from states $\mathbf{x} \in \mathcal{S}'_I$ are the same as that in J. The departures from the states on the threshold lines are modified such that a departure from one queue that might lead to a state out of $\mathcal{S}'$ will take away one more customer from the other queue also and, hence, keep the system state in $\mathcal{S}'$.

Let $\pi^{(u)}_{\mathbf{x}}$ and $\pi^{(l)}_{\mathbf{x}}$ be the stationary distributions of $\mathsf{J}^{(u)}$ and $\mathsf{J}^{(l)}$, respectively. For the following, we will assume that both of these systems are stable and, hence, that their stationary distributions exist. We will discuss the stability of these systems in Section 5.

Let $\{\mathbf{x}^{(u)}_n\}_{n\geq 0}$ and $\{\mathbf{x}^{(l)}_n\}_{n\geq 0}$ be the uniformized jump chains of the $\mathsf{J}^{(u)}$ and $\mathsf{J}^{(l)}$ systems obtained from a uniformizing Poisson process of rate $d := \mu_1 + \mu_2 + \lambda_1 + \lambda_2$. For these systems, we can write the revenue rate as

$$\mathcal{R}^{(u)} = \sum_{\mathbf{x}} c(\mathbf{x})\pi^{(u)}_{\mathbf{x}},$$

$$\mathcal{R}^{(l)} = \sum_{\mathbf{x}} c(\mathbf{x})\pi^{(l)}_{\mathbf{x}}. \tag{16}$$

In our truncation model, we observe that $\{\mathbf{x}^{(u)}_n\}_{n\geq 0}$ is obtained by redirecting transitions to preceding states and $\{\mathbf{x}^{(l)}_n\}_{n\geq 0}$ is obtained by redirecting transitions to succeeding states. Thus, from Theorem 1 in [20], we have the following result.

THEOREM 3: *If systems $\mathsf{J}^{(l)}$, J, and $\mathsf{J}^{(u)}$ start empty at $t = 0$, we have*

$$\mathcal{R}^{(u)} \geq \mathcal{R} \geq \mathcal{R}^{(l)}.$$

Now, consider the bounds on the mean queue lengths. Choosing $c(\mathbf{x}) = x_1$ in (10), we get the expression for the mean number in queue 1: $\bar{x}_1$. Similarly, choosing $c(\mathbf{x}) = x_2$ gives the expression for the mean number in queue 2: $\bar{x}_2$. Using induction, we can again prove that $\mathbf{m}$ precedes $\mathbf{m} + \mathbf{e_1}$ and $\mathbf{m} + \mathbf{e_2}$. Let $\bar{x}^{(u)}_i$ and $\bar{x}^{(l)}_i$ be the mean queue lengths in queue i for the $\mathsf{J}^{(u)}$ and $\mathsf{J}^{(l)}$ systems, respectively. Thus, we have the following result.

THEOREM 4: *Let $\bar{x}_i$, $\bar{x}^{(u)}_i$, and $\bar{x}^{(l)}_i$ exist for $i = 1,2$. Then,*

$$\bar{x}^{(l)}_i \leq \bar{x}_i \leq \bar{x}^{(u)}_i \quad \textit{for } i = 1,2.$$

https://doi.org/10.1017/S0269964804184027 Published online by Cambridge University Press

Further, if we take $c(\mathbf{x}) = 1$ for $x_1 > M$ and zero otherwise, where $M \in \mathcal{N}$, then $\sum_{\mathbf{x}} c(\mathbf{x})\pi_{\mathbf{x}}$ is the tail probability that the total number of customers in queue 1 exceeds M. As a result, the stationary number in queue 1 in the J system is stochastically bounded between the stationary number in queue 1 of the $\mathsf{J}^{(u)}$ and $\mathsf{J}^{(l)}$ systems. We have a similar result for the stationary number in queue 2 in the J system. In fact, we can show that if all systems start in the same feasible state, say (0,0), then at any jump epoch, the number in a queue of the J system is stochastically bounded by the number in the corresponding queues of the $\mathsf{J}^{(u)}$ and $\mathsf{J}^{(l)}$ systems.

Remark 1: We can also show that the number in each queue of $\mathsf{J}^{(u)}$ and $\mathsf{J}^{(l)}$ almost surely bound the number in the J system at every jump epoch. The proof technique is similar to that used to show such bounds for the work-conserving system, which we discuss next.

4.2. Work-Conserving System

We now obtain results similar to those in the previous subsection for the work-conserving system. We will use the same notation for the parameters and performance measures as in the previous subsection except that they will have a tilde to differentiate them from the corresponding variables of the non-work-conserving system. For example, $\tilde{x}_{i,k}$ will be the queue occupancy in queue i at t_k.

We first show that for the work-conserving system, we cannot obtain the bounds for the performance measures by proceeding exactly as in the previous section and applying the method of [20]. To see this, consider bounds for the revenue rate. As in the non-work-conserving system, we can write the revenue rate for the $\tilde{\mathsf{J}}$ system as

$$\tilde{\mathcal{R}} = \sum_{\tilde{\mathbf{x}}\in\mathcal{S}} [\tilde{\pi}_{\tilde{\mathbf{x}}}[\lambda_1 c_1 I(\delta^1_{\tilde{\mathbf{x}}} = 1) + \lambda_1 c_2 I(\delta^1_{\tilde{\mathbf{x}}} = 2) + \lambda_2 c_1 I(\delta^2_{\tilde{\mathbf{x}}} = 1) + \lambda_2 c_2 I(\delta^2_{\tilde{\mathbf{x}}} = 2)]]. \tag{17}$$

This can be written in terms of μ as

$$\tilde{\mathcal{R}} = \sum_{\tilde{\mathbf{x}}} c(\tilde{\mathbf{x}})\pi(\tilde{\mathbf{x}}), \tag{18}$$

where $c(\tilde{\mathbf{x}})$ represents the cost per period in state $\tilde{\mathbf{x}}$ given by

$$c([\tilde{x}_1, \tilde{x}_2]) = \begin{cases} c_1\mu_1 + c_2\mu_2 & \text{if } \tilde{x}_1 > 0 \text{ and } \tilde{x}_2 > 0 \\ c_1\mu & \text{if } \tilde{x}_1 > 0 \text{ and } \tilde{x}_2 = 0 \\ c_2\mu & \text{if } \tilde{x}_2 > 0 \text{ and } \tilde{x}_1 = 0 \\ 0 & \text{if } \tilde{x}_1 = 0 \text{ and } \tilde{x}_2 = 0. \end{cases}$$

In the work-conserving system, the cost function on the axes is forced to be $c_i\mu$, where $\mu = \mu_1 + \mu_2$. For our truncation model to yield bounds for the revenue rate, the precedence relation mentioned in the previous subsection must hold here also; that is, the state $\tilde{\mathbf{m}}$ must precede the states $\tilde{\mathbf{m}} + \tilde{\mathbf{e}}_1$ and $\tilde{\mathbf{m}} + \tilde{\mathbf{e}}_2$ for all $\tilde{\mathbf{m}} \in \mathcal{S}$. Suppose that the above-mentioned precedence relations hold. Let $\tilde{P}$ be the set of all $(\tilde{\mathbf{m}}, \tilde{\mathbf{m}} + \tilde{\mathbf{e}}_\mathbf{1})$ and $(\tilde{\mathbf{m}}, \tilde{\mathbf{m}} + \tilde{\mathbf{e}}_\mathbf{2})$. Then, for all $t = 0,1,2,\ldots,$

$$\tilde{v}_t(\tilde{\mathbf{m}}) \leq \tilde{v}_t(\tilde{\mathbf{n}}) \quad \text{for all } (\tilde{\mathbf{m}}, \tilde{\mathbf{n}}) \in \tilde{P}. \tag{19}$$

Specifically, this must hold for $t = 1$. Taking $t = 1$ in (19) leads to

$$c(\tilde{\mathbf{m}}) \leq c(\tilde{\mathbf{n}}) \quad \text{for all } (\tilde{\mathbf{m}}, \tilde{\mathbf{n}}) \in \tilde{P}. \tag{20}$$

However,

$$c([0,-T_l]) = c_2\mu > c_1\mu_1 + c_2\mu_2 = c([1,-T_l]),$$

contradicting (20). Thus, the assumed precedence relations are false and the proposed truncation model will not yield bounds using the techniques in [20]. We take an alternate approach to prove the bounds for the revenue rate of the $\tilde{\text{J}}$ system.

4.3. Sample-Path Approach

Let $\{\tilde{\mathbf{x}}(t)\}_{t\geq 0}$, $\{\mathbf{x}^{(u)}(t)\}_{t\geq 0}$, and $\{\mathbf{x}^{(l)}(t)\}_{t\geq 0}$ be the CTMC of the $\tilde{\text{J}}$, $\tilde{\text{J}}^{(u)}$, and $\tilde{\text{J}}^{(l)}$ systems, respectively. Let $\{\tilde{\mathbf{x}}_n\}_{n\geq 0}$, $\{\tilde{\mathbf{x}}_n^{(u)}\}_{n\geq 0}$, and $\{\tilde{\mathbf{x}}_n\}_{n\geq 0}^{(l)}$ be the corresponding uniformized jump chains of the $\tilde{\text{J}}$, $\tilde{\text{J}}^{(u)}$, and $\tilde{\text{J}}^{(l)}$ systems, respectively, obtained from a uniformizing Poisson process of rate $d := \mu_1 + \mu_2 + \lambda_1 + \lambda_2$ (see [5]). Now, consider the uniformized systems $\tilde{\text{J}}$, $\tilde{\text{J}}^{(u)}$, and $\tilde{\text{J}}^{(l)}$ evolving in parallel and driven by the same event sequence determined by the Poisson process. We now present a forward induction type of proof (see Walrand [22, Chap. 8]) to show that system $\tilde{\text{J}}^{(u)}$ (resp. $\tilde{\text{J}}$) componentwise dominates $\tilde{\text{J}}$ (resp. $\tilde{\text{J}}^{(l)}$) for all n.

Let $t_1 < t_2 < t_3 < \cdots$ be the event epochs of the uniformizing Poisson process. Every jump of this Poisson process corresponds to either an arrival of class j, $j = 1,2$ with probability λ_j/d, or a potential service completion from queue i, with probability μ_i/d. An arrival will join the queue that minimizes its cost, possibly different queues in different systems. The work-conserving property leads to the following queue dynamics at potential service completion instants. A potential service completion time from queue i is an actual service completion from that queue if it is nonempty, and it is an actual service completion in the "other" queue if queue i is empty and the other is nonempty. Because the service and interarrival times are exponentially distributed, we just disallow departures when actual departures are not possible at potential departure times. Let $\tilde{\mathbf{x}}_\mathbf{n}$ be the state of $\{\tilde{\mathbf{x}}_n\}_{n\geq 0}$ "just after" t_n and let $\tilde{\mathbf{x}}_\mathbf{1}, \tilde{\mathbf{x}}_\mathbf{2}, \ldots, \tilde{\mathbf{x}}_\mathbf{N}$ be the sample path of $\{\tilde{\mathbf{x}}_n\}_{n\geq 0}$ up to time t_N^+. Reference [22] has more details on the development of the sample paths. Denote the evolution paths by $\tilde{\mathbf{x}}_\mathbf{1}^{(\mathbf{u})}, \tilde{\mathbf{x}}_\mathbf{2}^{(\mathbf{u})}, \ldots, \tilde{\mathbf{x}}_\mathbf{N}^{(\mathbf{u})}$ in the $\tilde{\text{J}}^{(u)}$ system and by $\tilde{\mathbf{x}}_\mathbf{1}^{(\mathbf{l})}, \tilde{\mathbf{x}}_\mathbf{2}^{(\mathbf{l})}, \ldots, \tilde{\mathbf{x}}_\mathbf{N}^{(\mathbf{l})}$ in the $\tilde{\text{J}}^{(l)}$ system and let $\tilde{\mathbf{x}}_\mathbf{k} = [\tilde{x}_{1,k}, \tilde{x}_{2,k}]$, $\tilde{\mathbf{x}}_\mathbf{k}^{(\mathbf{u})} = [\tilde{x}_{1,k}^{(u)}, \tilde{x}_{2,k}^{(u)}]$, and $\tilde{\mathbf{x}}_\mathbf{k}^{(\mathbf{l})} = [\tilde{x}_{1,k}^{(l)}, \tilde{x}_{2,k}^{(l)}]$.

https://doi.org/10.1017/S0269964804184027 Published online by Cambridge University Press

THEOREM 5: *If the system starts empty at $k = 0$, then for any $k \geq 0$,*

(a) $\tilde{x}_{1,k} \leq \tilde{x}^{(u)}_{1,k}$
(b) $\tilde{x}_{2,k} \leq \tilde{x}^{(u)}_{2,k}$
(c) $\tilde{x}^{(l)}_{1,k} \leq \tilde{x}_{1,k}$
(d) $\tilde{x}^{(l)}_{2,k} \leq \tilde{x}_{2,k}$.

PROOF: Assume that (a) and (b) have not failed till t_k. Both (a) and (b) cannot fail for the first time simultaneously and we consider the events that might lead to them failing separately. We will consider (a) alone first. For (a) to fail before (b) at $t_{k'+1}$, we require that $\tilde{x}_{1,k'} = \tilde{x}^{(u)}_{1,k'}$ and $\tilde{x}_{2,k'} \leq \tilde{x}^{(u)}_{2,k'}$. We consider the possible events at $t_{k'}$ with this condition. If the event at $t_{k'}$ is a potential departure, the following subcases arise:

1. Potential departure from queue 1: Because $\tilde{x}_{1,k'} = \tilde{x}^{(u)}_{1,k'}$, $\tilde{\mathrm{J}}$ and $\tilde{\mathrm{J}}^{(u)}$ will behave identically with respect to queue 1 when $[\tilde{x}^{(u)}_{1,k'}, \tilde{x}^{(u)}_{2,k'}] \notin \mathcal{S}'_{T_l}$. If $[\tilde{x}^{(u)}_{1,k'}, \tilde{x}^{(u)}_{2,k'}] \in \mathcal{S}'_{T_l}$, a departure from queue 1 is disallowed in $\tilde{\mathrm{J}}^{(u)}$ and allowed in $\tilde{\mathrm{J}}$, and (a) is maintained.
2. Potential departure from queue 2: $\tilde{x}^{(u)}_{2,k'} = 0$ is necessary to cause an actual departure from queue 1. By assumption that (b) has not failed until $t_{k'}$, $\tilde{x}^{(u)}_{2,k'} \geq \tilde{x}_{2,k'}$ implying $\tilde{x}_{2,k'} = 0$. $\tilde{\mathrm{J}}^{(u)}$ and $\tilde{\mathrm{J}}$ will behave identically with respect to queue 1.

Now, consider the case when the event at $t_{k'+1}$ is an arrival. For (a) to fail, the arrival must join queue 1 in $\tilde{\mathrm{J}}$ and queue 2 in $\tilde{\mathrm{J}}^{(u)}$. For this to happen, $[\tilde{x}^{(u)}_{1,k'}, \tilde{x}^{(u)}_{2,k'}]$ must be on the right-hand side of the attractor line for the class of the arrival, and $[\tilde{x}_{1,k'}, \tilde{x}_{2,k'}]$ must be on its left-hand side. This is clearly not possible because $\tilde{x}_{1,k'} = \tilde{x}^{(u)}_{1,k'}$, $\tilde{x}_{2,k'} \leq \tilde{x}^{(u)}_{2,k'}$ and the attractor line has a positive slope.

Similar arguments show that (b) cannot fail before (a) at $t_{k'+1}$.

Now, consider (c) and (d). Once again, they cannot fail for the first time simultaneously. We proceed as above and look at events at $t_{k'+1}$ when $\tilde{x}_{1,k'} = \tilde{x}^{(l)}_{1,k'}$ and $\tilde{x}_{2,k'} \geq \tilde{x}^{(l)}_{2,k'}$, which is required for (c) to fail for the first time and before (d) at $t_{k'+1}$.

As previously, first consider a potential departure at $t_{k'}$:

1. Potential departure from queue 1: Because $\tilde{x}_{1,k'} = \tilde{x}^{(l)}_{1,k'}$, $\tilde{\mathrm{J}}$ and $\tilde{\mathrm{J}}^{(l)}$ will behave identically with respect to queue 1 when $[\tilde{x}^{(l)}_{1,k'}, \tilde{x}^{(l)}_{2,k'}] \notin \mathcal{S}'_{T_l}$. If $[\tilde{x}^{(l)}_{1,k'}, \tilde{x}^{(l)}_{2,k'}] \in \mathcal{S}'_{T_l}$, an actual departure takes place from both the queues in the $\tilde{\mathrm{J}}^{(l)}$ system and the inequality is maintained.
2. Potential departure from queue 2: $\tilde{x}_{2,k'} = 0$ is necessary to cause an actual departure from queue 1 in $\tilde{\mathrm{J}}$. However, by assumption that (d) had not failed until $t_{k'}$, $\tilde{x}_{2,k'} \geq \tilde{x}^{(l)}_{2,k'}$ and, hence, $\tilde{x}^{(l)}_{2,k'} = 0$. $\tilde{\mathrm{J}}$ and $\tilde{\mathrm{J}}^{(l)}$ will behave identically with respect to queue 1.

For an arrival of class j at $t_{k'+1}$, arguments identical to that from the first part are used to show that (c) does not fail at $t_{k'+1}$.

https://doi.org/10.1017/S0269964804184027 Published online by Cambridge University Press

Similar arguments are constructed to show that $\tilde{x}_{2,k'} < \tilde{x}^{(l)}_{2,k'}$ could not have happened at $t_{k'}$ and, hence, (d) could not have failed before (c) for the first time at $t_{k'+1}$. ■

Let $\bar{\tilde{x}}_i^{(u)}$ and $\bar{\tilde{x}}_i^{(l)}$ be the mean queue lengths in queue i for the $\tilde{\mathrm{J}}^{(u)}$ and $\tilde{\mathrm{J}}^{(l)}$ systems, respectively. Also, let $\bar{\tilde{w}}_i$, $\bar{\tilde{w}}_i^{(u)}$, and $\bar{\tilde{w}}_i^{(l)}$ be the mean waiting times in queue i of the $\tilde{\mathrm{J}}$, $\tilde{\mathrm{J}}^{(u)}$, and the $\tilde{\mathrm{J}}^{(l)}$ systems, respectively.

THEOREM 6: *Let $\bar{\tilde{x}}_i$, $\bar{\tilde{x}}_i^{(u)}$, and $\bar{\tilde{x}}_i^{(l)}$ exist for $i = 1,2$. Then,*

(a) $\bar{\tilde{x}}_i^{(l)} \le \bar{\tilde{x}}_i \le \bar{\tilde{x}}_i^{(u)}$
(b) $\bar{\tilde{x}}_i^{(l)}/\bar{\lambda}_i \le \bar{\tilde{w}}_i \le \bar{\tilde{x}}_i^{(u)}/\bar{\lambda}_i$

for $i = 1,2$, where $\bar{\lambda}_i$ is the long-run arrival rate into queue i in the $\tilde{\mathrm{J}}$ system.

PROOF: From Wolff [23], $\tilde{\mathrm{J}}$, $\tilde{\mathrm{J}}^{(u)}$, and $\tilde{\mathrm{J}}^{(l)}$ systems are regenerative and, hence, the mean queue lengths are also time averages. Hence,

$$\bar{\tilde{x}}_i = \lim_t \frac{1}{t}\int_0^t \tilde{x}_1(u)\,du \le \lim_t \frac{1}{t}\int_0^t \tilde{x}_1^{(u)}(u)\,du = \tilde{x}_i^{(u)}.$$

Similarly we can prove the left half of (a).

From Little's law, we write (b) as

$$\bar{\tilde{x}}_i^{(l)} \le \bar{\lambda}_i \bar{\tilde{w}}_i \le \bar{\tilde{x}}_i^{(u)}.$$

Dividing throughout by $\bar{\lambda}_i$, we get (b). ■

Now, we find the bounds on the revenue rate. In this case, we choose $T_l = 0$ and, hence, the left truncation threshold is the line $\tilde{x}_1 = \tilde{x}_2$ in $\mathcal{Z}_+^2$. The revenue processes in $\tilde{\mathrm{J}}^{(u)}$ and $\tilde{\mathrm{J}}^{(l)}$ are modified as follows. The $\tilde{\mathrm{J}}^{(u)}$ and $\tilde{\mathrm{J}}^{(l)}$ systems will earn revenue exactly like the $\tilde{\mathrm{J}}$ system, except for the following cases. We stipulate that when the system is in a state in $\mathcal{S}'_{T_r}$, $\tilde{\mathrm{J}}^{(u)}$ will "gain revenue" $(c_2 - c_1)$ and $\tilde{\mathrm{J}}^{(l)}$ will "lose revenue" c_1 according to the rate μ_2. When the system is in a state in $\mathcal{S}'_{T_l}$, only $\tilde{\mathrm{J}}^{(l)}$ will "lose revenue" c_2 according to the rate μ_1. Let $\tilde{\mathcal{R}}^{(u)}$ and $\tilde{\mathcal{R}}^{(l)}$ be the revenue rates so earned in systems $\tilde{\mathrm{J}}^{(u)}$ and $\tilde{\mathrm{J}}^{(l)}$, respectively. Then, from the Law of Large Numbers,

$$\begin{aligned}
\tilde{\mathcal{R}}^{(u)} &= \sum_{\tilde{\mathbf{x}}\in\mathcal{S}'} [\tilde{\pi}_{\tilde{\mathbf{x}}}^{(u)}[\lambda_1 c_1 I(\delta_{\tilde{\mathbf{x}}}^1 = 1) + \lambda_1 c_2 I(\delta_{\tilde{\mathbf{x}}}^1 = 2) + \lambda_2 c_1 I(\delta_{\tilde{\mathbf{x}}}^2 = 1) + \lambda_2 c_2 I(\delta_{\tilde{\mathbf{x}}}^2 = 2)]] \\
&\quad + \sum_{\tilde{\mathbf{x}}\in\mathcal{S}'_{T_r}} \tilde{\pi}_{\tilde{\mathbf{x}}}^{(u)} \mu_2 (c_2 - c_1), \\
\tilde{\mathcal{R}}^{(l)} &= \sum_{\tilde{\mathbf{x}}\in\mathcal{S}'} [\tilde{\pi}_{\tilde{\mathbf{x}}}^{(l)}[\lambda_1 c_1 I(\delta_{\tilde{\mathbf{x}}}^1 = 1) + \lambda_1 c_2 I(\delta_{\tilde{\mathbf{x}}}^1 = 2) + \lambda_2 c_1 I(\delta_{\tilde{\mathbf{x}}}^2 = 1) + \lambda_2 c_2 I(\delta_{\tilde{\mathbf{x}}}^2 = 2)]] \\
&\quad - \sum_{\tilde{\mathbf{x}}\in\mathcal{S}'_{T_l}} \tilde{\pi}_{\tilde{\mathbf{x}}}^{(l)} \mu_1 c_2 - \sum_{\tilde{\mathbf{x}}\in\mathcal{S}'_{T_r}} \tilde{\pi}_{\tilde{\mathbf{x}}}^{(l)} \mu_2 c_1. \qquad \textbf{(21)}
\end{aligned}$$

https://doi.org/10.1017/S0269964804184027 Published online by Cambridge University Press

Observe that the expressions for $\mathcal{R}^{(u)}$ and $\mathcal{R}^{(l)}$ are similar to that for $\mathcal{R}$ in (17) except that they are defined on the corresponding $\tilde{\mathrm{J}}^{(u)}$ and $\tilde{\mathrm{J}}^{(l)}$ systems, respectively, and the "revenue earnings" are modified for the states on the threshold line as discussed in the previous paragraph. Let $\tilde{R}_N$, $\tilde{R}_N^{(u)}$, and $\tilde{R}_N^{(l)}$ denote the cumulative revenue in $\tilde{\mathrm{J}}$, $\tilde{\mathrm{J}}^{(u)}$, and $\tilde{\mathrm{J}}^{(l)}$, respectively, until time t_N.

Theorem 7: *If the* $\tilde{\mathrm{J}}$ *and* $\tilde{\mathrm{J}}^{(l)}$ *systems start empty at time* $t = 0$, *then for all* $N \geq 0$,

$$\tilde{R}_N^{(l)} - \tilde{R}_N \leq c_1(\tilde{x}_{1N}^{(l)} - \tilde{x}_{1N}) + c_2(\tilde{x}_{2N}^{(l)} - \tilde{x}_{2N}) \leq 0. \tag{22}$$

Proof: Let $\tilde{R}_{t_k,t_{k+1}}$ ($\tilde{R}_{t_k,t_{k+1}}^{(u)}$) be the revenue "earned" in the transition at t_{k+1} from $[\tilde{x}_{1k}, \tilde{x}_{2k}]$ to $[\tilde{x}_{1(k+1)}, \tilde{x}_{2(k+1)}]$ ($[\tilde{x}_{1k}^{(u)}, \tilde{x}_{2k}^{(u)}]$ to $[\tilde{x}_{1(k+1)}^{(u)}, \tilde{x}_{2(k+1)}^{(u)}]$) in the $\tilde{\mathrm{J}}$ ($\tilde{\mathrm{J}}^{(u)}$) system. We can write the left-hand side of Eq. (22) as

$$\tilde{R}_N^{(l)} - \tilde{R}_N = \sum_{k=1}^{N-1} (\tilde{R}_{t_k,t_{k+1}}^{(l)} - \tilde{R}_{t_k,t_{k+1}}). \tag{23}$$

We will show that the following holds for all t_k, the epochs of the uniformizing process:

$$\begin{aligned}\tilde{R}_{t_k,t_{k+1}}^{(l)} - \tilde{R}_{t_k,t_{k+1}} \leq {} & [c_1(\tilde{x}_{1(k+1)}^{(l)} - \tilde{x}_{1k}^{(l)}) + c_2(\tilde{x}_{2(k+1)}^{(l)} - \tilde{x}_{2k}^{(l)})] \\ & - [c_1(\tilde{x}_{1(k+1)} - \tilde{x}_{1k}) + c_2(\tilde{x}_{2(k+1)} - \tilde{x}_{2k})].\end{aligned} \tag{24}$$

For an arrival at t_k, $\tilde{R}_{t_k,t_{k+1}}^{(l)} = c_1(\tilde{x}_{1(k+1)}^{(l)} - \tilde{x}_{1k}^{(l)}) + c_2(\tilde{x}_{2(k+1)}^{(l)} - \tilde{x}_{2k}^{(l)})$. A similar expression is written for $\tilde{R}_{t_k,t_{k+1}}$ and (24) is satisfied. Now, consider potential departures. We consider four cases corresponding to the state of the queues in $\tilde{\mathrm{J}}^{(l)}$.

Case 1: Both queues are empty. Irrespective of the state of $\tilde{\mathrm{J}}$, the first term on the right-hand side of (24) is zero, the second term is nonnegative, the left-hand side is zero, and (24) is satisfied.

Case 2: Queue 1 is empty and queue 2 is nonempty. This cannot happen because we choose $T_l = 0$ in our truncation of the state space.

Case 3: Queue 1 is nonempty and queue 2 is empty. For a potential departure from queue 1, queue 1 in $\tilde{\mathrm{J}}$ is also nonempty and there is an actual departure from queue 1 in both $\tilde{\mathrm{J}}$ and $\tilde{\mathrm{J}}^{(l)}$ and (24) is satisfied. For a potential departure from queue 2, the following subcases need to be considered:

1. Queue 2 in $\tilde{\mathrm{J}}$ is empty. By Theorem 5, queue 1 in $\tilde{\mathrm{J}}$ is necessarily nonempty. There will be an actual departure from queue 1 (because of work-conserving service) in both $\tilde{\mathrm{J}}$ and $\tilde{\mathrm{J}}^{(l)}$ and (24) is satisfied.
2. Queue 2 in $\tilde{\mathrm{J}}$ is nonempty. This means that there will be an actual departure from queue 1 in $\tilde{\mathrm{J}}^{(l)}$ and an actual departure from queue 2 in $\tilde{\mathrm{J}}$. In this case, the left-hand side in (24) is zero and the right-hand side is $c_2 - c_1$ and the inequality is satisfied.

https://doi.org/10.1017/S0269964804184027 Published online by Cambridge University Press

Case 4: Both queues are nonempty. In this case, both queues of $\tilde{J}$ will be nonempty by Theorem 5. First, consider a potential departure from queue 1 (resp. queue 2). If $\tilde{J}^{(l)}$ is not on the left (resp. right) truncation threshold, then in both the systems, there is an actual departure from queue 1 (resp. queue 2) and (24) is satisfied. If it is on the left (resp. right) threshold, left-hand and right-hand sides will both be $-c_2$ (resp. $-c_1$) and the inequality of (24) is satisfied.

Thus, the inequality of (24) holds, and substituting (24) in the right-hand side of (23), we get (22) if we start from an empty system. ■

To obtain upper bounds on the cumulative revenue, consider the $\tilde{J}$ and $\tilde{J}^{(u)}$ systems together. Let $\tilde{R}^{(u)}_{t_k,t_{k+1}}$ (resp. $\tilde{R}_{t_k,t_{k+1}}$) be the revenue earned in the transition from $[\tilde{x}^{(u)}_{1,k}, \tilde{x}^{(u)}_{2,k}]$ to $[\tilde{x}^{(u)}_{1,k+1}, \tilde{x}^{(u)}_{2,k+1}]$ (resp. $[\tilde{x}_{1,k}, \tilde{x}_{2,k}]$ to $[\tilde{x}_{1,k+1}, \tilde{x}_{2,k+1}]$) in the $\tilde{J}^{(u)}$ (resp. $\tilde{J}$) system at time t_{k+1}. For an event of the uniformizing process at time t_k, the difference in the behavior between the $\tilde{J}$ and $\tilde{J}^{(u)}$ systems depends on the slope of the line joining the points $[\tilde{x}^{(u)}_{1,k}, \tilde{x}^{(u)}_{2,k}]$ (state of $\tilde{J}^{(u)}$ at t_k) and $[\tilde{x}_{1,k}, \tilde{x}_{2,k}]$ (state of $\tilde{J}$ at t_k). To capture the dependence on this slope, let $\tau_k := [(\tilde{x}^{(u)}_{2,k} - \tilde{x}_{2,k}) - (\tilde{x}^{(u)}_{1,k} - \tilde{x}_{1,k})]$. The slope of the line joining the points $[\tilde{x}^{(u)}_{1,k}, \tilde{x}^{(u)}_{2,k}]$ and $[\tilde{x}_{1,k}, \tilde{x}_{2,k}]$ is less than (resp. greater or equal to) one if $\tau_k < 0$ (resp. $\tau_k \geq 0$).

A sequence $\tau_1, \ldots, \tau_N$ can be associated with a joint sample path in $\tilde{J}^{(u)}$ and $\tilde{J}$ for epochs $t_1, \ldots, t_N$. Let $G = (V, E)$ be the directed graph induced by this sequence, where the vertex set V is obtained from $\{\tau_k\}$ (τ_k takes values in $\mathcal{Z}$) and the directed edge set $E = \{e_k = (\tau_k, \tau_{k+1})\}$. In the following, our discussion will be based on a graph so obtained from a sample path. For every $e_k \in E$, define $l_k := \tau_{k+1} - \tau_k$ and $w_k := \tilde{R}^{(u)}_{t_k,t_{k+1}} - \tilde{R}_{t_k,t_{k+1}}$. We will call l_k the length of e_k and call w_k its weight. w_k is the excess revenue earned by $\tilde{J}^{(u)}$ over $\tilde{J}$ due to the event at time t_k. Also, define $S_{m,n} := \{e_k \in E \,|\, \tau_k = m, \tau_{k+1} = n\}$; that is, $S_{m,n}$ is the set of all directed edges from m to n, $m, n \in \mathcal{Z}$. Figure 5 shows an example of the directed graph induced by the τ_k from a sample path of the $\tilde{J}$ and $\tilde{J}^{(u)}$ systems.

Now, consider the possible combination of events in $\tilde{J}$ and $\tilde{J}^{(u)}$ at epoch t_k. They are listed in Table 1 along with the sign of τ_k and τ_{k+1} and the values of l_k and w_k. From Table 1, we see that $l_k \in \{-2,-1,0,1,2\}$, w_k is negative only due to events of type 2 (an arrival chooses queue 1 in $\tilde{J}^{(u)}$ and queue 2 in $\tilde{J}$), and if w_k is negative, then $\tau_k > 0$ and $l_k = -2$. Further, $l_k > 0$ and $\tau_{k+1} > 0$ are possible only from two events, those of type 3 and 12. A type 3 event is an arrival joining queue 2 in $\tilde{J}^{(u)}$ and queue 1 in $\tilde{J}$. A type 12 event is a potential departure from queue 2 when $\tilde{J}^{(u)}$ is on the right threshold and is, hence, disallowed and an actual departure from queue 2 in $\tilde{J}$. In this case, $w_k = c_2 - c_1$ and $l_k = 1$. We are now ready to state the following theorem.

THEOREM 8: *If $\tilde{J}$ and $\tilde{J}^{(u)}$ start empty at time $t = 0$, then for all $N \geq 0$,*

$$\tilde{R}_N \leq \tilde{R}^{(u)}_N.$$

PROOF: We can write

$$\tilde{R}^{(u)}_N - \tilde{R}_N = \sum_{k=1}^{N} w_k = \sum_{(m,n)\in V\times V} \sum_{e_k \in S_{m,n}} w_k. \tag{25}$$

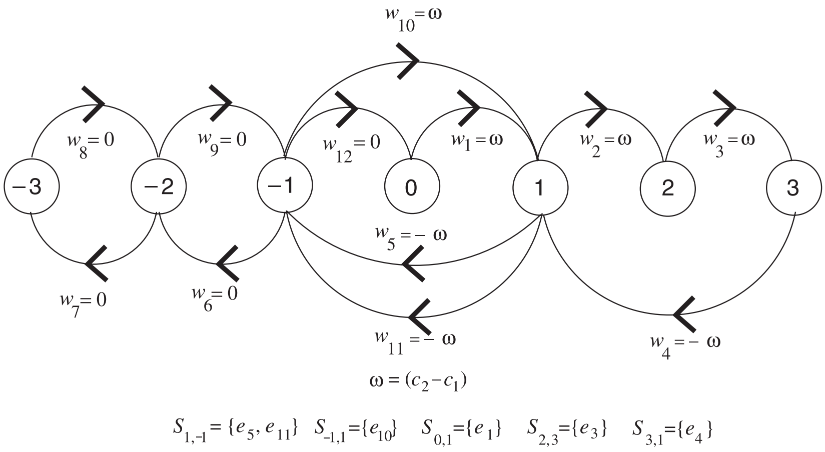


FIGURE 5. Example sequence of τ_k from a sample path represented as a directed graph. For clarity in the illustration, self loops are not shown, as these do not have negative weights. The remaining edges are renumbered such that their initial order is maintained. In the example shown, $m = 1$ and $m = 3$ have edges with negative weights. Observe that the terms in (25) corresponding to these states are "canceled" as follows: $\sum_{e_k \in S_{1,-1}} w_k + \sum_{e_k \in S_{-1,1}} w_k + \sum_{e_k \in S_{1,0}} w_k = 0$ and $\sum_{e_k \in S_{3,1}} w_k + \sum_{e_k \in S_{2,3}} w_k = 0$.

We will use the sample-path graph G obtained as described earlier. To prove the theorem, we show that those $S_{m,n}$ containing e_k for which $w_k = (c_1 - c_2) < 0$ are offset by edges with $w_k > 0$ in (25). From Table 1, these sets will be of the form $S_{m,m-2}$, with $m > 0$. Consider one such vertex, say m, and let there be r edges, $\{e_{k_1}, e_{k_2}, \ldots, e_{k_r}\}$, from m to $m - 2$ with negative weights. This means that $\sum_{e_k \in S_{m,m-2}} w_k \geq r(c_1 - c_2)$. Without loss of generality, we can assume that $k_1 < k_2 < \cdots < k_r$. Consider the two possibilities that arise.

Case 1: $m = 1$. Observe that $\tau_1 = 0$ and $\tau_{k_1} = 1$. This guarantees that there must be a right-directed edge e_k, an edge e_k with $l_k > 0$, with $0 < k < k_1$ into vertex 1. Now, consider the edge e_{k_l} from 1 to -1 with $l > 1$. Since $\tau_{k_l+1} = -1$ and $\tau_{k_{l+1}} = 1$, there must be a right-directed edge e_k into vertex 1 with $k_l < k < k_{l+1}$. The right-directed edges obtained above are due to transitions at different time epochs and, hence, are distinct. This shows the existence of at least r right-directed edges into vertex 1. There are two types of right-directed edge into vertex 1, edges due to events of types 3 and 12, each of which have $w_k = c_2 - c_1$. See Table 1. Thus,

$$\sum_{e_k \in S_{1,-1}} w_k + \sum_{e_k \in S_{-1,1}} w_k + \sum_{e_k \in S_{1,0}} w_k \geq 0. \tag{26}$$

Case 2: $m > 1$. Arguing as in Case 1, we can show that there are r right-directed edges into vertex m. The only possible right-directed edges into vertex $m, m > 1$, is

https://doi.org/10.1017/S0269964804184027 Published online by Cambridge University Press

TABLE 1. Combinations of Possible Events in $\tilde{J}^{(u)}$ and $\tilde{J}$ at Epoch t_{k+1} and the Corresponding τ_k, τ_{k+1}, l_k, and w_k

No.	Description of event at epoch t_k	τ_k	τ_{k+1}	l_k	w_k
1	Arrival joins same queue in both $\tilde{J}^{(u)}$ and $\tilde{J}$	*	*	0	0
2	Arrival joins queue 1 in $\tilde{J}^{(u)}$ and queue 2 in $\tilde{J}$	>0	*	-2	$-(c_2 - c_1)$
3	Arrival joins queue 2 in $\tilde{J}^{(u)}$ and queue 1 in $\tilde{J}$	<0	*	2	$(c_2 - c_1)$
4	Actual departure from same queue in $\tilde{J}^{(u)}$ and $\tilde{J}$	*	*	0	0
5	Actual departures from queue 1 in $\tilde{J}^{(u)}$ and queue 2 in $\tilde{J}$	<-1	≤ 0	2	0
6	Actual departure from queue 1 in $\tilde{J}^{(u)}$; $\tilde{J}$ is empty	<0	≤ 0	1	0
7	Actual departure from queue 2 in $\tilde{J}^{(u)}$; $\tilde{J}$ is empty	<0	<0	-1	0
8	Potential departure from queue 1 when $\tilde{J}^{(u)}$ is on the left threshold; actual departure from queue 1 in $\tilde{J}$	*	*	-1	0
9	Potential departure from queue 1 when $\tilde{J}^{(u)}$ is on the left threshold; actual departure from queue 2 in $\tilde{J}$	<0	≤ 0	1	0
10	Potential departure from queue 1 when $\tilde{J}^{(u)}$ is on the left threshold; $\tilde{J}$ is empty ([0,0])	0	0	0	0
11	Actual departures from queue 2 in $\tilde{J}^{(u)}$ and queue 1 in $\tilde{J}$	*	*	-2	0
12	Potential departure from queue 2 when $\tilde{J}^{(u)}$ is on the right threshold; actual departure from queue 2 in $\tilde{J}$	*	*	1	$c_2 - c_1$
13	Potential departure from queue 2 when $\tilde{J}^{(u)}$ is on the right threshold; actual departure from queue 1 in $\tilde{J}$	*	*	-1	$c_2 - c_1$
14	Potential departure from queue 2 when $\tilde{J}^{(u)}$ is on the right threshold; $\tilde{J}$ is empty	<0	<0	0	$c_2 - c_1$
15	Potential departure from either queue when both $\tilde{J}^{(u)}$ and $\tilde{J}$ are empty	0	0	0	0

Note: The asterisk denotes that these quantities can take either negative or positive values.

due to an event of type 12. So, $\sum_{e_k \in S_{m-1,m}} w_k \geq r(c_2 - c_1)$ and

$$\sum_{e_k \in S_{m,n-2}} w_k + \sum_{e_k \in S_{m-1,n}} w_k \geq 0. \tag{27}$$

See Figure 5 for an illustration of both cases.

For both of the cases, (25) can be written uniquely split into sums as in (26) and (27) and the theorem follows. ■

The stationary revenue rate $\tilde{\mathcal{R}}$ defined in (17) becomes, by the Law of Large Numbers,

$$\tilde{\mathcal{R}} = \lim_{N \to \infty} \frac{\tilde{R}_N}{N},$$

and from Theorems 7 and 8, we can now state the following theorem.

Theorem 9: *If systems $\tilde{\mathsf{J}}^{(l)}$, $\tilde{\mathsf{J}}$, and $\tilde{\mathsf{J}}^{(u)}$ start empty at $t = 0$, we have*

$$\tilde{\mathcal{R}}^{(u)} \geq \tilde{\mathcal{R}} \geq \tilde{\mathcal{R}}^{(l)}.$$

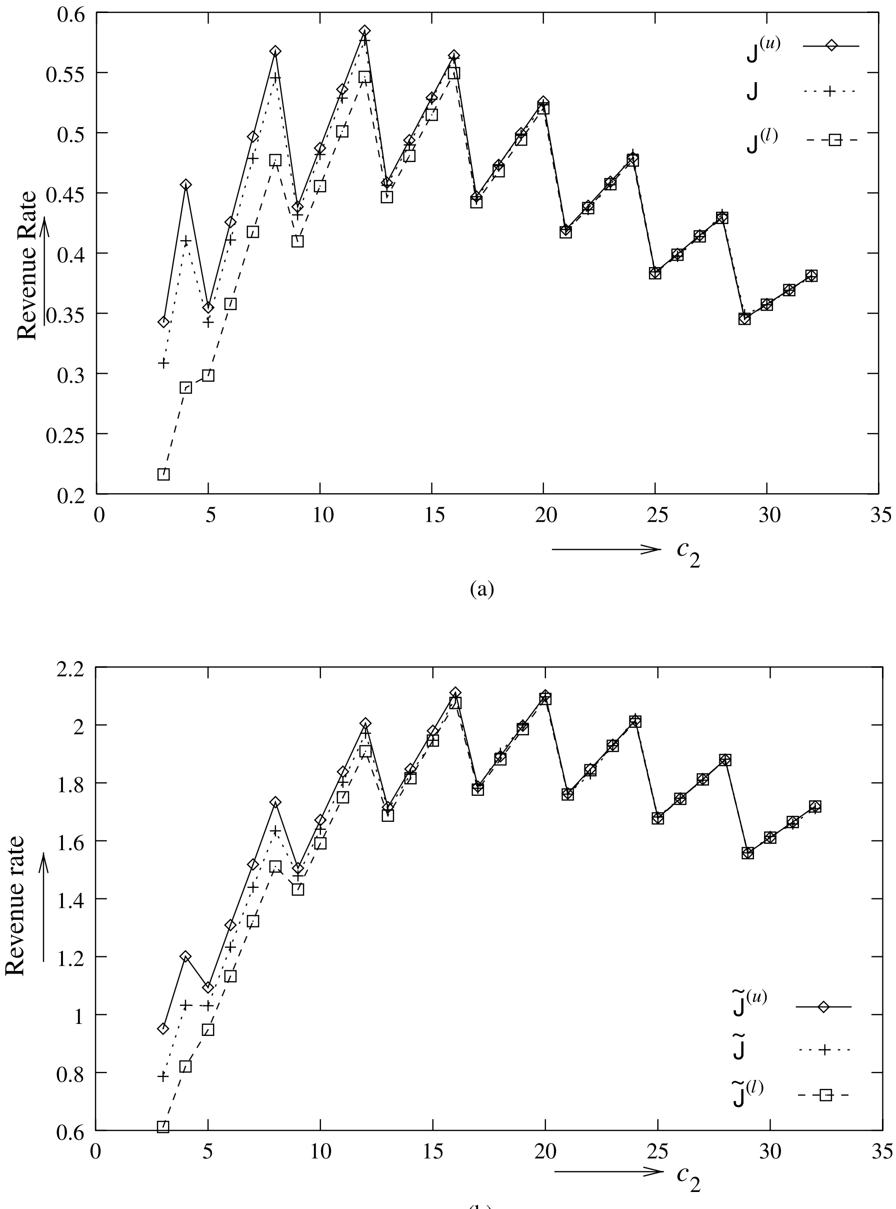
5. NUMERICAL EXAMPLES AND DISCUSSION

The primary motivation for the truncation method that we adopted was to allow us to numerically calculate the performance parameters for the $\mathsf{J}^{(u)}$, $\tilde{\mathsf{J}}^{(u)}$, $\mathsf{J}^{(l)}$, and $\tilde{\mathsf{J}}^{(l)}$ systems, which, in turn, allows us to obtain the bounds for the J and the $\tilde{\mathsf{J}}$ systems. As has been described in van Houtum et al. [21], the truncated model is a quasi-birth–death (QBD) process. The necessary and sufficient conditions for the stability of the truncated systems can be numerically computed from Theorem 3.1.1 of Neuts [14]. Further, by a proper choice of the thresholds, the bounds can be made fairly tight. The steady state distribution of the truncated system can be calculated using the method described in Theorem 3.1.1 of [14].

We present numerical results to show the tightness of the bounds. For the non-work-conserving system, we consider $\lambda_1 = \lambda_2 = 0.2$ and $\mu_1 = \mu_2 = 0.5$, $a_1 = 0.8$, $a_2 = 0.3$, $T_l = 0$, and $T_r = \lceil [(1 - a_2)/a_2](c_2 - c_1) \rceil + 2$. We compute the steady state distributions $\pi_{\mathbf{x}}^{(u)}$ and $\pi_{\mathbf{x}}^{(l)}$ and obtain the revenue rates $\mathcal{R}^{(u)}$ and $\mathcal{R}^{(l)}$ using (16). We also perform long-run simulations to obtain the steady state distribution $\pi_{\mathbf{x}}$ and $\mathcal{R}$ for the J system. Figure 6a shows $\mathcal{R}^{(u)}$, $\mathcal{R}$, and $\mathcal{R}^{(l)}$ as a function of c_2, the join price of the costly queue, with $c_1 = 0$. Similarly, for the work-conserving system, we plot $\tilde{\mathcal{R}}$, $\tilde{\mathcal{R}}^{(u)}$, and $\tilde{\mathcal{R}}^{(l)}$ as a function of c_2 for $\lambda_1 = \lambda_2 = 0.4$. $\mu_1 = \mu_2 = 0.5$, $a_1 = 0.8$, $a_2 = 0.3$, $T_l = 0$, and $T_r = \lceil [(1 - a_2)/a_2](c_2 - c_1) \rceil + 2$ in Figure 6b. Observe that the bounds are very good for both the work-conserving and non-work-conserving systems, especially for c_2 in the medium and high ranges.

An important observation is that the revenue is not an increasing or a convex function of the prices.

https://doi.org/10.1017/S0269964804184027 Published online by Cambridge University Press



(a)

(b)

Figure 6. Bounds on the revenue rates for the PMP and the Tirupati systems compared with results from a simulation model. (a) Revenue rates for the $\mathsf{J}^{(u)}$, J, and $\mathsf{J}^{(l)}$ systems for different values of c_2 with $c_1 = 0$, $\lambda_1 = \lambda_2 = 0.2$, $\mu_1 = \mu_2 = 0.5$, $a_1 = 0.8$, and $a_2 = 0.3$. (b) Revenue rates for the $\mathsf{J}^{(u)}$, $\tilde{\mathsf{J}}$, and $\tilde{\mathsf{J}}^{(l)}$ systems for different values of c_2 with $c_1 = 0$, $\lambda_1 = \lambda_2 = 0.4$, $\mu_1 = \mu_2 = 0.5$, $a_1 = 0.8$, and $a_2 = 0.3$.

https://doi.org/10.1017/S0269964804184027 Published online by Cambridge University Press

6. EXTENSIONS AND CONCLUSION

We now discuss some possible extensions of the results in the previous sections to $K, J > 2$. Consider the queue join process for an arriving customer of an arbitrary class, say class α. For any two queues i and j, the surface $x_j - x_i = [(1-a_\alpha)/a_\alpha](c_i - c_j)$ decides the preference among the queues i and j for class α customers; that is, if the state is on the "right" of this surface, then the cost of joining queue j is less than that of joining queue i, otherwise the cost for i is lower. For any two queues, there exists such a surface and it will be denoted by $\mathcal{S}_{ij}^{\alpha}$, where i and j are the queues and α is the customer class. For any customer class, only $K-1$ of $\mathcal{S}_{ij}^{l}$s are independent in the sense that they determine the remaining one. $\{\mathcal{S}_{K1}^1, \ldots, \mathcal{S}_{K(K-1)}^1\}$ will be the outermost $K-1$ surfaces and these will, in turn, determine the other surfaces. Also, all of these surfaces will be concurrent on a line given by

$$x_1 - \frac{(1-a_1)}{a_1}(c_K - c_1) = \cdots = x_{K-1} - \frac{(1-a_1)}{a_1}(c_K - c_{K-1}) = x_K.$$

This set of surfaces will together be called the attractors for customer class α. For J customer classes there will be J such parallel systems of surfaces. We first consider generalizing the stability results of Section 3 for $J, K > 2$.

6.1. Stability

First, consider the work-conserving system. The aggregated process $\{\tilde{Y}(t)\}$ defined earlier is a birth–death process. Arguing as earlier, it is stable if and only if $\sum_1^K \lambda_i < \sum_1^K \mu_i$ and transient if and only if $\sum_1^K \lambda_i > \sum_1^K \mu_i$. For the non-work-conserving JMCQ, the proof will require us to consider many cases and we conjecture that a similar result can be proved using quadratic Lyapunov functions.

6.2. Performance Bounds

As in Section 4.1, for a K-queue non-work-conserving JMCQ system, we consider the uniformized jump chain $\{\mathbf{x}_n\}_{n\geq 0}$. The revenue rate is given by

$$\mathcal{R} = \sum_{\mathbf{x}} c(\mathbf{x})\pi(\mathbf{x}),$$

where $c(\mathbf{x})$ is

$$c(\mathbf{x}) = \sum_{j:x_j\neq 0} \mu_j c_j.$$

We can verify that the entire state space has a precedence property; the state $\mathbf{m} \in \mathcal{Z}^K$ precedes state $\mathbf{m} + \mathbf{e_i}$ for $i = 1, \ldots, K$, where, as previously, $\mathbf{e_i}$ is a vector with 1 in the ith coordinate and 0 elsewhere. We use a truncated state space to obtain computable bounds for the revenue rate and other performance measures. The truncated

https://doi.org/10.1017/S0269964804184027 Published online by Cambridge University Press

state space $\mathcal{S}'$ is the set of all $\mathbf{x}(t) = [x_1, \ldots, x_K]$ such that $T_{il} \le x_i - x_K \le T_{ir}$ for $i = 1, 2, \ldots, K-1$ and $T_{il} \le \min_{j \in \{1,2,\ldots,j\}}[(1-a_j)/a_j](c_K - c_i)$ and $T_{ir} \ge \max_{j \in \{1,2,\ldots,j\}}[(1-a_j)/a_j](c_K - c_i)$. Let $\mathcal{S}'_{T_{il}}$ be the surface $x_i - x_K = T_{il}$ and let $\mathcal{S}'_{T_{ir}}$ be the surface $x_i - x_K = T_{ir}$. These are the "left" truncation surfaces and the "right" truncation surfaces, respectively. The upper and lower bounding systems like $\tilde{\mathsf{J}}^{(u)}$, $\tilde{\mathsf{J}}^{(l)}$, $\mathsf{J}^{(u)}$, and $\mathsf{J}^{(l)}$ are defined over this truncated state space as earlier; departures that cause the system to leave the $\mathcal{S}'$ are disallowed in $\tilde{\mathsf{J}}^{(u)}$ and $\mathsf{J}^{(u)}$, whereas in $\tilde{\mathsf{J}}^{(l)}$ and $\mathsf{J}^{(l)}$, they cause an additional simultaneous departure from queue K. By Theorem 1 of [21], $\mathcal{R}$ can be bounded by the revenue rates of $\mathsf{J}^{(u)}$ and $\mathsf{J}^{(l)}$ systems. We can also verify that the functions that capture the number in the system have the precedence property and, hence, we can find upper and lower bounds for the mean number in each queue.

The proof technique of obtaining performance bounds for the work-conserving JMCQ of Section 4.3 critically uses the fact that the state space is $\mathcal{Z}_+^2$. We believe that this methodology might not extend to models with more than two servers in a straightforward manner.

In conclusion, we have presented a generalization of the JSQ queuing system by allowing queues to prescribe join costs and customers to define cost functions in terms of the queue lengths seen on arrival and the join price. The stability results are discussed. We have also presented a technique to define truncated systems that will bound the original systems from above and below and are amenable to numerical calculations of the relevant performance measures using matrix geometric techniques developed for quasi-birth–death processes.

Acknowledgment

We thank the referee for very useful suggestions that helped in completely characterizing the stability conditions and also in the simplification of the proofs of our results on the non-work-conserving JMCQ.

References

1. Asmussen, S. (1987). *Applied probability and queues*. Chichester: Wiley.
2. Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., & Weiss, W. (1998). An architecture for differentiated services. *Internet Engineering Task Force, Request for Comments #2475*, Dec. 1998 (ftp://ftp.isi.edu/in-notes/rfc2475.txt).
3. Borkar, V.S. & Manjunath, D. (2002). Charge based control of DiffServ queues, submitted.
4. Boxma, O., Koole, G., & Liu, Z. (1996). Queueing theoretic solution methods for models of parallel and distributed systems. In O. Boxma & G. Koole (eds.), *Performance evaluation of parallel and distributed systems—Solution methods*, CWI Tract No. 105. Amsterdam, pp. 1–24.
5. Bremaud, P. (1998). *Markov chains, Gibbs fields, Monte Carlo simulation, and queues*. New York: Springer-Verlag.
6. Dube, P., Borkar, V.S., & Manjunath, D. (2002). Differential join prices for parallel queues: Social optimality, dynamic pricing algorithms and application to Internet pricing. In *Proceedings of IEEE INFOCOM 2002*.
7. Falkner, M., Devetsikiotis, M., & Lambadaris, I. (2000). An overview of pricing concepts for broadband IP networks. *IEEE Communications Surveys* 3: 2–13.
8. Foley, R.D. & McDonald, D.R. (2001). Join the shortest queue: Stability and exact asymptotics. *Annals of Applied Probability* 11(3): 569–607.

https://doi.org/10.1017/S0269964804184027 Published online by Cambridge University Press

9. Gibbens, R., Mason, R., & Steinberg, R. (2000). Internet service classes under competition. *IEEE Journal on Selected Areas in Communications* 18(12): 2490–2498.
10. Jain, R., Mullen, R., & Hausman, R. (2001). Analysis of Paris Metro pricing strategy for QoS with a single service provider. In *Proceedings of the Ninth International Workshop on Quality of Service (IWQoS 2001)*.
11. Kingman, J.F.C. (1962). Two queues in parallel. *Annals of Mathematical Statistics* 32: 1314–1323.
12. Manjunath, D., Goel, A., & Hemachandra, N. (2002). DiffServ node with join minimum cost queue policy: Analysis with multiclass traffic. In *Proceedings of IEEE Globecom 2002*, 3: 2573–2577.
13. Mertens, J.F., Samuel-Cahn, E., & Zamir, S. (1978). Necessary and sufficient conditions for recurrence and transience of Markov chains, in terms of inequalities. *Journal of Applied Probability* 15: 848–851.
14. Neuts, M.F. (1981). *Matrix geometric solutions in stochastic models*. Baltimore: Johns Hopkins University Press.
15. Norris, J.R. (1999). *Markov chains*. Cambridge: Cambridge University Press.
16. Odlyzko, A. (1999). Paris Metro pricing for the Internet. In *Proceedings of the ACM Conference on Electronic Commerce*, pp. 140–147.
17. Parekh, A.K. & Gallager, R.G. (1993). A generalized processor sharing approach to flow control in integrated services networks: The single node case. *IEEE/ACM Transactions on Networking* 1(3): 344–357.
18. Serfozo, R. (1999). *Introduction to stochastic networks*. New York: Springer-Verlag.
19. Tandra, R., Hemachandra, N., & Manjunath, D. (2004). DiffServ node with join minimum cost queue policy and multiclass traffic. *Performance Evaluation* 55: 69–91.
20. van Houtum, G.J., Zijm, W.H.M., Adan, I.J.B.F., & Wessels, J. (1998). Bounds for performance characteristics: A systematic approach via cost structures. *Stochastic models* 14: 205–224.
21. van Houtum, G.J., Adan, I.J.B.F., Wessels, J., & Zijm, W.H.M. (2000). Performance analysis of parallel identical machines with a generalized shortest queue arrival mechanism, *OR Spektrum* 23: 411–428.
22. Walrand, J. (1988). *Introduction to queueing networks*. Englewood Cliffs, NJ: Prentice-Hall.
23. Wolff, R. (1989). *Stochastic modeling and the theory of queues*. Englewood Cliffs, NJ: Prentice-Hall.

https://doi.org/10.1017/S0269964804184027 Published online by Cambridge University Press