Bounding blocking probabilities and throughput in queueing networks with buffer capacity constraints*

SUNIL KUMAR[1], R. SRIKANT[2] AND P. R. KUMAR[3]

[1] *Graduate School of Business, Stanford University, Stanford, CA 94305-5015.*
E-mail: skumar@leland.Stanford.edu

[2] *Department of General Engineering, and the Coordinated Science Laboratory,*
University of Illinois, 1308 West Main Street, Urbana, IL 61801.
E-mail: srikant@sambar.csl.uiuc.edu

[3] *Department of Electrical Engineering, and the Coordinated Science Laboratory,*
University of Illinois, 1308 West Main Street, Urbana, IL 61801.
E-mail: prkumar@decision.csl.uiuc.edu

We propose a new technique for upper and lower bounding the throughput and blocking probabilities in queueing networks with buffer capacity constraints, i.e, some buffers in the network have finite capacity. By studying the evolution of multinomials of the state of the system in its assumed steady state, we obtain constraints on the possible behavior of the system. Using these constraints, we obtain linear programs whose values upper and lower bound the performance measures of interest, namely throughputs or blocking probabilities. The main advantages of this new technique are that the computational complexity does not increase with the size of the finite buffers and that the technique is applicable to systems in which some buffers have infinite capacity. The technique is demonstrated on examples taken from both manufacturing systems and communication networks. As a special case, for the M/M/s/s queue, we establish the asymptotic exactness of the bounds, i.e, that the bounds on the blocking probability asymptotically approach the exact value as the degree of the multinomials considered is increased to infinity.

Subject classification: AMS(MOS) 60K25, 90B12, 68M20.

Keywords: Queueing Networks, Communication Networks, Performance Evaluation, Finite Buffers, Blocking.

*The research reported here has been supported in part by the U. S. Army Research Office under Grant No. DAAL 03-91-G-0182, the National Science Foundation under Grant No. ECS-94-05371, the Joint Services Electronics Program under Contract No. N00014-96-1-0129 and an NSF CAREER Award NCR 9701525.

1 Introduction

The study of queueing networks with buffer capacity constraints, i.e., with finite buffers, has quite a long history [15, 7]. Unfortunately, there is a dearth of results that are both definitive and useful in this area. Most of the proposed techniques suffer from one of the following two drawbacks. First, some of the techniques scale very poorly with the size of the buffers and thus suffer from high computational complexity. Second, techniques which attempt to avoid the first drawback tend to be approximations (mainly aggregations and decompositions) which are not always theoretically justified and whose reliability cannot always be established.

In this paper, we propose a new technique for estimating the performance of queueing networks with buffer capacity constraints i.e, queueing networks in which some of the buffers have finite capacity. Two performance measures of interest for such systems are the throughput and the blocking probabilities of various customer classes. Our technique allows us to obtain upper and lower bounds on such performance measures for a wide variety of systems.

The proposed technique extends the technique of obtaining bounds on the mean delay and throughput of queueing networks described in [10]. It can be described as follows. Assuming the existence of a steady state, we study the evolution of multinomials of the state of the system in steady state. Knowing that such multinomials, if integrable with respect to the invariant measure, cannot evolve in the mean, we obtain constraints on the possible behavior of the system. Using these constraints, with a suitable choice of variables, we obtain linear programs whose values upper and lower bound the performance measures of interest, namely throughput or blocking probabilities. This differs from [10] in that the performance measures on which bounds are obtained are probabilities, and in that multinomials rather than quadratics are used.

The main advantage of this new technique is that the computational complexity does not increase with the size of the finite buffers. Also, the technique is applicable to systems in which some buffers have infinite capacity. In some cases the bounds obtained are asymptotically exact, i.e., they approach the exact value as the degree of the multinomial considered increases. Another advantage of the method is that the main computation involved is the solution of a linear program which is rather routine with current software.

In Section 2, we present the model of the queueing network with capacity constraints which will be used in this paper. The model is sufficiently general to incorporate various problems of interest in both manufacturing systems and communication networks. In Section 3, we discuss the main idea behind the technique and obtain the basic linear programs. In Section 4, we illustrate the technique on some examples. First, we obtain bounds on the throughput of a two machine open reentrant line in which two of the buffers are infinite and one buffer has finite capacity. Second, we show that the the bounds are aymptotically exact for an $M/M/s/s$ loss model. While the blocking probability of $M/M/s/s$ system is given by the classical Erlang-B formula, the purpose of choosing this example is to illustrate how one might prove that the the bounds are asymptotically exact and how efficient recursions for the bounds can be constructed using the analytic solution of the LP's. Third, we consider tandem finite buffer queues with Poisson arrivals to illustrate the possibility that the bounds can give the exact throughput in a finite number of steps in some cases. We present a conjecture that in the case of the two stage tandem queue, the LP bounds achieve the exact value when some large but finite number of multinomials are considered. Next, we consider a single finite-buffer queue with Markov-modulated Poisson input to illustrate the fact that the bounds can be useful for non-Poisson arrival processes. Finally, we apply the technique to study a problem of interest in communication networks. We consider a multiserver link accessed by two call classes whose admission is controlled by a trunk reservation scheme, and obtain lower and upper bounds for the blocking probabilities of both classes. We present an example of a multirate communication link where the LP bounds out perform the commonly used approximation suggested by Bean, Gibbens and Zachary (BGZ) [1]. However, in most cases, the LP bounds do not perform better than the BGZ approximation, suggesting the need for further work on refining the LP bounds. Finally, in Section 5, we present some conclusions and directions for future work.

2 Model of a General Markovian Network

We now present a model of a Markovian queueing network which is sufficiently general incorporate various problems of interest in both manufacturing systems and communication networks. The network consists of S service stations. Each station $\sigma \in \{1, 2, ..., S\}$ has s_σ servers. Customer classes are considered to be distinguished by the buffer the customers reside in. Thus we need only describe the buffers. There are L buffers, $\{b_1, b_2, ..., b_L\}$. Each buffer b_j has a capacity $c_j \leq +\infty$. Customers in buffer b_j are served by servers at the service station $\sigma(j) \in \{1, 2, ..., S\}$. Each customer in b_j requires n_j servers simultaneously for an amount of time which is i.i.d. and exponentially distributed with mean $m_j = \frac{1}{\mu_j}$. We assume that n_j divides $s_{\sigma(j)}$ for all j. Upon completing service at buffer b_j, a customer wishes to move to buffer b_k with probability p_{jk} or wishes to leave the system with probability $1 - \sum_k p_{jk}$. (The customer wishing to move to b_k may however not be able to do so if it is blocked; this is described in the next paragraph.) Potential customers arrive from outside the system to buffer b_j as a Poisson process of rate γ_j. (They too can be blocked and lost.) All arrival streams, service times and customer wishes are assumed to be independent.

Let $x_j(t)$ denote the number of customers in buffer b_j at time t. Buffer b_j admits customers at time t only if the *state* of the system $x(t) = (x_1(t), ..., x_L(t))$ lies in an admission region $A_j \subseteq \mathcal{X}$, where the state space $\mathcal{X}$ is given by

$$\mathcal{X} = ([0, c_1] \times [0, c_2] \times \cdots \times [0, c_L]) \bigcap \mathcal{Z}^L.$$

Above, when $c_j = +\infty$, the interval $[0, c_j]$ is interpreted as $[0, +\infty)$. A_j can be used not only to capture the capacity of the buffers but also to model the admission control scheme employed. Any *stationary* admission control, i.e., one that depends only on the state $x(t)$ of the system, can be modeled in this fashion. If $x(t) \notin A_j$, an exogenous arrival at time t to buffer b_j is lost. If a customer completing service at time t at some buffer b_k wishes to move to buffer b_j, but $x(t) \notin A_j$, then there are two scenarios. In the first scenario, which we will call the *loss model*, the customer in b_k is lost, i.e, it exits the system. In the second scenario, which we will call the *blocking model*[1], the customer remains in buffer b_k and undergoes another service.

[1] This blocking mechanism is usually called "communication blocking."

The $s_{\sigma(j)}$ servers at service station $\sigma(j)$ are assigned to customers according to a scheduling policy in the following fashion.

1. The scheduling policy $\mathcal{S}$ employed is a map from $\mathcal{X}$ to $\mathcal{Z}_+^L$ which assigns $w_j(x)n_j$ servers to each buffer b_j when the state is x, that is, $w_j(x)$ customers in b_j are served. Thus, we restrict our attention to policies that only depend on the state x and, hence, are termed *stationary* policies.When x changes, it is assumed that the change in the number of servers being assigned to b_j is made in a preemptive fashion.

2. The scheduling policy is assumed to be *nonidling*. This means that every server that can be utilized is utilized. The decision $w(x) = (w_1(x), ..., w_L(x))$ on how the servers are employed satisfies the following constraints:

$$w_j(x) \leq x_j \text{ for all } x \text{ and } j, \text{ and}$$
$$0 \leq s_\sigma - \sum_{\{j:\sigma(j)=\sigma\}} w_j(x)n_j < \min_{\{j:\sigma(j)=\sigma \text{ and } x_j - w_j(x)>0\}} n_j. \tag{1}$$

Then $w_j(x)n_j$ servers are assigned to b_j, and are assumed to be working on the first $w_j(x)$ customers and the new arrivals to b_j are assumed to always be in the last position.

We note that under the assumptions made above the queueing network is a Markov chain with state space $\mathcal{X}$.

Last, we consider the performance measures of interest. For the queueing networks with capacity constraints, i.e., when one or more buffers have finite capacity, the natural performance measure of interest is λ_j, which is defined as the rate of departures from buffer b_j in steady state. Thus,

$$\lambda_j = E[\mu_j w_j(x)].$$

This is the performance measure which we will analyze using the technique outlined in the sequel. In some cases, we can relate the throughput to the blocking probability p_j of the buffer b_j. For example, if all arrivals to b_j were exogenous (i.e., no transfers from other buffers), then we would have

$$\lambda_j = (1 - p_j)\gamma_j.$$

3 Linear Program Bounds

We begin by describing the main idea of the technique. Assuming stability and examining the consequence of a steady state for general functionals of the state, we obtain a set of constraints on the possible behavior of the system. Then we attempt to obtains bounds on the performance measures by solving constrained optimization problems. With a suitable choice of variables, we obtain a set of linear equality constraints that determine the performance of the system. Further, additional knowledge about the particular system under consideration gives an augmenting set of linear equality and inequality constraints. Together, these allow us to bound the performance, either above or below, by solving a linear program. This method can be considered an extension of an idea first proposed by

Kumar and Kumar [10] and Bertsimas, Paschalidis and Tstisiklis [2], simultaneously and independently.

Consider the general queueing network of Section 2. Consider a "Lyapunov" functional $V : \mathcal{X} \to \mathcal{R}$. We make two assumptions about the network and V.

A1. The underlying Markov chain of the queueing network is positive recurrent.

A2. Under the unique[2] invariant probability measure, V is integrable.

The first assumption is not without loss of generality. In open queueing networks in which some of the buffers have infinite capacity, establishing that the Markov chain is positive recurrent is nontrivial and one usually attempts to do this through some Lyapunov argument. There has been considerable progress lately on this issue, see, for example, [3, 4, 11, 12].

The main idea of the Lyapunov approach to bounding performance is the following. If **A1, A2** hold, then in steady state we must have

$$E[V(X(t))] = E[V(X(t'))] \text{ for all } t, t', \tag{2}$$

where $X(t)$ denotes the state on the system at time t. Now, (2) is a statement about the infinitesimal generator of the underlying Markov chain. In order to use (2), we resort to a discretization technique known as uniformization [13]. Here, we use an exponential clock with rate no slower than the fastest exit rate out of a state in the Markov chain. For the Markov chain of the network described in Section 2, an acceptable rate is $\sum_{j=1}^{J}(\gamma_j + \frac{s_{\sigma(j)}}{n_j}\mu_j)$. We then sample the system at the random times $\{\tau_n\}$ when this clock goes off. This sampling scheme preserves the steady state information of the original chain. With uniformization, (2) becomes

$$E[V(X(\tau_{n+1}))] = E[V(X(\tau_n))] \text{ for all } n, \tag{3}$$

which then constrains the possible behavior of the system. The uniformization scheme and the Markov property of the discretized chain allow us to express $E[V(X(\tau_{n+1}))]$ in terms of $X(\tau_n)$. Now one tries to rewrite (3), using a suitable choice of variables, as a tractable constraint. One is free to use a family of functionals $\{V\}$ to obtain a set of constraints. If the performance measure of interest can also be expressed in terms of these variables, one can obtain lower and upper bounds on the performance measure by solving constrained optimization problems. This is the main idea behind our approach.

The key to implementing this idea is the choice of $\{V\}$ as well as the variables on which the constraints will be obtained. The choice is a trade-off between tractability of the resulting optimization problem and the quality of the bounds. Also, if the variables introduced have convenient probabilistic interpretation, we can refine the constraints by adding additional constraints which result from the nature of the policy, or any other usable information about system. We formalize these ideas in the subsequent sections.

In this paper we will concentrate on one particular family of functionals, namely, multinomials. Thus **A1, A2** translate to the existence of a unique steady state with

[2] If the state space does not have a single closed communicating class, we fix an initial condition.

finite multinomial moments. Consider a multinomial of degree d of the state, of the form

$$m_d(x) = \sum_{\{d_1,d_2,\ldots d_L : \sum_k d_k \le d\}} m_{(d_1,\ldots,d_L)} x_1^{d_1} x_2^{d_2} \cdots x_L^{d_L},$$

where $m_{(d_1,\ldots,d_J)}$ are the coefficients. Then from (3), we can conclude that

$$E[m_d(x(\tau_n))] = E[m_d(x(\tau_{n+1}))] \text{ for all } m_d(x), \text{ for all } d. \tag{4}$$

This is equivalent to

$$\begin{aligned} E[x_1^{d_1}(\tau_n) x_2^{d_1}(\tau_n) \cdots x_L^{d_L}(\tau_n)] = \quad & E[x_1^{d_1}(\tau_{n+1}) x_2^{d_2}(\tau_{n+1}) \cdots x_L^{d_L}(\tau_{n+1})] \\ & \text{for all } d_1, d_2, ..., d_J \in \mathcal{Z}_+. \end{aligned} \tag{5}$$

Let us now illustrate how one can use (5) to obtain constraints on the behavior of the system. We will describe this just for the *loss* model of Section 2. The *blocking* model can be handled in a similar fashion. Let us rescale time so that $\sum_{j=1}^{L}(\lambda_j + \frac{s_{\sigma(j)}}{n_j}\mu_j) = 1$. For simplicity, we assume that there is no self-feedback, i.e., $p_{ii} = 0$ for all i. Consider the special case of (5), where $d_k = 0$ for all $k \neq i$ and $k \neq j$, i.e.,

$$E[x_i^{d_i} \cdots x_j^{d_j}(\tau_n)] = E[x_i^{d_i} x_j^{d_j}(\tau_{n+1})] \text{ for some } i, j, d_i, d_j. \tag{6}$$

In order that we may exploit (6), we need to understand the evolution of $x_i^{d_i} x_j^{d_j}$ for each i, j and d_i, d_j. This can be obtained as (see also [10])

$$\begin{aligned} E[x_i^{d_i} x_j^{d_j}(\tau_{n+1}) \mid x(\tau_n) = x] \quad = \quad & x_i^{d_i}(\tau_n) x_j^{d_j}(\tau_n) + \gamma_i 1_{A_i}(x) \left[(x_i+1)^{d_i} x_j^{d_j} - x_i^{d_i} x_j^{d_j}\right] \\ + \quad & \mu_i w_i(x)(1-p_{ij}) \left[(x_i-1)^{d_i} x_j^{d_j} - x_i^{d_i} x_j^{d_j}\right] \\ + \quad & \mu_i w_i(x) p_{ij} 1_{A_j}(x) \left[(x_i-1)^{d_i} (x_j+1)^{d_j} - x_i^{d_i} x_j^{d_j}\right] \\ + \quad & \mu_i w_i(x) p_{ij} 1_{A_j^c}(x) \left[(x_i-1)^{d_i} x_j^{d_j} - x_i^{d_i} x_j^{d_j}\right] \\ + \quad & \gamma_j 1_{A_j}(x) \left[x_i^{d_i} (x_j+1)^{d_j} - x_i^{d_i} x_j^{d_j}\right] \\ + \quad & \mu_j w_j(x)(1-p_{ji}) \left[x_i^{d_i} (x_j-1)^{d_j} - x_i^{d_i} x_j^{d_j}\right] \\ + \quad & \mu_j w_j(x) p_{ji} 1_{A_i}(x) \left[(x_i+1)^{d_i} (x_j-1)^{d_j} - x_i^{d_i} x_j^{d_j}\right] \\ + \quad & \mu_j w_j(x) p_{ji} 1_{A_i^c}(x) \left[x_i^{d_i} (x_j-1)^{d_j} - x_i^{d_i} x_j^{d_j}\right]. \end{aligned} \tag{7}$$

Combining (7) with (6) and defining the variables

$$\begin{aligned} r_{d_i,d_j}^{i,j} \quad &:= \quad E[x_i^{d_i} x_j^{d_j}], \\ z_{k,d_i,d_j}^{i,j} \quad &:= \quad E[1_{A_k}(x) x_i^{d_i} x_j^{d_j}], \\ y_{k,d_i,d_j}^{i,j} \quad &:= \quad E[w_k(x) x_i^{d_i} x_j^{d_j}], \text{ and} \\ u_{k,l,d_i,d_j}^{i,j} \quad &:= \quad E[w_k(x) 1_{A_l}(x) x_i^{d_i} x_j^{d_j}], \end{aligned}$$

we obtain the *linear* constraint

$$\begin{aligned}
& \gamma_i \sum_{k=0}^{d_i-1} \binom{d_i}{k} z^{i,j}_{i,k,d_j} + \mu_i \sum_{k=0}^{d_i-1} (-1)^{d_i-k} \binom{d_i}{k} y^{i,j}_{i,k,d_j} \\
- \ & \mu_i p_{ij} \sum_{k=0}^{d_i-1} (-1)^{d_i-k} \binom{d_i}{k} u^{i,j}_{i,j,k,d_j} \\
+ \ & \mu_i p_{ij} \sum_{\{k=0;k+l<d_i+d_j\}}^{d_i} \sum_{l=0}^{d_j} (-1)^{d_i-k} \binom{d_i}{k} \binom{d_j}{l} u^{i,j}_{i,j,k,l} \\
+ \ & \gamma_j \sum_{k=0}^{d_j-1} \binom{d_j}{k} z^{i,j}_{j,d_i,k} + \mu_j \sum_{k=0}^{d_j-1} (-1)^{d_j-k} \binom{d_j}{k} y^{i,j}_{j,d_i,k} \\
- \ & \mu_j p_{ji} \sum_{k=0}^{d_j-1} (-1)^{d_j-k} \binom{d_j}{k} u^{i,j}_{i,j,d_i,k} \\
+ \ & \mu_i p_{ij} \sum_{\{k=0;k+l<d_i+d_j\}}^{d_i} \sum_{l=0}^{d_j} (-1)^{d_i-k} \binom{d_i}{k} \binom{d_j}{l} u^{i,j}_{i,j,k,l} = 0 \qquad (8)
\end{aligned}$$

Similar equality constraints can be obtained for every term of the form $x_1^{d_1} x_2^{d_2} \cdots x_L^{d_L}$. These equality constraints are linear with a suitable choice of variables, similar to z, y and w above. Note that the performance measure of interest is

$$\lambda_i = E[\mu_i w_i(x)] = \mu_i y^{i,j}_{i,0,0}. \qquad (9)$$

Thus, the performance measure is one of the constrained variables.

We can also obtain inequality constraints on the variables in various ways. Rather than providing a general list of all possible inequalities, we provide illustrations of the various kinds of inequalities which may be obtained. Specific knowledge in a particular application helps provide additional inequalities. In the examples considered in the subsequent section, we will highlight how such specific constraints can be obtained.

First, there are the obvious domination constraints. For example,

$$u^{i,j}_{k,l,d_i,d_j} \leq y^{i,j}_{k,d_i,d_j} \text{ for all } i, j, k, l, d_i, d_j. \qquad (10)$$

We also obtain inequality constraints due to the nonidling nature of the scheduling policy used. For example, consider the case when $s_\sigma = 1$ for some σ (which, of course, means that $n_j = 1$ for all j such that $\sigma(j) = \sigma$). Now consider $\sigma' \neq \sigma$ such that $s_{\sigma'} = 1$ also. Then we have, similar to the inequalities obtained in [10],

$$\sum_{\{k:\sigma(k)=\sigma'\}} y^{i,j}_{k,d_i,d_j} \leq \sum_{\{k:\sigma(k)=\sigma\}} y^{i,j}_{k,d_i,d_j} \text{ for all } i \text{ with } \sigma(i) = \sigma, \text{ for all } d_i \geq 1, j, d_j. \qquad (11)$$

We also obtain "Chebyshev" type inequalities. For example, suppose that for buffer b_i, we have an admission rule which specifies the set A_i such that $A_i^c \subseteq \{x : x_i \geq C\}$, where C is some constant. Also suppose that $w_i(x) = x_i$. Then, we must have

$$E[(1 - 1_{A_i}) w_i(x) x_i^{d_i}] \geq C^{d_i+1} P(x \in A_i^c).$$

But $P(x \in A_i^c) = 1 - z^{i,j}_{i,0,0}$ and hence we obtain

$$y^{i,j}_{i,d_i,0} - u^{i,j}_{i,i,d_i,0} \geq C^{d_i+1}(1 - z^{i,j}_{i,0,0}) \text{ for all } d_i. \tag{12}$$

Also, we have nonnegativity constraints

$$z^{i,j}_{k,d_i,d_j} \geq 0, y^{i,j}_{k,d_i,d_j} \geq 0, u^{i,j}_{k,l,d_i,d_j} \geq 0 \text{ for all } i,j,k,l,d_i,d_j. \tag{13}$$

Last, we note that there are some equality constraints obtained because of the way we have defined these variables. For example,

$$y^{i,j}_{k,0,0} = y^{l,m}_{k,0,0} \text{ for all } i,j,k,l,m. \tag{14}$$

Thus we have defined a set of variables and obtained a set of *linear* equality and inequality constraints on these variables. We have also established in (9) that the performance measure of interest is a linear combination of some of these variables. Thus, we can obtain bounds on the value of the performance measure by solving linear programs. This result is summarized in the theorem below.

Theorem 1
The throughput λ_i of any buffer b_i, is bounded above (below) by the solution to the linear program

$$\max\ (\min)\ \mu_i y^{i,j}_{i,0,0},$$

subject to the equality constraints of the form of (8) and (14), inequality constraints of the forms illustrated in (10), (11), (12), and the nonnegativity constraints (13).

In the next section, we consider the application of this theorem to various problems of interest in manufacturing systems and communication networks.

4 Applications

4.1 A Reentrant Line with a Finite Buffer

We now illustrate the application of the technique to a system containing both finite and infinite buffers. Methods involving finite Markov chain computations cannot be used because the resulting state space is infinite when some buffers have infinite capacity.

Consider the system shown in Figure 1. It is an open reentrant line [9], in which one of the buffers has finite capacity. It consists of two service stations ($S = 2$) and three buffers ($L = 3$). The number of servers at each machine is one ($s_1 = s_2 = 1$) and obviously each customer can ask for one server ($n_1 = n_2 = n_3 = 1$). Exogenous arrivals occur only to buffer b_1 at rate $\gamma_1 = \lambda = 0.3$. The mean service times are all 1 (i.e., $\mu_1 = \mu_2 = \mu_3 = 1$). Buffers b_1 and b_2 have infinite capacity, i.e., $c_1 = c_2 = +\infty$, but buffer b_3 is finite, $c_3 < \infty$. The only non-zero routing probabilities are $p_{12} = p_{23} = 1$. Buffer b_3 admits new customers if $x_3 < c_3$. The departures from b_2 which are not admitted to b_3 are assumed to exit the system. Thus, we use the loss model here[3]. We are interested in obtaining bounds on $\gamma_3 = E[w_3(x)]$, under all nonidling and stationary policies.

[3] This reinforces the assumption of stability of the system, see [11]

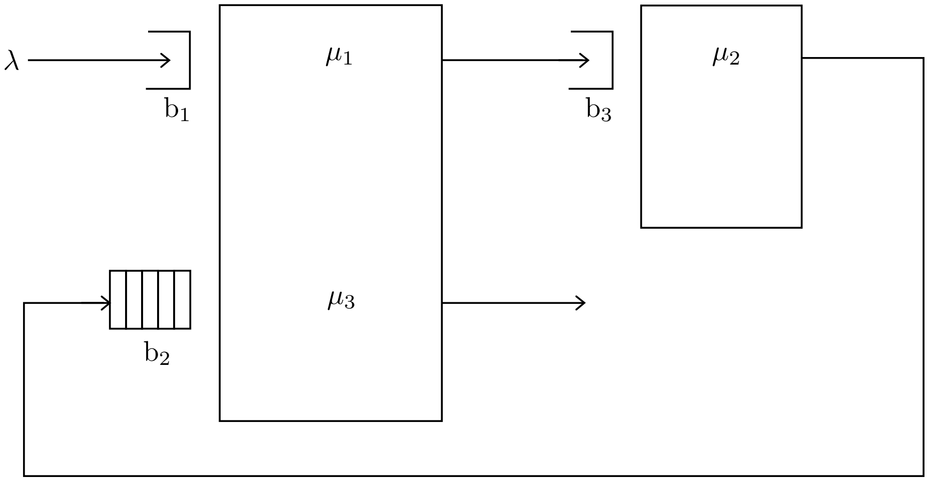

Figure 1: Reentrant Line with a Finite Buffer and Two Infinite Buffers

The equality constraints for this example are more general than (8). But, we will restrict ourselves to quadratics, in which case (8) is sufficient. The only z's that appear are of the form $z^{1,j}_{1,0,1}$. They can be written using the nonidling constraints as

$$z^{1,j}_{1,0,1} = \sum_{\{k:\sigma(k)=\sigma(j)\}} y^{1,j}_{k,0,1}.$$

Thus, we need not consider z's. We need only consider the y and u variables here.

The domination constraints (10) and nonnegativity constraints (13) obviously apply here.

Also, we have nonidling constraints of the form (11) here:

$$\sum_{\{k:\sigma(k)=\sigma(j)\}} y^{1,j}_{k,0,1} \geq \sum_{\{k:\sigma(k)\neq\sigma(j)\}} y^{1,j}_{k,0,1}.$$

The Chebyshev constraints of the form (12) are also applicable here. For example,

$$u^{2,3}_{2,3,0,1} + \left(\frac{\gamma_1 - \lambda_3}{\mu_2}\right) c_3 = y^{2,3}_{2,0,1}.$$

Note that the throughput which has to be estimated is λ_3, the rate of departures from the last buffer. Figure 2 plots the lower bound on the throughput as c_3 varies. From this we can infer, for example, that under any nonidling stationary policy, a buffer of size 10 guarantees a throughput of at least 0.27.

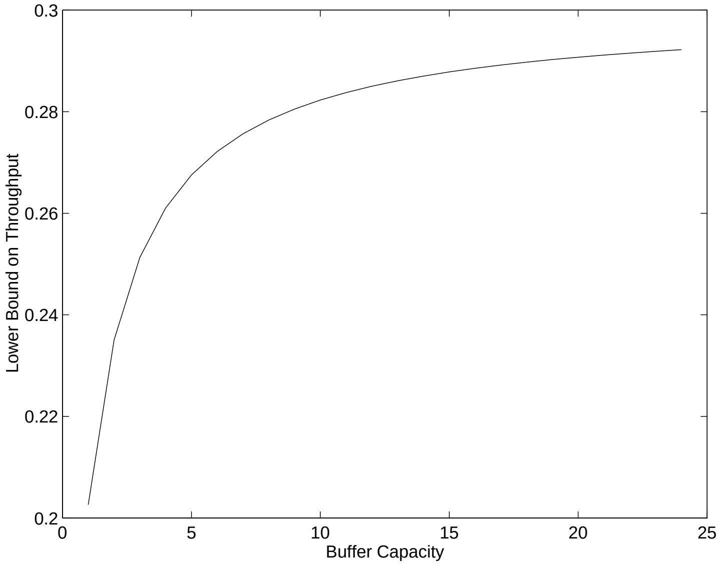

Figure 2: Lower Bound on Throughput of Reentrant Line

4.2 The M/M/s/s Queue

Here we consider a well–known example from classical queueing theory, the M/M/s/s queue. Its invariant measure is known and is given by the celebrated Erlang B–formula [8].

This corresponds to the model described in Section 2 with the following choice of parameters: $S = 1$ (one service station), $L = 1$ (one buffer), $p_{11} = 0$ (no self–feedback), $s_1 = s$ (s servers), $n_1 = 1$ (each customer asks for one server) and $c_1 = s$ (buffer has capacity s). Arrivals take place at rate γ and the service rate (at every server) is μ. The customers are admitted if there is room, i.e., $A_1 = \{x : x < s\}$. In this case, every customer admitted to and currently in the system is being served, and hence $w_1(x) = x_1 = x$. Note that the variables u of the previous section will not be required. Also,

$$y_{1,n} = E[x^{n+1}],$$

and

$$z_{1,n} = E[(1 - 1_{x=s})x^n] = E[x^n] - s^n P(x = s) = y_{1,n-1} - s^n P(x = s).$$

Thus in this case the linear program of Theorem 1 can be exclusively written in terms of the variables $y_{1,n}$ and the variable $p := Prob(x = s)$.

The equality constraints are obtained by substituting various values of d_i, keeping $d_j = 0$ and $p_{ij} = 0$ in (8). In order to keep the linear program finite dimensional, we use the first N equality constraints, which correspond to $d_i = 1, 2, ..., N$.

The family of equality constraints obtained from (8) corresponding to $d_i = n$ is

$$ny_{1,n-1} - \sum_{k=1}^{n-1} \left[\binom{n}{k} \rho - (-1)^{n-k} \binom{n}{k-1} \right] y_{1,k-1} = \qquad \rho - \rho \left[(s+1)^n - s^n \right] p,$$

$$\text{for } 1 \leq n \leq N, \tag{15}$$

where $\rho = \frac{\gamma}{\mu}$.

Consider the family of Equations (15). We have $N+1$ variables $y_{1,0}, y_{1,1}, ..., y_{1,N-1}$ and p, and N equations. Thus, we can parametrize the N variables $y_{1,0}, y_{1,1}, ..., y_{1,N-1}$ in terms of p. Then we obtain

$$y_{1,n-1} = \alpha_n - \beta_n p \text{ for } 1 \leq n \leq N, \tag{16}$$

where α_n and β_n, $n = 1, 2, ..., N$ are constants not depending on p which are easily determined from (15) as follows.

Consider $n = 1$. Substituting (16) in (15) gives

$$(\alpha_1 - \rho) - (\beta_1 - \rho)p = 0. \tag{17}$$

Since α_1 and β_1 do not depend on p, one has

$$\alpha_1 = \rho \text{ and } \beta_1 = \rho.$$

Now we continue this procedure of substituting (16) for $y_{1,0}$ and $y_{1,1}$ back in (15) for $n = 2$. This yields

$$(2\alpha_2 - (2\rho+1)\alpha_1 - \rho) - (2\beta_2 - (2\rho+1)\beta_1 - \rho(2s+1))\, p = 0.$$

As argued above, this yields

$$\begin{aligned} 2\alpha_2 &= (2\rho+1)\alpha_1 + \rho, \text{ and} \\ 2\beta_2 &= (2\rho+1)\beta_1 + \rho((s+1)^2 - s^2). \end{aligned}$$

Continuing this procedure for $n = 3, 4, ...N$, we obtain

$$n\alpha_n = \sum_{k=1}^{n-1} \left[\begin{pmatrix} n \\ k \end{pmatrix} \rho - (-1)^{n-k} \begin{pmatrix} n \\ k-1 \end{pmatrix} \right] \alpha_k + \rho \text{ and} \tag{18}$$

$$n\beta_n = \sum_{k=1}^{n-1} \left[\begin{pmatrix} n \\ k \end{pmatrix} \rho - (-1)^{n-k} \begin{pmatrix} n \\ k-1 \end{pmatrix} \right] \beta_k + \rho\left[(s+1)^n - s^n\right] \text{ for } 1 \leq n \leq N. \tag{19}$$

Thus, as our only equality constraint, we retain

$$y_{1,N-1} = \alpha_N - \beta_N p, \tag{20}$$

where α_N and β_N are computed using the recursions (18) and (19) respectively.

Also, we have a Chebyshev type inequality of the form of (12),

$$s^N p \leq y_{1,N-1}. \tag{21}$$

We also have a size constraint

$$y_{1,N-1} \leq s^N. \tag{22}$$

Last, we have the nonnegativity constraints of the form of (13),

$$p \geq 0, y_{1,N-1} \geq 0. \tag{23}$$

Thus we have obtained the following analog of Theorem 1.

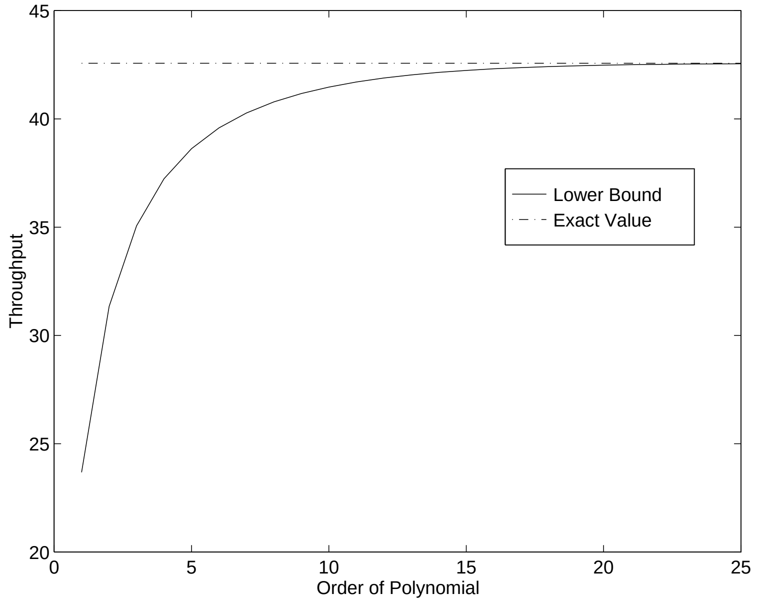

Figure 3: Lower Bound on Throughput of M/M/s/s Queue

Theorem 2
The throughput λ of a M/M/s/s queue is bounded below (above) by the solution to the linear program

$$\min(\max)\ \ \lambda = \gamma(1-p),$$

subject to (20), (21), (22) and (23), for each N.

Note that the upper bound on $p = Prob(x = s)$ provided by the LP is also an upper bound on the blocking probability by the PASTA property. Thus, in this example, we have obtained bounds on both throughput and blocking probabilities.

Figure 3 shows a plot of the lower bound on λ as N varies, for a system with $\gamma = 45$, $\mu_1 = 1$ and $s = 50$. Note that as N increases, the lower bound approaches the exact value of λ predicted by the Erlang B formula. This property, which we call *asymptotic exactness*, can be formalized as follows.

Theorem 3
The lower and upper bounds obtained by solving the LP involving $y_{1,N}$ and p converge to the exact value provided by the Erlang B formula as $N \uparrow \infty$.

Proof
First, we make a crucial observation. We note that (18) is the same as (8) with α_n denoting $E[Y^n]$, where Y is the number of customers in a $M/M/\infty/\infty$ system, i.e., Y is Poisson with parameter ρ. This can be verified easily. Thus, we obtain $\alpha_n > 0$ for all n

and

$$\lim_{N\to\infty} \frac{s^N}{\alpha_N} = 0. \tag{24}$$

Note that (20) must also hold for the actual moment $y_{1,N-1} = E[X^N]$ and blocking probability $p = \text{Prob }(X = s)$ of the M/M/s/s system. Thus

$$0 \leq \frac{E[X^N]}{\alpha_N} = 1 - \frac{\beta_N}{\alpha_N}\text{Prob }(X = s) \leq \frac{s^N}{\alpha_N}.$$

Taking limits above as $N \to \infty$, we obtain

$$\lim_{N\to\infty} \frac{\beta_N}{\alpha_N} = \frac{1}{\text{Prob }(X = s)}. \tag{25}$$

Thus, there exists a N' such that for all $N \geq N'$, $\beta_N \geq \alpha_N$, since Prob $(X = s) < 1$.

Now consider a relaxed version of the LP used in the statement of the theorem, in which the constraints are

$$\begin{array}{rcl} y_{1,N-1} & = & \alpha_N - \beta_N p, \\ y_{1,N-1} & \geq & s^N p, \\ y_{1,N-1} & \leq & s^N, \end{array}$$

for some $N \geq N'$. Then since $\beta_N \geq \alpha_N > 0$, the lower bound on p computed by the relaxed LP is

$$\underline{p} = \frac{\alpha_N - s^N}{\beta_N},$$

and the upper bound computed by the relaxed LP is

$$\bar{p} = \frac{\alpha_N}{s^N + \beta_N}.$$

The result then follows from (24) and (25). □

From Figure 3, we can see that the number of moments N which must be considered in order to obtain a reasonable approximation to γ_1 or p is much smaller than s. Also, we can obtain very efficient recursions for α_n and β_n as in (18,19) and obtain the bounds from them. Hence for large s, this scheme has $O(1)$ complexity (in s) as opposed to the Erlang formula which has $O(s)$ complexity. More generally, this illustrates the fact that the complexity of the LP approach does not grow with s. On the other hand, it has $O(N^2)$ complexity in N.

4.3 The Tandem Queue

In this section, we consider a problem which is very common in manufacturing, the tandem queue. Considerable effort has been expended in analyzing this system, see [15] for a collection of such works. The main problem in analyzing this system is that the state space of the Markov chain is of size $O(c_1c_2)$ and hence very large when c_1 and c_2 are large. We shall apply the LP approach to obtain bounds on the throughput and therefore the blocking probability of the system.

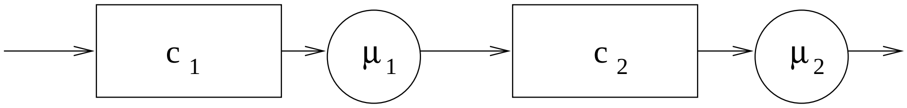

Figure 4: Tandem M/M/1/K Queue

Consider the tandem queue shown in Figure 4. This system is in the framework of Section 2. It corresponds to the case in which $S = L = 2$, $s_1 = s_2 = n_1 = n_2 = 1$, $\sigma(1) = 1$, $\sigma(2) = 1$, and the only nonzero routing probability is $p_{12} = 1$. Note that there is only one external arrival stream of rate $\gamma_1 = \lambda$. In this example we use the *blocking* model for transfers, i.e., if the second buffer is full when a customer in the first buffer completes service, then the customer undergoes another service. The performance measure of interest is the probability that an exogenous arrival to the system is not admitted. Note that

$$\text{Prob (Arrival Not Admitted)} = 1 - \frac{\lambda_1}{\gamma_1} = 1 - \frac{\lambda_2}{\gamma_1}.$$

Before we apply (8), we introduce additional variables as we have done in each of the previous examples. In this case, we define

$$p_{k_1,k_2} := \text{Prob } (X_1 = k_1, X = k_2).$$

Here p_{k_1,k_2} is the invariant measure of the underlying Markov chain. Obviously, to use all such p_{ij}'s would defeat the purpose of the LP approach since we would have more variables than states in the Markov chain. Instead, we use p_{0,k_2} and p_{c_1,k_2} for $k_2 = 0, 1, 2, ..., c_2$, as well as $p_{k_1,0}$ and p_{k_1,c_2} for $k_1 = 1, 2, ..., c_1 - 1$. Thus, we enumerate $2(c_1 + c_2)$ additional variables. It is easy to see that the equality constraints (8) can be rewritten linearly in terms of these variables and the variables $r_{ij} := E[X_1^i X_2^j]$. For example,

$$z_{1,i,j} = r_{i,j} - c_1^i \sum_{k_2=0}^{c_2} k_2^j p_{c_1,k_2},$$

and

$$y_{1,i,j} = r_{i,j} - 0^i \sum_{k_2=0}^{c_2} k_2^j p_{0,k_2}.$$

Also, we can rewrite the Chebyshev inequality constraints as

$$r_{i,j} \geq c_1^i \sum_{k_2=0}^{c_2} k_2^j p_{c_1,k_2} \quad + c_2^j \sum_{k_1=0}^{c_1-1} k_1^j p_{k_1,c_2}.$$

The usual nonnegativity constraints apply. Also note that the performance measure can be written using PASTA (which applies only to the first stage) as

$$\text{Prob (Arrival Blocked)} = \sum_{k_2=0}^{c_2} p_{c_1,k_2}.$$

Table 1
LP bounds on blocking probability for tandem $M/M/1/K$ queue

Moments	Lower Bound ($\times 10^{-2}$)	Upper Bound ($\times 10^{-2}$)	Error Bound%
6	2.3743	2.7523	15.92
8	2.4257	2.7240	8.46
10	2.6429	2.6629	0.76
12	2.6499	2.6556	0.21

Thus we can obtain bounds on the blocking probability using LP's in the variables r_{ij} and p_{k_1,k_2}. The bounds obtained using the LP involving the first N moments, i.e., using (8) for $d_i, d_j \leq N$ are tabulated in Table 1 for the case when $\gamma_1 = \lambda = 9$, $\mu_1 = 10$, $\mu_2 = 9$, $c_1 = 30$ and $c_2 = 25$. In this case, the Markov chain has 806 states, while the largest LP used has 500 variables. The table also tabulates a bound on the relative error computed as $\frac{\text{upper bound - lower bound}}{\text{lower bound}}$. If we only need an estimate with less than 10% error, for example, we could have used only eight moments, in which case the LP involves only 270 variables.

Note that the bounds are tight. Therefore, we are led to believe that an asymptotic exactness result like the one for the M/M/s/s queue may exist here as well. In fact, further analysis leads us to the following stronger conjecture.

Conjecture 1
The LP lower and upper bounds obtained using (8) for all $d_i + d_j \leq M$ are equal (and therefore equal to the exact value) when $M = 2(c_1 + c_2) - 1$.

The rationale for this conjecture is the following. Consider the equality constraints obtained using (8) with $d_i + d_j = n$ for some $n > 1$. We obtain $n + 1$ constraints, but introduce n new variables r_{pq} for $p = 0, 1, 2, ..., n - 1$, $q = n - 1 - p$. Also, we always have $2(c_1 + c_2)$ variables p_{k_1,k_2}. When $n = 1$, we obtain 2 constraints, but introduce no new variables (since $r_{00} = 1$). Thus using (8) for $n = 1, 2, 3, ..., 2(c_1 + c_2) - 1$, we have $(c_1 + c_2)[2(c_1 + c_2) + 1] - 1$ variables and $(c_1 + c_2)[2(c_1 + c_2) + 1] - 1$ equality constraints. Thus, if there is no loss of rank in this system of equations, we can exactly obtain all variables, thus exactly computing the probabilities. The open issue that has to be resolved with this argument is in establishing that there is no loss of rank, except for very specific choices of γ_1, μ_1 and μ_2.

The conjecture that we can obtain convergence of the bounds in a finite number of steps cannot hold in general. For example, it doesn't hold for the single $M/M/1/K$ queue, where the system of equations obtained from (8) never becomes square. Second, even when it is true, we have to solve a system of linear equations in approximately $2(c_1 + c_2)^2$ variables, whereas solving the Markov chain involves solving a system in only $(c_1 + 1)(c_2 + 1)$ variables.

4.4 The MMPP/M/1/s Queue

Consider a finite-buffer queue with $s - 1$ waiting spaces where customers arrive according to a Markov-modulated Poisson arrival process, and whose service times are exponentially distributed with mean $1/\mu$. Specifically, we assume that the arrivals are generated by an $(n+1)$-state MMPP obtained by a superposition of n identical, independent sources, each of which produces arrivals according to an interrupted Poisson process. (The technique outlined below can be easily extended to more general MMPP arrival processes.) For each source, the transition rate from the *on* to *off* state is denoted by q_1, the transition rate from the *off* to *on* is denoted by q_0, and the arrival rate in the *on* state is denoted by ν. The arrival rate in the *off* state is zero. Let $x(t)$ denote the number of customers in the system, $\alpha(t)$ denote the number of sources in the on state at time t, and t_{ij} denote the steady-state expectation $E(x^i\alpha^j)$. Then, (8) becomes

$$\begin{aligned} 0 &= \nu \sum_{k=0}^{i-1} \binom{i}{k} t_{k,j+1} - \nu \sum_{k=0}^{i-1} \binom{i}{k} s^k \left[\sum_{l=0}^{n} l^{j+1} p_{x=s,\alpha=l} \right] + nq_1 \sum_{j=0}^{j-1} t_{ik} \binom{j}{k} n^{k-j} \\ &+ q_0 \sum_{k=0}^{j-1} \binom{j}{k} (-1)^{j-k} t_{i,k+1} - q_1 \sum_{k=0}^{j-1} \binom{j}{k} t_{i,k+1} \\ &+ \mu \sum_{k=0}^{i-1} \binom{i}{k} (-1)^{i-k} t_{kj} - \mu(-1)^i \sum_{l=0}^{n} l^j p_{x=0,\alpha=l}, \end{aligned} \tag{26}$$

where we have introduced the additional variables $p_{x=0,\alpha=l}$, $l = 0, 1, \ldots, n$, and $p_{x=s,\alpha=l}$, $l = 0, 1, \ldots, n$, denoting $\text{Prob}(x = 0, \alpha = l)$ and $\text{Prob}(x = s, \alpha = l)$, respectively. The Chebyshev inequality constraints are

$$t_{ij} \geq s^i \sum_{l=0}^{n} p_{x=s,\alpha=l} l^j. \tag{27}$$

Also note that t_{0j} can be directly computed. Let P_k be the probability that there there are k sources in the on state. Then,

$$P_k = \binom{n}{k} \left(\frac{q_0}{q_0 + q_1} \right)^{n-k} \left(\frac{q_1}{q_0 + q_1} \right)^k,$$

and

$$t_{0j} = \sum_{k=0}^{n} k^j P_k.$$

The problem is to minimize the blocking probability denoted by p_b. As in (9), one can apply Little's law to the server and obtain

$$p_b = 1 - \frac{\mu}{\gamma}(1 - p_{x=0}) = 1 - \frac{\mu}{\gamma}(1 - \sum_{l=0}^{n} p_{x=0,\alpha=l}), \tag{28}$$

where $\gamma = \nu t_{01}$ is the overall arrival rate. An equivalent expression for p_b can be obtained by letting $i = 1$ and $j = 0$ in (26) which yields

$$\frac{\gamma}{\mu}(1 - \frac{\sum_{l=0}^{n} l p_{x=s,\alpha=l}(q_0 + q_1)}{nq_1}) = 1 - p_{x=0}. \tag{29}$$

Table 2
LP bounds on blocking probability for $MMPP/M/1/s$ queue

Moments	Lower Bound	Upper Bound
6	0.0	0.00105
8	0.0	0.00075
10	0.0	0.00067
15	0.00017	0.00065
20	0.00026	0.00065

Comparing this to Little's law for the server gives the following expression for p_b:

$$p_b = \frac{\sum_{l=0}^{n} l p_{x=s,\alpha=l}(q_0 + q_1)}{n q_1}. \tag{30}$$

Thus, the equivalent expressions for p_b in (28) or (30) can be used as the objective function for the LP. Note that the expression (30) can also be obtained using Palm theory.

As an example, let $s = 99$, $q_0 = q_1 = 0.1$, $\nu = 0.36$, $n = 5$ and $\mu = 1$. The upper and lower bounds on the blocking probability are presented in Table 2. The number of moments in Table 2 refers to the maximum value of $i + j$ in (26).

By solving the associated Markov chain, the exact blocking probability was determined to be 0.00037 with an accuracy of 10^{-5}. Comparing the LP bounds in Table 2 to the exact blocking probability, we notice that the lower bound requires a large number of moments to give an estimate of the same order as the exact value. By contrast, the upper bound provides a good estimate (i.e., of the same order as the exact value) with just 10 moments. Neither the upper nor lower bound converges to the exact value in Table 2. It is conceivable that the bounds will converge for a large number of moments but numerical problems do not allow us to verify this. In any case, it is not practically useful to increase the number of moments considered arbitrarily, since the size of the LP becomes prohibitive.

In addition to the examples reported here, we have performed many experiments with $MMPP/M/1/s$ models with blocking probabilities of the order of 10^{-4}. The numerical conclusion obtained in each of these experiments are the same as those of the previous paragraph: the upper bounds provides a better estimate of the blocking probability than the lower bounds, and further, in the numerical examples studied, the upper bound is of the same order of magnitude as the actual blocking probability.

The poor behavior of the lower bound may be explained as follows. For a fixed number of moments M, the extreme points of the LP are of the form $\frac{p_M(s)}{q_M(s)}$, where $p_M(.)$ and $q_M(.)$ are polynomials of degree $M-1$ in s. We note that one of the extreme points is the vector $\mathbf{0}$ indicating the point where all the variables are zero. Large deviations theory [18] tells us that the actual blocking probability is asymptotically of the form $e^{-\delta s}$ for large s and some $\delta > 0$. For large s, since a rational function cannot lower bound a decaying exponential function, the lower bound can be achieved only at the extreme point $\mathbf{0}$. Thus, we expect the lower bound to be zero when the number of moments, M, is small in comparison with the size of the buffer s. This is borne out by numerical experience. A similar argument applies for the upper bound also, and as s becomes very large, we do not expect the upper bound to be of the same magnitude as the exact loss probability.

However, as the numerical examples illustrate, the upper bound can serve as a useful estimate even when the lower bound does not.

4.5 Multirate Communication Link

We now illustrate how to analyze multirate communication links [17] using our technique. Consider a link of capacity s which is shared by two call classes. One class requires less bandwidth, and holds for less time, but arrives more frequently. The other class requires more bandwidth, and holds longer, but arrives less frequently. Thus, one class may represent audio calls, while the other video calls. The system operates under a trunk reservation admission control scheme. That is, calls of a class are admitted into the link when the total link capacity currently being utilized is less than a threshold. The problem of estimating the blocking probabilities (or equivalently the throughputs, since as in the previous example $\lambda_i = \gamma_i(1-p_i)$) has received attention due to applications in broadband networks [17].

This system can be modeled in the framework of Section 2. Since we consider only one link, $S = 1$ with $s_1 = s$. There are two buffers b_1 and b_2, which hold the respective call classes. Their capacities are $c_1 = s/n_1$ and $c_2 = s/n_2$, where n_i represents the bandwidth requirement of class i. The exogenous arrival rate to each buffer is λ_i and the service rate is μ_i. The set A_i is used to model both the capacity constraint as well as the trunk reservation policy for sharing of servers. Under a trunk reservation scheme, thresholds t_1 and t_2 are set for each buffer and the region in which customers are admitted to buffer b_i is specified as

$$A_i = \{x = (x_1, x_2) \mid n_1 x_1 + n_2 x_2 \leq s - n_i - t_i\} \text{ for } i = 1, 2.$$

With this choice of A_i, we can assign $w_i(x)n_i = x_i n_i$ servers for each buffer (since each admitted customer is served). As mentioned above, we are interested in obtaining bounds on $\lambda_i = \gamma_i(1-p_i)$, i.e., either on λ_i or on the blocking probability p_i, which, by the PASTA property, is given by $p_i = Prob(x \in A_i^c)$.

The equality constraints obtained are all of the form (8), with $p_{mn} = 0$ (since there is no feedback). Since there are only two buffers, we can omit the superscript $\{i, j\}$ from the variables. Also, note that we do not need the variables u. Here, since $w_i(x) = x_i$, we have

$$y_{1,d_1,d_2} = E[x_1^{d_1+1} x_2^{d_2}] \text{ and } y_{2,d_1,d_2} = E[x_1^{d_1} x_2^{d_2+1}] \text{ for all } d_1, d_2.$$

Thus we need only consider the variables $y_{d_1,d_2} := E[x_1^{d_1} x_2^{d_2}]$ and the variables z_{i,d_1,d_2} as before, for $i = 1, 2$. So, if we restrict attention to moments of order N or less, that is $d_1 + d_2 \leq N$, then we obtain $N(N+1)/2 - 1$ equality constraints on $3N(N+1)/2 - 1$ variables.

Let us now consider the inequality constraints. The nonnegativity constraints of (13) carry over without any modification, i.e,

$$z_{1,m,n} \geq 0, z_{2,m,n} \geq 0, y_{m,n} \geq 0, \text{ for all } m, n.$$

The domination constraints of (10) also carry over as

$$z_{1,m,n} \leq y_{m,n} \text{ and } z_{2,m,n} \leq y_{m,n} \text{ for all } m, n.$$

We also have domination constraints of the form

$$E[1_{A_i}(n_1x_1+n_2x_2)^n] \leq (s-n_i-t_i)^n Prob(x \in A_i) \text{ for all } n, \text{ and } i=1,2,$$

which yields

$$\sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) n_1^k n_2^{n-k} z_{i,k,n-k} \leq (s-n_i-t_i)^n z_{i,0,0} \text{ for all } n, \text{ and } i=1,2.$$

The nonidling constraints of (12) do not apply here. Last we need to obtain the Chebyshev type constraints of the form of (11). Note that $A_i^c = \{x : n_1x_1+n_2x_2 > s-n_i-t_i\}$. Hence we obtain

$$E[(1-1_{A_i})(n_1x_1+n_2x_2)^n] \geq (s-n_i-t_i+1)^n Prob(x \in A_i^c) \text{ for all } n, \text{ and } i=1,2,$$

which yields

$$\sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) n_1^k n_2^{n-k} (y_{k,n-k}-z_{i,k,n-k}) \geq (s-n_i-t_i+1)^n (1-z_{i,0,0}) \text{ for all } n, \text{ and } i=1,2.$$

Additional constraints can be obtained using the fact that either $A_1 \subseteq A_2$ or $A_2 \subseteq A_1$ (depending on whether $n_1+t_1 \geq n_2+t_2$). If $A_1 \subseteq A_2$, we have

$$z_{1,m,n} \leq z_{2,m,n} \text{ for all } m,n, \text{ and}$$

$$E[(1_{A_2}-1_{A_1})(n_1x_1+n_2x_2)^n] \geq (s-n_1-t_1+1)^n (Prob(x \in A_1^c) - Prob(x \in A_2^c)) \text{ for all } n,$$

which yields

$$\sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) n_1^k n_2^{n-k} (z_{2,k,n-k}-z_{1,k,n-k}) \geq (s-n_1-t_1+1)^n (z_{2,0,0}-z_{1,0,0}) \text{ for all } n.$$

We now introduce additional variables as well as constraints. Note that, without loss of generality, we can assume $n_1 = 1$. If we use the variables $q_i := Prob(n_1x_1+n_2x_2 = i)$, we can get additional equality constraints of the form[4]

$$E[(1-1_{A_2})(n_1x_1+n_2x_2)^n] = \sum_{i=0}^{n_2+t_2-1} (s-i)^n q_{s-i},$$

which yield

$$\sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) n_1^k n_2^{n-k} (y_{k,n-k}-z_{2,k,n-k}) = \sum_{i=0}^{n_2+t_2-1} (s-i)^n q_{s-i}. \quad (31)$$

Finally, we note that if $t_1 = t_2 = 0$, then the resulting Markov chain is product form. Hence the invariant measure depends only on the ratio γ_i/μ_i and not on the actual values

[4] The introduction of these variables increases the complexity of the LP, but as can be seen in (31), only a few q_i need be used, thus maintaining reasonable complexity.

Table 3
Comparison of results for a full sharing link

Acc. Prob.	Simulation	BGZ Approx.	LP Lower Bound	LP Upper Bound
Class 1	0.9964	0.9998	0.9733	1.0000
Class 2	0.9769	0.9982	0.9504	0.9998

Table 4
Comparison of results for a link with reservation

Acc. Prob.	Simulation	BGZ Approx.	LP Lower Bound	LP Upper Bound
Class 1	0.9383	0.9972	0.8196	0.9964
Class 2	0.9966	0.9999	0.8531	0.9986

of γ_i and μ_i. Thus, the terms involving only γ_1 and μ_1 and those involving only γ_2 and μ_2 must individually sum to zero in (8). This yields twice as many equality constraints as before.

Table 3 summarizes the results for a link with $\gamma_1 = 35$, $\mu_1 = 1$, $\gamma_2 = 0.07$, $\mu_2 = 0.01$, $n_1 = 1$, $n_2 = 5$, $s = 100$ and $t_1 = t_2 = 0$. The value for the acceptance probability ($Prob(x \in A_i)$) obtained by simulation from [5] and that obtained using the Bean, Gibbens and Zachary (BGZ) approximation [1] are compared to the bound obtained using the first ten moments in the LP. [5]

We now consider an example in which there is reservation against class 1. Here, $\gamma_1 = 70$, $\mu_1 = 2$, $n_1 = 1$, $\gamma_2 = 0.035$, $\mu_2 = 0.01$, $n_2 = 10$, $s = 100$ and $t_1 = 10$, $t_2 = 0$. It should also be noted that the only known way to exactly solve this problem is by finding the invariant measure of a Markov chain with about 1000 states. In Table 4, we present the results using the first thirteen moments for the LP bounds on the acceptance probabilities. Note that in this case, the steady-state distribution of Markov chain is not of product form, and so we cannot split the equality constraints as mentioned above. It can be seen that the bounds outperform the BGZ approximation but neither approximation is good.

This exposes a limitation of the LP approach. When dealing with multiclass loss models, we have to approximate a large subset of the state space by a simple representation. For example, we replace the probability that the system is in a blocking state for customer class i by $1 - P(x \in A_i)$, instead of using the expression $1 - \sum_{x \in A_i} P(x)$, which would lead to a large number of variables. The penalty we pay for this artificial reduction in the number of variables is reflected in the accuracy of the bounds.

[5] Note that in this case of *full sharing* the acceptance probability can be computed exactly using the recursions due to Kaufman [6] and Roberts [16].

5 Conclusions

The results presented here concerning blocking probabilities and throughput are interesting and merit further investigation.

On the theoretical front, the most important issue to resolve is when the LP bounds are asymptotically exact. It appears that in order to obtain this result we need to enumerate some probabilities of interest. An interesting question is what is the minimal enumeration scheme that ensures asymptotic exactness. Another interesting question is whether it is possible to find a "usable" LP bounding scheme, i.e., which has reasonable complexity, which is asymptotically exact for *all* networks with only finite buffers.

On the application front, we see that in some regimes for the multirate communication link problem, the bounds are not very tight. It would be nice if we could improve these bounds, enumerating some probabilities if need be, so as to get tighter bounds, without losing too much by the way of computational complexity. Finally, it would be good to have a technique for using the solution of the LP to design "good" scheduling and admission control policies.

References

[1] N. Bean, R. J. Gibbens, and S. Zachary, Analysis of large single resource loss systems under heavy traffic with applications to integrated networks, Advances in Applied Probability, 27, (1995) 273–292.

[2] D. Bertsimas, I. C. Paschalidis, and J. N. Tsitsiklis, Optimization of multiclass queueing networks: Polyhedral and nonlinear characterizations of achievable performance, Annals of Applied Probability, 4, (1994) 43–75.

[3] J. G. Dai, On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit model, Annals of Applied Probability, 5, (1995) 49–77.

[4] D. G. Down and S. P. Meyn, Piecewise linear test functions for stability of queueing networks, in *Proceedings of the IEEE 33th Conference on Decision and Control*, December 1994, pp. 2069–2074.

[5] A. G. Greenberg and R. Srikant, Computational technqiues for accurate performance evaluation of multirate, multihop communication networks, IEEE/ACM Transactions on Networking, 5, (1997) 266–277.

[6] J. S. Kaufman, Blocking in a shared resource environment, IEEE Transactions on Communication, 29, (1981) 1474–1481.

[7] F. P. Kelly, Loss networks, Annals of Applied Probability, 1, (1991) 319–378.

[8] L. Kleinrock, *Queueing Systems, Volume 1: Theory*, Wiley–Interscience, New York, 1975.

[9] P. R. Kumar, Re-entrant lines, Queueing Systems: Theory and Applications, 13, (1993) 87–110.

[10] S. Kumar and P. R. Kumar, Performance bounds for queueing networks and scheduling policies, IEEE Transactions on Automatic Control, 39, (1994) 1600–1611.

[11] P. R. Kumar and S. P. Meyn, Stability of queueing networks and scheduling policies, IEEE Transactions on Automatic Control, 40, (1995) 251–260.

[12] P. R. Kumar and S. P. Meyn, Duality and linear programs for stability and performance analysis of queueing networks and scheduling policies, IEEE Transactions on Automatic Control, 41, (1996) 4–17.

[13] S. Lippman, Applying a new device in the optimization of exponential queueing systems, Operations Research, 23, (1975) 687–710.

[14] S. H. Lu and P. R. Kumar, Distributed scheduling based on due dates and buffer priorities, IEEE Transactions on Automatic Control, 36, (1991) 1406–1416.

[15] H. G. Perros and E. T. Altiok, *Queueing Networks with Blocking*, North Holland, Amsterdam, 1989.

[16] J. W. Roberts, A service system with heterogeneous service requirements, Performance of Data Communication Systems and Their Applications, (1981) 423-431.

[17] K. W. Ross, *Multiservice Loss Models for Broadband Telecommunication Networks*, Springer, London, 1995.

[18] W. Whitt, Tail probabilities with statistical multiplexing and effective bandwidths in multi-class queues," Telecommunication Systems, 2, (1993) pp. 71–107.