Analyzing Queuing Systems with Coupled Processors through Semidefinite Programming

Balaji Rengarajan, Constantine Caramanis, and Gustavo de Veciana

Dept. of Electrical and Computer Engineering

The University of Texas at Austin

{balaji, cmcaram, gustavo}@ece.utexas.edu

June 30, 2008

Abstract

We consider queuing systems with coupled processors, where the service rate at each queue varies depending on the set of queues in the system with non-zero queue lengths. In general, such queuing systems are very difficult to analyze and steady state queue length distributions are known only for two-queue systems. The coupled-processors model arises naturally in the study of several systems where a resource is shared by several classes of customers. We study the stochastic recursive equations that govern such systems, and obtain lower and upper bounds on the moments of the queue length by formulating a moments problem and solving a semidefinite relaxation of the original problem. The resulting bounds are very close even at lower-order relaxations, and can be made progressively tighter at the expense of increasing the complexity of the associated semidefinite program. We demonstrate the effectiveness of our proposed methodology in studying and optimizing dynamic systems such as interference-coupled wireless networks and GPS servers.

1 Introduction

We study a queuing system with coupled processors, where the rate at which users in a queue are served depends on the lengths of the other queues in the system. In particular, we consider systems where the service rate at each queue varies depending on the set of queues in the system

with non-zero queue lengths. This coupled-processors model arises naturally in the study of systems where a resource is shared by several classes of customers, and has been studied for a long time in the queuing literature in the context of bandwidth sharing of elastic flows in packet networks.

Generalized Processor Sharing (GPS) [1] is a service discipline that was developed to allow capacity to be shared in a fair and flexible manner, and is an important mechanism for achieving differentiated quality of service. Unlike a strict priority scheme, the GPS discipline allows for service differentiation while preventing starvation. The following is a brief description of the GPS mechanism. Consider N classes of customers, each with some associated arrival rate and service requirements, sharing a server. Each class, n, also has an associated weight ϕ_n, with $\sum_{n=1}^{N} \phi_n = 1$ and can be served at rate R_n. If all the queues are non-empty, the fraction of time queue n is served is given by ϕ_n and the corresponding service rate is $\phi_n R_n$. However, if some of the queues are empty, their allotted time-fractions are distributed among the non-empty queues in proportion to their respective weights. Thus, the queues are coupled by the mechanism used to share excess capacity. While the GPS policy is not itself realistic to implement, disciplines that closely track the GPS mechanism such as Weighted Fair queuing (WFQ) [2] are used in practice. Understanding the performance of the system, and of the user classes is very important, for example, in order to optimize the choice of the class weights.

Another area where the coupled-processors model can be applied is in the study of interference-limited wireless networks. As network deployments with increased base station/access point densities are used to meet ever increasing demands for capacity, wireless systems are forced to operate in a highly dynamic, interference limited regime. There is a huge disparity in the data rates that interference-limited users perceive when neighboring access points/base stations are idle and when there are concurrent interfering transmissions in neighboring cells. However, wireless networks have mostly been studied assuming that they serve static user populations (a fixed set of infinitely backlogged users), and that adjacent transmitters are always interfering with each other. In a more realistic scenario, data requests from users are generated at random times, and the users leave when their service requirements have been met. Thus, it is not the case that cells always have active users; in fact base stations will be idle for some load-dependent fraction of time. Inter-cell

interference couples the service rates that users in different cells perceive, and thus their performance. This dynamic scenario has been examined in [3, 4, 5], and it was demonstrated that the performance perceived by users is very different from that predicted by the analysis of the static system. In this paper, we will consider (briefly) the effect of time-varying interference on delays experienced by best-effort downlink users. A further example for coupling in wireless networks, that we will study is the performance of a simple contention-based multiple-access scheme.

In general, coupled queuing systems are very difficult to analyze. Large deviations asymptotics of the workload in GPS systems have been derived in [6, 7, 8, 9], and the effect of serving customers with long-tailed service times has been analyzed in [10, 11]. However, steady state queue length distributions are known only in some special cases. The coupled-processors model with exponentially distributed service requests and the number of classes/queues restricted to 2 was analyzed in [12, 13, 14]. The uniformization technique was used in [13] to study the model with $\phi_1 = \phi_2$ and $R_1 = R_2$. The generating function of the steady-state queue length distribution was obtained in [12] by solving a Riemann-Hilbert boundary value problem, and closed-form expressions were obtained in [14] for the work-conserving case where $R_1 = R_2$, without resorting to the formulation of a Riemann-Hilbert problem. However, the techniques used in the above papers cannot be extended to systems with more than two queues.

In this paper, we obtain bounds on the moments of the steady-state queue-length in a system consisting of N coupled queues by studying the stochastic recursive equations that govern the system. We obtain lower and upper bounds on the moments of the queue length by formulating and solving a semidefinite relaxation of the original problem. These bounds can be made progressively tighter at the expense of increasing the complexity of the associated semidefinite program. Our model is motivated by the one used in [15] to analyze GI/GI queues, and draws on results obtained in [16, 17]. The derivation in [15] is unable to accommodate state-dependent service times and coupled processors. A main contribution of this paper is to extend the moment-based SDP approach to this broader class of problems. We then consider some example applications, to illustrate the importance of being able to accurately capture the impact of coupling.

The rest of the paper is organized as follows: The system model is described in Section 2, and the stochastic recursive equation governing the system is described. In Section 3, the

moments based approach to bounding the functions of interest is constructed, and the semidefinite relaxation of the moment problem is derived. We apply our proposed methodology to model, and optimize some coupled systems in Section 4. Section 4.1 considers the example of interference-limited wireless networks, Section 4.2 illustrates the problem of studying and optimizing GPS servers, and Section 4.3 examines the design of a multiple access system. Finally, Section 5 concludes the paper.

2 System Model and Notation

We consider a system of N queues. Users arrive at queue n as a Poisson process, with mean arrival rate λ_n. We assume that the users require exponential service times with mean 1. The users are served according to a first come first serve (FCFS) policy at each queue. We denote the queue length at each queue at time t by $Q_n(t)$, $n = 1,2,\ldots,N$. The service rates at the queues depends on the subset of queues in the system with non-zero queue length, where the number of possible subsets is 2^N. The status(busy or idle) of the queues in the system is captured by a vector $\vec{\Delta}(t)$ of length n that takes values δ_i, $i = 1,\ldots,2^N$. The n^{th} component of $\vec{\Delta}(t)$, $\Delta_n(t)$ is equal to 0 when $Q_n(t) = 0$, and 1 when $Q_n(t) > 0$. The rate at which queue n serves users when the state of the queues is $\vec{\Delta}(t)$ is denoted by $\mu_n^{\vec{\Delta}(t)}$. We assume that the queuing system is stable, and also that there is a well defined maximum rate μ^* that bounds the rate at which any queue can be served, irrespective of the state of the system.

The queue length process evolves as a continuous time Markov chain, parametrized by the rates defined above. In this paper, we will study the embedded discrete time Markov chain obtained after uniformizing this continuous time Markov chain. Since the maximum rate of transitions is bounded by $M = \sum_{n=1}^{N} \lambda_n + N\mu^*$, the continuous time Markov chain can be uniformized by introducing fictitious events that cause no change in the state of the Markov chain. We denote the state of the uniformized discrete time Markov chain at step k by $\vec{Q}(k) = \{Q_n(k), n = 1,2,\ldots,N\}$, and analogously, the state of the queues by $\vec{\Delta}(k)$. The transition probabilities for the uniformized

Markov chain when $Q_n(k) = q_n$ are as follows:

$$\begin{aligned}
\text{Pr.(Arrival into queue } n) &= \frac{\lambda_n}{M},\ n = 1,\ldots,N \\
\text{Pr.(Departure from queue } n) &= \frac{\mu_n^{\vec{\Delta}(k)}}{M} 1(q_n > 0),\ n = 1,\ldots,N \\
\text{Pr.(No change in state)} &= 1 - \frac{\sum_{n=1}^{N} (\lambda_n + \mu_n^{\vec{\Delta}(k)} 1(q_n > 0))}{M}
\end{aligned}$$

The steady state queue length distribution of this uniformized Markov chain is identical to that of the original. The evolution of the uniformized Markov chain can be expressed in the form of the following stochastic recursive model, where $\vec{X}(k) = \{X_n(k), n = 1,2,\ldots,N\}$ denotes the incremental change in the queue lengths that results in the Markov chain transitioning from $\vec{Q}(k)$ to $\vec{Q}(k+1)$.

$$\vec{Q}(k+1) = \vec{Q}(k) + \vec{X}(k),\ \ k = 0,1,\ldots \tag{1}$$

An arrival into queue n at iteration k is represented by $X_n(k) = 1$, a departure by $X_n(k) = -1$ and if the transition corresponds to the self-loop, $\vec{X}(k) = \vec{0}$. Note that $\vec{X}(k)$ and $\vec{Q}(k)$ are not independent. The distribution of $\vec{X}(k)$ clearly depends on $\vec{\Delta}(k)$, and thus on $\vec{Q}(k)$.

When the system is stable, and steady state exists, the stochastic recursive model converges [18, 19], and there exists a joint distribution for the steady state vectors $(\vec{Q}, \vec{X})$ such that

$$\vec{Q} \stackrel{d}{=} g(\vec{Q}, \vec{X}) = \vec{Q} + \vec{X}. \tag{2}$$

The equality here is in distribution. Our goal is to characterize the behavior of the queuing system by formulating a moments based approach to bound functions of the moments of $\vec{Q}$. As in [15], we only use information on the moments of $\vec{X}$ derived from the uniformized Markov chain in the formulation.

3 Problem Formulation

We denote the steady state marginal probability measure of $\vec{Q}$ by ψ_Q and that of $\vec{X}$ by ψ_X. These measures are supported on $\mathcal{S}_Q \subseteq \mathbb{R}^N$ and $\mathcal{S}_X \subseteq \mathbb{R}^N$ respectively. Let ψ denote the joint probability

measure for $(\vec{Q},\vec{X})$, supported on $\mathcal{S} = \mathcal{S}_Q \times \mathcal{S}_X$. Note that, ψ_Q and ψ_X are not necessarily independent and the joint distribution ψ cannot be expressed as a product form. Using this notation, the distribution of $g(\vec{Q},\vec{X})$ on the right hand side of Eq. (2) can be written as ψg^{-1}, and the steady state equation can be represented as

$$\psi_Q = \psi g^{-1}, \tag{3}$$

where $g(\vec{Q},\vec{X}) = \vec{Q}+\vec{X}$.

We partition $\mathcal{S}_Q$ into regions $\mathcal{S}_{\delta_i}$, such that the subset of queues with non-zero queue length is identical at all states within each region, i.e., if $\vec{Q} \in \mathcal{S}_{\delta_i}$, then $\Delta = \delta_i$. Let $\psi_Q^{\delta_i}$, $\psi_X^{\delta_i}$, ψ^{δ_i} be the conditional probability measures on $\vec{Q}$, $\vec{X}$, and $(\vec{Q},\vec{X})$ respectively, conditioned on $\vec{Q} \in \mathcal{S}_{\delta_i}$. Note that, given the status vector Δ, the service rates at the various queues are determined. Since specifying the region $\mathcal{S}_{\delta_i}$ that $\vec{Q}$ belongs to fixes Δ, the distributions $\psi_Q^{\delta_i}$ and $\psi_X^{\delta_i}$ are conditionally independent, and we can write

$$\psi^{\delta_i} = \psi_Q^{\delta_i}\psi_X^{\delta_i}. \tag{4}$$

We use the multi-index notation in the formulation of the optimization problem used to determine the upper and lower bounds. If $\vec{\alpha} = \{\alpha_n, n = 1,\ldots,N\} \in \mathbb{N}^N$ is a N-dimensional multi-index, and $\vec{Y} = \{Y_n, n = 1,\ldots,N\} \in \mathbb{R}^N$ a N-dimensional vector, we denote by $\vec{Y}^{\vec{\alpha}}$ the term $Y_1^{\alpha_1}\ldots Y_N^{\alpha_N}$, and we let $|\vec{\alpha}| = \sum_{n=1}^N \alpha_n$. We will use information about the conditional moments $\mathbf{E}_{\psi_X^{\delta_i}}\left[\vec{X}^{\vec{\beta}}\right]$, $|\vec{\beta}| \leq 2r$ and $i = 1,\ldots,2^N$, denoted $m_{\delta_i}^{\vec{\beta}}$. These conditional moments can be computed in a straightforward manner as the transition probabilities of the uniformized Markov chain do not change within regions $\mathcal{S}_{\delta_i}$. We then determine bounds on functions of the form $\mathbf{E}_{\psi_Q}\left[\sum_{|\vec{\gamma}|\leq 2r} w_{\vec{\gamma}}\vec{Q}^{\vec{\gamma}}\right]$, where $w_{\vec{\gamma}}$ are constant weights. The problem formulation follows:

PROBLEM 3.1 Given Borel measurable sets $\mathcal{S}_Q \subseteq \mathbb{R}^N$ and $\mathcal{S}_X \subseteq \mathbb{R}^N$, solve:

$$\sup / \inf_{\psi} \quad \mathbf{E}_{\psi_Q}\left[\sum_{|\vec{\gamma}|\leq 2r} w_{\vec{\gamma}}\vec{Q}^{\vec{\gamma}}\right] \tag{5}$$

s.t.

$$\psi_Q = \psi g^{-1} \tag{6}$$

$$\psi^{\delta_i} = \psi_Q^{\delta_i}\psi_X^{\delta_i}, \quad i = 1,\ldots,2^N \tag{7}$$

$$\mathbf{E}_{\psi_X^{\delta_i}}\left[\vec{X}^{\vec{\beta}}\right] = m_{\delta_i}^{\vec{\beta}}, \quad i = 1,\ldots,2^N, |\vec{\beta}| \leq 2r \tag{8}$$

$$\mathbf{E}_{\psi}[1] = \mathbf{E}_{\psi_Q}[1] = \mathbf{E}_{\psi_X}[1] = 1 \tag{9}$$

$$\psi \in \mathbb{M}(\mathcal{S}), \psi_Q \in \mathbb{M}(\mathcal{S}_Q), \psi_X \in \mathbb{M}(\mathcal{S}_X) \tag{10}$$

Here, $\mathbb{M}(\mathcal{S})$, $\mathbb{M}(\mathcal{S}_Q)$ and $\mathbb{M}(\mathcal{S}_X)$ are the sets of finite positive Borel measures supported by $\mathcal{S}$, $\mathcal{S}_Q$ and $\mathcal{S}_X$ respectively. Additionally, the constraints in Eq. (9) guarantee that the measures are valid probability measures. If we let $r \to \infty$, all the moments of $\vec{X}$ would be specified, thus specifying exactly the distribution of $\vec{X}$ which in turn uniquely determines the distribution of $\vec{Q}$, and the joint distribution of $(\vec{Q}, \vec{X})$. In this case, the bounds would converge to the exact solution.

3.1 The Moments Based Approach

We derive a moments-based relaxation for Problem 3.1, based on joint moments of degree no higher than $2r$. The decision variables for this problem are defined as follows:

$$x_k^{\vec{\alpha}\vec{\beta}} := \mathbf{E}[\vec{Q}^{\vec{\alpha}}\vec{X}^{\vec{\beta}}|\vec{Q} \in \mathcal{S}_{\delta_k}]\text{Pr.}[\vec{Q} \in \mathcal{S}_{\delta_k}], \ \forall|\vec{\alpha}| + |\vec{\beta}| \leq 2r, \ k = 1,\ldots,2^N = K$$

Now, we can express the objective function and the relaxed versions of the constraints in Problem 3.1 in terms of the above decision variables as described below:

3.1.1 Objective Function

The objective function in Eq. (5) can be written in terms of the conditional expectations as follows using the theorem of total expectation. This expression can then be expressed as a linear function of the decision variables.

$$\begin{aligned}\mathbf{E}_{\psi_Q}\left[\sum_{|\vec{\gamma}|\leq 2r} w_{\vec{\gamma}}\vec{Q}^{\vec{\gamma}}\right] &= \sum_{k=1}^{K}\mathbf{E}\left[\sum_{|\vec{\gamma}|\leq 2r} w_{\vec{\gamma}}\vec{Q}^{\vec{\gamma}}|\vec{Q} \in \mathcal{S}_{\delta_k}\right]\text{Pr.}[\vec{Q} \in \mathcal{S}_{\delta_k}] \\ &= \sum_{k=1}^{K}\sum_{|\vec{\gamma}|\leq 2r} w_{\vec{\gamma}} x_k^{\gamma,0}\end{aligned}$$

3.1.2 Constraints

The equality in distribution in the steady state constraint from Eq. (6) implies that the equality also holds for moments of all orders. We relax this constraint to the following constraint relating joint moments of order less than or equal to $2r$:

$$\begin{aligned} & \vec{Q} \stackrel{d}{=} g(\vec{Q},\vec{X}) = \vec{Q}+\vec{X} \\ \Rightarrow \quad & \mathbf{E}[\vec{Q}^{\vec{\alpha}}] = \mathbf{E}[(\vec{Q}+\vec{X})^{\alpha}], \quad \forall |\vec{\alpha}| \leq 2r. \end{aligned}$$

We can then break down the above equation in terms of the conditional expectations, again using the theorem of total expectation to get

$$\sum_{k=1}^{K} \mathbf{E}[\vec{Q}^{\vec{\alpha}} | \vec{Q} \in \mathcal{S}_{\delta_k}] \mathrm{Pr.}[\vec{Q} \in \mathcal{S}_{\delta_k}] = \sum_{k=1}^{K} \mathbf{E}[g(\vec{Q},\vec{X})^{\vec{\alpha}} | \vec{Q} \in \mathcal{S}_{\delta_k}] \mathrm{Pr.}[\vec{Q} \in \mathcal{S}_{\delta_k}], \; \forall |\vec{\alpha}| \leq 2r.$$

The term $g(\vec{Q},\vec{X})^{\vec{\alpha}}$ can be expanded using the binomial theorem as

$$g(\vec{Q},\vec{X})^{\vec{\alpha}} = (\vec{Q}+\vec{X})^{\vec{\alpha}} = \sum_{|\vec{\gamma_1}|+|\vec{\gamma_2}| \leq \alpha} g_{\vec{\alpha}}^{(\vec{\gamma_1},\vec{\gamma_2})} \vec{Q}^{\vec{\gamma_1}} \vec{X}^{\vec{\gamma_2}},$$

where $\{g_{\vec{\alpha}}^{(\vec{\gamma_1},\vec{\gamma_2})}\}$ are the coefficients resulting from the expansion. The constraint can now be written as a linear function of the decision variables as follows:

$$\begin{aligned} \sum_{k=1}^{K} \mathbf{E}[\vec{Q}^{\vec{\alpha}} | \vec{Q} \in \mathcal{S}_{\delta_k}] \mathrm{Pr.}[\vec{Q} \in \mathcal{S}_{\delta_k}] &= \sum_{k=1}^{K} \mathbf{E}\Big[\sum_{|\vec{\gamma_1}|+|\vec{\gamma_2}| \leq \alpha} g_{\vec{\alpha}}^{(\vec{\gamma_1},\vec{\gamma_2})} \vec{Q}^{\vec{\gamma_1}} \vec{X}^{\vec{\gamma_2}} | \vec{Q} \in \mathcal{S}_{\delta_k} \Big] \mathrm{Pr.}[\vec{Q} \in \mathcal{S}_{\delta_k}] \\ & \quad , \; \forall |\vec{\alpha}| \leq 2r \\ \Rightarrow \sum_{k=1}^{K} x_k^{\vec{\alpha},0} &= \sum_{k=1}^{K} \sum_{|\vec{\gamma_1}|+|\vec{\gamma_2}| \leq \alpha} g_{\vec{\alpha}}^{(\vec{\gamma_1},\vec{\gamma_2})} x_k^{\vec{\gamma_1}\vec{\gamma_1}}, \quad \forall |\vec{\alpha}| \leq 2r. \end{aligned}$$

The conditional independence of $\vec{Q}$ and $\vec{X}$ in constraint (7) is relaxed by equating the moments of the product of the random variables to the product of the moments.

$$\psi^{\delta_i} = \psi_Q^{\delta_i} \psi_X^{\delta_i} \Rightarrow \mathbf{E}[\vec{Q}^{\vec{\alpha}} \vec{X}^{\vec{\beta}} | \vec{Q} \in \mathcal{S}_{\delta_k}] = \mathbf{E}[\vec{Q}^{\vec{\alpha}} | \vec{Q} \in \mathcal{S}_{\delta_k}] \mathbf{E}[\vec{X}^{\vec{\beta}} | \vec{Q} \in \mathcal{S}_{\delta_k}], \; \forall |\vec{\alpha}| + |\vec{\beta}| \leq 2r, \; k = 1, \ldots, K.$$

This along with the given moments of X from constraint (8) can be used to constrain the moments

of the joint conditional distribution as follows:

$$\begin{aligned}
\mathbf{E}[\vec{Q}^{\vec{\alpha}}\vec{X}^{\vec{\beta}}|\vec{Q}\in\mathcal{S}_{\delta_k}] &= \mathbf{E}[\vec{Q}^{\vec{\alpha}}|\vec{Q}\in\mathcal{S}_{\delta_k}]\mathbf{E}[\vec{X}^{\vec{\beta}}|\vec{Q}\in\mathcal{S}_{\delta_k}] \\
&\quad ,\ \forall|\vec{\alpha}|+|\vec{\beta}|\leq 2r,\ k=1,\ldots,K \\
\Rightarrow \mathbf{E}[\vec{Q}^{\vec{\alpha}}\vec{X}^{\vec{\beta}}|\vec{Q}\in\mathcal{S}_{\delta_k}]\mathrm{Pr.}[\vec{Q}\in\mathcal{S}_{\delta_k}] &= \mathbf{E}[\vec{Q}^{\vec{\alpha}}|\vec{Q}\in\mathcal{S}_{\delta_k}]\mathbf{E}[\vec{X}^{\vec{\beta}}|\vec{Q}\in\mathcal{S}_{\delta_k}]\mathrm{Pr.}[\vec{Q}\in\mathcal{S}_{\delta_k}] \\
&\quad ,\forall|\vec{\alpha}|+|\vec{\beta}|\leq 2r,\ k=1,\ldots,K \\
\Rightarrow \mathbf{E}[\vec{Q}^{\vec{\alpha}}\vec{X}^{\vec{\beta}}|\vec{Q}\in\mathcal{S}_{\delta_k}]\mathrm{Pr.}[\vec{Q}\in\mathcal{S}_{\delta_k}] &= \mathbf{E}[\vec{Q}^{\vec{\alpha}}|\vec{Q}\in\mathcal{S}_{\delta_k}]m_{\delta_k}^{\vec{\beta}}\mathrm{Pr.}[\vec{Q}\in\mathcal{S}_{\delta_k}] \\
&\quad ,\forall|\vec{\alpha}|+|\vec{\beta}|\leq 2r,\ k=1,\ldots,K.
\end{aligned}$$

Rewriting the above constraint in terms of the decision variables, we get the following set of linear constraints:

$$x_k^{\vec{\alpha}\vec{\beta}} = m_{\delta_k}^{\vec{\beta}} x_k^{\vec{\alpha},0},\ \ \forall|\vec{\alpha}|+|\vec{\beta}|\leq 2r,\ \ k=1,\ldots,K.$$

In order to ensure that constraint (9) is satisfied, and that the joint measure is indeed, a probability measure, we need

$$\begin{aligned}
\sum_{k=1}^{K}\mathrm{Pr.}[\vec{Q}\in\mathcal{S}_{\delta_k}] &= 1 \\
\Rightarrow \sum_{k=1}^{K} x_k^{0,0} &= 1.
\end{aligned}$$

Finally, we need to ensure that constraint (10) is satisfied and $\{x_k^{\vec{\alpha}\vec{\beta}},\ |\vec{\alpha}|+|\vec{\beta}|\leq 2r\}$ represents a valid moment sequence for any $k=1,\ldots,K$. Partitioning $\mathcal{S}_Q$ into regions $\mathcal{S}_{\delta_k}$ induces a partition of $\mathcal{S}=\mathcal{S}_Q\times\mathcal{S}_X$ into the regions $\mathcal{S}_k=\mathcal{S}_{\delta_k}\times\mathcal{S}_X$. We denote the cone of moments supported on $\mathcal{S}_k$ by:

$$\mathcal{M}_{2r}(\mathcal{S}_k)=\left\{x_k|x_k^{\vec{\alpha}\vec{\beta}}=\mathbf{E}_{\psi_k}[\vec{Q}^{\alpha}\vec{X}^{\beta}],\forall|\vec{\alpha}|+|\vec{\beta}|\leq 2r \text{ and for some } \psi_k\in\mathbb{M}(\mathcal{S}_k)\right\}.$$

Letting $\overline{\mathcal{M}_{2r}(\mathcal{S}_k)}$ denotes the closure of the cone of moments, then, for constraint (10) to be true,

$$\{x_k\}\in\overline{\mathcal{M}_{2r}(\mathcal{S}_k)},\ \ \forall k=1,\ldots,K.$$

Now, the relaxed version of Problem 3.1 can be expressed in the form of the following conic

optimization problem based on joint moments of degree less than or equal to $2r$.

PROBLEM 3.2

$$\begin{aligned}
\sup / \inf_{x_k} \quad & \sum_{k=1}^{K} \sum_{|\vec{\gamma}| \leq 2r} w_{\vec{\gamma}} x_k^{\gamma,0} \\
\text{s.t.} \quad & \\
& \sum_{k=1}^{K} x_k^{\vec{\alpha},0} = \sum_{k=1}^{K} \sum_{|\vec{\gamma}_1|+|\vec{\gamma}_2| \leq \alpha} g_{\vec{\alpha}}^{(\vec{\gamma}_1,\vec{\gamma}_2)} x_k^{\vec{\gamma}_1 \vec{\gamma}_1}, \quad \forall |\vec{\alpha}| \leq 2r \\
& x_k^{\vec{\alpha}\vec{\beta}} = m_{\delta_k}^{\vec{\beta}} x_k^{\vec{\alpha},0}, \ \forall |\vec{\alpha}| + |\vec{\beta}| \leq 2r, \ k = 1, \ldots, K \\
& \sum_{k=1}^{K} x_k^{0,0} = 1 \\
& \{x_k\} \in \overline{\mathcal{M}_{2r}(\mathcal{S}_k)}, \ \forall k = 1, \ldots, K.
\end{aligned}$$

3.2 A Semidefinite Relaxation

The moment cone can be characterized using positive semidefinite matrices as in [15, 17, 20]. A necessary condition for the sequence $y = \{y^{\vec{\alpha}\vec{\beta}}, \ |\vec{\alpha}| + |\vec{\beta}| \leq 2r\}$ to be a valid truncated sequence of moments is that the associated moment matrix is positive semidefinite [17]. Let the moment matrix corresponding to the sequence y be denoted $\mathbf{M}_r(y)$. The moment matrix $\mathbf{M}_r(y)$ is the block matrix $\{M_r^{i,j}(y), 0 \leq i, j \leq r\}$ with rows and columns indexed in the basis of polynomials of degree less than or equal to r. The entries of the moment matrix satisfy the following condition:

$$\text{if } M_r^{1,j}(y) = y^{\vec{\alpha}_1 \vec{\beta}_1} \text{ and } M_r^{i,1}(y) = y^{\vec{\alpha}_2 \vec{\beta}_2}, \text{ then } M_r^{i,j}(y) = y^{(\vec{\alpha}_1 + \vec{\alpha}_1)(\vec{\beta}_1 + \vec{\beta}_2)}.$$

A necessary condition for y to be a truncated moment sequence is

$$\mathbf{M}_r(y) \succeq 0$$

In our formulation, this condition has to be satisfied by the decision variables $\{x_k\}, \forall k = 1, \ldots, K$.

Furthermore, the regions $\mathcal{S}_{\delta_k}$ can be defined as the intersection of a finite set of linear inequalities. If the n^{th} queue is constrained to be non-empty in the region $\mathcal{S}_{\delta_k}$, then the condition

$Q_n - 1 \geq 0$ holds in $\mathcal{S}_{\delta_k}$, and in the region $\mathcal{S}_k$. Thus, $\mathcal{S}_k$ can be defined by:

$$\mathcal{S}_k = \cap_{\theta \in \mathcal{S}_k} \theta(Q, X),$$

where θ is a polynomial of degree 1. Let the coefficients of θ be given by $\{\theta^\alpha, |\alpha| \leq 1\}$. Consider any region $\mathcal{S}_k$, a localizing matrix $\mathbf{M}_{r-1}(\theta, x_k)$ associated with one of its polynomials $\theta \in \mathcal{S}_k$ is defined as:

$$M_{r-1}^{(i,j)}(\theta, x_k) = \sum_{|\alpha| \leq 1} \theta^\alpha x_k^{\eta(i,j)+\gamma},$$

where $\eta(i,j)$ is the subscript of the entry $M_{r-1}^{i,j}(x_k)$ in matrix $\mathbf{M}_{r-1}(x_k)$. An additional condition which is necessary for the sequence x_k to be a valid moment sequence is that the localizing matrix corresponding to each $\theta \in \mathcal{S}_k$ be positive definite, i.e.

$$\mathbf{M}_{r-1}(\theta, x_k) \succeq 0.$$

Relaxing the constraint $\{x_k\} \in \overline{\mathcal{M}_{2r}(\mathcal{S}_k)}, \ \forall k = 1, \ldots, K$ in Problem 3.2 using the positive semidefiniteness constraints above, the semidefinite relaxation of the moment problem can be written as follows.

PROBLEM 3.3

$$\begin{aligned}
\sup / \inf_{x_k} \quad & \sum_{k=1}^{K} \sum_{|\vec{\gamma}| \leq 2r} w_{\vec{\gamma}} x_k^{\gamma,0} \\
\text{s.t.} \quad & \\
& \sum_{k=1}^{K} x_k^{\vec{\alpha},0} = \sum_{k=1}^{K} \sum_{|\vec{\gamma}_1| + |\vec{\gamma}_2| \leq \alpha} g_{\vec{\alpha}}^{(\vec{\gamma}_1, \vec{\gamma}_2)} x_k^{\vec{\gamma}_1 \vec{\gamma}_1}, \quad \forall |\vec{\alpha}| \leq 2r \\
& x_k^{\vec{\alpha}\vec{\beta}} = m_{\delta_k}^{\vec{\beta}} x_k^{\vec{\alpha},0}, \ \forall |\vec{\alpha}| + |\vec{\beta}| \leq 2r, \ k = 1, \ldots, K \\
& \sum_{k=1}^{K} x_k^{0,0} = 1 \\
& \mathbf{M}_r(x_k) \succeq 0, \ \forall k = 1, \ldots, K \\
& \mathbf{M}_{r-1}(\theta, x_k) \succeq 0, \ \forall \theta \in \mathcal{S}_k, k = 1, \ldots, K
\end{aligned}$$

This semidefinite problem can be solved to obtain the desired upper and lower bounds on the

objective function. As r is increased, and information about more moments of X are used in the semidefinite program, tighter bounds can be obtained at the cost of increased complexity of the optimization problem.

4 Illustrative Applications

In this section, we examine some simple scenarios where the dynamics of the system are properly modeled by the coupled processors. In all the computational results presented below, Gloptipoly [21] and Sedumi [22] were the tools used to solve the semidefinite program.

4.1 Modeling Mean User Delays in Cellular Networks

4.1.1 A Two Base Station System

Consider a system consisting of two neighboring base stations. A base station serves its associated users according to the FCFS discipline. Each base station serves a user at constant power, or is idle if there are no active users associated with it. We will focus in this case on the interference-limited setting, and examine a scenario where the users in the system are concentrated at two locations. The users who are served by base station 1 are concentrated at a location close to it, such that the received signal strength from base station 1 is very high while any interference from base station 2 is greatly attenuated. These users are immune to interference from base station 2, and can be served at a rate of 4 Mbps both when base station 2 is transmitting and when it is idle. The users associated with base station 2 are close to the cell edge and are served at a rate of 2.5 Mbps when base station 1 is transmitting, and a rate of 4 Mbps when base station 1 is idle. Users arrive at each location according to a Poisson process with rate λ and with service requirements that are exponentially distributed with mean 1 Mb. When a users' service requirement is met, the user leaves the system.

Fig. 1 depicts the bounds on the mean delay obtained by using progressively higher order relaxations, along with the simulated mean delay. The bounds on the delay are obtained by solving the semidefinite program with the mean sum queue length as the objective function, and then using Little's law. The bounds obtained from assuming that the neighboring base station is always

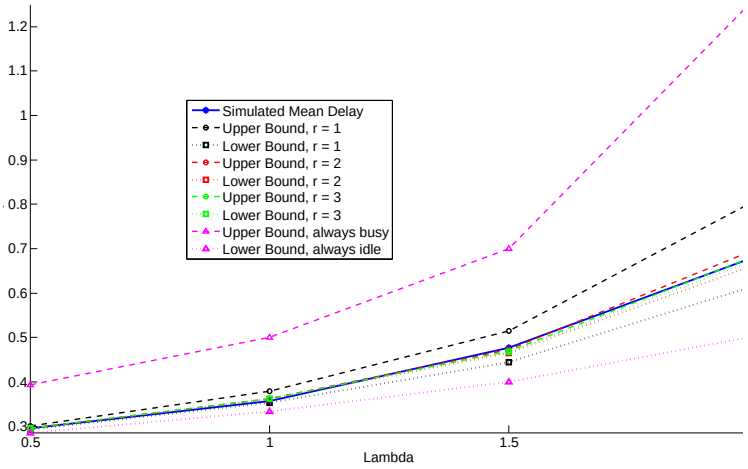
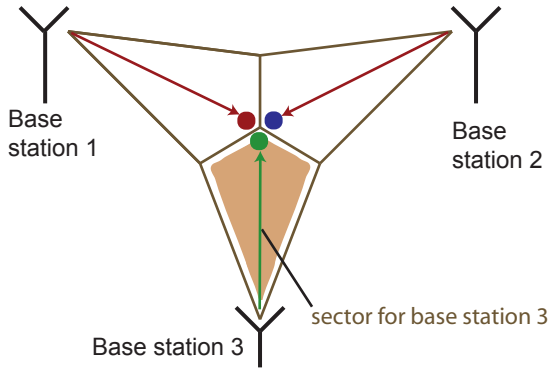


Figure 1: Mean delay for a two base station system

busy/idle are also plotted for comparison. Even when $r = 1$, the bounds computed through the optimization proposed above are much closer to the actual steady state mean delay. Further, as the relaxation order is increased, the upper and lower bounds are very close, and match the simulated results very well. The distance between the upper bound that assumes the system is saturated and the upper bounds obtained by solving the SDP clearly indicates the impact that coupling has on the mean delay may be more complex than simply assuming persistent interference, even under heavy loads. In this case, closed form expressions for the steady state delay are known, however, the degree of accuracy of our results is highly encouraging.

4.1.2 A Three Base Station System

Next, we consider a system with three base stations shown in Fig. 2(a). Since inter-cell interference is a local phenomenon, and users are not affected much by transmissions at base stations located far away, studying such systems with a small number of base stations can thus provide insight into the performance experienced by users in larger systems. Consider the scenario where users associated with each base station are nearly equidistant from all the base stations as in Fig. 2(a), and receive almost equal signals from each base station. Each base station serves its associated users at a rate of 5 Mbps when they are all busy, and at a rate of 7 Mbps when two of the base stations are busy, and a base station serves users at a rate of 10 Mbps if it is the only busy



(a) System configuration

Base station status			Service rates (Mbps)		
BS 1	BS 2	BS 3	BS 1	BS2	BS3
Idle	Idle	Idle	0	0	0
Idle	Idle	Busy	0	0	10
Idle	Busy	Idle	0	10	0
Idle	Busy	Busy	0	7	7
Busy	Idle	Idle	10	0	0
Busy	Idle	Busy	7	0	7
Busy	Busy	Idle	7	7	0
Busy	Busy	Busy	5	5	5

(b) Service Rates

Figure 2: Three base station system

base station. The state-dependent service rates are summarized in Table 2(b). The mean rate of arrival at each base station is again λ.

For this queuing system, no closed form expressions or tight bounds for the steady state delay are available to the best of our knowledge. As shown in Fig. 3, the upper and lower bounds are not far apart, and closely trail the simulated mean delay, and in particular capture the trends with increasing load. The bounds assuming the queues are saturated are much worse, demonstrating that the nature of the coupling plays a significant role in the performance experienced by the users in the system.

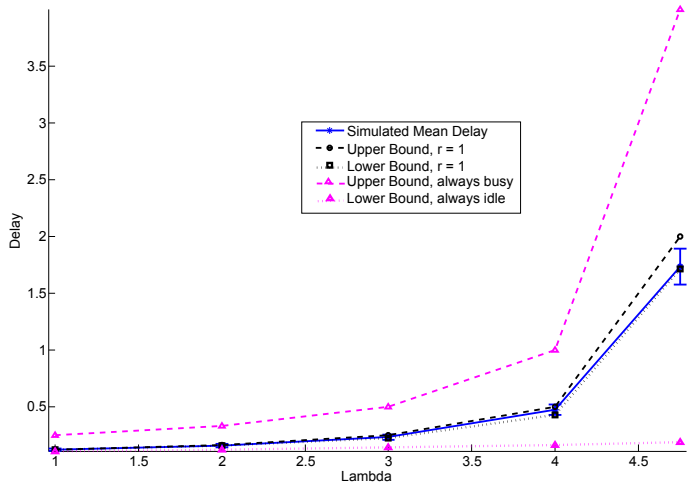


Figure 3: Mean delay of a three base station system.

4.2 A Generalized Processor Sharing System

4.2.1 Two Queue System

Consider a server with a service rate of 10 that is shared by two classes of packets with equal weights, i.e., $R_1 = R_2 = 10$ and $\phi_1 = \phi_2$. packets arrive at each class as a Poisson process of rate λ, with exponentially distributed service requirements with mean 1. In this case, since the sum service rate is a constant, independent of the system state, the sum queue length process is a Markov process, and the distribution of the sum queue length(but not the individual queue lengths) can be analytically determined. The bounds computed by solving the SDP with $r = 1, 2$ are plotted in Fig. 4, along with the simulated results. The lower and upper bounds match exactly, for all the values of λ. In addition, they precisely match the analytically computed mean delay (not shown). This is confirmed by the simulated mean delay matching the bound.

Fig. 5 exhibits the mean queue length at the queue associated with one of the classes as λ increases. While the system as a whole is work-conserving, the individual queues are not. The bounds computed by solving the SDP with $r = 1, 2$ are plotted along with the bounds computed assuming that the neighboring queue is always busy/idle. Note that, in this case closed-form expressions for the mean delay are known [12, 14]. However, for larger systems, analytic expressions for the mean delay are not known.

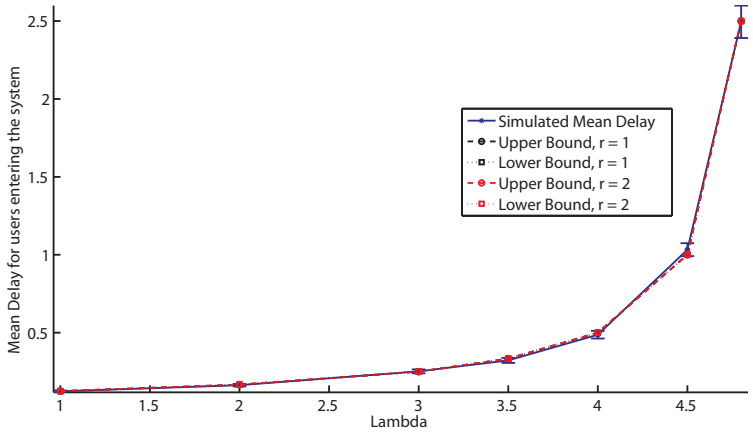


Figure 4: Mean overall system queue length for a 2 Queue GPS system

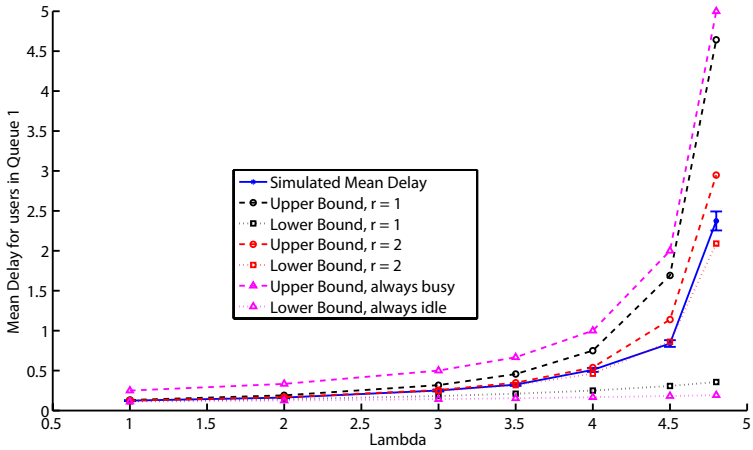


Figure 5: Mean queue length at one of the queues in the two queue GPS system

4.2.2 Three Queue System

Consider the case now, where the GPS server is shared by three classes of packets with weights $\phi_1 = 3, \phi_2 = 1, \phi_3 = 1$. packets arrive at each class as a Poisson process of rate λ, with exponentially distributed service requirements with mean 1. The sum queue length process is again a Markov process, and the mean sum queue length is plotted in Fig 6, along with the computed bounds.

Figs. 7 and 8 exhibit the mean queue lengths at queue 1 and at queue 2 respectively. The mean delay observed at queue 3 is identical to that observed at queue 2. In this case, note that trying to bound the delay at queues 2 and 3, assuming that the other queues are saturated and queue 1 is always busy leads to a predicted mean delay of ∞. This is because, the system is only stable

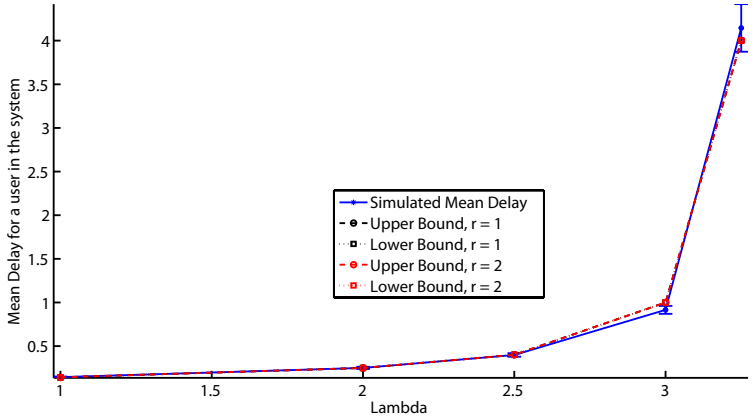


Figure 6: Mean system queue length

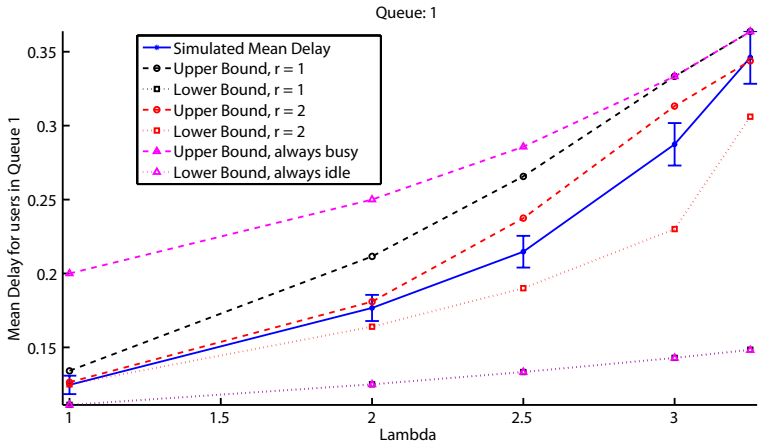


Figure 7: Mean queue length at queue 1

because queues 2 and 3 can borrow the excess capacity of the server when queue 1 is idle. Our simulations demonstrate that the SDP formulation provides reasonably tight bounds on the mean queue lengths of the individual queues when $r \geq 2$.

4.2.3 Choosing Queue Weights

The choice of class weights is a crucial factor affecting the performance experienced by packets in the system. In systems, such as wireless systems, where the maximum rates at which different classes can be served varies, a systematic method to choose the class weights is not apparent. Note that, in this case, even the sum queue length process is not a Markov process, as the system is non-work conserving. The close bounds provided by our formulation can be exploited to determine

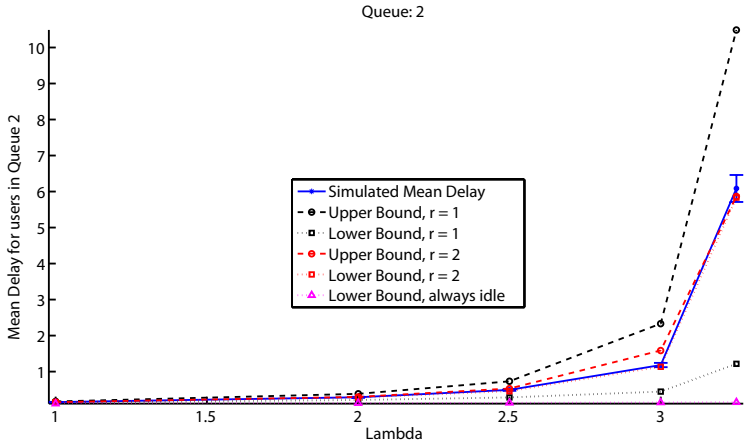


Figure 8: Mean queue length at queue 2

the weights that should be associated with each class, if the objective function is a weighted combination of the queue length moments.

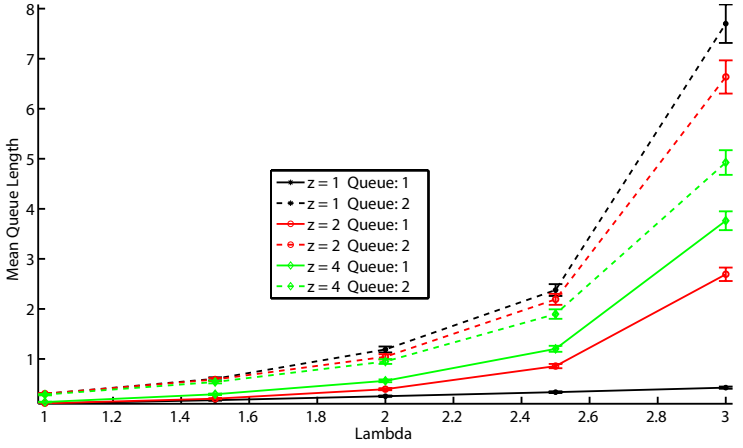


Figure 9: Mean queue lengths resulting from the different objective functions

Consider a two queue system, with $R_1 = 10, R_2 = 5$, where the arrival rate to each class is equal to λ. We consider an objective function of the form $\mathbf{E}[Q_1^z + Q_2^z]$, and use the lower bound from the SDP formulation to approximate the value of the objective function for various class weights. The weight associated with class 1 is assumed to be $\phi_1 = \phi$, the weight associated with class 2 is then $\phi_2 = 1 - \phi$. We do a simple line search to determine the weight to be associated with class 1 to minimize the objective function. As z is increased, the penalty associated with high queue lengths increases rapidly, and the objective function tends to balance the performance experienced by the queues.

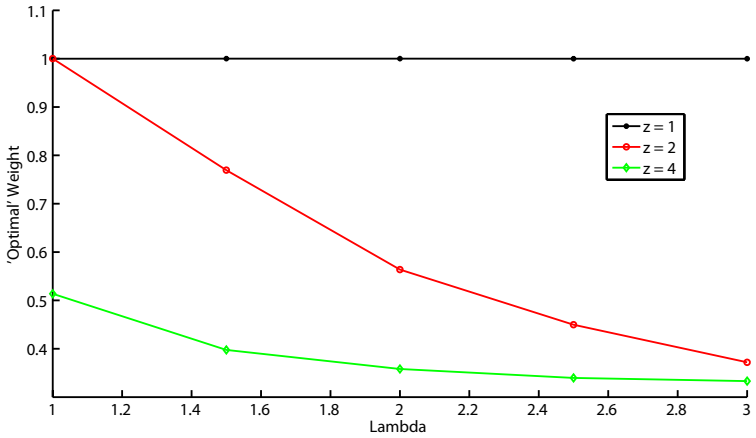


Figure 10: Optimal weights for the different objective functions

Fig. 9 exhibits the mean queue lengths of the two queues against the arrival rate λ, for $z = 1, 2, 4$, and Fig. 10 plots the weight determined using the SDP formulation. When $z = 1$, the objective is to minimize the system queue length. The optimal policy here is to always serve queue 1, whenever both queues are busy and $\phi = 1$ irrespective of the load, as shown in Fig. 10. However, this leads to very large queue lengths at queue 2. Increasing the value of z leads to larger weights being associated with queue 2, in order to balance the queue lengths at the two queues. Fig. 9 clearly shows the impact of the larger weights that are picked using the SDP formulation on the mean queue lengths of the two queues.

4.3 A Multiple Access System

The final example we consider is an ALOHA-like [23, 24] slotted multiple access system, where N users contend for access to a shared channel in order to communicate with a central server. We use a collision model to capture the effect of simultaneous transmissions, i.e., if a user is the only one contending for the channel in a slot, the transmission is successful and a 1Kb portion of the file is transmitted to the sever. If more than one user transmits at the same time, there is a collision and all transmissions are unsuccessful. Files with exponentially distributed sizes and mean 2Mb arrive for transmission at each user as a Poisson process with mean arrival rate λ. All users with files waiting for transmission contend at a time slot with probability P_C. Our aim is to find a contention probability that results in low file transfer delays and high user throughputs.

Clearly, the contention probability has to be chosen in a load-dependent manner. Intuitively, when the system load is low, the number of users with non-empty queues at any given time will be low and those users should contend with higher probability. However, at high loads, when a larger number of users are likely to have non-empty queues users should contend with a lower probability in order to avoid collisions. If all N users are saturated, the contention probability resulting in the highest throughput is $P_C = 1/N$. This can be seen easily by differentiating the probability of there being exactly one contending user,

$$\text{Pr.(exactly one contending user)} = \text{Pr.(successful transmission)} = P_C(1-P_C)^{N-1},$$

with respect to the contention probability P_C and equating to zero. However, in the dynamic system, the number of busy users varies as files arrive and are transmitted. In this case, it is not clear how the contention probability should be chosen.

We model this system using a fluid coupled processors model, such that each user transmits data at a rate of $1000P_C(1-P_C)^{n-1}$ when n of the N users have data to send. Our goal is to find, using the lower bound from the semidefinite formulation as an approximation, the contention probability that minimizes the mean sum system queue length. We consider a system with three users, and as before use a simple line search to find the common contention probability that minimizes the mean sum queue length of the system.

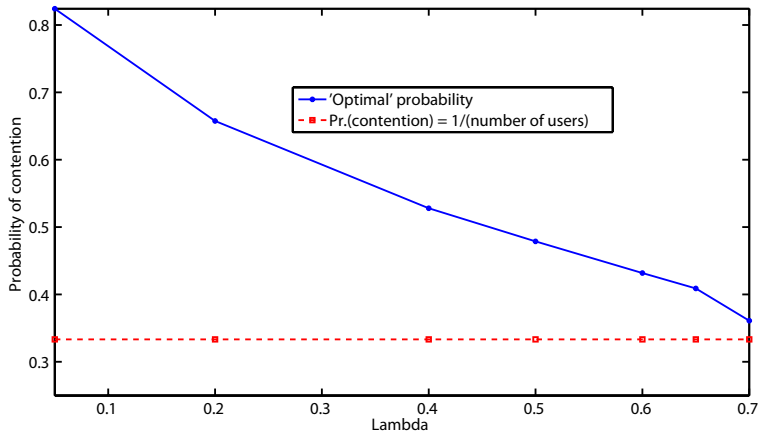


Figure 11: Optimal probability of contention

Fig. 11 shows how the contention probability determined using the SDP formulation varies

with the file arrival rate λ. As expected, the contention probability is high when load is low and collisions are rare, and decreases with increasing loads in order to avoid collisions. At very high loads, the contention probability approaches $\frac{1}{3}$, the optimal contention probability in the saturated system where all users have backlogged queues.

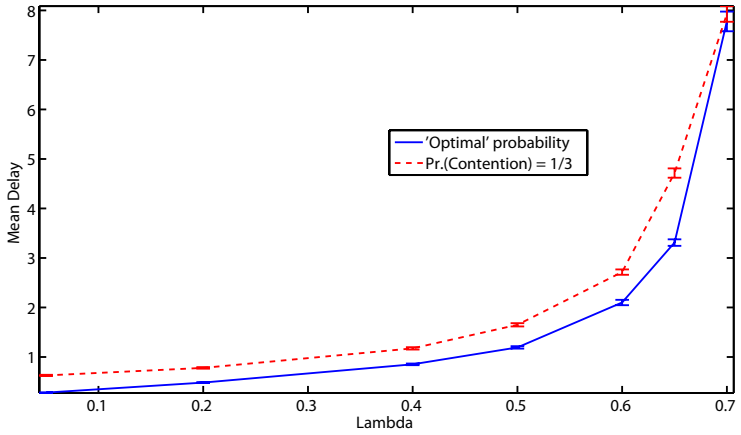


Figure 12: Mean delay comparison

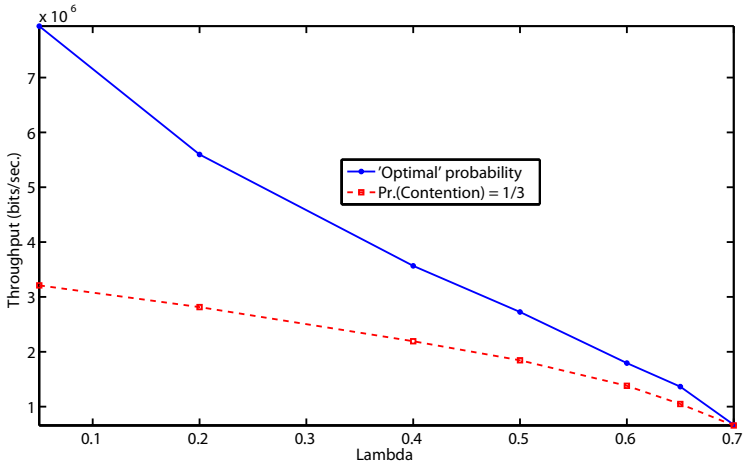


Figure 13: Mean throughput comparison

Figs. 12 and 13 exhibit the mean file transfer delay and mean user throughput for the system using the contention probability determined using our SDP formulation and the system with a contention probability of $\frac{1}{3}$. At very high loads, performance of the two schemes are close to identical. This is to be expected as the probability of contention chosen using the semidefinite formulation is very close to $\frac{1}{3}$. At lower loads, the mean file transfer delay and the mean user throughputs are much improved when using the contention probability from the SDP formulation.

This is particularly clear from the plot of the mean user throughputs against the user arrival rate. Thus, we can clearly see that the coupled processor model accurately reflects the dynamics of the system, and shows the potential of the semidefinite formulation in modeling and optimizing such systems.

5 Conclusion and Future Work

We proposed an approach to model coupled queuing systems using a semidefinite programming approach, and obtained bounds on linear functions of the moments of the queue lengths in the system at steady state. The resultant bounds were used to study some example scenarios, where they were found to closely approximate the actual values. The proposed approach can be extended, to arrive at bounds on the tail probabilities of the individual queue length distributions in a manner similar to that used in [15]. Another possible area of study is to examine systems with general service requirements, or those where the service requirements are modeled using phase type distributions. Another area that requires further study is when the queues in the system support multiple classes of users. Such a model could potentially be used to better model the dynamics of wireless networks coupled by interference. The various user classes could be used to capture the differences between users at different locations within the network. This would lead to a better understanding of the effect of interference on the delays perceived by users.

One drawback of the proposed formulation is the exponential scaling of the size of the semidefinite optimization as the number of queues increases. This is not surprising, as the size of the problem also increases exponentially with the number of queues. However, this is not the case in some relevant special cases like the three base station example in Section 4.1.2 where the service rate depends only on the number of active queues, and the multiple-access problem studied in Section 4.3. In such cases, we propose to explore formulations that exhibit a more graceful scaling with increasing problem size.

References

[1] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: the single-node case," *IEEE/ACM Trans. Netw.*, vol. 1, no. 3, pp. 344–357, 1993.

[2] A. Demers, S. Keshav, and S. Shenker, "Analysis and simulation of a fair queueing algorithm," in *Symposium proceedings on Communications architectures & protocols*. Austin, Texas, United States: ACM, 1989, pp. 1–12.

[3] T. Bonald, S. Borst, and A. Proutiere, "Inter-cell scheduling in wireless data networks," in *European Wireless Conference*, 2005.

[4] S. Borst, "User-level performance of channel-aware scheduling in wireless data networks," in *INFOCOM 2003*, vol. 1, March-April 2003, pp. 321 – 331.

[5] B. Rengarajan and G. de Veciana, "Architecture and abstractions for environment and traffic aware system-level coordination of wireless networks: The downlink case," in *INFOCOM 2008*, Aprol 2008, pp. 502–510.

[6] Z.-l. Zhang, "Large deviations and the generalized processor sharing scheduling for a two-queue system," *Queueing Syst. Theory Appl.*, vol. 26, no. 3-4, pp. 229–254, 1997.

[7] D. Bertsimas, I. C. Paschalidis, and J. N. Tsitsiklis, "Large deviations analysis of the generalized processor sharing policy," *Queueing Syst. Theory Appl.*, vol. 32, no. 4, pp. 319–349, 1999.

[8] P. Dupuis and K. Ramanan, "A skorokhod problem formulation and large deviation analysis of a processor sharing model," *Queueing Syst. Theory Appl.*, vol. 28, no. 1-3, pp. 109–124, 1998.

[9] Z.-l. Zhang, "Large deviations and the generalized processor sharing scheduling for a multiple-queue system," *Queueing Systems*, vol. 28, no. 4, pp. 349–376, 1998.

[10] S. Borst, O. Boxma, and P. Jelenkovic, "Coupled processors with regularly varying service times," in *IEEE INFOCOM 2000*, vol. 1, 2000, p. 157164.

[11] S. Borst, O. Boxma, and M. van Uitert, "The asymptotic workload behavior of two coupled queues," *Queueing Systems*, vol. 43, no. 1-2, pp. 81–102, January 2003.

[12] G. Fayolle and R. Lasnogorodski, "Two coupled processors: The reduction to a riemann–hilbert problem," *Wahrscheinlichkeitstheorie*, no. 3, pp. 1–27, Jan. 1979.

[13] A. G. Konheim, I. Meilijson, and A. Melkman, "Processor-sharing of two parallel lines," *J. Appl. Probab.*, vol. 18, no. 4, pp. 952–956, 1981.

[14] F. Guillemin and D. Pinchon, "Analysis of generalized processor-sharing systems with two classes of customers and exponential services," *Journal of Applied Probability*, vol. 41, no. 3, pp. 832–858, 2004.

[15] D. Bertsimas and K. Natarajan, "A semidefinite optimization approach to the steady-state analysis of queueing systems," *Queueing Syst. Theory Appl.*, vol. 56, no. 1, pp. 27–39, 2007.

[16] D. Bertsimas and I. Popescu, "Optimal inequalities in probability theory: A convex optimization approach," *SIAM J. on Optimization*, vol. 15, no. 3, pp. 780–804, 2005.

[17] J. Lasserre, "Bounds on measures satisfying moment conditions," *Annals of Applied Probability*, vol. 12, pp. 1114–1137, 2002.

[18] A. Borovkov and S. Foss, "Stochastically recursive sequences and their generalizations," *Siberian Advances in Mathematics*, vol. 2, no. 1, pp. 16–81, 1992.

[19] R. Loynes, "The stability of a queue with non-independent inter-arrival and service times," *Proc. Cambr. Phil. Soc.*, vol. 58, no. 3, pp. 497–520, 1962.

[20] L. Zuluaga and J. F. Pena, "A conic programming approach to generalized tchebycheff inequalities," *Mathematics of Operations Research*, vol. 30, no. 2, pp. 369–388, 2005.

[21] D. Henrion and J. . B. Lasserre, "Gloptipoly: global optimization over polynomials with matlab and sedumi," *ACM Transactions on Mathematical Software*, vol. 29, no. 2, pp. 165–194, 2003.

[22] J. F. Sturm and the Advanced Optimization Laboratory at McMaster University, "Sedumi version 1.1r3," 2006, see sedumi.mcmaster.ca.

[23] N. Abramson, "The aloha system-another alternative for computer communications," in *Proceedings of Fall Joint Computer Conference, AFIPS Conference*, vol. 36, 1970, pp. 295–298.

[24] L. G. Roberts, "Aloha packet system with and without slots and capture," *SIGCOMM Comput. Commun. Rev.*, vol. 5, no. 2, pp. 28–42, 1975.