Performance of Multiclass Markovian Queueing Networks

Prof. Dimitris Bertsimas
dbertsim@aris.mit.edu

Dr. David Gamarnik
gamarnik@watson.ibm.com

Prof. John Tsitsiklis
jnt@mit.edu

Abstract

We study the distribution of steady-state queue lengths in multiclass queueing networks under a stable policy. We propose a general methodology based on Lyapunov functions, for the performance analysis of infinite state Markov chains and apply it specifically to Markovian multiclass queueing networks. We establish a deeper connection between stability and performance of such networks by showing that if there exist linear and piecewise linear Lyapunov functions that show stability, then these Lyapunov functions can be used to establish geometric type lower and upper bounds on the tail probabilities, and thus bounds on the expectation of the queue lengths. As an example of our results, for a re-entrant line queueing network with two processing stations operating under a work-conserving policy we show that $E[L] = O\left(\frac{1}{(1-\rho^*)^2}\right)$, where L is the total number of customers in the system, and ρ^* is the maximal actual or virtual traffic intensity in the network. This extends a recent result by Dai and Vande-Vate, which states that a re-entrant line queueing network with two stations is globally stable if $\rho^* < 1$. We also present several results on the performance of multiclass queueing networks operating under general Markovian, and in particular, priority policies. The results in this paper are the first that establish explicit geometric type upper and lower bounds on tail probabilities of queue lengths, for networks of such generality. Previous results provide numerical bounds and only on the expectation, not the distribution, of queue lengths.

1 Introduction

The focus of this paper is performance analysis of multiclass queueing networks. Specifically, we are interested in estimating the steady-state queue lengths in the network, when interarrival and service times are exponentially distributed, assuming a stable scheduling policy is used.

The performance of queueing networks is largely an open research area. Some of the earlier and classical results include product form probability distributions for Jackson and BCMP type networks (see Gelenbe and Mitrani [13]). It was realized, however, that the presence of multiple classes does not allow, in general, for a product form distribution even if the interarrival and service times have exponential distributions and the First-In-First-Out policy is used. Several papers (Bertsimas, Paschalidis and Tsitsiklis [5], Kumar and Kumar [16], Kumar and Meyn [17], Jin, Ou and Kumar [14]) have analyzed the performance of multiclass queueing networks using quadratic Lyapunov functions. A certain linear program is constructed, which provides numerical bounds on the achievable performance region. The performance results obtained using quadratic Lyapunov functions were later analyzed and extended in a simpler and more intuitive way, using conservation laws (Bertsimas and Nino-Mora [4]).

The performance analysis of multiclass queueing networks is at least as hard as the stability problem for which no general conditions are available. It is known that the natural load condition $\rho_\sigma < 1$ for each station σ is necessary, but not sufficient, for stability; a variety of counterexamples have been constructed by Rybko and Stolyar [20], Lu and Kumar [18], Bramson [6], Seidman [21], Dai, Vande Vate and Hasenbein [8]. Sufficient conditions for stability have been found using Lyapunov functions by Dai and Weiss [10] and Down and Meyn [11]. Furthermore, fluid models were found to be a very useful tool for stability analysis. Dai's theorem [7] shows that the stability of a fluid model implies stability of a corresponding stochastic model. A complete characterization of fluid networks with two stations which are stable under any work-conserving policy ("globally stable") was obtained by Bertsimas, Gamarnik and Tsitsiklis [2] and subsequently by Dai and Vande Vate [9]. The second work used a very intuitive notion of virtual stations to explain instability in networks with two stations. Both works ([2] and [9]) prove that the existence of a piecewise linear Lyapunov function is both necessary and sufficient for global stability of fluid networks with two stations.

1.1 Our Results

The goal of this paper is to turn some of the stability analysis tools into useful performance analysis tools. We will show how linear and piecewise linear Lyapunov functions, and virtual stations can be used to obtain upper and lower bounds on the steady-state queue lengths. For many examples considered in this paper the upper bounds are finite if and only if the network is stable. Our contributions are summarized as follows. We start in Section 3 with an analysis of countably infi-

nite Markov chains. We show that if there exists a Lyapunov function proving the stability of the Markov chain, then certain computable upper and lower bounds hold on the steady-state queue length probability distribution as well as on its expectation. We then apply this methodology, in Sections 4 and 5, to the performance analysis of multiclass queueing networks with exponentially distributed interarrival and service times. Specifically, we use the notion of a *virtual station*, introduced by Dai and Vande-Vate in [9]. They showed that in networks with two stations, some priority policies lead to certain groups V of customer classes, called virtual stations, which cannot be served simultaneously. As a result, if the corresponding *virtual traffic intensity* $\rho(V) \equiv \sum_{i \in V} \rho_i$ is bigger than one, then the network is unstable. We prove that for networks with two stations, if V is a virtual station, with the corresponding virtual traffic intensity $\rho(V)$, then

$$E[L] \geq \frac{\rho(V)}{4(1-\rho(V))},$$

where L is the total number of customers in the network. These lower bounds are extended to networks with more than two stations.

It was also proven in [9] that queueing networks with two stations are globally stable if the maximum of all actual and virtual traffic intensities, denoted by ρ^*, is less than one for the original network and for a certain set of subnetworks. Also if $\rho^* > 1$, then the corresponding fluid network is not globally stable. Whether this holds true for stochastic Markovian networks is not known.

We show that ρ^* is a fundamental performance parameter. For re-entrant line networks with two stations, we show that if $\rho^* < 1$, then the following upper bound holds under any work-conserving policy

$$E[L] \leq \frac{C}{(1-\rho^*)^2}$$

where C is some constant, expressed explicitly in terms of the parameters of the network. An important implication of this result is that the performance region (the set of vectors of expected queue lengths obtained under different work-conserving scheduling policies) is bounded if and only if the corresponding fluid network is globally stable.

Our results show a deeper connection between stability and performance of multiclass queueing networks. Also the results in this paper are the first ones that use linear and piecewise linear Lyapunov functions for performance analysis. Previous methods for performance analysis have used quadratic Lyapunov functions, which have certain limitations. In particular, an example of a globally stable queueing network with two stations was constructed in [11] for which the quadratic Lyapunov function method leads to an infinite (inconclusive) upper bound, yet a piecewise linear Lyapunov function gives a finite upper bound. The methods developed here, on the other hand, match the sharpest known stability condition $\rho^* < 1$.

2 Queueing Model and Assumptions

We consider a network consisting of J single server stations, which are denoted by σ_j, $j = 1, 2, \ldots, J$. The network includes I types of customers, where customers of type $i = 1, 2, \ldots, I$ arrive to the network from an exogenous source. The arrival process corresponding to type i is assumed to be an independent Poisson process with rate λ_i. Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_I)$ denote the vector of arrival rates and let $\lambda_{\min} = \min_i\{\lambda_i\}$. Without loss of generality, we assume that $\lambda_{\min} > 0$. Similarly, we define $\lambda_{\max} = \max_i\{\lambda_i\}$. Customers of type i go through J_i stages, each of which corresponds to a service completion on a particular station. We denote these stations by $\sigma_{i,1}, \sigma_{i,2}, \ldots, \sigma_{i,J_i}$. The processing time of a type i customer at station $\sigma_{i,k}, k = 1, 2, \ldots, J_i$, is assumed to be exponentially distributed with rate $\mu_{i,k}$ and is independent from the processing times of all other stages of this type, from the processing times of the other types, and from the interarrival times. We let $\boldsymbol{\mu} = (\mu_{i,k})_{1 \leq i \leq I, 1 \leq k \leq J_i}$ denote the vector of service rates. Customers of type i receiving service at station $\sigma_{i,k}$ are called class (i,k) customers. Let $N = \sum_{i=1}^{I} J_i$ be the total number of classes. For convenience, we will also identify every station σ_j with the set of classes associated with this station. Let $C_{i,k} = j$ if class (i,k) customers are served at station σ_j. For $k \geq J_i$ we let $C_{i,k} = 0$. Let $\boldsymbol{C}$ denote the corresponding $I \times J_{\max}$ matrix, where $J_{\max} = \max_i J_i$. The matrix $\boldsymbol{C}$ defines the topology of the network. We assume that the buffers at each station have infinite capacity and no customers renege from the queue before receiving service. A queueing network of the form just described is called a Markovian multiclass queueing network with deterministic routing. The parameters $\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{C}$ constitute the primary parameters of the network and we denote the network by $(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{C})$. For each class (i,k), we let $\rho_{i,k} = \lambda_i/\mu_{i,k}$ be the *nominal load* of this class. For each station σ_j, $j = 1, 2, \ldots, J$, we define the *nominal load (traffic intensity)* as $\rho_{\sigma_j} \equiv \sum_{(i,k) \in \sigma_j} \rho_{i,k}$.

The evolution of a queueing network is fully specified only when a scheduling discipline is given. The scheduling discipline (policy) describes which customers (if any) are served at any moment at each station. Within each class, the customers are served in First-In-First-Out (FIFO) fashion. Therefore, the service discipline only specifies which *customer type* is served at any given moment. We will assume throughout the paper that the scheduling policies implemented are *Marko-*

vian, namely, scheduling decisions are purely a function of the system state, which in our case is the vector of all queue lengths. We also allow preemption. For example, preemptive priority policies are Markovian. We will be considering mostly *work-conserving* policies: each processing station is required to work on some customer, if there are any present at this station.

Given a multiclass queueing network $(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{C})$ and some scheduling policy, we let $\boldsymbol{Q}(t) = (Q_{i,k}(t))_{1 \le i \le I, 1 \le k \le J_i}$ denote the vector of queue lengths at time t. Our focus is on estimating the distribution of the random vector $\boldsymbol{Q}(t)$ in steady-state. A necessary condition for the existence of a steady-state is the load condition $\rho_{\sigma_j} < 1$ for each $j = 1, 2, \ldots, J$.

Definition 1 *A scheduling policy w is defined to be stable if the Markov process $\boldsymbol{Q}(t), t \ge 0$, admits a stationary probability distribution $\pi = \pi(w)$ satisfying*

$$\sum_{i,k} E[Q_{i,k}(t)] < \infty, \qquad \text{for all } t \ge 0. \tag{1}$$

A queueing network is defined to be globally stable if every work-conserving Markovian policy is stable.

3 Infinite Markov Chains and Lyapunov Functions

Let $\boldsymbol{X}(t)$, $t = 0, 1, 2, \ldots$, be a discrete time, discrete state Markov chain which takes values in some countable set $\mathcal{X}$. The transitions occur at integer times $t = 0, 1, 2, \ldots$. For any two vector $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, let $p(\boldsymbol{x}, \boldsymbol{x}')$ denote the transition probabilities

$$p(\boldsymbol{x}, \boldsymbol{x}') \equiv \mathrm{P}\{\boldsymbol{X}(t+1) = \boldsymbol{x}' | \boldsymbol{X}(t) = \boldsymbol{x}\}.$$

If a stationary probability distribution π on the state space $\mathcal{X}$ exists, it satisfies

$$\sum_{\boldsymbol{x} \in \mathcal{X}} \pi(\boldsymbol{x}) = 1,$$

and for all $\boldsymbol{x} \in \mathcal{X}$

$$\pi(\boldsymbol{x}) = \sum_{\boldsymbol{x}' \in \mathcal{X}} \pi(\boldsymbol{x}') p(\boldsymbol{x}', \boldsymbol{x}). \tag{2}$$

The existence of a stationary distribution is usually established by constructing a certain Lyapunov function. For a survey of Lyapunov methods for stability analysis of Markov chains, see [19].

Definition 2 *A nonnegative function $\Phi : \mathcal{X} \to \Re_+$ is said to be a Lyapunov function if there exist some $\gamma > 0$ and $B \ge 0$, such that for any $t = 1, 2, \ldots$ and any $\boldsymbol{x} \in \mathcal{X}$, with $\Phi(\boldsymbol{x}) > B$*

$$E[\Phi(\boldsymbol{X}(t+1)) | \boldsymbol{X}(t) = \boldsymbol{x}] \le \Phi(\boldsymbol{x}) - \gamma. \tag{3}$$

Also a nonnegative function $\Phi : \mathcal{X} \to \Re_+$ is said to be a lower Lyapunov function if there exists some $\gamma > 0$, such that for any $t = 1, 2, \ldots$ and any $\boldsymbol{x} \in \mathcal{X}$, with $\Phi(\boldsymbol{x}) > 0$

$$E[\Phi(\boldsymbol{X}(t+1)) | \boldsymbol{X}(t) = \boldsymbol{x}] \ge \Phi(\boldsymbol{x}) - \gamma.$$

We assume that the Markov chain $\boldsymbol{X}(t)$ is positive recurrent, and we denote by π the corresponding stationary distribution. Namely, $\pi(\boldsymbol{x})$ is the steady-state probability $\mathrm{P}_\pi\{\boldsymbol{X}(t) = \boldsymbol{x}\}$ that the chain is in a certain state $\boldsymbol{x} \in \mathcal{X}$. Also, we denote by $E_\pi[\cdot]$ the expectation with respect to the probability distribution π. For a given function $\Phi : \mathcal{X} \to \Re_+$, let

$$\nu_{\max} \equiv \sup_{\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X} : p(\boldsymbol{x}, \boldsymbol{x}') > 0} |\Phi(\boldsymbol{x}') - \Phi(\boldsymbol{x})|, \tag{4}$$

and

$$\nu_{\min} \equiv \inf_{\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X} : p(\boldsymbol{x}, \boldsymbol{x}') > 0, \Phi(\boldsymbol{x}) < \Phi(\boldsymbol{x}')} (\Phi(\boldsymbol{x}') - \Phi(\boldsymbol{x})). \tag{5}$$

Also let

$$p_{\max} = \sup_{\boldsymbol{x} \in \mathcal{X}} \sum_{\boldsymbol{x}' \in \mathcal{X}, \Phi(\boldsymbol{x}) < \Phi(\boldsymbol{x}')} p(\boldsymbol{x}, \boldsymbol{x}'), \tag{6}$$

and

$$p_{\min} = \inf_{\boldsymbol{x} \in \mathcal{X}} \sum_{\boldsymbol{x}' \in \mathcal{X}, \Phi(\boldsymbol{x}) < \Phi(\boldsymbol{x}')} p(\boldsymbol{x}, \boldsymbol{x}'). \tag{7}$$

We are interested in Lyapunov functions with finite $\nu_{\max}$, and lower Lyapunov functions with positive $\nu_{\min}$ and $p_{\min}$. The following theorem is a key result for the remainder of the paper.

Theorem 1 *Consider a Markov chain $\boldsymbol{X}(t)$ with a stationary probability distribution π such that $E_\pi[\Phi(\boldsymbol{X}(t))] < \infty$. If there exists a Lyapunov function Φ with drift $\gamma > 0$, and exception parameter $B \ge 0$, then for any $m = 0, 1, 2, \ldots$*

$$\mathrm{P}_\pi\{\Phi(\boldsymbol{X}(t)) > B + 2\nu_{\max} m\} \le \left(\frac{p_{\max}\nu_{\max}}{p_{\max}\nu_{\max} + \gamma}\right)^{m+1}. \tag{8}$$

As a result,

$$E_\pi[\Phi(\boldsymbol{X}(t))] \le B + \frac{2p_{\max}(\nu_{\max})^2}{\gamma}. \tag{9}$$

If there exists a lower Lyapunov function Φ with drift $\gamma > 0$, then for any $m = 0, 1, 2, \ldots$

$$\mathrm{P}_\pi\{\Phi(\boldsymbol{X}(t)) \ge (1/2)\nu_{\min} m\} \ge \left(\frac{(1/2)p_{\min}\nu_{\min}}{(1/2)p_{\min}\nu_{\min} + \gamma}\right)^{m}. \tag{10}$$

As a result,

$$E_\pi[\Phi(\boldsymbol{X}(t))] \ge \frac{p_{\min}(\nu_{\min})^2}{4\gamma}. \tag{11}$$

4 Lower Bounds on Queue Lengths Using Linear Lower Lyapunov Functions

In this section, we use linear lower Lyapunov functions to find closed form lower bounds on the distribution and expectation of steady-state queue lengths, which hold when an arbitrary stable scheduling policy is implemented.

4.1 Closed Form Lower Bounds Under an Arbitrary Work-Conserving Policies

Given a stable scheduling policy w, let $\pi = \pi(w)$ denote the corresponding stationary distribution. In order to apply the results of Section 3 on discrete time Markov chains we consider a uniformized embedded Markov chain $Q(\tau_s), s = 0, 1, 2, \ldots,$ instead of the original process $Q(t)$. For details see [3]. For each station σ_j, we now construct a lower Lyapunov function. For any class (i, k), let

$$\rho_{i,k}^{\sigma_j +} = \sum_{k':(i,k')\in\sigma_j, k'\geq k} \rho_{i,k'}. \tag{12}$$

Let

$$\Phi_j(\boldsymbol{Q}) = \sum_{i,k} \frac{\rho_{i,k}^{\sigma_j +}}{\lambda_i} Q_{i,k}. \tag{13}$$

Proposition 1 *Let w be an arbitrary Markovian policy. Then, Φ_j is a lower Lyapunov function with drift $\gamma_j = 1 - \rho_{\sigma_j}$ and $p_{\min} = \sum_i \lambda_i, \nu_{\min} \geq \rho_{\sigma_j}/\lambda_{\max}$.*

Proof: see [3].

We now are ready to state the main result of this section. The result is presented for *re-entrant line* type queueing networks, that is $I = 1$. In this case, all customers follow the same route in the network. The extension of this result for networks with multiple routes can be found in [3]. We denote by $Q_k(t)$ the queue length at the k-th stage in the network. The parameters $\rho_{i,k}, \rho_{i,k}^{\sigma_j +}$ are denoted simply by ρ_k and $\rho_k^{\sigma_j +}$.

Theorem 2 *Given a re-entrant line type queueing network $(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{C})$, operating under any stable Markovian policy, the following lower bounds hold on the number of customers in the network in steady-state. For each $j = 1, 2, \ldots, J$, and $m = 0, 1, 2, \ldots$*

$$\mathrm{P}\left\{\sum_k \rho_k^{\sigma_j +} Q_k(t) \geq \frac{\rho_{\sigma_j}}{2} m\right\} \geq \left(\frac{\rho_{\sigma_j}}{2 - \rho_{\sigma_j}}\right)^m,$$

and

$$E\left[\sum_k \rho_k^{\sigma_j +} Q_k(t)\right] \geq \frac{\rho_{\sigma_j}^2}{4(1 - \rho_{\sigma_j})}.$$

4.2 Closed Form Lower Bounds Under a Priority Policy

In this section, we derive lower bounds on the tail probabilities and the expected number of customers in a multiclass queueing network operating under a priority policy w_θ that is described by a permutation θ of the set of classes $\{(i, k)\}_{1\leq i\leq I, 1\leq k\leq J_i}$. For two classes (i, k), (i', k') associated with the same station σ_j, we say that class (i', k') has a higher priority than class (i, k) if $\theta(i', k') < \theta(i, k)$. A corresponding priority policy w_θ is a policy which for every station σ_j works on a class with the highest priority.

The lower bounds to be presented in this section are based on the concept of a virtual station and virtual traffic intensity introduced by Dai and Vande-Vate in [9], where the virtual station concept is used for the stability analysis. The definition of a virtual station for networks with two stations and its extension to networks with K stations (called K virtual station) is given in [1],[8],[3] and is omitted here for the lack of space. Intuitively, it is shown in [8] and [3] that if a set of K classes V is a K-virtual stations and priority is given to classes in V then in steady state only $K-1$ classes of V can be served simultaneously. Using this result and by constructing a lower Lyapunov function very similar to the one given by (13) we obtain the following bounds. The result is again stated for re-entrant line type networks. Its proof and extension to general networks can be found in ([3]).

Theorem 3 *Suppose that $(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{C})$ is a re-entrant line type queueing network and that a set of classes V is a K-virtual station. If a stable priority policy w_θ gives priority to classes in V over classes outside V, then the following lower bound holds on the number of customers in the network in steady-state. For each $m = 0, 1, 2, \ldots,$*

$$\mathrm{P}\left\{\sum_k \rho_k^{V+} Q_k(t) \geq \frac{\rho(V)}{2} m\right\} \geq \left(\frac{\rho(V)}{2(K-1) - \rho(V)}\right)^m,$$

and

$$E\left[\sum_k \rho_k^{V+} Q_k(t)\right] \geq \frac{\rho^2(V)}{4(K - 1 - \rho(V))},$$

where $\rho(V) \equiv \sum_{(i,k)\in V} \rho_{i,k}$ and $\rho_{i,k}^{V+} \equiv \sum_{(i,k')\in V, k'\geq k} \rho_{i,k'}$, $1 \leq i \leq I$, $1 \leq k \leq J_i$.

4.3 Example

Consider the Lu-Kumar network on Figure 1. This re-entrant line network was introduced first by Lu and Kumar in [18]. For this network $\rho_i = \lambda/\mu_i$, $i = 1, 2, 3, 4$, $\rho_{\sigma_1} = \rho_1 + \rho_4$, $\rho_{\sigma_2} = \rho_2 + \rho_3$. We have $\rho_1^{\sigma_1 +} = \rho_1 + \rho_4$ and $\rho_i^{\sigma_1 +} = \rho_4$ for $i = 2, 3, 4$. Also, $\rho_i^{\sigma_2 +} = \rho_2 + \rho_3$ for $i = 1, 2$, $\rho_3^{\sigma_2 +} = \rho_3$ and $\rho_4^{\sigma_2 +} = 0$.

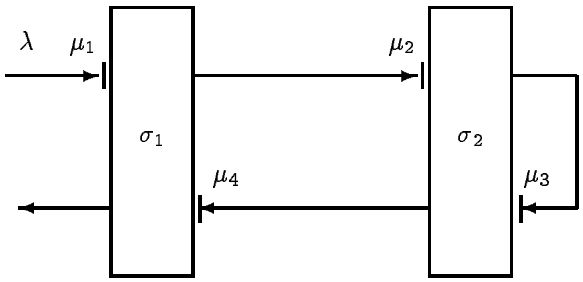

Figure 1: Lu-Kumar network.

Proposition 2 *If the network on Figure 1 operates under priority policy w_θ with priority rule $\theta(4) < \theta(1), \theta(2) < \theta(3)$, then the following bounds hold*

$$\mathrm{P}_{\pi_\theta}\left\{2\frac{(\rho_2+\rho_4)(Q_1(t)+Q_2(t))+\rho_4(Q_3(t)+Q_4(t))}{\rho_2+\rho_4} \geq m\right\}$$

$$\geq \left(\frac{\rho_2+\rho_4}{2-\rho_2-\rho_4}\right)^m,$$

for all $m = 0, 1, 2, \ldots$, and

$$E_\pi[(\rho_2+\rho_4)(Q_1(t)+Q_2(t))+\rho_4(Q_3(t)+Q_2(t))] \geq$$

$$\frac{1}{4}\frac{(\rho_2+\rho_4)^2}{(1-\rho_2-\rho_4)}. \qquad (14)$$

Proof: The proof is obtained by applying Theorem 3 to the virtual station $V = \{2, 4\}$. ■

5 Upper Bounds for Networks With Two Stations

In this section, we provide explicit performance bounds for queueing networks with two stations. We will consider only re-entrant line queueing networks. The Poisson arrival rate is denoted by λ. An explicit and tight characterization of global stability of fluid networks with two stations is given in [9]. Specifically, it is proven that a fluid queueing network with two stations is globally stable if and only if the maximal of all the real and virtual traffic intensities ρ^* is smaller than one. From this result and Dai's theorem [7] connecting fluid and stochastic stability, the condition $\rho^* < 1$ is also sufficient for global stability of the stochastic network (with arbitrary and not necessarily exponential service distribution). In this section, we describe a matching performance result: whenever $\rho^* < 1$, we construct a finite upper bound on the tail probabilities and the expectation of queue lengths in the network. We show that ρ^* is a fundamental performance parameter of the network.

An outline of our approach is as follows. We consider a certain linear program that was considered by Dai and Vande Vate [9] and which is a witness of stability: whenever a feasible solution to this linear program exists a piecewise linear Lyapunov function can be constructed and as a result the network is globally stable. We use the results in [9] to show that if $\rho^* < 1$, then this modified linear program has a feasible solution with positive γ, construct a piecewise linear Lyapunov function out of this solution and use the result of Theorem 1. In addition, by analyzing the linear program we obtain explicit bounds on the solution variables and specifically on the drift γ. The latter allows us to obtain the explicit dependence of the drift on the maximal traffic intensity ρ^*. Our main result is the following theorem, the proof of which can be found in [3].

Theorem 4 *We consider a re-entrant line queueing network with two stations σ_1, σ_2, arrival rate λ and service rates $\mu_1, \mu_2, \ldots, \mu_N$. Class 1 is assumed to belong to station σ_1. If $\rho^* < 1$, then the following upper bounds hold on the steady-state number of customers in the network.*

$$\mathrm{P}\left\{\sum_{i=1}^N \rho_i^{\sigma_1+}Q_i(t) - B \geq m\frac{1+\rho^*+2\rho^*\sum_{i=1}^N \rho_i^{-1}}{1+\sum_{i=1}^N \rho_i^{-1}}\right\} \leq$$

$$\left(\frac{\frac{1}{2}+\frac{1}{2}\rho^*+\rho^*\sum_{i=1}^N \rho_i^{-1}}{\frac{3}{4}+\frac{1}{4}\rho^*+\rho^*\sum_{i=1}^N \rho_i^{-1}}\right)^m$$

and

$$\mathrm{P}\left\{\sum_{i=1}^N \rho_{l(e_2)+1}^{\sigma_1+}\rho_i^{\sigma_2+}Q_i(t) - B \geq m\frac{1+\rho^*+2\rho^*\sum_{i=1}^N \rho_i^{-1}}{1+\sum_{i=1}^n \rho_i^{-1}}\right\} \leq$$

$$\left(\frac{\frac{1}{2}+\frac{1}{2}\rho^*+\rho^*\sum_{i=1}^N \rho_i^{-1}}{\frac{3}{4}+\frac{1}{4}\rho^*+\rho^*\sum_{i=1}^N \rho_i^{-1}}\right)^m$$

for all $m = 0, 1, 2, \ldots,$ where

$$B = \frac{64N(\rho^*\sum_{i=1}^N \rho_i^{-1})^3}{(1+\sum_{i=1}^N \rho_i^{-1})(1-\rho^*)^2}.$$

Also

$$E\left[\sum_{i=1}^N \rho_i^{\sigma_1+}Q_i(t)\right] \leq$$

$$\frac{64N(\rho^*\sum_{i=1}^N \rho_i^{-1})^3}{(1+\sum_{i=1}^N \rho_i^{-1})(1-\rho^*)^2} + \frac{2(1+\rho^*+2\rho^*\sum_{i=1}^N \rho_i^{-1})^2}{(1+\sum_{i=1}^N \rho_i^{-1})(1-\rho^*)},$$

and

$$E\left[\sum_{i=1}^N \rho_{l(e_2)+1}^{\sigma_1+}\rho_i^{\sigma_1+}Q_i(t)\right] \leq$$

$$\frac{64N(\rho^*\sum_{i=1}^N \rho_i^{-1})^3}{(1+\sum_{i=1}^N \rho_i^{-1})(1-\rho^*)^2} + \frac{2(1+\rho^*+2\rho^*\sum_{i=1}^N \rho_i^{-1})^2}{(1+\sum_{i=1}^N \rho_i^{-1})(1-\rho^*)}.$$

In particular,

$$E\left[\sum_{i=1}^N Q_i(t)\right] = O\left(\frac{1}{(1-\rho^*)^2}\right). \qquad (15)$$

6 Conclusions

We have proposed a general methodology based on Lyapunov functions for the performance analysis of infinite state Markov chains and applied it specifically to multiclass queueing networks with exponentially distributed interarrival and service times.

In the full version of this paper [3] it is shown that whenever some piecewise linear Lyapunov function is a witness for the global stability of the network, certain finite upper bounds can be derived on the probability distribution and expectation of queue lengths.

Since piecewise linear Lyapunov functions provide an exact test for stability of fluid networks with two stations, our bounds for two-station networks are finite if and only if the corresponding fluid network is globally stable. Whether this remains true for the original stochastic network remains to be seen.

For re-entrant line type queueing networks with two processing stations closed form bounds were constructed on the distribution and expectation of the number of customers in the network:

$$E[L] = O\left(\frac{1}{(1-\rho^*)^2}\right)$$

where ρ^* is a maximal (actual or virtual) traffic intensity.

The results obtained here are the first ones that establish exponential upper and lower bounds on the distribution of queue lengths in networks of such generality. Previous results on performance analysis of multiclass queueing networks can in general achieve only numerical bounds and only on the expectation of queue lengths.

References

[1] D. Bertsimas. Lecture notes on stability of multiclass queueing networks, 1996.

[2] D. Bertsimas, D. Gamarnik and J. Tsitsiklis. Stability Conditions for Multiclass Fluid Queueing Networks. *IEEE Trans. Automat. Control*, 41, 1618-1631, 1996.

[3] D. Bertsimas, D. Gamarnik and J. Tsitsiklis. Performance of Multiclass Markovian Queueing Networks via Piecewise Linear Lyapunov Functions. Submitted.

[4] D. Bertsimas and J. Nino-Mora. Optimization of multiclass queueing networks with changeover times via the achievable region approach: Part II, the multiple station case. *Math. Oper. Res.*, 24, 2, 331-361, 1999.

[5] D. Bertsimas, I. Paschalidis and J. Tsitsiklis. Optimization of multiclass queueing networks: Polyhedral and nonlinear characterization of achievable performance. *Ann. Appl. Probab.*, 4, 43-75, 1994.

[6] M. Bramson. Instability of FIFO queueing networks. *Ann. Appl. Probab.*, 2, 414-431, 1994.

[7] J.G. Dai. On the positive Harris recurrence for multiclass queueing networks: A unified approach via fluid models. *Ann. Appl. Probab.*, 5, 49-77, 1995.

[8] J. G. Dai, J. J. Hasenbein and J. H. Vande Vate. Stability of a Three-Station Fluid Network. *Queueing Systems*, 33, 293-325, 1999.

[9] J. G. Dai and J. H. Vande Vate. The Stability of Two-Station Multi-Type Fluid Networks. To appear in *Operations Research.*

[10] J. G. Dai and G. Weiss. Stability and instability of fluid models for certain re-entrant lines. *Math. Oper. Res.*, 21, 115-134, 1996.

[11] D. D. Down and S. P. Meyn. Piecewise linear test functions for stability and instability of queueing networks. *Queueing Systems*, 1998.

[12] P. Dupuis and R. J. Williams. Lyapunov functions for semimartingale reflecting Brownian motions. *Annals of Probability*, 22, 680-702, 1994.

[13] E. Gelenbe and L. Mitrani. *Analysis and Synthesis of Computer Systems.* Academic, London, 1980.

[14] C. Humes, Jr., J. Ou and P. R. Kumar. The delay of open Markovian queueing networks: Uniform functional bounds, heavy traffic pole multiplicities, and stability. *Math. Oper. Res.* 22, 921-954, 1997.

[15] P.R. Kumar. Re-entrant lines. *Queueing Systems*, 13, 87-110, 1993.

[16] S. Kumar and P.R. Kumar. Performance bounds for queueing networks and scheduling policies. *IEEE Transactions on Automatic Control* AC-39, 1600-1611, 8, 1994.

[17] P. R. Kumar and S. P. Meyn. Stability of queueing networks and scheduling policies. *IEEE Trans. Autom. Control*, 40, 2, 251-261, 1995.

[18] S.H. Lu and P.R. Kumar. Distributed scheduling based on due dates and buffer priorities. *IEEE Trans. Autom. Control*, 36, 12, 1406-1416, 1991.

[19] S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability.* London: Springer-Verlag, 1993.

[20] A. N. Rybko and A.L. Stolyar. On the ergodicity of stochastic processes describing open queueing networks. *Problemy Peredachi Informatsii*, 28, 3, 3-26, 1992.

[21] T.I. Seidman. First come first serve can be unstable. *IEEE Trans. Autom. Control*, 39, 10, 2166-2170, 1994.