Stability and Performance of Multiclass Queueing Networks

by

David Gamarnik

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 1998

Author ....
Sloan School of Management
September 2, 1997

Certified by ....
Dimitris Bertsimas
Boeing Professor of Operations Research
Thesis Supervisor

Certified by ....
John Tsitsiklis
Professor of Electrical Engineering
Thesis Supervisor

Accepted by ....
Robert Freund
Seley Professor of Operations Research; Codirector, Operations
Research Center

Stability and Performance of Multiclass Queueing Networks

by

David Gamarnik

Submitted to the Sloan School of Management
on September 2, 1997, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Operations Research

Abstract

Multiclass queueing networks have been used to model manufacturing, communication, and computer systems. Questions of stability and performance arise naturally in such systems.

In the first part of the thesis, we address the question of stability of multiclass queueing networks operating under any work-conserving scheduling policy (global stability). It was recently discovered that there are examples of such networks that are unstable, i.e., the total number of customers goes to infinity, even though all processing stations have enough processing capacity. It has also been established that the stability of queueing networks is essentially equivalent to the stability of the underlying fluid model and several sufficient conditions for stability have been found by means of the Lyapunov function approach. However, the exact stability conditions for multiclass queueing networks are unknown. We propose a new and direct approach for stability analysis based on linear programming. We construct a linear program that provides an upper bound on the duration of the time until the fluid network empties. As a result, linear programs with finite optimal cost correspond to stable networks. We prove that this criterion for stability is both necessary and sufficient for networks with two stations. This is the first sharp stability result in such generality. This approach also gives sufficient conditions for stability of networks with more than two processing stations. We then modify this approach specifically for priority policies in multiclass fluid queueing networks. Again a certain linear program is constructed which bounds the duration of the time until the fluid network empties under a specific priority policy. Many early stability results, which were derived using various techniques, are reestablished using our trajectory decomposition approach.

In the second part of the thesis we concentrate on the performance analysis of multiclass queueing networks. The objective is to estimate steady-state queue lengths in queueing networks assuming a priori that the scheduling policy implemented is stable. We propose a very general methodology based on Lyapunov functions, for the performance analysis of infinite state Markov chains and apply it specifically to multiclass exponential type queueing networks. We show, in particular, that whenever some piece-wise linear Lyapunov function witnesses global stability of the network,

certain finite upper bounds hold on the probability distribution and expectation of the queue lengths. Lower bounds are also constructed by means of a linear Lyapunov function. Specifically we show that whenever some piece-wise linear function witnesses global stability of the network, bounds of the form

$$c_1^m \leq \Pr\{\text{total number of customers} \geq m\} \leq c_2^m$$

hold for some computable constants $c_1 < c_2 < 1$. For re-entrant line type networks with two processing stations the constants are expressed explicitly in terms of traffic intensities of the network and the following upper bound is proven on the total expected number of customers in the network

$$E[\text{total number of customers}] = O\Big(\frac{1}{(1-\rho^*)^2}\Big),$$

where ρ^* is the maximal actual or virtual traffic intensity of the network. This extends a recent result by Dai and Vande-Vate stating that re-entrant line fluid queueing networks with two stations are globally stable if and only if $\rho^* < 1$.

We then concentrate on the performance analysis of queueing networks operating under some priority policy. Piece-wise linear and linear functions are used again to construct upper and lower bounds on steady-state queue lengths in such networks. The proposed lower bounds are closed form and are expressed exclusively in terms of traffic intensities of the network. The upper bounds are numerical and are constructed using the solution of linear program which establishes the existence of a piece-wise linear Lyapunov function.

The results proposed here are the first that establish geometric type upper and lower bounds on tail probabilities of queue lengths, for networks of such generality. The previous results on performance analysis can in general achieve only numerical bounds and only on expectation and not the distribution of queue lengths.

Thesis Supervisor: Dimitris Bertsimas
Title: Boeing Professor of Operations Research

Thesis Supervisor: John Tsitsiklis
Title: Professor of Electrical Engineering

To the memory of my father

Acknowledgments [1]

First and foremost I am indebted to my advisors, Prof. Dimitris Bertsimas and Prof. John Tsitsiklis. I have been very fortunate to work with these two wonderful scholars and individuals. They helped me to develop as a researcher and personality and showed a tremendous amount of care and support. Dimitris and John have taught me the mechanics, "nuts" and "bolts" of the research process, insisted on intuition and precision in problem solving, and encouraged high standards of research performance. For all of this I express my deepest appreciation.

I would like to thank Prof. Lawrence Wein for serving on my doctoral committee and providing me with feedback and criticism on my dissertation.

I would like to thank the whole ORC community for being such a fun group of people to spend four years of my life with. Special thanks to staff members Paulette and Laura for being helpful in every way. Thanks to my fellow students Rafael, Elaine, Thalia, Jim, Charu, Manuel, Beril, Jay, Jose, Gina, Iannis, Marina, Leonid, John and Sarah. Thanks to Georgia for genuine involvement and encouragement throughout my years as a doctoral student.

I would like to thank my advisors from my undergraduate school, Prof. Michael Megrelishvili and Prof. Josef Koshiashvili who introduced me into the subject of mathematics, mentored me at my first steps as a researcher and taught me mathematical rigor and taste.

I would like to thank all of my friends outside the ORC community. Special thanks to Irene for being so understanding and encouraging, for not feeling "preempted" by my thesis work, for her love and friendship. Thanks to Denis, Berg, Alma, Efim, Ksenia, Boris, Natali, Alex, Simon, Maria, Dmitri, Laura, Kate and Eugene who have enriched so much my life in Boston. Thanks to my friends from the KVN club with whom I have spent so many happy hours.

Most of all, I am indebted to the members of my family, who have constantly

[1] Research partially supported by the ARO under grant DAAL-03-92-G-0115 and by the NSF under grant DDM-9158118.

encouraged me in my academic pursuits. Thanks to my mother and my grandmother who showed so much love, affection and support. Thanks to my brother Alex, my sister in law Irene and to my cousins Boris and Bella for being very helping and supportive all of these years. My family has always been with me in happy or difficult moments of my life, and to them I express my deepest gratitude.

Contents

List of Figures

Chapter 1

Introduction

1.1 Multiclass Queueing Networks

Queueing networks have been used to model manufacturing, communication, and computer systems. Because of their modeling power, queueing networks have attracted a lot of research activity. The issues of stability, performance and optimization of various queueing systems have been addressed.

The behavior of queueing systems becomes increasingly complex as more generality is added to the system. For example, whereas the behavior of queueing systems with one processing station is relatively well understood, the theory of queueing networks is far from being complete, and a lot of problems are still not resolved. Also, the analysis of queueing systems with exponentially distributed interarrival and service times, and with Bernoulli type routing has been much more successful than the analysis of systems with general assumptions on the distributions of arrival, service and routing processes.

Recently queueing networks with multiple customer classes have become a focus of research. In mutliclass queueing networks customers served at the same station have different service requirements and follow different paths through the network. Multiclass queueing networks are used to model, for example, wafer fabrication facilities. In these facilities, there is a single stream of jobs arriving into a production floor. Jobs follow a deterministic route and revisit the same station multiple time (see

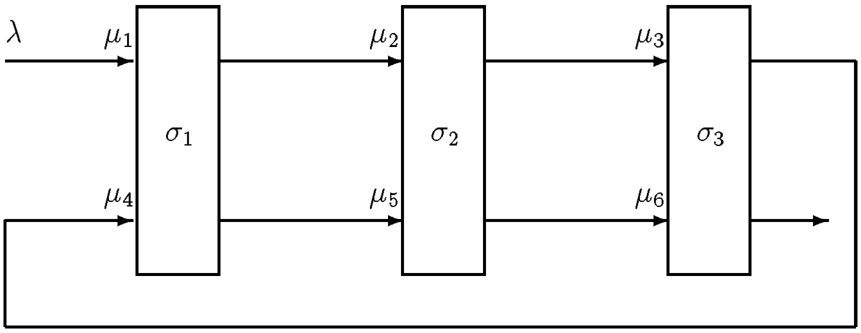

Figure 1-1: Re-entrant Line Queueing Network.

Figure 1-1). Such multiclass queueing networks are called *re-entrant line* queueing networks (see [36], [19]).

Exact analysis of multiclass queueing networks, even of the Markovian type, seems to be an intractable problem. As a result, many practitioners prefer to operate such systems in a decentralized manner - servers in a network are considered in isolation and locally optimal policies are used. It was recently discovered, however, that in some systems such decentralized policies can make the system unstable. This poses the question of stability - which queueing systems have enough processing capacity to avoid instability. Given that a queueing system is stable, understanding its steady-state behavior (performance) is important. Some interesting performance measures are the number of customers in the queueing system, waiting times and throughput times. The problems of stability and performance of multiclass queueing networks are the primary focus of this thesis.

1.2 Model Formulation and Assumptions

In this section we give the formal definition of a multiclass queueing network and provide the main assumptions for the models of interest.

We consider a network composed of J single server stations, which are denoted by σ_j, $j = 1, 2, \ldots, J$. The network includes n classes of customers, where customers of class k arrive to the network from an exogenous source with random independent and identically distributed interarrival times $a_k(n), n \geq 1$. We allow $a_k(n) = \infty$ for some n. This simply means that there are no external customers coming into class k. Class k customers are associated with and require service at a unique station σ_j, $1 \leq j \leq J$ which we denote by $\sigma(k)$. For convenience, we will also assume that σ_j is a set, the set of classes associated with the j–the station. Thus $k \in \sigma_j$ if and only if $\sigma(k) = j$. The service times of class k customers are assumed to be random i.i.d. and are denoted by $s_k(n)$. We assume that buffers at each station have infinite capacity and no customers renege from the queue before receiving service.

Routing is assumed to be of the Bernoulli type. Upon completion of service at station $\sigma(k)$, a class k customer becomes a customer of type l with probability p_{kl}, and exits the network with probability $1 - \sum_l p_{kl}$, independent of all previous history. The transition matrix P with components p_{kl} is assumed to have spectral radius less than unity. As a consequence, the matrix

$$[I - P]^{-1} = I + P + P^2 + \ldots \tag{1.1}$$

exists and is nonnegative. The spectral radius requirement plus condition (1.1) together imply that all the customers eventually leave the network. For each class k we define the arrival rate $\alpha_k \equiv 1/E[a_k(1)]$ and service rate $\mu_k \equiv 1/E[s_k(1)]$. Put $\alpha = (\alpha_1, \ldots, \alpha_n)$, $\mu = (\mu_1, \ldots, \mu_n)$. Let C be a $J \times n$ constituency matrix, defined as follows

$$C_{jk} = \begin{cases} 1, & \text{if } k \in \sigma_j, \\ 0, & \text{otherwise.} \end{cases}$$

Parameters α, μ, P, C constitute the primary parameters of a multiclass queueing network. We will therefore denote multiclass queueing networks by (α, μ, P, C) (assuming, of course, that the distributions of interarrival and service times are given as well).

For each class $k = 1, 2, \ldots,$ we introduce effective arrival rates λ_k, $k = 1, 2, \ldots, n$, as a solution to the system of equations

$$\lambda_k = \alpha_k + \sum_{l=1}^{n} p_{lk}\lambda_l, \qquad k = 1, 2, \ldots, n,$$

which are called *traffic equations.* Equivalently, $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n)$ is defined as

$$\lambda = [I - P']^{-1}\alpha. \tag{1.2}$$

Finally, for each station σ_j, $j = 1, 2, \ldots$, we define the *nominal load (traffic intensity)* per unit of time as

$$\rho_{\sigma_j} \equiv \sum_{k \in \sigma_j} \lambda_k/\mu_k. \tag{1.3}$$

In vector form $\rho = CM^{-1}\lambda$, where $M = diag(\mu_1, \mu_2, \ldots, \mu_n)$.

In re-entrant line queueing networks, we assume that customer classes are numbered consistently with the path they follow in the network (see Figure 1-1). In particular, $\alpha_2 = \ldots = \alpha_n = 0$ and $p_{k,k+1} = 1,\ k = 1, 2, \ldots, n-1,\ p_{kl} = 0$ otherwise. It is easy to check that this implies $\lambda_1 = \lambda_2 = \ldots = \lambda_n = \alpha_1$ As a result, for each station σ_j in a re-entrant line network

$$\rho_{\sigma_j} \equiv \sum_{k \in \sigma_j} \alpha_1/\mu_k. \tag{1.4}$$

The evolution of a queueing network is fully specified only when a scheduling discipline is given. The scheduling (service) 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 be interested primarily in *work conserving* scheduling disciplines - each service station is required to work on at least some customer, if there are any present at the station. Many of the well-known policies, like FIFO or priority policy are work conserving.

1.3 Problem Formulation: Stability and Performance of Mutliclass Queueing Networks

In this section we formally pose the problems of stability and performance in multiclass queueing networks.

Given a multiclass queueing network (α, μ, P, C) and some scheduling policy, for each $k = 1, 2, \ldots, n$, let $T_k(t)$ denote the cumulative amount time that server $\sigma(k)$ has spent on serving class k customers in $[0, t]$ and let $T(t) = (T_1(t), T_2(t), \ldots, T_n(t))$. We will sometimes identify this stochastic process with the scheduling policy. Let also $Q(t) = (Q_1(t), Q_2(t), \ldots, Q_n(t))$ denote a vector of queue lengths in this network at time t.

We now introduce a formal definition of stability - *positive Harris recurrence.* This definition of stability has been used by many researchers (see [17], [34], [48]).

Let $X(t)$ denote the state of the queueing network at time t. The state of the system incorporates all the information that needs to be known in order to describe the future evolution of a queueing system and it depends on the scheduling policy used. For example, under priority policies, the state of system includes the vector of queue lengths $Q(t)$ and the vector of residual interarrival and service times. Under the FIFO policy, the state must also include the order at which customers, currently present at each server, have arrived into this server.

Let $\mathcal{X}$ denote the state space for the state process, which is by definition equal to the set of possible values of the state $X(t)$. It was shown in [14] that for a wide class of service disciplines $X(t)$ is a strong Markov process. Let $\mathcal{B}(\mathcal{X})$ be a family of Borel sets on $\mathcal{X}$.

Definition 1.1 *The Markov process $X(t)$ is called positive Harris recurrent if there exists some probability measure ν on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, such that for each $x \in \mathcal{X}$, whenever $D \in \mathcal{B}(\mathcal{X})$ and $\nu(D) > 0$ then*

$$\Pr\{\tau_D < \infty | X(0) = x\} \equiv 1,$$

and

$$E[\tau_D | X(0) = x] < +\infty$$

where $\tau_D \equiv \inf\{t \geq 0 : X(t) \in D\}$. *Namely, starting from any initial state* x *the process* $X(t)$ *eventually gets into the set* D, *in finite (in expectation) time, whenever the set has a positive measure.*

A scheduling policy $T(t)$ *is defined to be stable, if the underlying Markov process* $X(t)$ *is positive Harris recurrent.*

As we see, stability is a policy specific property. We will be interested in networks in which all *work conserving* scheduling policies are stable.

Definition 1.2 *A scheduling policy is defined to be work conserving if whenever there are customers present at some station* $\sigma_j, 1 \leq j \leq J$, *this station serves at least one of them. Formally, for any* $t_2 > t_1 \geq 0$

$$\sum_{k \in \sigma_j} Q_k(t) > 0, \qquad \forall t \in (t_1, t_2)$$

implies

$$\sum_{k \in \sigma_j} \Big(T_k(t_2) - T_k(t_1) \Big) = t_2 - t_1.$$

Definition 1.3 *A multiclass queueing network* (α, μ, P, C) *is called globally stable if all work conserving policies are stable.*

Clearly, global stability of a queueing network depends on its parameters: interarrival and service time distributions, the routing probabilities and the network topology. One of the interesting conclusions obtained via the *fluid model* technique (see Chapter 2) is that global stability is solely a function of the parameters α, μ, P and C and does not depend on higher order information like variances of interarrival or service times (as long as certain mild assumptions are satisfied, for example variances are finite).

A fundamental problem arising in the theory of queueing networks is understanding which scheduling policies are stable and which networks are globally stable. The questions of global stability and stability of priority policies is the focus of this thesis.

Performance analysis is the next problem addressed in this thesis. Given that a queueing network is stable (or a specific scheduling policy is stable) we wish to find steady-state expected queue lengths $E_\pi[Q_k(0)]$ and the steady-state probability distribution of queue lengths $Q_k(0)$. Exact performance analysis seems to be an intractable problem. As a result we concentrate on obtaining bounds on these performance measures.

1.4 Literature Review

The problem of establishing conditions under which a multiclass queueing network (α, μ, P, C) is globally stable or is stable under a particular policy has attracted a lot of attention in recent years. It is known that for a single station queueing system (Dai [14]), single class (Borovkov [9], Sigman [54], Meyn and Down [47]) and multiclass acyclic queueing networks (Meyn and Down [22]) a necessary and sufficient condition for stability of all work-conserving policies (assuming non-deterministic interarrival and service times) is given by

$$\rho_{\sigma_j} < 1, \tag{1.5}$$

for all stations σ_j, $j = 1, 2, \ldots, J$, where ρ_{σ_j} is defined by (1.3).

Condition (1.5) means that the traffic intensity at each station of the network is less than one. It was also established that the Last-Buffer-First-Serve (LBFS) and the First-Buffer-First-Serve (FBFS) priority policies in re-entrant lines are stable as long as condition (1.5) holds. The FBFS priority rule gives highest priority to class 1, next highest to class 2, then class 3 and so on. The LBFS is the reverse priority rule that gives highest priority to class n, then to class $n-1$, and so on. For multiclass networks with feedback, Kumar and Seidman [40] (see also Lu and Kumar [42], and Rybko and Stolyar [50]) have identified particular priority policies that lead to instability even if condition (1.5) holds. More surprisingly, Bramson [10] and Seidman [52] have shown that these instability phenomena are present even for the standard FIFO policy. It is therefore a rather interesting problem to identify the right set of necessary and sufficient conditions for the stability of multiclass queueing networks under work-conserving policies.

In recent years, researchers have identified progressively sharper sufficient conditions for stability of all work-conserving policies through the use of Lyapunov functions. Kumar and Meyn [38] used quadratic Lyapunov functions, while Botvich and Zamyatin [11], Dai and Weiss [19], and Down and Meyn [21] used piecewise linear convex Lyapunov functions. Chen and Zhang [13] have found some closed form sufficient (but not necessary) conditions for the stability of multiclass queueing networks under FIFO. In all other cases, it was established that a multiclass network is stable if certain linear programming problems are feasible. For some specific examples (for example in Botvich and Zamyatin [11]), the conditions obtained are indeed sharp. In general, however, the problem of establishing the exact stability region, i.e., sharp necessary and sufficient conditions for stability, is open. Furthermore, it is not known whether the Lyapunov function method with piecewise linear convex functions (or with any convex function) has the power of establishing the exact stability region.

Fluid networks (to be discussed in Chapter 2) have proven to be a very effective tool for stability analysis. It was shown by Dai [14] that a scheduling policy is stable if a corresponding fluid model is stable (see Chapter 2). Dai [15] and Meyn also [46] have proven that a partial converse result holds. Motivated by these results, most of the stability analysis in this thesis is conducted within the framework of fluid queueing networks.

There is a vast literature on performance of queueing systems. The performance of single server queueing systems has been actively studied during the two last decades. It was shown (Gelenbe and Mitrani [26]) that the performance region of a multiclass queue can be described as a polyhedron. Federgruen and Groenvelt [25] advanced the theory by observing that in certain special cases of multiclass queues the polyhedron has a very special (polymatroid) structure that gives rise to very simple optimal policies (the $c\mu$ rule). Shantikumar and Yao [53] generalized the theory further by observing that if a system satisfies conservation laws, then the underlying performance space is necessarily a *polymatroid polytope* and the optimal policy is a strict priority rule. These results extend to some rather restricted queueing networks. Tsoucas [55] derived the achievable region for scheduling a multiclass nonpreemtive M/G/1 queue with Bernoulli feedback introduced by Klimov [35]. Bertsimas, Paschalidis and Tsitsiklis [4] extended these results by considering the average cost branching bandit problem and fully characterizing the achievable region. Bertsimas and Nino-Mora [2] generalized the idea of conservation laws and showed that for all systems that satisfy these generalized conservation laws, the underlying performance space is a polyhedron with very strong structural properties, called extended polymatroid in Bhattacharaya, Georgiadis and Tsoucas [8]. The framework of Bertsimas and Nino-Mora includes all the cases above as well as the multiarmed bandit problem (Gittins [27]), branching bandits (Weiss [58]) and some deterministic scheduling problems. A multiclass single server queueing system in *heavy-traffic* (meaning that ρ defined by (1.3) is closed to unity) was analyzed by Bertsimas and Mortzinou [1] by means of distributional laws. The resulting closed form approximations of expected queue lengths coincide with ones obtained via diffusion approximations by Dai and Kurtz [16].

The performance of queueing networks is a much less understood area. Some of the earlier and classical results include product form probability distributions for Jackson and BCMP type networks (see Gelenbe and Mitrani [26]). It was realized however that the multiclass structure in queueing networks does not allow, in general, for a product form distribution even if interarrival and service times have exponential distributions and FIFO is used. There are several types of approaches in the area of performance of multiclass queueing networks. We outline here some of them.

- *Simulation.* Simulations remains a widely used and practical approach for performance analysis of multiclass queueing networks.

- *Diffusion approximations.* One of the most successful approaches to performance analysis as well as optimization of queueing networks in *heavy-traffic* involves Brownian Networks models. The approach works especially well for closed queueing networks (Harrison and Wein [30]) and networks with controllable input (Wein [57]) and helps to build heuristic suboptimal policies that typically outperform more traditional ones. However, this approach is not that powerful for open queueing networks. It is not known which multiclass queueing networks (open or closed) exhibit diffusion type of behavior in heavy traffic (see Harrison and Nguen [29], Dai and Wang [18]). Some networks exhibit an unconventional type of heavy-traffic behavior (see Williams [31]), but in general, building approximation of multiclass queueing networks in heavy-traffic is an open problem. The other difficulty in heavy-traffic theory is in deriving the steady-state distribution of the corresponding Brownian motion and relating it to the steady-state distribution of the underlying queueing network. Even for the case of a single-class queueing network with generally distributed interarrival and service times (Generalized Jackson Network), the steady-state distribution of the corresponding reflected Brownian motion is not known, and furthermore, apart from special cases, it is not known whether it approximates correctly the steady-state distribution of the underlying queueing network.

- *Quadratic Lyapunov function and Conservation Laws.* Recently, several researches (Bertsimas, Paschalidis and Tsitsiklis [4], and Kumar and Kumar [37], Jin, Ou and Kumar [33]) have analyzed certain types of multiclass queueing network via the quadratic Lyapunov function approach. A certain quadratic functional of a vector of queue lengths is formed which leads to set of valid linear equalities and inequalities for the underlying stochastic process in steady-state. The resulting linear program provides polyhedral type of bounds on the achievable performance region. A dual of this linear program is shown to be related to stability analysis (Kumar and Meyn [38]). This approach has however limitations: a quadratic Lyapunov function approach to stability is proven not to be tight in some networks (Down and Meyn [21]). As a result, the corresponding approximating polyhedron is unbounded, even though the underlying queueing network is globally stable. The quadratic Lyapunov function approach was also used to build some polyhedral type performance bounds in closed exponential queueing networks (Jin, Ou and Kumar [32]). The performance results obtained via quadratic Lyapunov function were later analyzed and extended by means of the conservation laws approach (Bertsimas and Nino-Mora [3]). This approach is much simpler and more intuitive.

- *Large deviation approach for communication networks.* Large deviation theory has been applied to the analysis of queueing networks, arising in communications (see Paschalidis [49], Schwartz and Weiss [51]). By means of this approach, upper bounds on tail probabilities of queue lengths are constructed. The advantage of this approach is that it allows for very general assumptions on distributions of arrival processes (usually the arrival process is just required to be stationary mixing and not necessarily i.i.d.) However, this approach has been successful only for special (acyclic) types of queueing networks.

1.5 Thesis Contribution

1.5.1 Global Stability

We introduce in Chapter 2 a new method to investigate the global stability of multiclass fluid networks. The method looks at the detailed structure of possible trajectories. Given a multiclass fluid queueing network, a certain linear program is constructed with the following properties:

- If the objective value of this linear program is finite, then the queueing (stochastic and fluid) network is globally stable.

- For multiclass fluid networks with two stations, if the objective value of this linear program is infinite, then a certain unstable fluid scheduling policy can be constructed.

As a result, the boundedness of this linear program is a necessary and sufficient condition for global stability of fluid networks with two stations.

By means of this linear programming approach, we prove that multiclass acyclic queueing networks are globally stable.

For networks with two stations we also demonstrate a duality relationship between our linear program and the Lyapunov function linear program proposed in [21]. As a corollary of strong duality, our linear program has a bounded objective value if and only if a linear program yielding to a piecewise linear Lyapunov function, has a feasible solution. We, therefore, establish that piecewise linear, convex Lyapunov functions have the power of checking stability exactly for networks with two stations.

1.5.2 Stability of Priority Policies

Stability of priority policies in multiclass fluid networks is discussed in Chapter 3. The trajectory decomposition method from Chapter 2 is tailored to priority policies. The method again leads to a construction of a certain linear program. If this linear program has a finite objective value, then a priority policy is stable. By virtue of the result of Dai [14], this implies that this priority policy is stable for the stochastic network as well. Using linear programming, we then prove stability of Last-Buffer-First-Serve and First-Buffer-First-Serve priority policies. We extend the result on stability of FBFS policy to networks with Bernoulli routing and upper triangular routing matrix. We apply this result to a specific (Rybko-Stolyar) network, to show that all but one priority policies are stable in this network.

1.5.3 Performance of Multiclass Queueing Networks

Chapters 4, 5 and 6 are devoted to the performance analysis of multiclass queueing networks in steady-state. The main objective of the performance analysis is calculating various performance measures like queue lengths, waiting times or cycle times. The performance analysis of multiclass queueing networks seems to be a highly non-trivial task since even the conditions for existence of steady-state (stability) are not known in general. However, with emerging methods of checking stability for specific classes of networks, it is natural to expect that the same methods may become useful for the performance analysis as well. The recent advances in stability analysis have demonstrated the power of Lyapunov functions as a method of establishing (global) stability of queueing networks. Our main goal is to develop a general methodology of using Lyapunov functions for the performance analysis.

We start in Chapter 4 with an analysis of infinite state Markov chains with state space in R^n_+. We consider a class of Markov chains, for which there exists a Lyapunov functions, proving the ergodicity (stability) of the underlying stochastic process. We prove that the existence of a Lyapunov function simultaneously implies that certain computable upper bounds hold on the steady-state probability distribution as well as expectations of the process (Theorem 4.1). We also introduce the notion of a lower Lyapunov function as a way of getting lower bounds on the steady-state distribution and expectations of the Markov chain. We apply these results to homogeneous random walks in the nonnegative orthant Z^n_+. In particular, for a large class of such random walks we construct certain constants $0 < c_1 < c_2 < 1$ for which

$$c_1^m \le \Pr_\pi\{\sum_{i=1}^n Q_i(t) \ge m\} \le c_2^m,$$

where $Q(t) = (Q_1(t), Q_2(t), \ldots, Q_n(t))$ is the vector description of the random walk in Z^n_+ and π is the steady-state probability distribution (see Theorems 4.2-4.3). As a result, for such random walks, tail probabilities decay exponentially fast in m.

We then apply this methodology in Chapters 5 and 6 for the performance analysis of multiclass exponential type queueing networks - networks with exponentially distributed interarrival and service times.

In Chapter 5, we concentrate on getting performance bounds which hold under all work conserving policies. Our results are summarized as follows:

- If a multiclass queueing network admits a piecewise linear Lyapunov function, which establishes the global stability of the network, then certain upper bounds hold on the tail distribution and the expectation of queue lengths in steady-state under any work conserving policy. Specifically, if a certain linear program (GLP[md], Chapter 5) has a feasible solution, then certain constants $0 < c <$

$1, C > 0$ can be computed, for which

$$\Pr_\pi\{\text{total number of customers} \ge m\} \le c^m,$$

and

$$E_\pi[\text{total number of customers}] \le C,$$

where π is a steady-state probability distribution corresponding to any work conserving policy (Theorem 5.4). Thus, in such networks, the tail probabilities of queue lengths decay exponentially fast in m.

- Dai and Vande-Vate introduced in [20] an important notion of *virtual station* and *virtual traffic intensity.* They have proven that re-entrant line type fluid queueing networks with two stations are globally stable if and only if the maximum of all actual and virtual traffic intensities ρ^* is less than one. We show that the virtual traffic intensity is a fundamental performance parameter. For re-entrant line type networks with two stations we construct explicit upper bounds on steady-state queue lengths, in terms of actual and virtual traffic intensities (Theorem 5.7). Specifically, we show that whenever the queueing network is globally stable (equivalently $\rho^* < 1$), the following upper bound holds under any work conserving policy

$$E_\pi[\text{total number of customers}] = O\Big(\frac{1}{(1-\rho^*)^2}\Big).$$

An important implication of this result is that the performance region (the set of vectors of expected queue lengths) is bounded if and only if fluid the network is globally stable, or specifically, if and only if the maximal (actual or virtual) traffic intensity ρ^* is less than one.

We do not know whether this result holds true for networks with more than two stations. If piecewise linear functions were proven to be an exact test of global stability, then the result would follow automatically from Theorem 5.4.

- Using a simple closed form Lower Lyapunov function we construct lower bounds on the steady-state number of customers in the network, which hold under any stable policy. In particular, the following simple bounds are proven for each processing station σ_j

$$\Pr_\pi\{\text{total number of customers} \geq \frac{1}{2}m\} \geq \Big(\frac{\rho_{\sigma_j}}{2-\rho_{\sigma_j}}\Big)^m$$

  and

$$E_\pi[\text{total number of customers}] \geq \frac{\rho_{\sigma_j}}{4(1-\rho_{\sigma_j})},$$

  where ρ_{σ_j} is the traffic intensity of station σ_j.

- We construct closed form upper bounds on the steady-state queue lengths in feedforward (acyclic) multiclass queueing networks. The bounds are finite whenever the traffic intensities of all the stations are less than one. As we have mentioned above, this is sufficient for global stability of feedforward networks. Specifically, we show that for all such networks the following bound holds under any work conserving policy.

$$E_\pi[\text{total number of customers}] = O\Big(\frac{1}{(1-\rho)^2}\Big),$$

  where $\rho = \max\{\rho_{\sigma_j}\}$ is the largest (bottleneck) traffic intensity among all the processing stations σ_j. This is in general an improvement upon the following bound obtained using quadratic Lyapunov function in [33]

$$E_\pi[\text{total number of customers}] = O\Big(\frac{1}{(1-\rho)^B}\Big),$$

  where B is the number of stations σ_j which achieve the bottleneck traffic intensity $\rho_{\sigma_j} = \rho$.

In Chapter 6 we further extend these results to multiclass exponential type queueing networks operating under priority policies. The results of this chapter are as follows.

- We utilize the notion of a virtual station to derive lower bounds on the steady-state queue lengths when a particular priority policy is implemented. As shown in [20], some priority policies naturally lead to certain groups of classes, which cannot be served simultaneously. These group of classes are called *K-virtual stations*. We show that for each K-virtual station V, the corresponding *virtual traffic intensity* $\rho(V) \equiv \sum_{i\in V} \rho_i$ defines a bottleneck behavior of the queueing network under this priority policy. Specifically (Theorem 6.2) we prove that under any priority policy θ and for the corresponding K-virtual station V there holds

$$\Pr_{\pi}\{\text{total number of customers} \geq \frac{1}{2}m\} \geq \Big(\frac{\rho(V)}{2(K-1)-\rho(V)}\Big)^m$$

$$E_{\pi}[\text{total number of customers}] \geq \frac{\rho(V)}{4(K-1-\rho(V))},$$

where π is a stationary probability distribution corresponding to the priority policy θ.

- We build numerical upper bounds on the steady-state number of customers in the network whenever some piecewise linear function establishes stability of a particular priority policy. Specifically, we show that certain constants $0 < c < 1, C > 0$ can be constructed, for which

$$\Pr_{\pi}\{\text{total number of customers} \geq m\} \leq c^m, \tag{1.6}$$

and

$$E_{\pi}[\text{total number of customers}] \leq C,$$

where π is a steady-state probability distribution corresponding to the underlying priority policy (Theorem 6.3).

The results above are the first ones that use linear and piecewise linear functions for the performance analysis. Previous methods for stability analysis involve quadratic Lyapunov functions (see [4],[36],[37], [39], [32], [33]) or conservations laws of queueing networks (see [3]). It turns out that the two methods lead to the same set of polyhedral type bounds, although the second approach is simpler and more intuitive. The limitation of these approaches is two-fold. First, the methods are proven not to be exact tests of global stability. An example of a globally stable (stochastic and fluid) queueing network with two stations was constructed in [21] for which the methods above give an infinite (inconclusive) upper bound. The methods developed here, on the other hand, guarantee to construct finite upper bounds for networks with two stations, whenever corresponding fluid network is globally stable.

The second limitation of previous methods is that the bounds constructed are in most cases only numerical and are provided only for *expectations* of queue lengths. We propose here the first bounds on the *distribution* of steady-state queue lengths, proving exponential decay of the tail probabilities.

Chapter 2

Stability of Mutliclass Queueing Networks

Fluid networks have been studied by numerous authors and have proven to be a very effective tool for stability analysis. In this chapter, we derive conditions for the global stability of multiclass fluid networks by means of a trajectory decomposition approach. We show that this approach provides exact stability conditions for networks with two stations. We also relate our approach to stability analysis based on Lyapunov functions.

In Section 2.1, we define fluid models and outline the Lyapunov function approach for the stability analysis which has been widely used in the literature. Section 2.2 demonstrates the power of fluid models and Lyapunov functions for queueing systems with one station.

Or main contributions are included in Sections 2.3 and 2.4. In Section 2.3 we describe the trajectory decomposition approach to stability analysis. By means of this approach we derive linear programming conditions for stability of multiclass fluid networks. We prove that the approach is exact for networks with two stations. Also for networks with two stations, we show that our approach involves the linear programming dual to the stability conditions derived via Lyapunov function approach. As a corollary, we prove that the Lyapunov function stability conditions are exact for fluid networks with two stations. We conclude the chapter with some examples in

Section 2.5.

2.1 Background: Fluid Models

2.1.1 Notation

We introduce a fluid model (α, μ, P, C) consisting of n classes and J service stations $\sigma_1, \ldots, \sigma_J$ as follows. Each class is served at a particular station. Let σ_j be the set of classes that are served in station j. The external arrival rate for class i is α_i and the service rate is μ_i. Let $\alpha = (\alpha_1, \ldots, \alpha_n)'$ and $\mu = (\mu_1, \ldots, \mu_n)'$. After a service completion, a fraction p_{ij} of class i customers becomes class j and a fraction $1 - \sum_j p_{ij}$ exits the system. Let P be the substochastic matrix $P = (P_{ij})_{1 \leq i,j \leq n}$. Finally, we define the $J \times n$ matrix C as follows: $c_{jk} = 1$ if class k is served at station j and $c_{jk} = 0$ otherwise. We let $M = \text{diag}\{\mu_1, \ldots, \mu_n\}$ and assume that the matrix P has spectral radius less that one.

Any scheduling policy can be described in terms of the variables $T_k(t)$ defined as the amount of time class k is being served in the interval $[0, t]$, and $Q_k(t)$ defined as the queue length for class k at time t. We let $T(t) = (T_1(t), \ldots, T_n(t))'$ and $Q(t) = (Q_1(t), \ldots, Q_n(t))'$.

Throughout this chapter we call $Q(t)$ the trajectory of the fluid process under the allocation process $T(t)$. Given the initial condition $Q(0)$, the dynamics of the queue length process are as follows:

$$Q_k(t) = Q_k(0) + \alpha_k t + \sum_{i=1}^{n} \mu_i T_i(t) p_{ik} - \mu_k T_k(t) \geq 0, \quad k = 1, \ldots, n, \tag{2.1}$$

or in matrix form:

$$Q(t) = Q(0) + \alpha t + [P' - I]MT(t) \geq 0. \tag{2.2}$$

In addition to the inequality (2.2), we assume that the allocation process satisfies the following conditions:

1. $T(0) = 0$.

2. (Feasibility) For any $t_2 > t_1 \geq 0$ and any station i:

$$\sum_{k \in \sigma_i} [T_k(t_2) - T_k(t_1)] \leq t_2 - t_1, \tag{2.3}$$

   and $T_k(t)$ is nondecreasing.

3. (Work-conservation) If for all $t \in [t_1, t_2]$ we have $\sum_{k \in \sigma_i} Q_k(t) > 0$ for some station i, then

$$\sum_{k \in \sigma_i} [T_k(t_2) - T_k(t_1)] = t_2 - t_1. \tag{2.4}$$

Any scheduling policy satisfying all the above properties is called a (feasible) work-conserving policy.

An alternative characterization of the above requirements is to introduce for any station i, the cumulative idling process:

$$U_i(t) = t - \sum_{k \in \sigma_i} T_k(t).$$

The feasibility condition (2.3) then requires that $U_i(t)$ be nonnegative and nondecreasing, while the work-conservation condition is rewritten as follows: if for all $t \in [t_1, t_2]$ we have $\sum_{k \in \sigma_i} Q_k(t) > 0$, then

$$U_i(t_1) = U_i(t_2). \tag{2.5}$$

The following definitions are adopted from Chen [12]:

Definition 2.1 *A work-conserving policy $T(t)$ is defined to be stable if there exists a time t_0 such that $Q(t) = 0$ for all $t \geq t_0$.*

A fluid network is defined to be (globally) stable if all work-conserving policies are stable. More formally

Definition 2.2 *A fluid network (α, μ, P, C) is said to be (globally) stable for all work conserving policies if for every work-conserving allocation process $T(t)$ and every initial condition $Q(0)$, there exists a finite time t_0 such that $Q(t) = 0$ for all $t \geq t_0$.*

We focus exclusively on global stability in this chapter, therefore we will suppress the notion "global".

Rybko and Stolyar [50] show that this definition of stability is equivalent to the weaker condition: for every work-conserving allocation process $T(t)$ and every initial condition $Q(0)$, there exists a finite time t_0 such that $Q(t_0) = 0$. We will use this as our working definition of stability.

A necessary condition for stability (see Chen [12]) is that the traffic intensity vector ρ defined by $\rho = CM^{-1}[I - P']^{-1}\alpha$, satisfies

$$\rho < e, \tag{2.6}$$

where $e = (1, \ldots, 1)'$. As mentioned in the introduction, for general multiclass networks with feedback, this condition is not sufficient. The goal of the Section 2.3 is to establish necessary and sufficient conditions for the stability of a multiclass fluid network with two stations, given that $\rho < e$.

2.1.2 Fluid and Stochastic Stability. Equivalence Theorems

The fluid models and their stability have been studied by several authors. The most fundamental theorem stating that stability of fluid models implies stability of the underlying stochastic queueing network was proved by Dai [14]. Theorem 2.1 below is stronger version of this result.

The following theorem was proved by Dai and Meyn ([17], theorem 4.1)

Theorem 2.1 *Let (α, μ, P, C) be a multiclass queueing network. Suppose a fluid model for a service discipline $T(t)$ is stable (see Definition 2.1). Then under very mild assumptions on interarrival and service times, there exists an invariant probability distribution π for the Markov process $X(t)$, such that*

1. *The Markov process $X(t)$ corresponding to the service discipline $T(t)$ is positive Harris recurrent.*

2. *The transient average queue lengths converge to their steady-state values:*

$$\lim_{t\to\infty} E[Q_k(t)] = E_\pi[Q_k(0)]$$

3. *The strong law of large numbers holds: for all $k = 1, 2, \ldots$*

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t Q_k(\tau)d\tau = E_\pi[Q_k(0)] \qquad \text{almost surely}$$

Partial converse results were proven by Dai [15] and Meyn [46].

However, whether stochastic and fluid stability are equivalent to each other in general, remains an open problem.

2.1.3 Lyapunov Function Approach

The Lyapunov function method has been used to attack the problem of stability of fluid networks.

Definition 2.3 *A nonnegative function*

$$\Phi : R_+^n \to R_+$$

with the property that for some $\gamma_0 > 0$, under any work-conserving policy $T(t)$,

$$\dot{\Phi}(Q(t)) \le -\gamma_0,$$

for all t such that $Q(t) \neq 0$, is called a Lyapunov (potential) function

The following theorem was proven by Down and Meyn ([21], Theorem 3.4)

Theorem 2.2 *Suppose that $\Phi : R_+^n \to R_+$ is a C^1 function, and that for some $\gamma_0 > 0$, the fluid model satisfies*

$$\dot{\Phi}(Q(t)) = \bigtriangledown\Phi(Q(t))\dot{Q}(t) < -\gamma_0$$

for all work conserving policies $T(t)$ and for all times such that $Q(t) \neq 0$. Then the fluid model is stable.

Motivated by this, several sufficient conditions for stability were found by means of quadratic and piece-wise linear Lyapunov functions (Kumar and Meyn [38], Botvich and Zamyatin [11], Dai and Weiss [19]). In particular, it was shown by Down and Meyn ([21], Theorem 4.2) that if a certain linear program is feasible then a (fluid and stochastic) network is positive Harris recurrent for all work conserving policies. It was also shown in [21] that quadratic potential functions do not capture stability exactly. The corresponding question for piece-wise linear functions remained open. In Section 2.3, we show (Corollary 2.1) that in two-station networks piece-wise linear potential functions lead to a complete method for checking stability.

2.2 Single Station Queueing System: Linear Lyapunov Functions and Duality

Theorem 2.1 motivates searching for stability conditions for multiclass fluid networks. In Sections 2.3 and 2.4 we propose a new approach for stability analysis based on *trajectory decomposition*. Section 2.3.4 relates our method with the method of piece-wise linear potential function, by means of a linear programming duality. To motivate the discussion, we show how trajectory decomposition method and duality manifest themselves in a simple case of a single station fluid queueing system.

Suppose we are given a one-station fluid system (α, μ, P) with n classes. Note that in this system, the $1 \times n$ matrix C is trivially equal to $e' = (1, 1, \ldots, 1)'$. The utilization factor ρ of the server is calculated as

$$\rho = e'[I - P']^{-1}M^{-1}\alpha$$

Consider a work-conserving policy $T(t)$ starting with a non-zero vector of fluid levels $Q(0)' = (Q_1(0), Q_2(0), \ldots, Q_n(0))'$. Let τ^* be the emptying time (possibly $+\infty$) of this fluid system:

$$\tau^* \equiv \inf\{t : Q(t) = 0\}$$

Proposition 2.1 *The following holds:*

1. *If $\rho \geq 1$, then $\tau^* = +\infty$ and the system is unstable.*

2. *If $\rho < 1$, then*
$$\tau^* = \frac{e'M^{-1}[I-P']^{-1}Q(0)}{1-\rho} \tag{2.7}$$
and does not depend on the particular work-conserving policy $T(t)$ used,

3. *The emptying time τ^* is equal to the optimal value of the following linear program LP[Q(0)]:*
$$\text{maximize } \tau$$
subject to
$$\tau = \sum_{k=1}^{n} \tau_n$$
$$\alpha_k \tau + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i - \mu_k \tau_k \geq -Q_k(0), \qquad \tau_k \geq 0, k = 1, 2, \ldots, n$$

Proof:

Let us multiply the dynamic equation (2.2) by the nonnegative vector $e'M^{-1}[I - P']^{-1}$ from both sides. We obtain

$$e'M^{-1}[I-P']^{-1}Q(t) = e'M^{-1}[I-P']^{-1}Q(0) + \rho t - \sum_{i=1}^{n} T_k(t) \geq 0 \tag{2.8}$$

For all $t \geq 0$ by the feasibility constraint (2.3), $t \geq \sum_{i=1}^{n} T_k(t)$. If $\rho \geq 1$, then

$$e'M^{-1}[I-P']^{-1}Q(t) \geq e'M^{-1}[I-P']^{-1}Q(0) \neq 0$$

Namely, for all $t \geq 0$, $Q(t) \neq 0$ and $\tau^* = +\infty$.

Observe that if $\rho < 1$, then from (2.3), for all $t < \frac{e'M^{-1}[I-P']^{-1}Q(0)}{1-\rho}$, we have

$$e'M^{-1}[I-P']^{-1}Q(t) = e'M^{-1}[I-P']^{-1}Q(0) + \rho t - \sum_{i=1}^{n} T_i(t) > 0$$

and $Q(t) \neq 0$, implying $\sum_{i=1}^{n} T_i(t) = t$. Also, for $t = \frac{e'M^{-1}[I-P']^{-1}Q(0)}{1-\rho}$

$$e'M^{-1}[I-P']^{-1}Q(t) = 0$$

so $Q(t) = 0$. We conclude $\tau^* = \frac{e'M^{-1}[I-P']^{-1}Q(0)}{1-\rho}$.

For any feasible work-conserving policy $T(t)$ and $t \leq \tau^*$ set $\tau_k = T_k(t)$ and $\tau = t$. From dynamic equations (2.1) and work-conservation (2.4), τ, τ_k, $k = 1, 2, \ldots, n$ is a feasible solution to the linear program. Therefore the objective value of our linear program is not smaller than τ^*. On the other hand, for any feasible solution τ, τ_k, $k = 1, 2, \ldots, n$ to our linear program, consider the policy

$$T_k(t) = \frac{\tau_k}{\tau} t, \qquad k = 1, 2, \ldots n$$

For this policy, for all $k = 1, 2, \ldots, n$ and $t < \tau$, we have

$$Q_k(t) = Q_k(0) + \frac{t}{\tau}\Big(\alpha_k \tau + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i - \mu_k \tau_k\Big) > 0$$

Therefore, this policy is feasible and satisfies $Q(t) \neq 0$ for all $t < \tau \leq \tau^*$. $\square$.

In the proof of the last part of the theorem, we have decomposed the time axis $[0, \infty)$ into parts $[0, \tau^*)$ and $[\tau^*, \infty)$, and further decomposed $[0, \tau^*)$ into time intervals (variables) τ_k which stand for the amount of time the server spent serving class k fluids. In the next section we elaborate this idea further, and apply it to multiclass fluid networks with several stations.

From Definition 2.3, a linear function $\Phi(Q(t)) = L'Q(t) = \sum_{k=1}^{n} L_k Q_k(t)$ is a potential function if for any work-conserving policy $T(t)$, $L'\dot{Q}(t) \leq -1$. This condition is met if the following system of inequalities in the variables L_k, $k = 1, 2, \ldots, n$, is feasible

$$\sum_{i=1}^{n} L_i \alpha_i + \sum_{i=1}^{n} L_i \mu_i p_{ki} \mu_k - L_k \mu_k + 1 \leq 0,$$
$$L_k \geq 0 \qquad k = 1, 2, \ldots n. \tag{2.9}$$

The following proposition relates the polyhedron defined by (2.9) with the polyhedron of the linear program LP[Q(0)].

Proposition 2.2 *The dual of the linear program LP[Q(0)] is a linear program DLP[Q(0)] with constraints given by (2.9) and objective function $Q(0)'L$. DLP[Q(0)] is feasible (consequently a linear potential function exists) if and only if $\rho < 1$.*

The duality is checked trivially. By the strong duality of linear programming, the linear program LP[Q(0)] has a finite optimum if and only if DLP[Q(0)] is feasible. From Proposition 2.1 LP[Q(0)] has a finite optimum if and only if $\rho < 1$.□

We see that linear potential functions have the power of checking stability in fluid systems with one station. Notice that the variables and the objective function of $LP[Q(0)]$ have a clear physical meaning - they correspond to the amount of time the server spent serving different fluid types and the emptying time of the system, respectively. The variables and objective function of DLP[Q(0)] on the other hand, do not have an obvious physical interpretation. In Section 2.3.4 we show that the same duality relationship holds between the trajectory decomposition method and the piecewise linear Lyapunov function method in fluid networks with two stations

2.3 Exact Stability Conditions for Multiclass Two-Station Fluid Networks

In this section we establish necessary and sufficient conditions for stability, for the case where $J = 2$, i.e., for multiclass networks with two stations. Throughout this section, we assume that $\rho < e$, since otherwise the system is unstable.

2.3.1 Trajectory Decomposition

We partition the set $R_+^n - \{0\}$ of nonzero states into the following finite family of subspaces. For any non-empty set of service stations $S \subset \{1,2\}$, we let

$$R_S = \{x \in R_+^n : \forall i \in S,\ \sum_{k\in\sigma_i} x_k > 0,\ \text{and } \forall i \notin S,\ \sum_{k\in\sigma_i} x_k = 0\},$$

i.e., R_S corresponds to states for which all stations in S are busy, while all other stations have empty buffers. We denote by R_1, R_2 and R_{12} the subspaces corresponding to $S = \{1\},\{2\},\{1,2\}$, respectively. In particular, for $Q \in R_1$ station 2 has no customers, for $Q \in R_2$ station 1 has no customers, while for $Q \in R_{12}$ both stations have customers in queue. The proposition that follows states that a trajectory can be broken down into subtrajectories of four different types.

Proposition 2.3 *Consider a stable work-conserving trajectory $Q(t)$ and let τ be the smallest time such that $Q(\tau) = 0$. There exists a (finite or infinite) nondecreasing sequence t_i such that $\sup_i t_i = \tau$ and such that for all times less than τ the following hold:*

$Q(t_{4m+1}) \in R_1$ *and for* $t \in [t_{4m+1}, t_{4m+2}]$, $Q(t) \in R_1 \cup R_{12}$;

$Q(t_{4m+2}) \in R_1$ *and for* $t \in (t_{4m+2}, t_{4m+3})$, $Q(t) \in R_{12}$;

$Q(t_{4m+3}) \in R_2$ *and for* $t \in [t_{4m+3}, t_{4m+4}]$, $Q(t) \in R_2 \cup R_{12}$;

$Q(t_{4m+4}) \in R_2$ *and for* $t \in (t_{4m+4}, t_{4m+5})$, $Q(t) \in R_{12}$.

Proof:

This is a simple consequence of the fact that starting in R_1, the system can get to R_2 only by first going through R_{12}, and vice versa; see Figure 2-1.

In particular, once t_{4m+1} has been defined, we may let $t_{4m+3} = \min\{t > t_{4m+1} \mid Q(t) \in R_2\}$ and $t_{4m+2} = \max\{t < t_{4m+3} \mid Q(t) \in R_1\}$. [In case $Q(t)$ never enters R_2 after time t_{4m+1}, then the preceding definition of t_{4m+3} is inapplicable; however, in this case, the system gets to $Q(\tau) = 0$ without ever leaving $R_1 \cup R_{12}$. Thus, $[t_{4m+1}, \tau)$ can be taken as the last interval.] Having thus defined t_{4m+3}, the times t_{4m+4} and t_{4m+5} are defined similarly. □

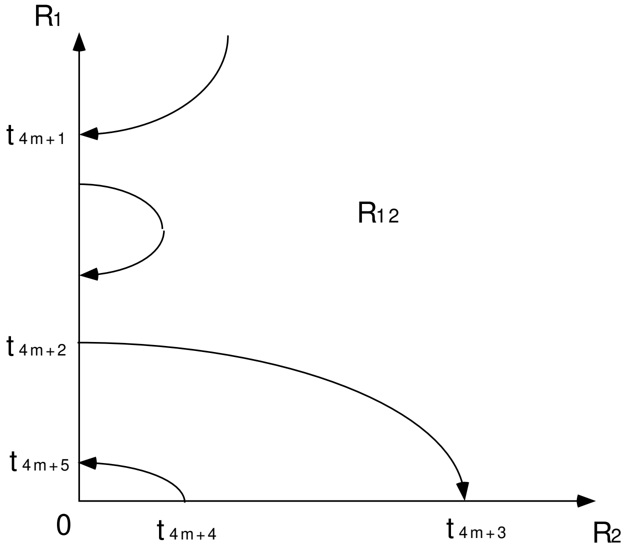

Figure 2-1: The times t_i for a typical trajectory.

We now find an upper bound on the time that stable work-conserving policies take to empty the fluid network starting with an initial condition $Q(0)$. This time is usually called the strong busy period. This result is of independent interest, as it contributes to our understanding of the performance of the network; it is also the key to our stability analysis in the next subsection.

Proposition 2.4 *Consider a stable work-conserving policy $T(t)$ starting with initial condition $Q(0) \neq 0$. Let τ be the smallest time such that $Q(\tau) = 0$. Then, τ is bounded above by the optimal value of the following linear program to be called $LP[Q(0)]$:*

$$\text{maximize } \sum_{j=1}^{4} \tau_j$$

subject to

$$\tau_1 = \sum_{k \in \sigma_1} \tau_k^1, \quad \tau_1 \geq \sum_{k \in \sigma_2} \tau_k^1,$$

$$\tau_2 = \sum_{k \in \sigma_1} \tau_k^2, \quad \tau_2 = \sum_{k \in \sigma_2} \tau_k^2,$$

$$\tau_3 \geq \sum_{k \in \sigma_1} \tau_k^3, \quad \tau_3 = \sum_{k \in \sigma_3} \tau_k^3,$$

$$\tau_4 = \sum_{k \in \sigma_1} \tau_k^4, \quad \tau_4 = \sum_{k \in \sigma_2} \tau_k^4,$$

$\forall\, k \in \sigma_2$:

$$\alpha_k \tau_1 + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i^1 - \mu_k \tau_k^1 = 0,$$

$$\alpha_k \tau_2 + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i^2 - \mu_k \tau_k^2 \geq 0,$$

$$\alpha_k \tau_4 + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i^4 - \mu_k \tau_k^4 \leq 0,$$

$\forall\, k \in \sigma_1$:

$$\alpha_k \tau_3 + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i^3 - \mu_k \tau_k^3 = 0,$$

$$\alpha_k \tau_4 + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i^4 - \mu_k \tau_k^4 \geq 0,$$

$$\alpha_k \tau_2 + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i^2 - \mu_k \tau_k^2 \leq 0,$$

$\forall k \in \{1, \ldots, n\}$:

$$\alpha_k \sum_{j=1}^{4} \tau_j + \sum_{i=1}^{n} \mu_i p_{ik} \sum_{j=1}^{4} \tau_i^j - \mu_k \sum_{j=1}^{4} \tau_k^j = -Q_k(0), \tag{2.10}$$

$$\tau_j \geq 0, \quad \tau_k^j \geq 0.$$

Proof:

Consider a stable work conserving policy with initial condition $Q(0) \neq 0$. Without loss of generality, we only provide the proof for the case $Q(0) \in R_1$; the proof for the other cases is essentially identical. Let $t_1 = 0$ and let the times t_j be as in the statement of Proposition 2.3. For $j = 1, \ldots, 4$ we introduce the following variables:

$$\tau_j = \sum_{m=0}^{\infty} (t_{4m+j+1} - t_{4m+j}) \tag{2.11}$$

and

$$\tau_k^j = \sum_{m=0}^{\infty} (T_k(t_{4m+j+1}) - T_k(t_{4m+j})). \tag{2.12}$$

Intuitively, τ_1 is the total amount of time the trajectory spends in R_1 as well as in excursions from R_1 into R_{12} and back into R_1; τ_2 is the total amount of time the trajectory spends in R_{12} coming from R_1 and going to R_2; τ_3 is the total amount of time the trajectory spends in R_2 as well as in excursions from R_2 into R_{12} and back into R_2; finally, τ_4 is the total amount of time the trajectory spends in R_{12} coming from R_2 and going to R_1. Clearly $\tau_j \geq 0$ and the first time that $Q(t)$ becomes zero is given by $\tau = \tau_1 + \tau_2 + \tau_3 + \tau_4$. Note that for every class k, τ_k^1, τ_k^2, τ_k^2 and τ_k^4 is the total work allocated to class k, during the time intervals that enter in the definitions of $\tau_1, \tau_2, \tau_3, \tau_4$, respectively.

For all $t \in [t_{4m+1}, t_{4m+2}]$, we have $Q(t) \in R_1 \cup R_{12}$, and therefore $\sum_{k \in \sigma_1} Q_k(t) > 0$. Because the policy is work-conserving,

$$t_{4m+2} - t_{4m+1} = \sum_{k \in \sigma_1} (T_k(t_{4m+2}) - T_k(t_{4m+1})). \tag{2.13}$$

By summing over $m \geq 0$ we obtain that

$$\tau_1 = \sum_{k \in \sigma_1} \tau_k^1,$$

which simply expresses the work conservation in station 1, while the trajectory is in $R_1 \cup R_{12}$ (station 1 busy). Similarly, work conservation for station 2, while the trajectory is in $R_2 \cup R_{12}$ (station 2 busy) leads to

$$\tau_3 = \sum_{k \in \sigma_2} \tau_k^2.$$

Moreover, for $t \in (t_{4m+2}, t_{4m+3}) \cup (t_{4m+4}, t_{4m+5})$, we have $Q(t) \in R_{12}$, and work conservation for both stations leads to

$$\tau_2 = \sum_{k \in \sigma_1} \tau_k^2 = \sum_{k \in \sigma_2} \tau_k^2, \qquad \tau_4 = \sum_{k \in \sigma_1} \tau_k^4 = \sum_{k \in \sigma_2} \tau_k^4.$$

For every station j, we have

$$\sum_{k \in \sigma_j} (T_k(t_{i+1}) - T_k(t_i)) \leq t_{i+1} - t_i,$$

leading to

$$\tau_1 \geq \sum_{k\in\sigma_2} \tau_k^1, \quad \tau_3 \geq \sum_{k\in\sigma_1} \tau_k^2.$$

By definition of the times t_i, we have $Q(t_{4m+1}) \in R_1$ and $Q(t_{4m+2}) \in R_1$. Thus, for all $k \in \sigma_2$ we have

$$Q_k(t_{4m+1}) = Q_k(t_{4m+2}) = 0,$$

which leads to

$$\alpha_k(t_{4m+2}-t_{4m+1})+\sum_{i=1}^{n}\mu_i p_{ik}(T_i(t_{4m+2})-T_i(t_{4m+1}))-\mu_k(T_k(t_{4m+2})-T_k(t_{4m+1})) = 0,\ k\in\sigma_2.$$

Summing over all $m \geq 0$, we obtain

$$\alpha_k\tau_1 + \sum_{i=1}^{n}\mu_i p_{ik}\tau_i^1 - \mu_k\tau_k^1 = 0,\ k\in\sigma_2.$$

Similarly, for $k \in \sigma_1$, we have $Q_k(t_{4m+3}) = Q_k(t_{4m+4}) = 0$, which yields

$$\alpha_k(t_{4m+4}-t_{4m+3})+\sum_{i=1}^{n}\mu_i p_{ik}(T_i(t_{4m+4})-T_i(t_{4m+3}))-\mu_k(T_k(t_{4m+4})-T_k(t_{4m+3})) = 0,\ k\in\sigma_1,$$

and leads to

$$\alpha_k\tau_3 + \sum_{i=1}^{n}\mu_i p_{ik}\tau_i^3 - \mu_k\tau_k^3 = 0,\ k\in\sigma_1.$$

Since $Q(t_{4m+2}) \in R_1$ and $Q(t_{4m+3}) \in R_2$, we obtain

$$0 = Q_k(t_{4m+2}),\ k\in\sigma_2$$

and

$$0 \leq Q_k(t_{4m+3}),\ k\in\sigma_2,$$

which implies that for all $k \in \sigma_2$, $Q_k(t_{4m+3}) - Q_k(t_{4m+2}) \geq 0$, leading to

$$\alpha_k(t_{4m+3}-t_{4m+2})+\sum_{i=1}^{n}\mu_i p_{ik}(T_i(t_{4m+3})-T_i(t_{4m+2}))-\mu_k(T_k(t_{4m+3})-T_k(t_{4m+2})) \geq 0,\ k\in\sigma_2.$$

Summing over all $m \geq 0$, we obtain

$$\alpha_k \tau_2 + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i^2 - \mu_k \tau_k^2 \geq 0,\ \ k \in \sigma_2.$$

Similarly, for all $k \in \sigma_1$, $Q_k(t_{4m+3}) - Q_k(t_{4m+2}) \leq 0$, leading to

$$\alpha_k(t_{4m+3} - t_{4m+2}) + \sum_{i=1}^{n} \mu_i p_{ik}(T_i(t_{4m+3}) - T_i(t_{4m+2})) - \mu_k(T_k(t_{4m+3}) - T_k(t_{4m+2})) \leq 0,\ k \in \sigma_1,$$

and therefore,

$$\alpha_k \tau_2 + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i^2 - \mu_k \tau_k^2 \leq 0,\ \ k \in \sigma_1.$$

Finally, since $Q(t_{4m+4}) \in R_2$ and $Q(t_{4m+5}) \in R_1$, we obtain:

$$\alpha_k(t_{4m+5} - t_{4m+4}) + \sum_{i=1}^{n} \mu_i p_{ik}(T_i(t_{4m+5}) - T_i(t_{4m+4})) - \mu_k(T_k(t_{4m+5}) - T_k(t_{4m+4})) \geq 0,\ k \in \sigma_1,$$

$$\alpha_k(t_{4m+5} - t_{4m+4}) + \sum_{i=1}^{n} \mu_i p_{ik}(T_i(t_{4m+5}) - T_i(t_{4m+4})) - \mu_k(T_k(t_{4m+5}) - T_k(t_{4m+4})) \leq 0,\ k \in \sigma_2,$$

leading respectively to

$$\alpha_k \tau_4 + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i^4 - \mu_k \tau_k^4 \geq 0,\ \ k \in \sigma_1,$$

$$\alpha_k \tau_4 + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i^4 - \mu_k \tau_k^4 \leq 0,\ \ k \in \sigma_2.$$

Recall that $\tau = \sum_{j=1}^{4} \tau_j$. Then, from the dynamics of the network

$$Q_k(\tau) = Q_k(0) + \alpha_k \tau + \sum_{i=1}^{n} \mu_i p_{ik} \sum_{j=1}^{4} \tau_i^j - \mu_k \sum_{j=1}^{4} \tau_k^j.$$

Since $Q(\tau) = 0$, we obtain

$$\alpha_k \tau + \sum_{i=1}^{n} \mu_i p_{ik} \sum_{j=1}^{4} \tau_i^j - \mu_k \sum_{j=1}^{4} \tau_k^j = -Q_k(0), \qquad k = 1, \ldots, n.$$

We have shown that all of the constraints of the linear program $LP[Q(0)]$ must be satisfied and therefore τ must be bounded above by the value of this linear program. $\square$

The linear program LP[Q(0)] gives an upper bound on the strong busy period of all stable work-conserving policies. Similarly, if we minimize $\sum_{i=1}^{4} \tau_i$ we find a lower bound on the time it takes for the network to empty using a work-conserving policy starting from an initial condition $Q(0)$. The lower bound is particularly interesting as it gives information on the least possible emptying time.

2.3.2 LP Formulation. Sufficient Conditions for Stability

In this subsection, we derive sufficient conditions for stability of the fluid network. The sufficient conditions involve the linear program LP[0] which is defined exactly as the linear program LP[Q(0)] of the preceding subsection, except that the right–hand side variables $Q_k(0)$ in the constraints (2.10) are set to zero.

Theorem 2.3 (Sufficient Conditions for stability) *Consider the following set of linear inequalities in $4(n+1)$ variables*

$$\tau_1 = \sum_{k\in\sigma_1} \tau_k^1, \quad \tau_1 \geq \sum_{k\in\sigma_2} \tau_k^1, \tag{2.14}$$

$$\tau_2 = \sum_{k\in\sigma_1} \tau_k^2, \quad \tau_2 = \sum_{k\in\sigma_2} \tau_k^2, \tag{2.15}$$

$$\tau_3 \geq \sum_{k\in\sigma_1} \tau_k^2, \quad \tau_3 = \sum_{k\in\sigma_2} \tau_k^3, \tag{2.16}$$

$$\tau_4 = \sum_{k\in\sigma_1} \tau_k^4, \quad \tau_4 = \sum_{k\in\sigma_2} \tau_k^4, \tag{2.17}$$

$\forall\, k \in \sigma_2$:

$$\alpha_k \tau_1 + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i^1 - \mu_k \tau_k^1 = 0, \tag{2.18}$$

$$\alpha_k \tau_2 + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i^2 - \mu_k \tau_k^2 \geq 0, \tag{2.19}$$

$$\alpha_k \tau_4 + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i^4 - \mu_k \tau_k^4 \leq 0, \tag{2.20}$$

$\forall\, k \in \sigma_1$:

$$\alpha_k \tau_3 + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i^3 - \mu_k \tau_k^3 = 0, \tag{2.21}$$

$$\alpha_k \tau_4 + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i^4 - \mu_k \tau_k^4 \geq 0, \tag{2.22}$$

$$\alpha_k \tau_2 + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i^2 - \mu_k \tau_k^2 \leq 0, \tag{2.23}$$

$\forall k \in \{1, \ldots, n\}$:

$$\alpha_k \sum_{j=1}^{4} \tau_j + \sum_{i=1}^{n} \mu_i p_{ik} \sum_{j=1}^{4} \tau_i^j - \mu_k \sum_{j=1}^{4} \tau_k^j = 0, \tag{2.24}$$

$$\tau_j \geq 0, \quad \tau_k^j \geq 0,$$

to be referred to as $LP[0]$. If $LP[0]$ has zero as the only feasible solution, then the multiclass fluid network (α, μ, P, C) is stable for all work conserving policies.

Proof:

Let us assume that zero is the only feasible solution of $LP[0]$. Let us also assume that there exists an initial condition $Q(0) \neq 0$ and a work-conserving policy such that $Q(t)$ never becomes zero. We will derive a contradiction.

Recall that the constraints in $LP[0]$ and in $LP[Q(0)]$ are the same except that the right hand-side in (2.10) is changed from $-Q_k(0)$ to zero. Using linear programming theory ([6]) and since 0 is the only feasible solution of $LP[0]$, it follows that the feasible set of $LP[Q(0)]$ is bounded. Let Z be the optimal value of the objective function in $LP[Q(0)]$, which is finite.

Let us now consider the unstable policy starting from $Q(0)$. Let us follow this policy up to time Z; from then on, let us switch to some stable work-conserving policy (under our standing assumption that $\rho < e$, it is known that such a policy exists.) We then obtain a work-conserving policy that, starting from $Q(0)$, eventually leads the state to zero, say at some time τ. By construction $\tau > Z$. On the other hand, Proposition 2.4 asserts that $\tau \leq Z$. This is a contradiction and the proof is complete.

□

2.3.3 Necessary Conditions for Stability of Two-Station Networks

In this subsection we show that the conditions of Theorem 2.3 are also necessary. In particular, we show that if the linear program $LP[0]$ has a nonzero solution $(\tau_j, \tau_k^j),\ j = 1, \ldots, 4,\ k = 1, \ldots, n$, then there exists a work-conserving policy and an initial condition $Q(0) \neq 0$, such that for some time $\tau > 0$, $Q(\tau) = Q(0)$. By repeating the same policy each time that the state $Q(0)$ is revisited, the system never empties and therefore the fluid network is unstable. In preparation for the instability theorem we prove the following proposition.

Proposition 2.5 *If* $(\tau_j, \tau_k^j),\ j = 1, \ldots, 4,\ k = 1, \ldots, n$, *is a nonzero solution of* $LP[0]$, *then* $\tau_j > 0$ *for all* $j = 1, \ldots, 4$.

Proof

Suppose $\tau_1 = 0$. Then from (2.14) $\tau_k^1 = 0$ for all $k = 1, \ldots, n$ and therefore, from (2.24) we obtain for all $k = 1, \ldots, n$,

$$\alpha_k(\tau_2 + \tau_3 + \tau_4) + \sum_{i=1}^{n} \mu_i p_{ik}(\tau_i^2 + \tau_i^3 + \tau_i^4) - \mu_k(\tau_k^2 + \tau_k^2 + \tau_k^4) = 0$$

or in matrix form, with $\tau^j = (\tau_1^j, \ldots, \tau_n^j)'$,

$$\alpha(\tau_2 + \tau_3 + \tau_4) + [P' - I]M[\tau^2 + \tau^3 + \tau^4] = 0$$

Multiplying both sides from the left by $CM^{-1}[I - P']^{-1}$ we obtain

$$\begin{pmatrix} \rho_1 - 1 \\ \rho_2 - 1 \end{pmatrix} (\tau_2 + \tau_3 + \tau_4) + \begin{pmatrix} \tau_2 + \tau_3 + \tau_4 - \sum_{k\in\sigma_1}(\tau_k^2 + \tau_k^3 + \tau_k^4) \\ \tau_2 + \tau_3 + \tau_4 - \sum_{k\in\sigma_2}(\tau_k^2 + \tau_k^2 + \tau_k^4) \end{pmatrix} = 0.$$

But from (2.15), (2.16) and (2.17) we obtain

$$\tau_2 + \tau_3 + \tau_4 = \sum_{k \in \sigma_2} (\tau_k^2 + \tau_k^3 + \tau_k^4).$$

Since $\tau_2 + \tau_3 + \tau_4 > 0$, we obtain that $\rho_2 = 1$, a contradiction. A similar argument shows that $\tau_3 > 0$.

Suppose now that $\tau_2 = 0$. From (2.15), $\tau^2 = (\tau_1^2, \ldots, \tau_n^2) = 0$, while from (2.18), (2.20), and (2.24), we obtain that

$$\alpha_k \tau_3 + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i^3 - \mu_k \tau_k^3 \geq 0, \qquad k \in \sigma_2.$$

From (2.21) we obtain

$$\alpha_k \tau_3 + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i^3 - \mu_k \tau_k^3 = 0, \qquad k \in \sigma_1.$$

Combining these two equations in matrix form, we obtain

$$\alpha \tau_3 + [P' - I] M \tau^3 \geq 0.$$

Multiplying both sides of the inequality by $CM^{-1}[I - P']^{-1}$, we obtain

$$\begin{pmatrix} \rho_1 - 1 \\ \rho_2 - 1 \end{pmatrix} \tau_3 + \begin{pmatrix} \tau_3 - \sum_{k \in \sigma_1} \tau_k^3 \\ \tau_3 - \sum_{k \in \sigma_2} \tau_k^3 \end{pmatrix} \geq 0.$$

Since from (2.16), $\tau_3 = \sum_{k \in \sigma_2} \tau_k^3$ and $\tau_3 > 0$, we obtain that $\rho_2 = 1$, a contradiction. By a similar argument $\tau_4 > 0$. $\square$

We next prove that the condition of Theorem 2.3 is also necessary.

Theorem 2.4 (Necessary Conditions for stability) *If the linear program $LP[0]$ has a nonzero solution, then there exists a work-conserving policy under which the multiclass fluid network (α, μ, P, C) is unstable.*

Proof:

Let (τ_j, τ_k^j) be a nonzero solution of the linear program $LP[0]$. We will construct an initial condition $Q(0) \in R_1$ and a work-conserving policy, such that for some time $\tau > 0$, $Q(\tau) = Q(0)$. It will follow that there exists a work-conserving policy under which the system never empties and therefore the fluid network is unstable.

Let

$$Q_k(0) = -(\alpha_k \tau_2 + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i^2 - \mu_k \tau_k^2), \qquad k \in \sigma_1 \tag{2.25}$$

and

$$Q_k(0) = 0, \qquad k \in \sigma_2.$$

Constraint (2.23) guarantees that $Q(0) \geq 0$. We next show that $\sum_{k\in\sigma_1} Q_k(0) > 0$, i.e., $Q(0) \in R_1$. If $Q(0) = 0$, then, for all $k \in \sigma_1$

$$\alpha_k \tau_2 + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i^2 - \mu_k \tau_k^2 = 0$$

Moreover, from (2.19) for all $k \in \sigma_2$

$$\alpha_k \tau_2 + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i^2 - \mu_k \tau_k^2 \geq 0.$$

In matrix form, with $\tau^i = (\tau_1^i, \ldots, \tau_n^i)'$, the previous equations become

$$\alpha \tau_2 + [P' - I] M \tau^2 \geq 0.$$

Multiplying by $CM^{-1}[I - P']^{-1}$, we obtain

$$\begin{pmatrix} \rho_1 - 1 \\ \rho_2 - 1 \end{pmatrix} \tau_2 + \begin{pmatrix} \tau_2 - \sum_{k\in\sigma_1} \tau_k^2 \\ \tau_2 - \sum_{k\in\sigma_2} \tau_k^2 \end{pmatrix} \geq 0.$$

From (2.15), we have $\tau_2 = \sum_{k\in\sigma_1} \tau_k^2 = \sum_{k\in\sigma_2} \tau_k^2$. From Proposition 2.5, $\tau_2 > 0$, so $\rho_1, \rho_2 \geq 1$, a contradiction and therefore, $Q(0) \neq 0$.

We next construct the following allocation process for $k = 1, \ldots, n$:

$$T_k(t) = \begin{cases} \frac{t}{\tau_2}\tau_k^2 & t \in [0, \tau_2]; \\ \tau_k^2 + \frac{t-\tau_2}{\tau_3}\tau_k^2 & t \in (\tau_2, \tau_2 + \tau_3]; \\ \tau_k^2 + \tau_k^2 + \frac{t-\tau_2-\tau_3}{\tau_4}\tau_k^4 & t \in (\tau_2 + \tau_3, \tau_2 + \tau_3 + \tau_4]; \\ \tau_k^2 + \tau_k^2 + \tau_k^4 + \frac{t-\tau_2-\tau_3-\tau_4}{\tau_1}\tau_k^1 & t \in (\tau_2 + \tau_3 + \tau_4, \tau_2 + \tau_3 + \tau_4 + \tau_1]. \end{cases}$$

We show that the above allocation process is both feasible and work-conserving.

We first consider the first interval $[0, \tau_2]$. By the dynamics of the fluid network for this allocation process and starting from the initial condition given above we obtain from (2.25) and (2.19)

$$Q_k(\tau_2) = 0, \;\; k \in \sigma_1$$

$$Q_k(\tau_2) = \alpha_k \tau_2 + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i^2 - \mu_k \tau_k^2 \geq 0, \;\; k \in \sigma_2.$$

We next show that

$$\sum_{k \in \sigma_2} Q_k(\tau_2) > 0,$$

so $Q(\tau_2) \in R_2$. If not, then

$$Q_k(\tau_2) = 0, \quad k \in \sigma_2$$

or

$$\alpha_k \tau_2 + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i^2 - \mu_k \tau_k^2 = 0, \quad k \in \sigma_2,$$

then from (2.18) and (2.24) we obtain that

$$\alpha_k(\tau_3 + \tau_4) + \sum_{i=1}^{n} \mu_i p_{ik}(\tau_i^3 + \tau_i^4) - \mu_k(\tau_k^3 + \tau_k^4) = 0, \quad k \in \sigma_2.$$

Also from (2.21) and (2.22) we obtain that

$$\alpha_k(\tau_3 + \tau_4) + \sum_{i=1}^{n} \mu_i p_{ik}(\tau_i^3 + \tau_i^4) - \mu_k(\tau_k^3 + \tau_k^4) \geq 0, \quad k \in \sigma_1.$$

Written in matrix from, the two previous relations become

$$\alpha(\tau_3+\tau_4)+[P'-I]M(\tau^3+\tau^4)\geq 0.$$

Multiplying by $CM^{-1}[I-P']^{-1}$, we obtain

$$\begin{pmatrix}\rho_1-1\\ \rho_2-1\end{pmatrix}(\tau_3+\tau_4)+\begin{pmatrix}\tau_3+\tau_4-\sum_{k\in\sigma_1}(\tau_k^2+\tau_k^4)\\ \tau_3+\tau_4-\sum_{k\in\sigma_2}(\tau_k^2+\tau_k^4)\end{pmatrix}\geq 0.$$

Since $\tau_3+\tau_4=\sum_{k\in\sigma_2}(\tau_k^2+\tau_k^4)$ and $\tau_3+\tau_4>0$, we obtain $\rho_2\geq 1$, a contradiction, and therefore $\sum_{k\in\sigma_2}Q_k(\tau_2)>0$.

Since the allocation process is linear, we obtain:

$$\forall t\in[0,\tau_2],\qquad Q(t)\geq 0,$$

and

$$\forall t\in(0,\tau_2),\qquad Q(t)\in R_{12},$$

i.e., the allocation process is feasible. We next show that it is also work-conserving. From (2.15)

$$t=\sum_{k\in\sigma_1}\frac{t}{\tau_2}\tau_k^2=\sum_{k\in\sigma_2}\frac{t}{\tau_2}\tau_k^2$$

or equivalently

$$\forall t\in[0,\tau_2]:\ U_1(t)=U_2(t)=U_1(0)=U_2(0)=0,$$

and the process is indeed work-conserving.

In the interval $(\tau_2,\tau_2+\tau_3]$, we prove similarly that for $k\in\sigma_2$ we have $Q_k(\tau_2+\tau_3)\geq 0$ and $\sum_{k\in\sigma_2}Q_k(\tau_2+\tau_3)>0$. Therefore, $Q(\tau_2+\tau_3)\in R_2$, and since $Q(\tau_2)\in R_2$, we obtain by linearity that

$$\forall t\in[\tau_2,\tau_2+\tau_3],\ Q(t)\in R_2.$$

Work-conservation is shown similarly.

Similarly, we show that in the interval $t \in (\tau_2 + \tau_3, \tau_2 + \tau_3 + \tau_4]$, $Q(t) \in R_{12}$ and in the interval $t \in [\tau_2 + \tau_3 + \tau_4, \tau_2 + \tau_3 + \tau_4 + \tau_1]$, $Q(t) \in R_1$, while the process is work-conserving.

In addition, because of (2.24), $Q(\tau_1 + \tau_2 + \tau_3 + \tau_4) = Q(0)$. It follows that the fluid network never empties for this work-conserving feasible policy, and is unstable. □

The necessity proof has identified a particular way that an unstable work-conserving trajectory materializes, leading to some insight as to how instability may be reached. In particular, we have shown that if there exists an unstable trajectory, then there exists a periodic trajectory with a particular structure.

Combining Theorems 2.3 and 2.4 we obtain the main theorem of this section.

Theorem 2.5 *A two–station multiclass fluid network* (α, μ, P, C) *is stable for all work conserving policies if and only if the load condition* $\rho < e$ *holds and the linear program* $LP[0]$ *has zero as the only feasible solution.*

2.3.4 On the Power of Piecewise Linear Lyapunov Functions

In this subsection we will show that a two-station network is stable if and only if the linear program constructed by Meyn and Down in [21] has a feasible solution. The solution to the linear program by Meyn and Down (as discussed in [21]), if it exists, provides a certain piecewise linear Lyapunov function which guarantees stability. We will demonstrate that the dual of this linear program is a projection of the linear program LP[0] constructed in the previous section. Finally we will significantly simplify LP[0] and construct a linear program with only $2n$ variables that exactly characterizes stability.

Consider a multiclass fluid network (α, μ, P, C) with two stations, which is a reentrant line. Namely, there is only a single arrival stream of customers, i.e., $\alpha_1 = \lambda$, $\alpha_2 = \ldots = \alpha_n = 0$. These customers are processed deterministically from class k to class $k+1$ ($p_{k,k+1} = 1$ for $k = 1, 2, \ldots, n-1$, $p_{ij} = 0$ otherwise). Meyn and

Down [21] proved that if the following linear program

$$
\begin{aligned}
\lambda L_1 + \mu_i(L_{i+1} - L_i) &\le -1 && i \in \sigma_1 \\
\lambda Q_1 + \mu_j(Q_{j+1} - Q_j) &\le -1 && j \in \sigma_2 \\
\lambda L_1 + \mu_i(L_{i+1} - L_i) + \mu_j(L_{j+1} - L_j) &\le -1 && i \in \sigma_1,\, j \in \sigma_2 \\
\lambda Q_1 + \mu_i(Q_{i+1} - Q_i) + \mu_j(Q_{j+1} - Q_j) &\le -1 && i \in \sigma_1,\, j \in \sigma_2 \\
L_i &\ge Q_i && i \in \sigma_1 \\
L_j &\le Q_j && j \in \sigma_2 \\
L \ge 0,\ Q &\ge 0,
\end{aligned}
$$

is feasible, then the piecewise linear function $\Phi(x) = \max(L'x, Q'x)$, for $x \ge 0$, is a Lyapunov function and therefore the network is stable for all work conserving policies.

We can easily extend this linear program to a general multiclass two-station fluid network (α, μ, P, C), i.e., not necessarily a reentrant line. If the following linear program (we call it LP[md])

$$
\begin{array}{llcll}
(X_i) & \sum_{k=1}^n L_k\alpha_k + \sum_{k=1}^n L_k p_{ik}\mu_i - L_i\mu_i + V & \le & -1 & i \in \sigma_1 \\
(X_j) & \sum_{k=1}^n L_k p_{jk}\mu_j - L_j\mu_j & \le & V & j \in \sigma_2 \\
(Y_j) & \sum_{k=1}^n Q_k\alpha_k + \sum_{k=1}^n Q_k p_{jk}\mu_j - Q_j\mu_j + W & \le & -1 & j \in \sigma_2 \\
(Y_i) & \sum_{k=1}^n Q_k p_{ik}\mu_i - Q_i\mu_i & \le & W & i \in \sigma_1 \\
(m_i) & L_i & \ge & Q_i & i \in \sigma_1 \\
(n_j) & L_j & \le & Q_j & j \in \sigma_2 \\
 & L,\ Q,\ V,\ W & \ge & 0 &
\end{array}
$$

is feasible, then a piecewise linear function $\Phi(x) = \max(L'x, Q'x)$, is a Lyapunov function and therefore the network is stable for all work conserving policies (the associated dual variables are indicated in parenthesis).

Let the objective function in LP[md] be to maximize $0L + 0Q + 0V + 0W$ and consider the dual LP. It is a homogeneous LP in the variables $X_k, Y_k,\ k = 1, 2, \ldots, n,\ \ m_k,\ k \in$

$\sigma_1,\ n_k,\ k \in \sigma_2$ which has the following form:

$$\begin{array}{lll}
\text{maximize} & -\sum_{i\in\sigma_1} X_i - \sum_{j\in\sigma_2} Y_j & \\
\text{subject to} & \alpha_k \sum_{i\in\sigma_1} X_i + \sum_{i=1}^{n} \mu_i p_{ik} X_i - \mu_k X_k - m_k \leq 0 & k \in \sigma_1 \\
 & \alpha_k \sum_{j\in\sigma_2} Y_j + \sum_{i=1}^{n} \mu_i p_{ik} Y_i - \mu_k Y_k + m_k \leq 0 & k \in \sigma_1 \\
 & \alpha_k \sum_{i\in\sigma_1} X_i + \sum_{i=1}^{n} \mu_i p_{ik} X_i - \mu_k X_k + n_k \leq 0 & k \in \sigma_2 \\
 & \alpha_k \sum_{j\in\sigma_2} Y_j + \sum_{i=1}^{n} \mu_i p_{ik} Y_i - \mu_k Y_k - n_k \leq 0 & k \in \sigma_2 \\
 & \sum_{i\in\sigma_1} X_i \leq \sum_{j\in\sigma_2} X_j & \\
 & \sum_{j\in\sigma_2} Y_j \leq \sum_{i\in\sigma_1} Y_i & \\
 & X, Y, m, n \leq 0 &
\end{array}$$

By multiplying objective function, constraints and variables by -1 we obtain the following equivalent linear program:

$$\begin{array}{lll}
\text{maximize} & \sum_{i\in\sigma_1} X_i + \sum_{j\in\sigma_2} Y_j & \\
\text{subject to} & \alpha_k \sum_{i\in\sigma_1} X_i + \sum_{i=1}^{n} \mu_i p_{ik} X_i - \mu_k X_k - m_k \geq 0 & k \in \sigma_1 \\
 & \alpha_k \sum_{j\in\sigma_2} Y_j + \sum_{i=1}^{n} \mu_i p_{ik} Y_i - \mu_k Y_k + m_k \geq 0 & k \in \sigma_1 \\
 & \alpha_k \sum_{i\in\sigma_1} X_i + \sum_{i=1}^{n} \mu_i p_{ik} X_i - \mu_k X_k + n_k \geq 0 & k \in \sigma_2 \\
 & \alpha_k \sum_{j\in\sigma_2} Y_j + \sum_{i=1}^{n} \mu_i p_{ik} Y_i - \mu_k Y_k - n_k \geq 0 & k \in \sigma_2 \\
 & \sum_{i\in\sigma_1} X_i \geq \sum_{j\in\sigma_2} X_j & \\
 & \sum_{j\in\sigma_2} Y_j \geq \sum_{i\in\sigma_1} Y_i & \\
 & X, Y, m, n \geq 0 &
\end{array}$$

which we call DLP[md].

Lemma 2.1 *LP[md] is feasible if and only if DLP[md] has zero as the only feasible solution.*

Proof:

follows immediately from strong duality of linear programming.□

We will gradually simplify DLP[md]. We start with the following lemma:

Lemma 2.2 *DLP[md] has a nonzero feasible solution if and only if the following linear program, called DLP[1], has a nonzero feasible solution.*

$$\begin{array}{lll} \text{maximize} & \sum_{i\in\sigma_1} X_i + \sum_{j\in\sigma_2} Y_j & \\ \text{subject to} & \alpha_k \sum_{i\in\sigma_1} X_i + \sum_{i=1}^{n} \mu_i p_{ik} X_i - \mu_k X_k \geq 0 \quad k \in \sigma_1 & (2.26) \\ & \alpha_k \sum_{j\in\sigma_1} Y_j + \sum_{i=1}^{n} \mu_i p_{ik} Y_i - \mu_k Y_k \geq 0 \quad k \in \sigma_2 & (2.27) \\ & \alpha_k(\sum_{i\in\sigma_1} X_i + \sum_{j\in\sigma_2} Y_j) & \\ & + \sum_{i=1}^{n} \mu_i p_{ik}(X_i + Y_i) - \mu_k(X_k + Y_k) \geq 0 \quad k = 1, 2, \ldots, n & (2.28) \\ & \sum_{k\in\sigma_1} X_k \geq \sum_{k\in\sigma_2} X_k & (2.29) \\ & \sum_{k\in\sigma_2} Y_k \geq \sum_{k\in\sigma_1} Y_k & (2.30) \\ & X, Y \geq 0 & \end{array}$$

Proof:

Let X_k, Y_k, m_k, n_k be a feasible nonzero solution to DLP[md]. Since,

$$\alpha_k \sum_{i\in\sigma_1} X_i + \sum_{i=1}^{n} \mu_i p_{ik} X_i - \mu_k X_k - m_k \geq 0, \;\; m_k \geq 0$$

(2.26) follows. Similarly, (2.27) follows. By adding inequalities in DLP[md] corresponding to stations σ_1 and σ_2 separately, we obtain that X_k, Y_k is a feasible nonzero solution to DLP[1].

Conversely, if X_k, Y_k is a nonzero solution to DLP[1], then by setting

$$\forall\, k \in \sigma_1 : \;\; \alpha_k \sum_{i\in\sigma_1} X_i + \sum_{i=1}^{n} \mu_i p_{ik} X_i - \mu_k X_k = m_k$$

and

$$\forall\, k \in \sigma_2 : \;\; \alpha_k \sum_{j\in\sigma_1} Y_j + \sum_{i=1}^{n} \mu_i p_{ik} Y_i - \mu_k Y_k = n_k$$

we obtain that X_k, Y_k, m_k, n_k is a nonzero solution to DLP[md]. $\square$

The next lemma shows that we can change inequality (2.28) to an equality.

Lemma 2.3 *Let DLP[2] be a linear program obtained from DLP[1] by replacing inequality (2.28) with equality. Then, if the condition $\rho < e$ holds , DLP[2] has a nonzero feasible solution if and only if DLP[1] has a nonzero feasible solution.*

Proof:

Trivially, if X, Y is a nonzero solution to DLP[2], then it is also a nonzero solution to DLP[1]. For the converse part, let X, Y be a nonzero solution to DLP[1]. We will construct a nonzero solution to DLP[2].

Let us rewrite (2.28) in matrix form as follows:

$$\alpha(x+y) + [P' - I]M(X+Y) \geq 0 \tag{2.31}$$

where we define

$$x = \sum_{i \in \sigma_1} X_i, \; y = \sum_{j \in \sigma_2} Y_j, \; X = (X_1, \ldots, X_n), \; Y = (Y_1, \ldots, Y_n) \tag{2.32}$$

Since $[I - P']^{-1}$ and M^{-1} exist and are non-negative, (2.31) is equivalent to

$$M^{-1}[I - P']^{-1}\alpha(x+y) - (X+Y) \geq 0$$

or simply

$$\rho(x+y) - (X+Y) \geq 0.$$

Now we will increase X_k to $\hat{X}_k$ for all $k \in \sigma_2$ so that for all $k \in \sigma_2$:

$$\rho_k(x+y) - (\hat{X}_k + Y_k) = 0$$

This is possible to do because x is not affected by X_k for $k \in \sigma_2$. Notice also that this change can only increase the LHS of (2.26).

Similarly we construct $\hat{Y}_k$ for all $k \in \sigma_1$ such that for all $k \in \sigma_1$

$$\rho_k(x+y) - (X_k + \hat{Y}_k) = 0$$

and (2.27) is still satisfied. Finally we show that (2.29) and (2.30) are still satisfied. We have, by construction,

$$\sum_{k\in\sigma_2} \hat{X}_k + \sum_{k\in\sigma_2} Y_k = \sum_{k\in\sigma_2} \rho_k(x+y) = \rho_{\sigma_2}(x+y) \le x+y$$

Since, by definition, $y = \sum_{k\in\sigma_2} Y_k$ we obtain that

$$\sum_{k\in\sigma_2} \hat{X}_k \le x,$$

i.e., (2.29) holds. By a similar reason (2.30) holds, i.e.,

$$\sum_{k\in\sigma_1} \hat{Y}_k \le y.$$

The new solution $\hat{X}, \hat{Y}$ satisfies $\hat{X} \ge X, \hat{Y} \ge Y$ and therefore it is nonzero. By construction, it is a feasible solution to DLP[2]. □

In the remaining part of this section we will show that DLP[2] has a nonzero solution if and only if LP[0] (from Section 3) has a nonzero solution. We show first that DLP[2] is a projection of LP[0]:

Lemma 2.4 *Let* $(\tau_1, \tau_2, \tau_3, \tau_4, \tau_k^1, \tau_k^2, \tau_k^3, \tau_k^4)$, $k = 1, 2, \dots, n$ *be a nonzero feasible solution to LP[0] . Let* $X_k = \tau_k^4 + \tau_k^1$, $Y_k = \tau_k^2 + \tau_k^3$, $k = 1, 2, \dots, n$. *Then* (X_k, Y_k) *is a nonzero feasible solution to DLP[2].*

Proof:

Combining (2.14) with (2.17) we obtain (2.29). Combining (2.15) with (2.16) we obtain (2.30). (2.24) shows that (2.28) (with equality) holds. Combining (2.21) with

(2.23) we obtain that

$$\forall\, k \in \sigma_1 : \ \alpha_k \sum_{j\in\sigma_2} Y_j + \sum_{i=1}^{n} \mu_i p_{ik} Y_i - \mu_k Y_k \leq 0.$$

By subtracting this from (2.28) (with equality) we obtain (2.26). (2.27) is obtained similarly. By construction, if

$$(\tau_1, \tau_2, \tau_3, \tau_4, \tau_k^1, \tau_k^2, \tau_k^3, \tau_k^4)$$

is nonzero, then the solution (X_k, Y_k) is nonzero as well. □

We now intend to prove the more involved converse part.

Lemma 2.5 *If there exists a nonzero solution to DLP[2], then there exists a nonzero solution to LP[0].*

Proof:

Let (X_k, Y_k), $k = 1, 2, \ldots, n$ be a nonzero solution to DLP[2]. Let $x = \sum_{i\in\sigma_1} X_i$ and $y = \sum_{j\in\sigma_2} Y_j$.

We will construct a nonzero solution to LP[0].

We select a number $\gamma \in [0, 1]$; we specify how γ is selected later. Combining (2.27) and (2.28) (with equality, cf. Lemma 2.3) we obtain

$$\alpha_k x + \sum_{i=1}^{n} \mu_i p_{ik} X_i - \mu_k X_k \leq 0, \quad k \in \sigma_2.$$

Then

$$\alpha_k \gamma x + \sum_{i=1}^{n} \mu_i p_{ik} \gamma X_i - \mu_k \gamma X_k \leq 0, \quad k \in \sigma_2. \tag{2.33}$$

Let us rewrite this as follows

$$\alpha_k \gamma x + \sum_{i\in\sigma_1} \mu_i p_{ik} \gamma X_i + \sum_{j\in\sigma_2} \mu_i p_{ik} \gamma X_i - \mu_k \gamma X_k \leq 0, \quad k \in \sigma_2. \tag{2.34}$$

We introduce the following notation. For any vector $W \in R_+^n$ let W_{σ_1} and W_{σ_2} be the portion of the vector W corresponding to the indices in σ_1 and σ_2 respectively.

Also let $P_{ij},\ i,j=1,2$ be the portion of the matrix P corresponding to the pairs of indices in $\sigma_i \times \sigma_j$:

$$P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}.$$

The matrices P_{12} and P_{22} are just the portions of the matrix P corresponding to flows going from station σ_1 to σ_2 and from σ_2 to itself.

We rewrite (2.34) in matrix form:

$$\alpha_{\sigma_2}\gamma x + P_{12}M_{\sigma_1}\gamma X_{\sigma_1} + [P_{22} - I_{\sigma_2}]M_{\sigma_2}\gamma X_{\sigma_2} \le 0, \tag{2.35}$$

The matrix P_{22} is non-negative and has spectral radius less than one. Therefore the matrix $[I_{\sigma_2} - P_{22}]^{-1}$ exists and is non-negative. We rewrite (2.35) as follows

$$M_{\sigma_2}^{-1}[I_{\sigma_2} - P_{22}]^{-1}\alpha_{\sigma_2}\gamma x + M_{\sigma_2}^{-1}[I_{\sigma_2} - P_{22}]^{-1}P_{12}M_{\sigma_1}\gamma X_{\sigma_1} - \gamma X_{\sigma_2} \le 0 \tag{2.36}$$

We next introduce a $|\sigma_2|$-dimensional vector $\tau_{\sigma_2}^1$

$$\tau_{\sigma_2}^1 = M_{\sigma_2}^{-1}[I_{\sigma_2} - P_{22}]^{-1}\alpha_{\sigma_2}\gamma x + M_{\sigma_2}^{-1}[I_{\sigma_2} - P_{22}]^{-1}P_{12}M_{\sigma_1}\gamma X_{\sigma_1} = \gamma Z_{\sigma_2} \ge 0. \tag{2.37}$$

From (2.36) it follows that

$$\tau_{\sigma_2}^1 = \gamma Z_{\sigma_2} \le \gamma X_{\sigma_2}. \tag{2.38}$$

We have defined thus variables τ_k^1 for $k \in \sigma_2$. For all $k \in \sigma_1$ let $\tau_k^1 = \gamma X_k$. Thus we have defined variables $\tau_k^1,\ k = 1,2,\ldots,n$. By construction (2.37), we have (2.18) is satisfied.

We have from (2.29)

$$\gamma x = \sum_{k\in\sigma_1} \gamma X_k \ge \sum_{k\in\sigma_2} \gamma X_k$$

Then from (2.38) it follows that (2.14) is satisfied (where we set $\tau_1 = \gamma x$).

We next let $\tau_k^4 = X_k - \tau_k^1 = (1-\gamma)X_k$ for $k \in \sigma_1$, $\tau_k^4 = X_k - \tau_k^1$ for $k \in \sigma_2$ and $\tau_4 = (1-\gamma)x$. It follows from (2.38) that τ_k^4 are non-negative for $k \in \sigma_2$ and therefore all the new variables τ_k^4 are non-negative. From the way we introduced x in (2.32) it follows that the first part of (2.17) is satisfied. We next show that we can select $\gamma \in [0,1]$ so that the second part of (2.17) is satisfied as well. Recall that $\tau_k^1 = \gamma X_k,\ k \in \sigma_1,\ \tau_k^1 = \gamma Z_k,\ k \in \sigma_2$(from (2.37)), $\tau_k^4 = X_k - \tau_k^1 = (1-\gamma)X_k,\ k \in \sigma_1,\ \tau_k^4 = X_k - \tau_k^1,\ k \in \sigma_2$. Then

$$\sum_{k\in\sigma_2} \tau_k^4 = \sum_{k\in\sigma_2} (X_k - \tau_k^1)$$

and

$$\sum_{k\in\sigma_1} \tau_k^4 = (1-\gamma) \sum_{k\in\sigma_1} X_k$$

From (2.38) $Z_k \leq X_k,\ k \in \sigma_2$ and from (2.29)

$$\sum_{k\in\sigma_1} X_k \geq \sum_{k\in\sigma_2} X_k$$

Therefore

$$\sum_{k\in\sigma_2} Z_k \leq \sum_{k\in\sigma_2} X_k \leq \sum_{k\in\sigma_1} X_k$$

In case the first sum is strictly less than the third sum, we take γ to be

$$\gamma = \frac{\sum_{k\in\sigma_1} X_k - \sum_{k\in\sigma_2} X_k}{\sum_{k\in\sigma_1} X_k - \sum_{k\in\sigma_2} Z_k}$$

This guarantees

$$(1-\gamma) \sum_{k\in\sigma_1} X_k = \sum_{k\in\sigma_2} (X_k - \gamma Z_k)$$

or

$$\sum_{k\in\sigma_1} \tau_k^4 = \sum_{k\in\sigma_2} \tau_k^4$$

From the inequalities above, this value of γ satisfies $\gamma \in [0,1]$. If, on the other hand, all sums are equal, then we take γ to be any number in $[0,1]$ and (2.17) is still satisfied.

We thus have satisfied (2.14), (2.17) and (2.18). We next prove that (2.20) and (2.22) are satisfied as well.

Subtracting (2.27) from (2.28) (with equality) we obtain

$$\forall\, k \in \sigma_2 : \;\; \alpha_k x + \sum_{i=1}^{n} \mu_i p_{ik} X_i - \mu_k X_k \leq 0,$$

which in terms of the variables $\tau_1^1, \ldots, \tau_n^1, \tau_1^4, \ldots, \tau_n^4$ means

$$\forall\, k \in \sigma_2 : \;\; \alpha_k(\tau_4 + \tau_1) + \sum_{i=1}^{n} \mu_i p_{ik}(\tau_i^4 + \tau_i^1) - \mu_k(\tau_k^4 + \tau_k^1) \leq 0.$$

This combined with (2.18) proves (2.20). Also from (2.26)

$$\forall\, k \in \sigma_1 : \;\; \alpha_k(1-\gamma)x + \sum_{i=1}^{n} \mu_i p_{ik}(1-\gamma)X_i - \mu_k(1-\gamma)X_k \geq 0.$$

From (2.38) we obtain for $k \in \sigma_2$

$$\tau_k^4 = X_k - \tau_k^1 \geq X_k - \gamma X_k = (1-\gamma)X_k.$$

Therefore,

$$\forall k \in \sigma_1 : \;\; \alpha_k(1-\gamma)x + \sum_{i\in\sigma_1} \mu_i p_{ik}(1-\gamma)X_i + \sum_{i\in\sigma_2} \mu_i p_{ik}\tau_i^4 - \mu_k(1-\gamma)X_k \geq$$

$$\alpha_k(1-\gamma)x + \sum_{i=1}^{n} \mu_i p_{ik}(1-\gamma)X_i - \mu_k(1-\gamma)X_k \geq 0,$$

or, equivalently

$$\forall\, k \in \sigma_1 : \;\; \alpha_k \tau^4 + \sum_{i=1}^{n} \mu_i p_{ik}\tau_i^4 - \mu_k \tau_k^4 \geq 0$$

which is (2.22).

We thus constructed $\tau_1, \tau_4, \tau_k^1, \tau_k^4,\; k = 1, 2, \ldots n$ which satisfy (2.14), (2.17), (2.18), (2.20) and (2.22). The construction of $\tau_2, \tau_3, \tau_k^2, \tau_k^3,\; k = 1, 2, \ldots n$ is symmetrical. Finally (2.24) is a simple implication of (2.28) (with equality). If the initial solution (x, y, X_k, Y_k) is nonzero then the solution $\tau_1, \tau_2, \ldots, \tau_n^4$ is also nonzero. This concludes the proof of the lemma. □

We now summarize the results obtained in this and the previous subsections.

Corollary 2.1 *A multiclass fluid network (α, μ, P, C) with two stations is stable for all work conserving policies if and only if one of the following equivalent conditions holds:*

1. *Linear program LP[md] constructed by Meyn and Down in [21] is feasible.*

2. *Linear program DLP[2] constructed in this section has zero as the only feasible solution.*

3. *Linear program LP[0] constructed in the previous section has zero as the only feasible solution.*

It seems that from all three equivalent tests for stability, the second one - DLP[2] is the most economical. Unlike LP[md], it can be interpreted physically, with variables corresponding to times arising from a decomposition of trajectories. On the other hand, it has half as many variables compared to LP[0].

2.4 Sufficient Conditions for Stability of Multiclass Fluid Networks with Three or More Stations

In this section we extend the sufficient conditions for stability from Section 2.3 to networks with three or more stations.

Given a multiclass fluid network (α, μ, P, C), let Λ be any non-empty subset of the set of stations: $\Lambda \subset \{\sigma_1, \ldots, \sigma_J\}$. Define R_Λ as follows:

$$R_\Lambda = \{x \in R_+^n : \forall\, j \in \Lambda,\ \sum_{k \in \sigma_j} x_k > 0,\ \ \forall\, j \notin \Lambda,\ \sum_{k \in \sigma_j} x_k = 0\}$$

i.e. R_Λ corresponds to states where all the stations in Λ are busy (have customers present), and all the other stations do not have any customers present. Then we may partition the entire state space R_+^n as

$$R_+^n - \{0\} = \bigcup_{\Lambda} R_\Lambda$$

2.4.1 Trajectory Decomposition

Consider any feasible stable work-conserving allocation process $T(t)$ and its associated fluid process $Q(t)$. Let $\tau = \inf\{t : Q(t) = 0\}$. We may divide the time interval $[0, \tau]$ into the finite or countably infinite set of nonoverlapping intervals $[t_0, t_1], [t_2, t_3], \ldots, [t_m, t_{m+1}]$, such that

$$[t_0, t_1] \cup [t_2, t_3] \cup \ldots \cup [t_m, t_{m+1}] = [0, \tau)$$

and within each interval (t_m, t_{m+1}) the process stays within a single R_Λ part of the state space. Namely, for each $m = 0, 2, 4, \ldots$ there exist non-empty $\Lambda_m \subset \{1, 2, \ldots, J\}$ such that for all $t \in (t_m, t_{m+1})$, $Q(t) \in R_{\Lambda_m}$. Let us consider any such interval. For each $\sigma \in \Lambda_m$ from the work-conservation constraint (2.4) we obtain

$$\sum_{k \in \sigma} (T_k(t_{m+1}) - T_k(t_m)) = t_{m+1} - t_m \tag{2.39}$$

For each $\sigma \notin \Lambda_m$ from the feasibility constraint (2.3) we obtain:

$$\sum_{k \in \sigma} (T_k(t_{m+1}) - T_k(t_m)) \leq t_{m+1} - t_m \tag{2.40}$$

Suppose for some $\sigma \in \{\sigma_1, \ldots, \sigma_J\}$, and positive integer m we have $\sigma \notin \Lambda_m$. Then for all $k \in \sigma$ we have $Q_k(t_{m+1}) = Q_k(t_m) = 0$. therefore

$$\forall\, \sigma \notin \Lambda_m,\ \forall\, k \in \sigma :$$

$$Q_k(t_{m+1}) - Q_k(t_m) = \alpha_k(t_{m+1} - t_m) + \sum_{i=1}^{n} \mu_i p_{ik}(T_i(t_{m+1} - T_i(t_m)) - \mu_k(T_k(t_{m+1} - T_k(t_m)) = 0 \tag{2.41}$$

2.4.2 LP Formulation. Sufficient Conditions for Stability

Properties (2.39)-(2.41) are crucial in establishing the conditions for stability of fluid networks. We used them in Section 2.3 to construct a linear program LP[0] which gives the exact stability region for two-station fluid networks. We will derive here a version of this LP for networks with more than two stations. Let $[t_0, t_1], [t_2, t_3], \ldots, [t_{2m}, t_{2m+1}]$, $\Lambda_1, \Lambda_2, \ldots, \Lambda_m, \ldots$ be the partition of $[0, \tau]$ time interval and state space, respectively such that $Q(t) \in \Lambda_m$ for all $t \in (t_{2m}, t_{2m+1})$. Introduce

$$\tau_\Lambda = \sum_{m:\Lambda_m=\Lambda} (t_{2m+1} - t_{2m})$$

and

$$\tau_k^\Lambda = \sum_{m:\Lambda_m=\Lambda} (T_k(t_{2m+1}) - T_k(t_{2m}))$$

for $k = 1, 2, \ldots, n$. Suppose now some station σ belongs to the set Λ. By summing equation (2.39) over all m such that $\Lambda_m = \Lambda$ we obtain:

$$\forall\, \Lambda \neq \emptyset,\ \forall \sigma,\ s.t.\, \sigma \in \Lambda: \ \sum_{k\in\sigma} \tau_k^\Lambda = \tau_\Lambda \tag{2.42}$$

For a similar reason, from (2.40) we obtain:

$$\forall\, \Lambda \neq \emptyset,\ \forall \sigma,\ s.t.\, \sigma \notin \Lambda: \ \sum_{k\in\sigma} \tau_k^\Lambda \leq \tau_\Lambda \tag{2.43}$$

Also, by summing (2.41) over all m we obtain

$$\forall\, \Lambda \neq \emptyset,\ \forall\, \sigma \notin \Lambda: \quad \alpha_k \tau_\Lambda + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i^\Lambda - \mu_k \tau_k^\Lambda = 0 \tag{2.44}$$

Introduce also

$$\tau_k = \sum_\Lambda \tau_k^\Lambda$$

Clearly $T_k(\tau) = \tau_k$. Then by the definition of τ

$$\forall\, k = 1, 2, \ldots, n: \ \ Q_k(\tau) = Q_k(0) + \alpha_k \tau + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i - \mu_k \tau_k = 0 \tag{2.45}$$

Observe that (2.42)-(2.45) specify some linear program in non-negative variables $\tau_\Lambda, \tau_k^\Lambda, \tau, \tau_k$, and parameters $Q_k(0), \alpha_k, \mu_k, p_{ik}, c_{ij}$. The following theorem holds

Theorem 2.6 *Let* (α, μ, P, C) *be a multiclass fluid network with some initial state* $Q(0)$*. If the linear program*

$$\text{Max}\, \tau \ \textit{subject to}\ (2.42) - (2.45),$$

is bounded then the network is stable and the optimal value of $LP[Q(0)]$ *provides an upper bound on the time it takes for the network to empty. Also, linear program*

$$\text{Min}\, \tau, \ \textit{subject to}\ (2.42) - (2.45)$$

gives a lower bound on time it takes for the network to empty.

Proof:

Identical to the proof of the Theorem 2.3. □

We denote the linear program with Max τ objective function and constraints (2.42)-(2.45) by GLP[Q(0)] (general linear program). Notice, that $Q(0)$ is the only non-zero part in the right hand side of our linear program. The boundedness of linear program does not depend on the right hand side. So, let us replace $Q_k(0)$ by 0 in constraint (2.45) and obtain

$$\forall\, k = 1, 2, \ldots, n: \ \ \alpha_k \tau + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i - \mu_k \tau_k = 0 \tag{2.46}$$

The following version of Theorem 2.6 then holds :

Theorem 2.7 *Let* (α, μ, P, C) *be a multiclass fluid network.*

If the linear program (which we denote by GLP[0])

$$Max\ \tau \ \textit{subject to}\ (2.42) - (2.44),\ \textit{and}\, (2.46) \tag{2.47}$$

has zero as the only feasible solution, then the network is stable for all initial states $Q(0)$*.*

2.5 Examples

In this section, we apply the results obtained in Sections 2.3 and 2.4 to establish stability of some fluid networks. In all of the examples, the stability is established by showing that the corresponding homogeneous linear programs have zero as the only feasible solution.

2.5.1 A Two Station Network

We prove that a two-station fluid network, in which one of the two stations has only one class, is stable provided that the load condition (2.6) is satisfied. This generalizes previous results obtained by Kumar [36] and Meyn and Down [21] for a three–class, two–station network.

Theorem 2.8 *A fluid network satisfying the load condition $\rho < e$ with two stations and such that only one class is served by station 2 ($|\sigma_2| = 1$) is stable.*

Proof:

We show that the corresponding linear program $LP[0]$ cannot have a nonzero solution. For the purposes of contradiction suppose that (τ_j, τ_k^j) is a nonzero solution to $LP[0]$. Let $\sigma_2 = \{l\}$. We distinguish two cases:

Case 1: $\alpha_l \tau_3 + \sum_{i=1}^n \mu_i p_{il} \tau_i^3 - \mu_l \tau_l^3 \geq 0$.

From (2.21):

$$\alpha_k \tau_3 + \sum_{i=1} \mu_i p_{ik} \tau_i^3 - \mu_k \tau_k^2 = 0, \ \forall k \in \sigma_1.$$

We combine the previous relations in matrix form as follows:

$$\alpha \tau_3 + [P' - I] M \tau^3 \geq 0.$$

We multiply both sides by $CM^{-1}[I - P']^{-1}$ to obtain:

$$\begin{pmatrix} \rho_1 - 1 \\ \rho_2 - 1 \end{pmatrix} \tau_3 + \begin{pmatrix} \tau_3 - \sum_{k \in \sigma_1} \tau_k^2 \\ \tau_3 - \tau_l^3 \end{pmatrix} \geq 0.$$

But from (2.16) we obtain $\tau_3 = \tau_l^3$ and from Proposition 2.5, we obtain $\tau_3 > 0$, leading to $\rho_2 = 1$, a contradiction.

Case 2 : $\alpha_l \tau_3 + \sum_{i=1}^n \mu_i p_{il} \tau_i^3 - \mu_l \tau_l^3 \leq 0$.

From (2.24), we obtain

$$\alpha_l(\tau_4 + \tau_1 + \tau_2) + \sum_{i=1}^n \mu_i p_{il}(\tau_i^4 + \tau_i^1 + \tau_i^2) - \mu_l(\tau_l^4 + \tau_l^1 + \tau_l^2) \geq 0.$$

Moreover, from (2.21) and (2.24) we obtain

$$\alpha_k(\tau_4 + \tau_1 + \tau_2) + \sum_{i=1}^n \mu_i p_{ik}(\tau_i^4 + \tau_i^1 + \tau_i^2) - \mu_k(\tau_k^4 + \tau_k^1 + \tau_k^2) = 0, \;\; k \in \sigma_1,$$

which, in matrix form, becomes

$$\alpha(\tau_4 + \tau_1 + \tau_2) + [P' - I]M(\tau^4 + \tau^1 + \tau^2) \geq 0.$$

Multiplying both sides by $CM^{-1}[I - P']^{-1}$ we obtain:

$$\begin{pmatrix} \rho_1 - 1 \\ \rho_2 - 1 \end{pmatrix} (\tau_4 + \tau_1 + \tau_2) + \begin{pmatrix} \tau_4 + \tau_1 + \tau_2 - \sum_{k \in \sigma_1}(\tau_k^4 + \tau_k^1 + \tau_k^2) \\ \tau_4 + \tau_1 + \tau_2 - (\tau_l^4 + \tau_l^1 + \tau_l^2) \end{pmatrix} \geq 0$$

From (2.14), (2.15), and (2.17)we obtain

$$\tau_4 + \tau_1 + \tau_2 = \sum_{k \in \sigma_1} (\tau_k^4 + \tau_k^1 + \tau_k^2),$$

and since $\tau_4 + \tau_1 + \tau_2 > 0$, then $\rho_1 = 1$, a contradiction. $\square$

2.5.2 Feedforward (acyclic) networks

Definition 2.4 *A queueing network* (α, μ, P, C) *with stations* $\sigma_1, \ldots, \sigma_J$ *is called a feedforward network, if for any two stations* σ_i, σ_j *with* $i < j$ *and for any* $k \in \sigma_i$, $l \in \sigma_j$ *we have* $p_{lk} = 0$. *Namely, higher order stations do not have flows going into lower order stations (Figure 2-2).*

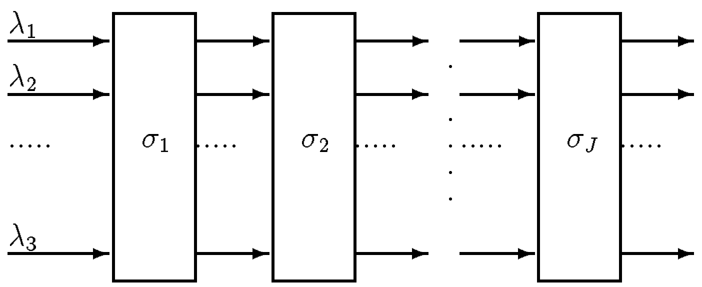

Figure 2-2: Multiclass Feedforward Network

Theorem 2.9 *Let* (α, μ, P, C) *be a multiclass feedforward fluid network. Under the load condition (2.6) the associated linear program GLP[0] has zero as the only feasible solution and, as a result, the network is stable for all initial conditions* $Q(0)$.

Remark. This is a known result ([14],[22]). However, we not only prove the result, but show that our linear program has the power to detect the stability.

Proof:

Suppose, for the purposes of contradiction, there exists a non-zero solution $\tau_\Lambda, \tau_k^\Lambda, \tau, \tau_k,$ to GLP[0].

Let

$$l = \min\{j : \exists\, \Lambda, \text{ with } \tau_\Lambda > 0, \text{ such that } \sigma_j \in \Lambda\}$$

Introduce

$$\tau_0 = \sum_{\Lambda:\sigma_l \in \Lambda, \tau_\Lambda > 0} \tau_\Lambda$$

and

$$\tau_k^0 = \sum_{\Lambda:\sigma_l \in \Lambda, \tau^\Lambda > 0} \tau_k^\Lambda$$

for $k = 1, 2, \ldots, n$. From the constraint (2.42) of LP[0]:

$$\tau_0 = \sum_{k \in \sigma_l} \tau_k^0 \tag{2.48}$$

For all $j < l$, and all Λ with $\tau_\Lambda > 0$ we must have $\sigma_j \notin \Lambda$. This implies from (2.44)

$$\forall\, j < l,\ \forall\, k \in \sigma_j : \qquad \alpha_k \tau_0 + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i^0 - \mu_k \tau_k^0 = 0 \tag{2.49}$$

From (2.45) and (2.44) we have

$$\forall\, k \in \sigma_l : \qquad \alpha_k \tau_0 + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i^0 - \mu_k \tau_k^0 \geq 0 \tag{2.50}$$

Combine (2.49) and (2.50) as

$$\forall\, k \in \sigma_1 \cup \sigma_2 \cup \ldots \cup \sigma_l : \ \ \alpha_k \tau_0 + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i^0 - \mu_k \tau_k^0 \geq 0 \tag{2.51}$$

Let W be the collection of all the classes in station $\sigma_1, \ldots, \sigma_l$:

$$W = \bigcup_{j=1}^{l} \{\sigma_j\}$$

Without loss of generality, we may assume $W = \{1, 2, \ldots, r\}$ where

$$r = |\sigma_1| + \ldots + |\sigma_l|$$

Notice, that for all $k > r, i \leq r$ we have $p_{ki} = 0$ from the definition of a feedforward network. So we may rewrite (2.51) as

$$\forall\, k = 1, 2, \ldots, r : \ \ \alpha_k \tau_0 + \sum_{i=1}^{r} \mu_i p_{ik} \tau_i^0 - \mu_k \tau_k^0 \geq 0 \tag{2.52}$$

For any n-dimensional vector $x = (x_1, \ldots, x_n)$ let x_W be $|W|$-dimensional portion of the vector $x : x_W = (x_i)_{i \in W}$. For any $n \times n$ matrix A, let A_W be the $|W| \times |W|$ dimensional portion of matrix A with elements $(a_{ij})_{i,j \in W}$. We write (2.52) as

$$\alpha_W \tau_0 + [P'_W - I_W] M_W \tau_W^0 \geq 0 \tag{2.53}$$

Notice, that the lower $(n - r) \times r$ part of the matrix P is zero :

$$
P = \begin{pmatrix}
p_{11} & p_{12} & \cdots & p_{1r} & p_{1r+1} & \cdots & p_{1n} \\
p_{21} & p_{22} & \cdots & p_{1r} & p_{2r+1} & \cdots & p_{2n} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
p_{r1} & p_{r2} & \cdots & p_{rr} & p_{rr+1} & \cdots & p_{rn} \\
0 & 0 & \cdots & 0 & p_{r+1r+1} & \cdots & p_{r+1n} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 0 & p_{nr+1} & \cdots & p_{nn}
\end{pmatrix}
$$

$$
P_W = \begin{pmatrix}
p_{11} & p_{12} & \cdots & p_{1r} \\
p_{21} & p_{22} & \cdots & p_{1r} \\
\vdots & \vdots & \ddots & \vdots \\
p_{r1} & p_{r2} & \cdots & p_{rr}
\end{pmatrix}
$$

Then the upper $(n-r) \times r$ part of $I - P'$ is the zero matrix. But then, the matrix $I_W - P'_W$ is invertible and its inverse is the W portion of the total inverse matrix $[I_W - P'_W]^{-1}$:

$$[I_W - P'_W]^{-1} = ([I - P']^{-1})_W$$

In particular, we obtain that it is non-negative. Let us multiply both sides of (2.53) by $M_W^{-1}([I - P']^{-1})_W$.

$$M_W^{-1}([I - P']^{-1})_W \alpha_W \tau_0 - \tau_W^0 \geq 0$$

But

$$M_W^{-1}([I - P']^{-1})_W \alpha_W = \rho_W$$

So

$$\rho_W \tau_0 - \tau_W^0 \geq 0 \tag{2.54}$$

Let us add, in this vector inequality, all the components from the set $\{\sigma_l\}$. We obtain

$$\rho_{\sigma_l} \tau_0 - \sum_{k \in \sigma_l} \tau_k^0 \geq 0$$

But from (2.48) this implies

$$\rho_{\sigma_l} - 1 \geq 0$$

-contradiction. This proves that LP[0] cannot have a non-zero solution and the feed-forward network is stable. $\square$

Chapter 3

Stability of Priority Policies in Multiclass Queueing Networks

We now turn to a linear programming approach for establishing sufficient conditions for stability of priority policies. For a given multiclass fluid network (α, μ, P, C) with some initial condition $Q(0)$ and given priority policy θ we will build a linear program $LP_\theta[Q(0)]$, which gives upper and lower bounds on the time it takes for the system to empty under the priority rule θ. The boundedness of $LP_\theta[Q(0)]$ for a given network would then imply stability.

Some of the existing results on the stability of fluid networks under given priority policies are based on the Lyapunov function method. The advantage of our method is that it is more intuitive, in the sense that it deals with both stability and performance. Not only it establishes stability without a Lyapunov function argument, but it also gives estimates for the time it takes to empty the network.

3.1 Background: Priority Policies in Fluid Models

We now describe priority policies in fluid networks. As usual, consider a multiclass fluid network (α, μ, P, C) with n classes and J stations $\sigma_1, \ldots, \sigma_J$. Let θ be some specified priority order. Namely, θ is a one-to-one function on the set $\{1, 2, \ldots, n\}$. For two classes $i, k \in \sigma_j$ we will say class i has priority over class k if $\theta(i) < \theta(k)$.

Let $T(t)$ and $Q(t)$ be a feasible work conserving policy and the corresponding fluid process in the (α, μ, P, C) network. Namely, $T(t), Q(t)$ satisfy (2.1)-(2.4). We will say that $T(t)$ satisfies the θ priority constraint, if for any station σ any $k \in \sigma$ and any $t_2 > t_1 \geq 0$,

$$Q_k(t) > 0 \ \forall \ t \in (t_1, t_2) \text{ implies } \sum_{i \in \sigma : \theta(i) > \theta(k)} (T_i(t_2) - T_i(t_1)) = 0. \tag{3.1}$$

In other words, whenever there is some positive amount of fluid present in a high priority class, then lower priority classes are not processed. We may rewrite (3.1) equivalently as

$$\int_0^\infty Q_k(t) d\Big\{ \sum_{i \in \sigma(k), \, \theta(i) > \theta(k)} T_i(t) \Big\} = 0. \tag{3.2}$$

In general, the priority constraint (3.1) does not uniquely determine the fluid trajectory $Q(t)$, as opposed to stochastic networks, where the priority policy probabilistically uniquely specifies the evolution of the process. However, Dai [14] proved the following theorem

Theorem 3.1 *If every fluid process $(T(t), Q(t))$ satisfying the θ priority constraint is stable in the sense of Definition 2.1, then the θ priority policy is stable in a corresponding stochastic network.*

This theorem motivates searching for stability conditions of fluid solutions satisfying the priority constraint (3.1) for some priority rule θ.

3.2 Sufficient Conditions for Stability of Priority Policies in Multiclass Fluid Networks

3.2.1 Trajectory Decomposition

We will proceed similarly as in Chapter 2. We will decompose the entire trajectory into finite parts. But the decomposition will be class-wise and not station-wise as in Chapter 2.

To be precise, let Λ be any non-empty subset of the set of classes, $\Lambda \subset \{1, 2, \ldots, n\}$, such that for any station σ_j, $j = 1, 2, \ldots, J$, there holds

$$|\Lambda \ \cap \sigma_j| \leq 1,$$

namely, at most one class per station belongs to set Λ. For a given set Λ and station σ_j any such class is denoted by

$$f = f(\Lambda, \sigma_j). \tag{3.3}$$

If $\Lambda \cap \sigma_j = \emptyset$, we set $f(\Lambda, \sigma_j) = 0$. The collection of all such Λ is denoted by Ω. Clearly

$$|\Omega| \leq \prod_{j=1}^{J}(|\sigma_j| + 1).$$

For any $\Lambda \in \Omega$, denote by R_Λ the following set:

$$R_\Lambda = \{x \in R_+^n : x_k > 0, \ \forall k \in \Lambda\} \cap$$

$$\{x_j = 0, \ \forall j, \text{ such that } \theta(j) < \theta(f(\Lambda, \sigma(j)) \text{ or } \Lambda \cap \sigma(j) = \emptyset\}. \tag{3.4}$$

In words, when $x \in R_\Lambda$, class k belongs to Λ if and only if the fluid content in class k is positive ($x_k > 0$) and all the classes in $\sigma(k)$ with priority higher than k have an empty fluid content.

Consider any feasible work conserving fluid process $(T(t), Q(t))$ satisfying the priority constraint (3.1) for some priority rule θ. Let $\tau_0 = \inf\{t \geq 0 : Q(t) = 0\}$. Here, we define $\tau_0 = \infty$ if the underlying process is unstable. Let τ be any positive finite value less than or equal to τ_0.

We divide the interval $[0, \tau]$ into a finite or countably infinite number of intervals $[t_0, t_1], [t_2, t_3], \ldots, [t_m, t_{m+1}], \ldots$ (with disjoint interiors), such that within each interval (t_m, t_{m+1}), $m = 0, 2, 4, \ldots$, the process stays within a single part R_Λ of the state space. Namely, for each $m = 0, 2, 4, \ldots$, there exist a non-empty $\Lambda_m \in \Omega$ such that for all $t \in (t_m, t_{m+1})$, $Q(t) \in R_{\Lambda_m}$.

Consider now any of these intervals $[t_m, t_{m+1}]$ and the associated Λ_m. Suppose, some station σ satisfies

$$\sigma \cap \Lambda_m \neq \emptyset.$$

Since by definition $Q(t) \in R_{\Lambda_m}$ for $t \in (t_m, t_{m+1})$, then by the work-conservation constraint (2.4):

$$\sum_{k \in \sigma} (T_k(t_{m+1}) - T_k(t_m)) = t_{m+1} - t_m. \tag{3.5}$$

Also, by the feasibility constraint (2.3), for an arbitrary station σ' :

$$\sum_{k \in \sigma'} (T_k(t_{m+1}) - T_k(t_m)) \leq t_{m+1} - t_m. \tag{3.6}$$

These constraints hold for all work conserving service disciplines. The following constraint, however, is special to priority disciplines only. Take an arbitrary $\Lambda \in \Omega$ and any $\sigma \in \{\sigma_1, \sigma_2, \ldots, \sigma_J\}$ satisfying

$$\sigma \cap \Lambda_m \neq \emptyset.$$

Since by Definition (3.3)

$$\forall\, t \in (t_m, t_{m+1}),\ Q_{f(\Lambda_m, \sigma)}(t) > 0,$$

then by the constraint (3.1), for every $k \in \sigma$ satisfying $\theta(k) > \theta(f(\Lambda_m, \sigma))$ we must have:

$$T_k(t_{m+1}) - T_k(t_m) = 0. \tag{3.7}$$

In words, since $f(\Lambda_m, \sigma)$ is the highest priority *busy* class in station σ during the interval (t_m, t_{m+1}), then no classes of lower priority than $f(\Lambda_m, \sigma)$ are worked on.

Suppose, now, some $k = 1, 2, \ldots, n$ satisfies $\theta(k) < \theta(f(\Lambda_m, \sigma(k))$ or satisfies $\Lambda_m \cap \sigma(k) = \emptyset$. From (3.4), this means $Q_l(t_{m+1}) = Q_l(t_m) = 0$. Therefore:

$$Q_k(t_{m+1}) - Q_l(t_m) = \alpha_k(t_{m+1} - t_m) +$$

$$\sum_{i=1}^{n} \mu_i p_{ik}(T_i(t_{m+1}) - T_i(t_m)) - \mu_k(T_k(t_{m+1}) - T_k(t_m)) = 0. \tag{3.8}$$

In words, all classes k with priority higher than $f(\Lambda_m, \sigma(k))$ have zero fluid level during the interval $[t_m, t_{m+1}]$. Also, all classes k with $\sigma(k) \cap \Lambda_m = \emptyset$ have zero fluid level during the interval $[t_m, t_{m+1}]$.

3.2.2 LP Formulation

We will use now (3.5)-(3.8) to build an associated linear program. Let $\Lambda \in \Omega$. We introduce

$$\tau_\Lambda = \sum_{m:\Lambda_m=\Lambda} (t_{m+1} - t_m),$$

and

$$\tau_k^\Lambda = \sum_{m:\Lambda_m=\Lambda} (T_k(t_{m+1}) - T_k(t_m)),$$

for $k = 1, 2, \ldots, n$. Since we chose τ to be finite, τ_Λ and τ_k^Λ are well defined. Trivially then

$$\tau = \sum_{\Lambda \in \Omega} \tau_\Lambda.$$

By summing Equation (3.5) over all m such that $\Lambda_m = \Lambda$ we obtain:

$$\forall\, \Lambda \in \Omega,\ \forall \sigma, \text{ such that } \sigma \cap \Lambda \neq \emptyset : \ \sum_{k \in \sigma} \tau_k^\Lambda = \tau_\Lambda. \tag{3.9}$$

Similarly from (3.6) we obtain:

$$\forall\, \Lambda \in \Omega,\ \forall \sigma, \text{ such that } \sigma \cap \Lambda = \emptyset : \ \sum_{k \in \sigma} \tau_k^\Lambda \leq \tau_\Lambda. \tag{3.10}$$

From (3.7) we have:

$$\forall\, \Lambda \in \Omega,\ \forall \sigma, \text{ such that } \sigma \cap \Lambda \neq \emptyset \text{ and } \forall\, k \in \sigma \text{ such that } \theta(k) > \theta(f(\Lambda, \sigma)) :$$

$$\tau_k^\Lambda = 0. \tag{3.11}$$

By summing constraint (3.8) over all m such that $\Lambda_m = \Lambda$ we obtain for all $\Lambda \in \Omega$ and for all $k = 1, 2, \ldots, n$, such that $\theta(k) < \theta(f(\Lambda, \sigma(k))$ or such that $\Lambda \cap \sigma(k) = \emptyset$:

$$\alpha_k \tau_\Lambda + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i^\Lambda - \mu_k \tau_k^\Lambda = 0. \tag{3.12}$$

We have from the dynamics of the system:

$$\forall\, k = 1, 2, \ldots, n: \;\; Q_k(\tau) = Q_k(0) + \alpha_k \tau + \sum_{i=1}^{n} \mu_i p_{ik} T_i(\tau) - \mu_k T_k(\tau) \geq 0.$$

We introduce

$$\tau_k = \sum_{\Lambda \in \Omega} \tau_k^\Lambda.$$

Clearly, $T_k(\tau) = \tau_k$ for all $k = 1, 2, \ldots, n$. Then from the previous inequality

$$\forall\, k = 1, 2, \ldots, n: \;\; Q_k(\tau) = Q_k(0) + \alpha_k \tau + \sum_{i=1}^{n} \mu_i p_{ik} \tau_i - \mu_k \tau_k \geq 0. \tag{3.13}$$

Notice, that (3.9)-(3.13) specify some linear program in non-negative variables $\tau, \tau_k, \tau_\Lambda, \tau_k^\Lambda$ with parameters $Q_k(0), \alpha_k, \mu_k, p_{ik}$. The value of τ is upper bounded by the maximum value of τ subject to the constraints (3.9)-(3.13). Since τ was an arbitrary finite number less than or equal to τ_0 we conclude that the value of τ_0 is bounded by the maximum value of τ subject to the constraints (3.9)-(3.13). In particular, if the linear program is has finite objective value, then $\tau_0 < +\infty$ and the system is stable. The following theorem then holds

Theorem 3.2 *Let (α, μ, P, C) be a multiclass fluid network with some initial state $Q(0)$, and let θ be some priority rule. If the linear program $LP_\theta[Q(0)]$:*

$$Max\ \tau\ subject\ to\ (3.9) - (3.13),$$

and $\tau, \tau_k, \tau_\Lambda, \tau_k^\Lambda \geq 0$ has a finite objective value, then the priority policy θ is stable and the optimal value of $LP_\theta[Q(0)]$ provides an upper bound on the time it takes for the network to empty under θ.

Also, if the linear program $LP_\theta[Q(0)]$ has a finite objective value, then linear program

$$Min\,\tau, \quad subject\,to\;(3.9)-(3.12)$$

and

$$\forall\, k = 1,2,\ldots,n: \quad Q_k(\tau) = Q_k(0) + \alpha_k\tau + \sum_{i=1}^{n} \mu_i p_{ik}\tau_i - \mu_k\tau_k = 0 \tag{3.14}$$

gives a lower bound on the time it takes for the network to empty under the priority rule θ.

Remark. Note that the constraint (3.13) is an inequality constraint as opposed to the analogous constraint (2.24) in Chapter 2, where we considered a stable general policy with a very long busy period. We cannot use the same argument here since we are restricted to priority policies only.

Proof:

We have proved that boundedness of $LP_\theta[Q(0)]$ implies stability. To prove the second part, observe, that if $LP_\theta[Q(0)]$ is bounded, then from the stability of the network under the priority rule θ, it follows that $\tau_0 < +\infty$. Then by setting $\tau = \tau_0$ we obtain that constraint (3.13) is satisfied with equality and we may replace it with (3.14). □

Notice, that $Q(0)$ is the only non-zero part in the right hand side of the linear program. The boundedness of a linear program does not depend on the right hand side. So let us replace $Q_k(0)$ by 0 in constraint (3.13) and obtain

$$\forall\, k = 1,2,\ldots,n: \quad \alpha_k\tau + \sum_{i=1}^{n} \mu_i p_{ik}\tau_i - \mu_k\tau_k \geq 0. \tag{3.15}$$

The following version of Theorem 3.2 then holds :

Theorem 3.3 *Let (α,μ,P,C) be a multiclass fluid network and let θ be some priority* *rule. If the linear program $LP_\theta[0]$:*

$$Max\ \tau\ subject\ to\ (3.9)-(3.12),(3.15),$$

and $\tau, \tau_k, \tau_\Lambda, \tau_k^\Lambda \geq 0$ has zero as the only feasible solution, then the priority policy θ is stable for all initial states $Q(0)$.

Remark. Theorem 3.3 states that our stability criterion is in fact insensitive to the initial condition Q(0). Also observe that the number of variables and the number of constraints in LP[0] are both of the order $O(n|\Omega|) = O(n\prod_{j=1}^{J}(|\sigma_j|+1))$. Thus our stability checking algorithm is exponential in the number of stations.

3.3 Examples

In this section, we apply Theorem 3.2 to address the stability of particular priority policies.

3.3.1 Last-Buffer-First-Serve Priority Rule in Re-entrant Line Queueing Networks

Recall that a re-entrant line is a queueing network in which there is only a single external arrival stream with rate λ, and flows from each class $k = 1, 2, \ldots, n-1$ go entirely to the next class $k+1$, with flow from class n leaving the system.

Recall also that for each class k and each station σ in a re-entrant line, we have defined

$$\rho_k = \lambda/\mu_k$$

and

$$\rho_\sigma = \sum_{i\in\sigma}\rho_i$$

The Last-Buffer-First-Serve (LBFS) priority rule θ_{LBFS} is the priority policy θ for which, given $i, k \in \sigma$,

$$\theta(i) < \theta(k) \text{ iff } i > k.$$

Namely, higher order classes have higher priority.

Theorem 3.4 *Let* θ_{LBFS} *be the LBFS priority rule in a re-entrant line network* (α, μ, P, C). *Then, the corresponding linear program* $LP_{\theta_{LBFS}}[0]$ *has zero as the only feasible solution. As a result, the LBFS policy is stable.*

Remark. The stability of LBFS policy was proven in [37] and [19]. However, our theorem not only proves the stability of the LBFS policy, but shows that stability is captured by the linear programming sufficient conditions of the Theorem 3.3.

Proof:

Suppose there exists a non-zero solution $\tau, \tau_k, \tau_\Lambda, \tau_k^\Lambda$ to $LP_{\theta_{LBFS}}[0]$. Consider an arbitrary $\Lambda \in \Omega$, such that $\tau_\Lambda > 0$. Let

$$k = \max\{f(\Lambda, \sigma) : \sigma \cap \Lambda \neq \emptyset\}.$$

Namely, among the highest priority busy classes in all the stations, k is the largest (in order). Since for all $j > k$ we have $j \notin \Lambda$ then from (3.12) and since $\theta(j) < \theta(k)$:

$$\forall\, j > k : \quad \alpha_j \tau_\Lambda + \sum_{i=1}^{n} \mu_i p_{ij} \tau_i^\Lambda - \mu_j \tau_j^\Lambda =$$

$$= \mu_{j-1} \tau_{j-1}^\Lambda - \mu_j \tau_j^\Lambda = 0. \tag{3.16}$$

From (3.11):

$$\forall\, i \in \sigma(k),\, i < k : \; \tau_i^\Lambda = 0.$$

So from (3.9)

$$\sum_{i \in \sigma(k), i \geq k} \tau_i^\Lambda = \tau_\Lambda. \tag{3.17}$$

But from (3.16) for each $i > k$:

$$\tau_i^\Lambda = \frac{\mu_k}{\mu_i}\tau_k^\Lambda.$$

Combining with (3.17):

$$\mu_k \tau_k^\Lambda = \frac{\tau_\Lambda}{\sum_{i\in\sigma(k),\, i\geq k} 1/\mu_i}.$$

So from (3.16)

$$\mu_n \tau_n^\Lambda = \mu_k \tau_k^\Lambda = \frac{\tau_\Lambda}{\sum_{i\in\sigma(k),\, i\geq k} 1/\mu_i}.$$

But

$$1 > \rho_{\sigma(k)} = \lambda \sum_{i\in\sigma(k)} 1/\mu_i \geq \lambda \sum_{i\in\sigma(k),\, i\geq k} 1/\mu_i.$$

Therefore, since $\tau_\Lambda > 0$, then

$$\lambda\tau_\Lambda - \mu_n \tau_n^\Lambda = \lambda\tau_\Lambda - \frac{\tau_\Lambda}{\sum_{i\in\sigma(k),\, i\geq k} 1/\mu_i} < 0. \tag{3.18}$$

Recall, that this holds for for any $\Lambda \in \Omega$ satisfying $\tau_\Lambda > 0$.

Notice that from the re-entrant line structure, we have

$$\sum_{k=1}^{n} \alpha_k \tau_\Lambda + \sum_{k=1}^{n}\sum_{i=1}^{n} \mu_i p_{ik} \tau_i^\Lambda - \mu_k \tau_k^\Lambda = \lambda\tau_\Lambda - \mu_n \tau_n^\Lambda.$$

From (3.15):

$$\sum_{\Lambda\subset\Omega} \lambda\tau_\Lambda - \mu_n \tau_n^\Lambda = \sum_{\Lambda\in\Omega,\tau_\Lambda>0} (\lambda\tau_\Lambda - \mu_n \tau_n^\Lambda) \geq 0.$$

This contradicts (3.18) and completes the proof.□

3.3.2 First-Buffer-First-Serve Priority Rule

The First-Buffer-First-Serve (FBFS) is a priority rule in which lower order classes have higher priority. Namely,

$$\forall\ i,k \in \sigma:\ \theta(i) < \theta(k) \text{ iff } i < k$$

We now prove that FBFS is stable in networks with upper triangular routing probability matrix P (i.e, P satisfying: for all $i > k$, $p_{ik} = 0$).

Theorem 3.5 *Let θ_{FBFS} be a FBFS priority rule in a multiclass fluid network (α, μ, P, C) with upper triangular routing matrix P. Then the linear program $LP_{\theta_{FBFS}}[0]$ has zero as the only feasible solution, and FBFS policy is stable.*

Remark. The stability of FBFS in re-entrant lines was proved in [19]. We prove it for more general case - networks with upper triangular matrix P, by showing that the linear programming test will detect the stability of FBFS in such networks.

Proof:

Suppose there exists a non-zero solution $\tau, \tau_k, \tau_\Lambda, \tau_k^\Lambda$ to $LP_{\theta FBFS}[0]$. For any $\Lambda \in \Omega$ let

$$L(\Lambda) = \min\{j : j \in \Lambda\}.$$

By assumption, there exists $\Lambda \in \Omega$ such that $\tau_\Lambda > 0$.

Let

$$l = \min\{L(\Lambda) : \tau_\Lambda > 0\}. \tag{3.19}$$

If for some $\Lambda \in \Omega$ with $\tau_\Lambda > 0$, there holds $l \notin \Lambda$, then from (3.19) $l < i$ for all $i \in \Lambda$. So, from (3.12), there holds

$$\alpha_l \tau_\Lambda + \sum_{i=1}^{n} \mu_i p_{il} \tau_i^\Lambda - \mu_l \tau_l^\Lambda = 0.$$

Then, from (3.15),

$$\sum_{\Lambda : l \in \Lambda, \tau_\Lambda > 0} \alpha_l \tau_\Lambda + \sum_{\Lambda : l \in \Lambda, \tau_\Lambda > 0} \sum_{i=1}^{n} \mu_i p_{il} \tau_i^\Lambda - \sum_{\Lambda : l \in \Lambda, \tau_\Lambda > 0} \mu_l \tau_l^\Lambda \geq 0$$

So, there exists $\Lambda_0 \in \Omega$ for which $l \in \Lambda_0$, $\tau_{\Lambda_0} > 0$ and

$$\alpha_l \tau_{\Lambda_0} + \sum_{i=1}^{n} \mu_i p_{il} \tau_i^{\Lambda_0} - \mu_l \tau_l^{\Lambda_0} \geq 0. \tag{3.20}$$

For all $i > l$, $p_{il} = 0$, so (3.20) is rewritten as

$$\alpha_l \tau_{\Lambda_0} + \sum_{i=1}^{l} \mu_i p_{il} \tau_i^{\Lambda_0} - \mu_l \tau_l^{\Lambda_0} \geq 0. \tag{3.21}$$

From (3.19) for all $k < l$ we must have $k \notin \Lambda_0$. Then from (3.12), since $\theta(k) < \theta(l)$

$$\alpha_k \tau_{\Lambda_0} + \sum_{i=1}^{k} \mu_i p_{ik} \tau_i^{\Lambda_0} - \mu_k \tau_k^{\Lambda_0} = \alpha_k \tau_{\Lambda_0} + \sum_{i=1}^{l} \mu_i p_{ik} \tau_i^{\Lambda_0} - \mu_k \tau_k^{\Lambda_0} = 0. \tag{3.22}$$

Let $W = \{1, 2, \ldots, l\}$. For any n-dimensional vector $x = (x_1, \ldots, x_n)$ let x_W be the $|W|$ -dimensional portion of the vector $x : x_W = (x_i)_{i \in W}$. For an $n \times n$ matrix A, let A_W be the $|W| \times |W|$ dimensional portion of the matrix A with elements $(a_{ij})_{i,j \in W}$.

We may then combine (3.21) and (3.22) in matrix form as

$$\alpha_W \tau_{\Lambda_0} + [I_W - P'_W] M_W \tau_W^{\Lambda_0} \geq 0, \tag{3.23}$$

where, $\tau_W^{\Lambda_0} = (\tau_1^{\Lambda_0}, \tau_2^{\Lambda_0}, \ldots, \tau_l^{\Lambda_0})$.

Notice, that the lower left $(n - l) \times l$ submatrix of the matrix P is zero, since it lies completely below the diagonal:

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1l} & p_{1l+1} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{1l} & p_{2l+1} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ p_{l1} & p_{l2} & \cdots & p_{ll} & p_{ll+1} & \cdots & p_{ln} \\ 0 & 0 & \cdots & 0 & p_{l+1l+1} & \cdots & p_{l+1n} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & p_{nl+1} & \cdots & p_{nn} \end{pmatrix}.$$

Notice also that

$$P_W = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1l} \\ p_{21} & p_{22} & \cdots & p_{1l} \\ \vdots & \vdots & \ddots & \vdots \\ p_{l1} & p_{l2} & \cdots & p_{ll} \end{pmatrix},$$

(of course, $p_{21}, p_{31}, p_{32}, \ldots, p_{l1}, \ldots, p_{ll-1}$ are all zeros as well).

Then the upper $l \times (n-l)$ submatrix of $I - P'$ is the zero matrix. But then, the matrix $I_W - P'_W$ is invertible and its inverse is the W portion of the total inverse matrix $[I_W - P'_W]^{-1}$:

$$[I_W - P'_W]^{-1} = [I - P']^{-1}_W$$

In particular, it is non-negative.

Let us multiply both sides of (3.23) by $M_W^{-1}[I_W - P'_W]^{-1}$ to obtain

$$M_W^{-1}[I - P']^{-1}_W \alpha_W \tau_{\Lambda_0} - \tau_W^{\Lambda_0} \geq 0.$$

Since $[I - P']^{-1}$ is a nonnegative matrix, then

$$[I - P']^{-1}_W \alpha_W \leq ([I - P']^{-1}\alpha)_W,$$

and

$$M_W^{-1}[I - P']^{-1}_W \alpha_W \leq (M^{-1}[I - P']^{-1}\alpha)_W = \rho_W.$$

It follows that

$$\rho_W \tau_{\Lambda_0} - \tau_W^{\Lambda_0} \geq 0.$$

By summing over all $i \in \sigma(l)$, $i \leq l$

$$\sum_{i \in \sigma(l), i \leq l} \rho_i \tau_{\Lambda_0} - \sum_{i \in \sigma(l), i \leq l} \tau_i^{\Lambda_0} \geq 0. \tag{3.24}$$

We have from (3.19) $l = f(\Lambda_0), \sigma(l)$ (with $f(\Lambda, \sigma)$ defined by (3.3)). Then, from (3.11), for all $i \in \sigma(l)$ such that $i > l$ we have $\tau_i^{\Lambda_0} = 0$. Also, from the work-

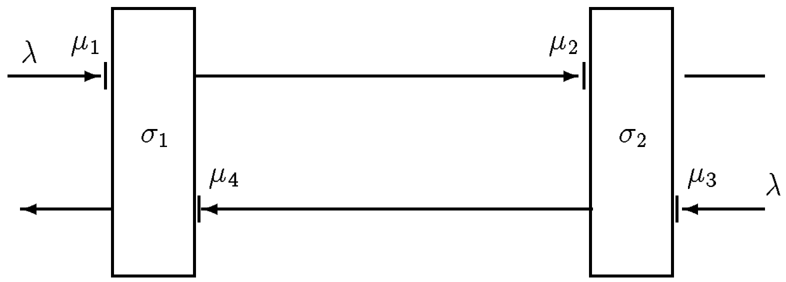

Figure 3-1: Rybko-Stolyar network

conservation constraint (3.9)

$$\sum_{i\in\sigma(l)} \tau_i^{\Lambda_0} = \tau_{\Lambda_0}.$$

Combining,

$$\sum_{i\in\sigma(l),i\le l} \tau_i^{\Lambda_0} = \tau_{\Lambda_0} > 0.$$

But then, from (3.24) we obtain:

$$\rho_{\sigma(l)} \ge \sum_{i\in\sigma(l),i\le l} \rho_i \ge 1$$

- contradiction. Therefore, our linear program cannot have a non-zero solution and the FBFS policy is stable. □

3.3.3 The Rybko-Stolyar Network

The Rybko-Stolyar network is the queueing network described on Figure 3-1

It is known ([19]), that all work conserving policies are stable in this network, if the load condition (2.6) holds and

$$\lambda/\mu_2 + \lambda/\mu_4 < 1. \tag{3.25}$$

This can also be established using the linear program LP[0] developed in the previous chapter for networks with two stations. Also, it is known that if

$$\theta(4) < \theta(1), \theta(2) < \theta(3), \tag{3.26}$$

and (3.25) is violated, then the priority policy is unstable.

In this example we will show, using the previous result, that all priority policies other than (3.26) are stable just under load condition (2.6). Notice that there are only three priority policies $\theta_1, \theta_2, \theta_3$ other than the priority policy given by (3.26). These are:

$$\theta_1(1) < \theta_1(4),\ \theta_1(2) < \theta_1(3),$$

$$\theta_2(1) < \theta_2(4),\ \theta_2(3) < \theta_2(2),$$

$$\theta_3(4) < \theta_2(1),\ \theta_3(3) < \theta_3(2).$$

Observe, that the policy θ_1 is the FBFS priority policy and the routing matrix is upper triangular:

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Therefore, by Theorem 3.5, policy θ_1 is stable. For the policies θ_2 and θ_3, we will use a different arrangement of the classes.

Let C_1 be class 1, C_2 be class 3, C_3 be class 2 and C_4 be class 4. Then, we have $\theta_2(C_1) < \theta_2(C_2) < \theta_2(C_3) < \theta_2(C_4)$, and the priority policy θ_2 is FBFS. The routing matrix for classes C_1, C_2, C_3, C_4 is

$$P = \begin{pmatrix} p_{C_1C_1} & p_{C_1C_2} & p_{C_1C_3} & p_{C_1C_4} \\ p_{C_2C_1} & p_{C_2C_2} & p_{C_2C_3} & p_{C_2C_4} \\ p_{C_3C_1} & p_{C_3C_2} & p_{C_3C_3} & p_{C_3C_4} \\ p_{C_4C_1} & p_{C_4C_2} & p_{C_4C_3} & p_{C_4C_4} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

which is upper triangular. In conclusion, the priority policy θ_2 is stable.

For the policy θ_3, let $C_1 = 3, C_2 = 4, C_3 = 1, C_4 = 2$. Again, for this ordering, the priority policy θ_3 is FBFS and the routing matrix

$$P = \begin{pmatrix} p_{C_1C_1} & p_{C_1C_2} & p_{C_1C_3} & p_{C_1C_4} \\ p_{C_2C_1} & p_{C_2C_2} & p_{C_2C_3} & p_{C_2C_4} \\ p_{C_3C_1} & p_{C_3C_2} & p_{C_3C_3} & p_{C_3C_4} \\ p_{C_4C_1} & p_{C_4C_2} & p_{C_4C_3} & p_{C_4C_4} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

is upper triangular. Thus, the priority policy θ_3 is stable.

Chapter 4

Bounds on Stationary Probability Distributions of Infinite State Markov Chains using Lyapunov and lower Lyapunov Functions

In this chapter we develop tools for the performance analysis of exponential type multiclass queueing networks. Our main object is positive recurrent infinite state Markov chains and specifically random walks in the nonnegative orthant R^n_+.

Numerous researchers have studied random walks on the nonnegative orthant. Most of the research focused on conditions for positive recurrence of random walks (see Meyn and Tweedie [48], Fayolle [23], Fayolle, Malyshev, Menshikov and Sidorenko [24], Malyshev [43], Malyshev and Menshikov[44], Menshikov [45]). Necessary and sufficient conditions for the positive recurrence of random walks in Z^2_+ and Z^3_+ with bounded steps or with finite second moments of the steps are obtained in [43], [45] and [23]. In particular, quadratic Lyapunov functions are used in [23] to obtain exact conditions for positive recurrence, for the cases Z^2_+, Z^3_+. Sufficient conditions are also obtained for the case $Z^n_+, n \geq 4$.

Nevertheless, none of the aforementioned studies have addressed the performance problem - estimation of the stationary probability distributions of such random walks. In this chapter we provide a general methodology of using Lyapunov functions for the performance analysis of infinite state Markov chains. We will show that whenever an infinite state Markov chain has a Lyapunov function satisfying a particular drift condition, certain geometric type bounds hold on the stationary distribution of the Markov chain. We then apply these results to random walks in the nonnegative orthant R_+^n.

In [28], Hajek constructed bounds on stationary and transient probability distribution for systems more general than Markov chains. In particular, supermartingales with a constant drift, subexponential steps and a continuous state space are considered. Using the results from this paper one can obtain geometric upper bounds on the stationary tail probabilities for Markov chains with bounded steps. The limitation of the approach is that it leads only to upper bounds and that even for a simple random walk in Z_+ defined by

$$\Pr\Big\{Q(t+1) = m+1 | Q(t) = m\Big\} = p, \qquad m = 0, 1, 2, \dots,$$

$$\Pr\Big\{Q(t+1) = m-1 | Q(t) = m\Big\} = 1-p, \qquad m = 1, 2, \dots,$$

and

$$\Pr\Big\{Q(t+1) = 0 | Q(t) = 0\Big\} = 1-p,$$

the corresponding bounds on tail probabilities are greater than a certain positive value, even when the parameter p converges to zero. As we will see this does not occur for the bounds constructed in this chapter.

The structure of this chapter is as follows. In the next section we define a model of interest - Markov chains in R_+^n with a countably infinite state space and bounded step sizes. We will introduce the notions of Lyapunov and lower Lyapunov functions together with key parameters of interest. In Section 4.2, we prove the main results of this chapter. We show that whenever a positive recurrent Markov chain admits a Lyapunov function with bounded values of change during the transitions, certain geometric upper bounds hold on the stationary tail probabilities. Similarly, the existence of a lower Lyapunov function (to be defined) implies geometric lower bounds on the tail probabilities. In Section 4.3 we apply these results to a special case of Markov chains - random walks in the nonnegative orthant R^n_+, and we discuss some possible extension of the results of this chapter.

4.1 Stationary Infinite State Markov Chains

Let $Q(t),\ t = 0, 1, 2, \ldots$, be a discrete time discrete state Markov chain which takes values in some countable set $\mathcal{X} \subset R^n_+$, where again R^n_+ denotes a set of n-dimensional vectors with nonnegative components. The transitions occur at integer times $t = 0, 1, 2, \ldots$. For any two vector $x, x' \in \mathcal{X}$, let $p(x, x')$ denote the transition probabilities

$$p(x, x') \equiv \Pr\Big\{Q(t+1) = x' | Q(t) = x\Big\}.$$

We now introduce the definitions of Lyapunov and lower Lyapunov functions. Lyapunov functions have been used widely for establishing the ergodicity (stability) of certain Markov chains (see for example [48],[38], [23]). However, we will use Lyapunov functions for the performance analysis of Markov chains, assuming a priori that the Markov chain is stable. In particular, upper bounds will be constructed on the stationary distribution of Markov chains. We also introduce here the notion of a *lower Lyapunov function* exclusively as a means of getting the lower bounds on the stationary distribution of a Markov chain.

Definition 4.1 *A nonnegative function*

$$\Phi : R^n_+ \to R_+$$

is said to be a Lyapunov function if there exist some $\gamma > 0$ and $B \geq 0$, such that for any $t = 1, 2, \ldots$ and any $x \in R^n_+$, with $\Phi(x) > B$ there holds

$$E[\Phi(Q(t+1))|Q(t) = x] \leq \Phi(x) - \gamma.$$

Also a nonnegative function

$$\Phi : R_+^n \to R_+$$

is said to be a lower Lyapunov function if there exists some γ *such that for any* $t = 1, 2, \ldots$ *and any* $x \in R_+^n$ *there holds*

$$E[\Phi(Q(t+1))|Q(t) = x] \geq \Phi(x) - \gamma.$$

Remarks:

1. The terms γ and B will be called drift and exception parameter respectively.

2. The definition of a Lyapunov function is similar to Definition 2.3 where the differential operator is used instead of expectation. Usually, for stability analysis it is required that the set $\Big\{x : \Phi(x) \leq B\Big\}$ is bounded. We do not require it here, (although in all of the cases considered it will be) and assume a priori that a stationary distribution exists. So in principle, our Lyapunov function Φ may not be a witness of stability.

3. The definition of the lower Lyapunov function could also include the exception parameter. But for the purposes of the future analysis, the given (more restrictive) definition suffices.

From now on we will assume that our Markov chain $Q(t)$ is positive recurrent. Denote by π the corresponding unique stationary distribution. Namely, $\pi(x)$ is the steady-state probability $\Pr_\pi\Big\{Q(t) = x\Big\}$ that the chain is in a certain state $x \in \mathcal{X}$. Also by $E_\pi[\cdot]$ we denote the expectation with respect to the probability distribution π.

For any function $\Phi : \mathcal{X} \to R_+$ let

$$\nu_{max}^{\Phi} \equiv \sup_{x,x' \in \mathcal{X}: p(x,x') > 0} |\Phi(x') - \Phi(x)|, \tag{4.1}$$

and

$$\nu_{min}^{\Phi} \equiv \inf_{x,x' \in \mathcal{X}: p(x,x') > 0, \Phi(x) < \Phi(x')} (\Phi(x') - \Phi(x)). \tag{4.2}$$

Namely, ν^{Φ}_{max} is the largest possible change of the function Φ during an arbitrary transition, and ν^{Φ}_{min} is the smallest possible increase of the function Φ. Also let

$$p^{\Phi}_{max} = \sup_{x\in\mathcal{X}} \sum_{x'\in\mathcal{X},\Phi(x)<\Phi(x')} p(x,x') \tag{4.3}$$

and

$$p^{\Phi}_{min} = \inf_{x\in\mathcal{X}} \sum_{x'\in\mathcal{X},\Phi(x)<\Phi(x')} p(x,x') \tag{4.4}$$

Namely, p^{Φ}_{max} and p^{Φ}_{min} are upper and lower bounds on the probability that the value of Φ is increasing during an arbitrary transition.

Consider a simple example of an one dimensional random walk with unit step lengths on the set of nonnegative integers Z_+. Set $\mathcal{X} = Z_+$ and

$$\Pr\Big\{Q(t+1) = m+1|Q(t) = m\Big\} = p,$$

$$\Pr\Big\{Q(t+1) = m-1|Q(t) = m\Big\} = 1-p,$$

for $m > 0$ and

$$\Pr\Big\{Q(t+1) = 1|Q(t) = 0\Big\} = 1.$$

If $p < 1/2$, then $\Phi(x) = x$ is both Lyapunov and lower Lyapunov function with drift $\gamma = 1-2p$ and exception parameter $B = 0$. In fact, for any positive integer x

$$E[\Phi(Q(t+1))|Q(t) = x] - \Phi(x) = -(1-2p) < 0.$$

Then $\nu^{\Phi}_{max} = \nu^{\Phi}_{min} = 1$ and $p^{\Phi}_{max} = 1, p^{\Phi}_{min} = p$. Another example is a multidimensional random walk $Q(t)$ on a nonnegative lattice Z^n_+ with unit (coordinate-wise) step lengths. For any $z = (z_1,\ldots,z_n) \in Z^n_+$ set $\Phi(z) = \sum_{i=1}^{n} z_i$. Then Φ is a lower Lyapunov function with $\nu^{\Phi}_{min} = 1$. We will consider this type of Markov chains when we analyze the performance of exponential type queueing networks in Chapter 5.

4.2 Upper and Lower Bounds on the Stationary Probability Distribution

In this section we derive upper and lower bounds on the stationary distributions of Markov chains, which admit Lyapunov and/or lower Lyapunov function. We will show that whenever a Markov chain $Q(t)$ has a Lyapunov function Φ with a finite maximal step size ν^{Φ}_{max}, certain geometric type upper bounds hold on the stationary distribution π. Similarly, whenever a Markov chain $Q(t)$ has a lower Lyapunov function Φ with a positive minimal step size ν^{Φ}_{min} and positive p^{Φ}_{min} certain geometric type lower bounds hold on the stationary distribution π.

We start by proving two key lemmas.

Lemma 4.1 *Given a Markov chain $Q(t)$ with stationary probability distribution π, if Φ is a Lyapunov function with drift γ and exception parameter B, then for any (possibly negative) $c_0 \geq B - \nu^{\Phi}_{max}$ there holds*

$$\mathrm{Pr}_{\pi}\Big\{c_0 + \nu^{\Phi}_{max} < \Phi(Q(t))\Big\} \leq \frac{p^{\Phi}_{max}\nu^{\Phi}_{max}}{p^{\Phi}_{max}\nu^{\Phi}_{max} + \gamma}\mathrm{Pr}_{\pi}\Big\{c_0 - \nu^{\Phi}_{max} < \Phi(Q(t))\Big\},$$

where ν^{Φ}_{max} and p^{Φ}_{max} are defined in (4.1) and (4.3).

Proof:

We introduce the function

$$\hat{\Phi}(c_0, x) = \max\Big\{c_0, \Phi(x)\Big\} \tag{4.5}$$

From stationarity we have the following equilibrium equation

$$E_{\pi}[\hat{\Phi}(c_0, Q(t))] = E_{\pi}[\hat{\Phi}(c_0, Q(t+1))] \tag{4.6}$$

which we rewrite as

$$\sum_x \pi(x)\Big(E[\hat{\Phi}(c_0, Q(t+1))|Q(t) = x] - \hat{\Phi}(c_0, x)\Big) = 0 \tag{4.7}$$

We decompose the left-hand side of the equation above into three parts and obtain

$$0 = \sum_{x:\Phi(x)\leq c_0-\nu^{\Phi}_{max}} \pi(x)\Big(E[\hat{\Phi}(c_0, Q(t+1))|Q(t)=x] - \hat{\Phi}(c_0,x)\Big)+ \tag{4.8}$$

$$\sum_{x:c_0-\nu^{\Phi}_{max}<\Phi(x)\leq c_0+\nu^{\Phi}_{max}} \pi(x)\Big(E[\hat{\Phi}(c_0, Q(t+1))|Q(t)=x] - \hat{\Phi}(c_0,x)\Big)+$$

$$\sum_{x:c_0+\nu^{\Phi}_{max}<\Phi(x)} \pi(x)\Big(E[\hat{\Phi}(c_0, Q(t+1))|Q(t)=x] - \hat{\Phi}(c_0,x)\Big).$$

We will analyze each of the summands above separately. In particular, we will show that the first summand is equal to zero, and the expression in parenthesis of the third summand is upper bounded by $-\gamma$, from the assumption that Φ is a Lyapunov function with drift γ. Thus, the advantage of using modified function $\max\Big\{c_0, \Phi\Big\}$ is that while it allows us to exclude the states $x \in \mathcal{X}$ with low values of $\Phi(x)$ from the equilibrium equation (4.7), it does not affect the states x with high value of $\Phi(x)$.

Now we analyze in detail each of the summands in the equation (4.8)

1. To analyze the first summand in Equation (4.8), observe that for any x, x' with $\Phi(x) \leq c_0 - \nu^{\Phi}_{max} < c_0$ and $p(x,x') > 0$, there holds from the definition of ν^{Φ}_{max}

$$\Phi(x') \leq \Phi(x) + \nu^{\Phi}_{max} \leq c_0 - \nu^{\Phi}_{max} + \nu^{\Phi}_{max} = c_0.$$

   Therefore, $\hat{\Phi}(c_0,x) = \hat{\Phi}(c_0,x') = c_0$. We conclude that for all x with $\Phi(x) \leq c_0 - \nu^{\Phi}_{max}$,

$$E[\hat{\Phi}(c_0, Q(t+1))|Q(t)=x] - \hat{\Phi}(c_0,x) = c_0 - c_0 = 0.$$

2. We now analyze the third summand in Equation (4.8). For any x, x' with $\Phi(x) > c_0 + \nu^{\Phi}_{max}$ and $p(x,x') > 0$, we again obtain from the definition of ν^{Φ}_{max}, that

$$\Phi(x') \geq \Phi(x) - \nu^{\Phi}_{max} > c_0 + \nu^{\Phi}_{max} - \nu^{\Phi}_{max} = c_0.$$

   Therefore, $\hat{\Phi}(c_0,x) = \Phi(x)$ and $\hat{\Phi}(c_0,x') = \Phi(x')$. Also, by assumption, $c_0 +$

$\nu^{\Phi}_{max} \geq B$. We conclude that for all x with $\Phi(x) > c_0 + \nu^{\Phi}_{max} \geq B$

$$E[\hat{\Phi}(c_0, Q(t+1))|Q(t) = x] - \hat{\Phi}(c_0, x) = E_{\pi}[\Phi(Q(t+1)|Q(t) = x] - \Phi(x) \leq -\gamma,$$

where the last inequality holds since $\Phi(x) > B$ and since Φ is a Lyapunov function with drift γ.

3. To analyze the second summand in Equation (4.8) we fix an arbitrary $x \in \mathcal{X}$.

   If x' is such that $p(x, x') > 0$ and $\hat{\Phi}(c_0, x') > \hat{\Phi}(c_0, x)$, then since $\hat{\Phi}(c_0, x) \geq c_0$ then $\hat{\Phi}(c_0, x') > c_0$ and $\hat{\Phi}(c_0, x') = \Phi(x)$. We obtain

$$\hat{\Phi}(c_0, x') = \Phi(x) > \hat{\Phi}(c_0, x) \geq \Phi(x) \tag{4.9}$$

$$\hat{\Phi}(c_0, x') - \hat{\Phi}(c_0, x) = \Phi(x') - \hat{\Phi}(c_0, x) \leq \Phi(x') - \Phi(x) \leq \nu^{\Phi}_{max} \tag{4.10}$$

   Therefore, for any x, in particular for all x satisfying $c_0 - \nu^{\Phi}_{max} < \Phi(x) \leq c_0 + \nu^{\Phi}_{max}$ we have that

$$E[\hat{\Phi}(c_0, Q(t+1))|Q(t) = x] - \hat{\Phi}(c_0, x) = \sum_{x': \hat{\Phi}(c_0, x') > \hat{\Phi}(c_0, x)} p(x, x')(\hat{\Phi}(c_0, x') - \hat{\Phi}(c_0, x)) +$$

$$\sum_{x': \hat{\Phi}(c_0, x') \leq \hat{\Phi}(c_0, x)} p(x, x')(\hat{\Phi}(c_0, x') - \hat{\Phi}(c_0, x)) \leq \sum_{x': \hat{\Phi}(c_0, x') > \hat{\Phi}(c_0, x)} p(x, x')\nu^{\Phi}_{max} \leq$$

$$p^{\Phi}_{max}\nu^{\Phi}_{max},$$

   where the first inequality follows from (4.10) and the second follows from (4.9).

   We conclude that for all $x \in \mathcal{X}$, in particular, for all x satisfying $c_0 - \nu^{\Phi}_{max} < \Phi(x) \leq c_0 - \nu^{\Phi}_{max}$ there holds

$$E[\hat{\Phi}(c_0, Q(t+1))|Q(t) = x] - \hat{\Phi}(c_0, x) \leq p^{\Phi}_{max}\nu^{\Phi}_{max}.$$

We now incorporate all of this into Equation (4.8). We obtain

$$0 \leq 0 + \sum_{x: c_0 - \nu^{\Phi}_{max} < \Phi(x) \leq c_0 + \nu^{\Phi}_{max}} \pi(x) p^{\Phi}_{max} \nu^{\Phi}_{max} + (-\gamma) \sum_{x: c_0 + \nu^{\Phi}_{max} < \Phi(x)} \pi(x).$$

From

$$\sum_{x: c_0 - \nu^{\Phi}_{max} < \Phi(x) \leq c_0 + \nu^{\Phi}_{max}} \pi(x) = \sum_{x: c_0 - \nu^{\Phi}_{max} < \Phi(x)} \pi(x) - \sum_{x: c_0 + \nu^{\Phi}_{max} < \Phi(x)} \pi(x),$$

and by dividing the inequality above by $p^{\Phi}_{max} \nu^{\Phi}_{max} + \gamma$ we obtain the result. $\square$

We now state and prove a similar result for lower Lyapunov functions.

Lemma 4.2 *Given a Markov chain $Q(t)$ with stationary probability distribution π, if Φ is a lower Lyapunov function with drift $\gamma > 0$, then for any $c_0 \geq 0$ there holds*

$$\Pr_{\pi}\Big\{c_0 \leq \Phi(Q(t))\Big\} \geq \frac{p^{\Phi}_{min} \nu^{\Phi}_{min}}{p^{\Phi}_{min} \nu^{\Phi}_{min} + 2\gamma} \Pr_{\pi}\Big\{c_0 - \frac{1}{2}\nu^{\Phi}_{min} \leq \Phi(Q(t))\Big\},$$

where ν^{Φ}_{min} and p^{Φ}_{min} are defined by (4.2) and (4.4).

Proof:

Let us select an arbitrary $c_0 \geq 0$. We again consider the function $\hat{\Phi}(c_0, x)$ defined by (4.5) and the corresponding identities (4.6) and (4.7). We rewrite this identity as follows

$$0 = \sum_{x: \Phi(x) < c_0 - (1/2)\nu^{\Phi}_{min}} \pi(x)\Big(E[\hat{\Phi}(c_0, Q(t+1))|Q(t) = x] - \hat{\Phi}(c_0, x)\Big) + \tag{4.11}$$

$$\sum_{x: c_0 - (1/2)\nu^{\Phi}_{min} \leq \Phi(x) < c_0} \pi(x)\Big(E[\hat{\Phi}(c_0, Q(t+1))|Q(t) = x] - \hat{\Phi}(c_0, x)\Big) +$$

$$\sum_{x: c_0 \leq \Phi(x)} \pi(x)\Big(E[\hat{\Phi}(c_0, Q(t+1))|Q(t) = x] - \hat{\Phi}(c_0, x)\Big).$$

We now analyze each of the summands in the identity above

1. For any x with $\Phi(x) \geq c_0$, we have $\hat{\Phi}(c_0, x) = \Phi(x)$. So,

$$\sum_{x:\Phi(x)\geq c_0} \pi(x)\Big(E[\hat{\Phi}(c_0, Q(t+1))|Q(t) = x] - \hat{\Phi}(c_0, x)\Big) =$$

$$\sum_{x:\Phi(x)\geq c_0} \pi(x)\Big(E[\hat{\Phi}(c_0, Q(t+1))|Q(t) = x] - \Phi(x)\Big) \geq$$

$$\sum_{x:\Phi(x)\geq c_0} \pi(x)\Big(E[\Phi(Q(t+1))|Q(t) = x] - \Phi(x)\Big) \geq (-\gamma) \sum_{x:\Phi(x)\geq c_0} \pi(x).$$

2. For any x with $\Phi(x) < c_0 - (1/2)\nu^{\Phi}_{min}$, $\hat{\Phi}(c_0, x) = c_0$. So,

$$\sum_{x:\Phi(x)<c_0-(1/2)\nu^{\Phi}_{min}} \pi(x)\Big(E[\hat{\Phi}(c_0, Q(t+1))|Q(t) = x] - \hat{\Phi}(c_0, x)\Big) =$$

$$\sum_{x:\Phi(x)<c_0-(1/2)\nu^{\Phi}_{min}} \pi(x)\Big(E[\hat{\Phi}(c_0, Q(t+1))|Q(t) = x] - c_0\Big) \geq 0.$$

3. We next consider x satisfying

$$c_0 - (1/2)\nu^{\Phi}_{min} \leq \Phi(x) < c_0 \tag{4.12}$$

We have $\hat{\Phi}(c_0, x) = c_0$. Now if x' is such that $p(x, x') > 0$ and $\Phi(x') > \Phi(x)$ then $\Phi(x') - \Phi(x) \geq \nu^{\Phi}_{min}$. Therefore, $\Phi(x') \geq \Phi(x) + \nu^{\Phi}_{min} \geq c_0 + (1/2)\nu^{\Phi}_{min}$. In particular, $\hat{\Phi}(c_0, x') = \Phi(x')$ and

$$\hat{\Phi}(c_0, x') - \hat{\Phi}(c_0, x) \geq c_0 + (1/2)\nu^{\Phi}_{min} - c_0 = (1/2)\nu^{\Phi}_{min}.$$

Also if $p(x, x') > 0$ and $\Phi(x') \leq \Phi(x) < c_0$ then $\hat{\Phi}(c_0, x') = c_0$ and $\hat{\Phi}(c_0, x') - \hat{\Phi}(c_0, x) = 0$. Then we obtain, for any x satisfying (4.12)

$$E[\hat{\Phi}(c_0, Q(t+1))|Q(t) = x] - \hat{\Phi}(c_0, x) = \sum_{x':\Phi(x')>\Phi(x)} p(x, x')\Big(\hat{\Phi}(c_0, x') - \hat{\Phi}(c_0, x)\Big) +$$

$$\sum_{x':\Phi(x')\leq\Phi(x)} p(x, x')\Big(\hat{\Phi}(c_0, x') - \hat{\Phi}(c_0, x)\Big) = \sum_{x':\Phi(x')>\Phi(x)} p(x, x')\Big(\hat{\Phi}(c_0, x') - \hat{\Phi}(c_0, x)\Big) \geq$$

$$p^{\Phi}_{min}(1/2)\nu^{\Phi}_{min}.$$

We incorporate this into (4.11) to obtain

$$0 \geq \sum_{x: c_0-(1/2)\nu^{\Phi}_{min}<\Phi(x)\leq c_0} \pi(x)(1/2)p^{\Phi}_{min}\nu^{\Phi}_{min} + (-\gamma)\sum_{x: c_0<\Phi(x)} \pi(x).$$

We have

$$\sum_{x: c_0-(1/2)\nu^{\Phi}_{min}<\Phi(x)\leq c_0} \pi(x) = \sum_{x: c_0-(1/2)\nu^{\Phi}_{min}<\Phi(x)} \pi(x) - \sum_{x:\Phi(x)\leq c_0} \pi(x).$$

Using this and dividing the previous inequality by $(1/2)p^{\Phi}_{min}\nu^{\Phi}_{min}+\gamma$, we obtain the result. □

We now are ready to state and prove the main result of this chapter.

Theorem 4.1 *Consider a Markov chain $Q(t)$ with stationary probability distribution π.*

1. *If there exist a Lyapunov function Φ with drift term $\gamma > 0$, and exception parameter $B \geq 0$, then for any $m = 0, 1, \ldots$*

$$\Pr_{\pi}\Big\{\Phi(Q(t)) > B + 2\nu^{\Phi}_{max}m\Big\} \leq \Big(\frac{p^{\Phi}_{max}\nu^{\Phi}_{max}}{p^{\Phi}_{max}\nu^{\Phi}_{max}+\gamma}\Big)^{m+1}.$$

   As a result,

$$E_{\pi}[\Phi(Q(t))] \leq B + \frac{2p^{\Phi}_{max}(\nu^{\Phi}_{max})^2}{\gamma}.$$

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

$$\Pr_{\pi}\Big\{\Phi(Q(t)) \geq (1/2)\nu^{\Phi}_{min}m\Big\} \geq \Big(\frac{(1/2)p^{\Phi}_{min}\nu^{\Phi}_{min}}{(1/2)p^{\Phi}_{min}\nu^{\Phi}_{min}+\gamma}\Big)^{m}.$$

   As a result,

$$E_{\pi}[\Phi(Q(t))] \geq \frac{p^{\Phi}_{min}(\nu^{\Phi}_{min})^2}{4\gamma}.$$

Proof:

To prove the first part put $c_0 = B - \nu^{\Phi}_{max}$. By applying Lemma 4.1, we obtain

$$\Pr_{\pi}\Big\{B < \Phi(Q(t))\Big\} \leq \frac{p^{\Phi}_{max}\nu^{\Phi}_{max}}{p^{\Phi}_{max}\nu^{\Phi}_{max} + \gamma}\Pr_{\pi}\Big\{B - 2\nu^{\Phi}_{max} < \Phi(Q(t))\Big\} \leq$$

$$\frac{p^{\Phi}_{max}\nu^{\Phi}_{max}}{p^{\Phi}_{max}\nu^{\Phi}_{max} + \gamma}.$$

Now we subsequently put $c_0 = B + \nu^{\Phi}_{max}, B + 3\nu^{\Phi}_{max}, B + 5\nu^{\Phi}_{max}, \ldots$. Applying again Lemma 4.1 we obtain the needed upper bound on the tail distribution.

For the expectation part note that

$$E_{\pi}[\Phi(Q(t))] \leq B\Pr_{\pi}\Big\{\Phi(Q(t)) \leq B\Big\}+$$

$$\sum_{m=0}^{\infty}(B + 2\nu^{\Phi}_{max}(m+1))\Pr_{\pi}\Big\{B + 2\nu^{\Phi}_{max}m < \Phi(Q(t)) \leq B + 2\nu^{\Phi}_{max}(m+1)\Big\} =$$

$$B\Pr_{\pi}\Big\{\Phi(Q(t)) \leq B\Big\} + B\sum_{m=0}^{\infty}\Pr_{\pi}\Big\{B + 2\nu^{\Phi}_{max}m < \Phi(Q(t)) \leq B + 2\nu^{\Phi}_{max}(m+1)\Big\}+$$

$$2\nu^{\Phi}_{max}\sum_{m=0}^{\infty}(m+1)\Pr_{\pi}\Big\{B + 2\nu^{\Phi}_{max}m < \Phi(Q(t)) \leq B + 2\nu^{\Phi}_{max}(m+1)\Big\}.$$

But

$$\sum_{m=0}^{\infty}(m+1)\Pr_{\pi}\Big\{B + 2\nu^{\Phi}_{max}m < \Phi(Q(t)) \leq B + 2\nu^{\Phi}_{max}(m+1)\Big\} =$$

$$\sum_{m=0}^{\infty}\Pr_{\pi}\Big\{B + 2\nu^{\Phi}_{max}m < \Phi(Q(t))\Big\}.$$

Applying the bounds on the tail distribution we obtain

$$E_{\pi}[\Phi(Q(t))] \leq B + 2\nu^{\Phi}_{max}\sum_{m=0}^{\infty}\Big(\frac{p^{\Phi}_{max}\nu^{\Phi}_{max}}{p^{\Phi}_{max}\nu^{\Phi}_{max} + \gamma}\Big)^{m+1} =$$

$$B + \frac{2p^{\Phi}_{max}(\nu^{\Phi}_{max})^2}{\gamma}.$$

To prove the second part put $c_0 = (1/2)\nu^{\Phi}_{min}, c_0 = \nu^{\Phi}_{min}, c_0 = 3/2\nu^{\Phi}_{min}, \ldots$. Then, by applying Lemma 4.2, we obtain

$$\Pr_\pi\Big\{\Phi(Q(t)) \geq (1/2)\nu^{\Phi}_{min}m\Big\} \geq \Big(\frac{(1/2)p^{\Phi}_{min}\nu^{\Phi}_{min}}{(1/2)p^{\Phi}_{min}\nu^{\Phi}_{min}+\gamma}\Big)^m \Pr_\pi\Big\{\Phi(Q(t)) \geq 0\Big\} =$$

$$\Big(\frac{(1/2)p^{\Phi}_{min}\nu^{\Phi}_{min}}{(1/2)p^{\Phi}_{min}\nu^{\Phi}_{min}+\gamma}\Big)^m.$$

For the expectation part, note that

$$E_\pi[\Phi(Q(t))] \geq \sum_{m=0}^{\infty}(1/2)\nu^{\Phi}_{min}m\Pr_\pi\Big\{(1/2)\nu^{\Phi}_{min}m \leq \Phi(Q(t)) < (1/2)\nu^{\Phi}_{min}(m+1)\Big\}.$$

But

$$\sum_{m=0}^{\infty}(1/2)\nu^{\Phi}_{min}m\Pr_\pi\Big\{(1/2)\nu^{\Phi}_{min}m \leq \Phi(Q(t)) < (1/2)\nu^{\Phi}_{min}(m+1)\Big\} =$$

$$(1/2)\nu^{\Phi}_{min}\sum_{m=1}^{\infty}\Pr_\pi\Big\{(1/2)\nu^{\Phi}_{min}m \leq \Phi(Q(t))\Big\}.$$

Using the lower bounds on the tail probability distribution we obtain

$$E_\pi[\Phi(Q(t))] \geq (1/2)\nu^{\Phi}_{min}\sum_{m=0}^{\infty}\Big(\frac{(1/2)p^{\Phi}_{min}\nu^{\Phi}_{min}}{(1/2)p^{\Phi}_{min}\nu^{\Phi}_{min}+\gamma}\Big)^m = \frac{p^{\Phi}_{min}(\nu^{\Phi}_{min})^2}{4\gamma}.$$

This completes the proof of the theorem. □

4.3 Random Walk in the Nonnegative Orthant. Examples and Discussion

Theorem 4.1 from the previous section provides a simple tool for the analysis of stationary probability distributions of a particular class of Markov chains - random walks in the nonnegative orthant Z^n_+.

- To illustrate the method we start with a simple case of a one-dimensional random walk in Z_+.

Consider the following random walk

$$\Pr\Big\{Q(t+1)=m+1|Q(t)=m\Big\}=p, \qquad m=0,1,2,\ldots,$$

$$\Pr\Big\{Q(t+1)=m-1|Q(t)=m\Big\}=1-p, \qquad m=1,2,\ldots,$$

and

$$\Pr\Big\{Q(t+1)=0|Q(t)=0\Big\}=1-p.$$

For this Markov chain we obtain

$$\Pr_\pi\Big\{Q(t)\geq m\Big\}\leq \Big(\big(\frac{p}{p+(1-p)}\big)^m=p^m.$$

If $p\to 0$ then the bound converges to zero. Our bounds dominate the bounds constructed by Hajek in [28].

- **State-homogeneous random walk in Z_+^n with bounded step sizes.** Suppose we are given a random walk $Q(t)=(Q_1(t),Q_2(t),\ldots,Q_n(t))\in Z_+^n$. For any $\Lambda\subset\Big\{1,2,\ldots,n\Big\}$ let

$$Z_\Lambda=\Big\{x\in Z_+^n: x_i>0,\ \text{for}\ i\in\Lambda,\ x_i=0,\ \text{for}\ i\notin\Lambda\Big\}.$$

As in [24] we call these sets *faces.*

We assume that our random walk is *state-homogeneous* in the following sense. For each $\Lambda\subset\Big\{1,2,\ldots,n\Big\}$ there exists an n-dimensional vector $\theta_\Lambda\in R^n$ such that for all $x\in Z_\Lambda$

$$E[Q(t+1)|Q(t)=x]=x+\theta_\Lambda.$$

Namely, the *expected step* of the random walk depends only on the face the walk is currently on. This includes a more specific case, when the *distribution of the step* during a transition depends only on the face (for example, a two-dimensional random walk above). The latter random walks will be called *state-homogeneous in distribution.*

Assume also, that our random walk has bounded steps. Namely, for ν defined by

$$\nu \equiv \max\Big\{|x'-x| : \Pr\Big\{Q(t+1)=x'|Q(t)=x\Big\}>0\Big\},$$

there holds $\nu < \infty$, Here $|x| = \sum_{i=1}^{n}|x_i|$ for each $x=(x_1,x_2,\ldots,x_n)\in Z^n$. State-homogeneous random walks with bounded steps can model, in particular, single-class exponential type queueing networks (classical Jackson networks) and multiclass exponential type queueing networks, operating under a priority scheduling policy. This motivates our interest in such random walks. In [24] the authors considered state-homogeneous random walks to establish ergodicity of Jackson networks. We will study closely such Markov chains in Chapter 6, when we analyze queueing networks with priority scheduling policies. For now, we are interested in obtaining results for general state-homogeneous random walks with bounded steps. We establish that for such random walks, whenever there exists some linear Lyapunov function, certain geometric bounds hold on the stationary probability distribution.

Theorem 4.2 *Suppose we are given a space-homogeneous random walk in Z_+^n with the set of expected steps $\Big\{\theta_\Lambda : \Lambda\subset\{1,2,\ldots,n\}\Big\}$. If there exists a nonnegative vector $L=(L_1,L_2,\ldots,L_n)\in R_+^n$ and $\gamma>0$, such that for all non-empty $\Lambda\subset\{1,2,\ldots,n\}$ there holds*

$$L'\theta_\Lambda \le -\gamma,$$

then the random walk is positive recurrent. For the corresponding stationary probability distribution π, there holds

$$\Pr_\pi\Big\{L'Q(t)\ge 2\nu m\Big\} \le \Big(\frac{L_{max}\nu}{L_{max}\nu+\gamma}\Big)^m,$$

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

$$E_\pi[L'Q(t)] \le \frac{2(L_{max}\nu)^2}{\gamma},$$

where $L_{max} = \max_{1 \le i \le n}\{L_i\}$ *and*

$$\nu \equiv \max\Big\{|x' - x| : \Pr\{Q(t+1) = x' | Q(t) = x\} > 0\Big\}.$$

The bounds are non-trivial if $\nu < \infty$.

Proof:

From the conditions of the theorem, $\Phi(x) = L'x$ is a Lyapunov function with drift γ. Note that for this Lyapunov function $\nu_{max}^{\Phi} \le L_{max}\nu$. Also $p_{max}^{\Phi} \le 1$. The result then follows directly from the first part of Theorem 4.1.$\Box$

We now show how to use the second part of Theorem 4.1 to get lower bounds for state-homogeneous in distribution random walks (possibly with $\nu < \infty$), using practically *any* nonnegative linear function with integral components. Recall that state-homogeneous in distribution random walks are such that the probabilities of transitions depend only on the face the walk is currently on. Formally, for every $\Lambda \subset \{1, 2, \ldots, n\}$, for every $x_1, x_2 \in Z_\Lambda$ and for every $\triangle x \in Z_+^n$, the probabilities of transitions from x_1 to $x_1 + \triangle x$ and from x_2 to $x_2 + \triangle x$ are the same

$$p(x_1, x_1 + \triangle x) = p(x_2, x_2 + \triangle x).$$

In words, the distribution of the step vector $\triangle x$ depends only on the face that the walk is currently on.

Theorem 4.3 *Suppose we are given a positive recurrent space-homogeneous in distribution random walk in* Z_+^n *with the set of expected steps* $\Big\{\theta_\Lambda : \Lambda \subset \{1, 2, \ldots, n\}\Big\}$ *and a stationary probability distribution* π.

For any nonnegative vector with integral components $L = (L_1, L_2, \ldots, L_n) \in Z_+^n$, *we define*

$$\gamma = -\min_{\Lambda}\{L'\theta_\Lambda < 0\}.$$

Then

$$\Pr_\pi\Big\{L'Q(t) \geq \frac{m}{2}\Big\} \geq \Big(\frac{(1/2)p^L_{min}}{(1/2)p^L_{min}+\gamma}\Big)^m,$$

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

$$E_\pi[L'Q(t)] \geq \frac{p^L_{min}}{4\gamma},$$

where

$$p^L_{min} = \min_\Lambda \Big\{\Pr\{L'Q(t+1) - L'Q(t) > 0 | Q(t) \in Z_\Lambda\}\Big\}.$$

Remark. Note that p^L_{min} depends only on the distributions of 2^n different step vectors $\triangle x$, which can be calculated in finite time by considering all 2^n faces Z_Λ, $\Lambda \subset \{1, 2, \ldots, n\}$.

Proof:

We consider a linear lower Lyapunov function $\Phi(x) = L'x$. By definition, γ is a drift of this function. The minimal increase of the value of this Lyapunov function is not smaller than one since both L and $Q(t)$ have integral components. Also for this lower Lyapunov function Φ the corresponding p^Φ_{min} (defined by (4.4)) is equal to p^L_{min}. The result is obtained then by applying the second part of Theorem 4.1. $\square$

It would be interesting to investigate whether Theorem 4.1 can be extended from Markov chains admitting Lyapunov functions to more general processes like supermartingales. Also, it would be interesting to investigate whether bounded step size condition can be relaxed, so that, for example, quadratic Lyapunov functions could be used for the analysis of stationary probability distributions of random walks.

Meyn and Tweedie in [48] have proven that whenever some Lyapunov function establishes the positive recurrence of the Markov chain, the transient probability distribution converges to the stationary with polynomial rate. In particular, let $Q(t)$ be a Markov chain in $\mathcal{X} \subset R^n_+$ for which there exists some Lyapunov function. Let

$p^t_{x_0}(x)$ denote the t step transition probability starting from the state x_0

$$p^t_{x_0}(x) = \Pr\Big\{Q(t) = x | Q(0) = x_0\Big\}.$$

Then for a certain integer $m > 0$ and a certain constant $C(x_0) > 0$ which depends on initial state x_0, there holds

$$\Big|p^t_{x_0}(x) - \pi(x)\Big| \leq \frac{C(x_0)}{t^m}.$$

where π is the corresponding stationary distribution π. It would be interesting to use the results obtained in this chapter in conjunction with the result above to obtain bounds on the transient probability distributions.

Chapter 5

Upper and Lower Bounds on Steady-State Queue Lengths in Multiclass Exponential Type Queueing Networks

In this chapter we develop performance results for multiclass exponential type queueing networks. The model under consideration is a multiclass queueing network (α, μ, P, C) with n classes and J processing stations $\sigma_1, \ldots, \sigma_J$ as in Chapter 2. Interarrival and service times are exponentially distributed, with parameters $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ and $\mu = (\mu_1, \mu_2, \ldots, \mu_n)$ respectively. We assume a Bernoulli type routing given by $n \times n$ matrix $P = (p_{ij})_{1 \leq i,j \leq n}$. As in Chapter, 2 denote $1 - \sum_{k=1}^{n} p_{ik}$ by $p_{k,n+1}$ for each $k = 1, 2, \ldots, n$. We adopt orders of magnitude notations $O(\cdot)$ and $\Omega(\cdot)$ as follows. Given two real valued functions $f(s)$ and $g(s)$ we write $f = O(g)$ or $f = \Omega(g)$ if there exists some constant C, such that $f(s) \leq \frac{C}{g(s)}$ or $f(s) \geq \frac{C}{g(s)}$ respectively. If $f = O(g)$ and $f = \Omega(g)$, then we will write $f = \Theta(g)$.

Exponential type multiclass queueing networks have been studied by numerous authors (see [4],[37],[38],[39],[32],[33]).

Using the results of the previous chapter, we will show that whenever there exists a piecewise linear Lyapunov function proving the global stability of the network, certain geometric type upper bounds hold on the stationary probability distribution of the queue lengths for all work conserving policies. Also for policies, which are assumed a priori to be stable, we will construct lower bounds on the stationary probability distribution of the queue lengths, by means of a linear lower Lyapunov function, which was introduced in the previous chapter. We will also present upper and lower bounds on the expectation of queue lengths in steady-state.

We have shown in Chapter 2 that a two-station fluid network is globally stable if and only if there exists a suitable piecewise linear Lyapunov function. Dai and Vande-Vate in [20] introduced the notion of *virtual* traffic intensity and showed that the existence of a piecewise linear Lyapunov function in two-station networks is equivalent to all actual and *virtual traffic intensities* being less than one. We will show that the traffic intensity of virtual stations is a fundamental performance parameter. Our result can be summarized as follows: in two-station networks, the total expected queue length $E[\sum_{i=1}^{n} Q_i]$ under any work conserving policy satisfies the following inequalities

$$\Omega\Big(\frac{1}{1-\rho}\Big) = E[\sum_{i=1}^{n} Q_i] = O\Big(\frac{1}{(1-\rho^*)^2}\Big) \tag{5.1}$$

where ρ is a maximal traffic intensity among all the stations $\rho = \max_{1\le j\le J}\{\rho_{\sigma_j}\}$ and ρ^* is a maximal actual or *virtual* traffic intensity (to be defined in Section 5.4). Namely, the largest traffic intensity factor ρ^* is the bottleneck factor for the total expected queue length. It is not clear whether the upper bound in Equation (5.1) is tight, but it nevertheless is stronger than all available results.

Several researchers have used quadratic Lyapunov function for the performance analysis of multiclass exponential type queueing networks (see [4],[36],[37], [39], [32], [33]). Upper and lower bounds on weighted sums of the expected queue lengths are obtained by solving a certain linear program. The method of quadratic Lyapunov function is inferior to the method presented in this chapter in several respects.

- Apart from some specific cases, the bounds constructed by quadratic Lyapunov function are only numerical. In this chapter we will construct closed form lower bounds for all multiclass queueing networks and closed form upper bounds for re-entrant line networks with two stations.

- For a specific queueing network with two stations, the upper bounds provided by quadratic Lyapunov function can be infinite even if the queueing (fluid and stochastic) network is globally stable (see [21] on a comparison of piecewise linear and quadratic Lyapunov functions for stability analysis). We will show that this is not the case with the bounds constructed in this chapter. Our bounds for networks with two stations are finite *whenever* the corresponding (fluid and stochastic) queueing network is globally stable.

- The method proposed in this chapter is the first one that provides bounds on distribution of the number of customers in the network. It establishes an exponential decay of the tail probabilities of the following form

$$c_1^m \leq \Pr_\pi\{\text{total number of customers} \geq r\} \leq c_2^m,$$

  for some positive constants $c_1 < c_2 < 1$, which hold for all work conserving policies, whenever some piecewise linear function witnesses global stability. In particular, for networks with two stations they hold if and only if the corresponding fluid network is globally stable. In contrast, the quadratic Lyapunov function approach only provides bounds on expectation of queue lengths in steady state.

5.1 Multiclass Exponential Type Queueing Networks. Markov Chain Structure

In this section we introduce the model under consideration. We are given a multiclass queueing network (α, μ, P, C) with n classes and J processing stations $\sigma_1, \ldots, \sigma_J$ as in Chapter 2. Interarrival and service times are exponentially distributed, with parameters $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ and $\mu = (\mu_1, \mu_2, \ldots, \mu_n)$ respectively. Exponential type multiclass queueing networks have been studied by numerous authors (see [4],[37],[38],[39],[32],[33]). As in [4] and [37], we restrict ourselves to Markovian scheduling policies. Specifically:

Markovian policy assumption: The control decisions are purely a function of the system state, which is the vector of all queue lengths.

From now on all the policies under consideration are assumed to be Markovian. The exponentiality assumption together with the Markovian assumption allow us to resort to *uniformization* (see Lippman [41]). This is a method for sampling a continuous time system to obtain a discrete time system with the same steady-state probability distribution. We rescale the parameters, so that $\sum_{i=1}^{n} \alpha_i + \sum_{i=1}^{n} \mu_i = 1$ and assume that we have a superposition of $2n$ Poisson processes with rates $\alpha_1, \ldots, \alpha_n, \mu_1, \ldots, \mu_n$. The arrivals of the first n Poisson processes correspond to external arrivals into the network. The arrivals of the process with rate μ_k correspond to service completions of class k customers, if a server actually worked on a class k customer, or they correspond to "imaginary" service completions of an "imaginary" customer, if the server was idle or worked on customers from other classes. Let $\tau_s, s = 1, 2, \ldots$ be the sequence of such random sampling times. Let $Q(t)$ denote a vector of (random) queue lengths at time t and let π denote a stationary distribution, corresponding to some stable work conserving policy w. Then, as a result of this construction, for any $x \in Z_+^n$

$$\Pr_\pi\{Q(t) = x\} = \Pr_\pi\{Q(\tau_s) = x\} \tag{5.2}$$

Note, that since all policies are assumed to be Markovian, $Q(\tau_s)$ is a Markov chain. Identity (5.2) allows us to concentrate exclusively on this embedded Markov chain for the purpose of performance analysis.

Any policy w satisfying the Markovian assumption can be described as a function $w : Z_+^n \to \{0,1\}^n$ where for any $q \in Z_+^n$, $w(q) = (w_1, w_2, \ldots, w_n)$ and $w_i = 1$ if station $\sigma(i)$ works on a customer of class i and $w_i = 0$ otherwise. Of course, $w_i = 1$ only if $q_i > 0$. We can specify the transitions of the Markov chain $Q(\tau_s)$ as follows. For $i = 1, 2, \ldots, n$, let e_i be a unit vector with the i-th coordinate equal to one and all other coordinates equal to zero. We adopt the convention $e_0 = e_{n+1} = 0$. The following proposition holds.

Proposition 5.1 *Given a multiclass queueing network* (α, μ, P, C) *rescaled so that* $\sum_{i=1}^{n} \alpha_i + \sum_{i=1}^{n} \mu_i = 1$ *and given a Markovian policy* w*, the transition probabilities of the corresponding embedded Markov chain* $Q(\tau_s),\ s = 0, 1, 2, \ldots$ *are given as*

$$Q(\tau_{s+1}) = \begin{cases} Q(\tau_s) + e_i & \text{with probability } \alpha_i \\ Q(\tau_s) - e_j + e_i & \text{with probability } \mu_j w_j(Q(\tau_s)) p_{ji} \\ Q(\tau_s) & \text{with probability } \sum_j \mu_j (1 - w_j(Q(\tau_s)) \end{cases} \tag{5.3}$$

Proof:

Note that the change $Q(\tau_{s+1}) - Q(\tau_s)$ of our embedded Markov chain corresponds to either arrival of class i customers $i = 1, 2, \ldots, n$, or it corresponds to "actual" service completion of class i customers and this customer joining class j. The first event has a probability α_i and corresponds to a change e_i. The second event has probability $\mu_j w_j(Q(\tau_s)) p_{ji}$ and corresponds to a change $e_j - e_i$.□

5.2 Multiclass Single Station Queueing System

In this section we analyze the performance of an exponential type multiclass queueing system with a single processing station. The data of this system is given by a triplet (α, μ, P). We will show that the average steady-state workload in the system $E_\pi[W(t)]$ is exactly of the order $\Theta\left(\frac{1}{1-\rho}\right)$ and we will derive geometric type bounds on the tail probability $\Pr_\pi\{W(t) > m\},\ m = 0, 1, 2, \ldots$. Then we will consider the examples of an M/M/1 queue, a re-entrant line single station queueing system, and a multiclass single station queueing system without feedback.

5.2.1 Upper and Lower Bounds via (Lower) Lyapunov Function

Let w be any Markovian scheduling policy. As we have discussed in Section 2.2 of Chapter 2, any such policy is stable (the underlying Markov chain $Q(\tau_s)$ is positive recurrent) if and only if the spectral radius of P is less than 1 and the traffic intensity satisfies

$$\rho = e'M^{-1}[I - P']^{-1}\alpha < 1, \tag{5.4}$$

where $M = \mathrm{diag}\{\mu_1, \ldots, \mu_n\}$ and $e = (1, 1, \ldots, 1)$. Let $\pi = \pi(w)$ denote the corresponding stationary distribution. Consider a vector $L = (L_1, L_2, \ldots, L_n)$ in R^n_+ and the function $L'Q(\tau_s)$ of the Markov chain $Q(\tau_s)$. Using Proposition 5.1, after rescaling the parameters so that $\sum_{i=1}^n \alpha_i + \sum_{i=1}^n \mu_i = 1$ we obtain

$$E[L'Q(\tau_{s+1})|Q(\tau_s)] = L'Q(\tau_s) + L'\alpha + L'[P' - I]Mw(Q(\tau_s)).$$

In particular, let us set

$$L' = e'M^{-1}[I - P']^{-1}. \tag{5.5}$$

Then

$$E[L'Q(\tau_{s+1})|Q(\tau_s)] = L'Q(\tau_s) + \rho - e'w(Q(\tau_s)).$$

Suppose now that in addition the scheduling policy is work conserving. Namely, $e'w(Q(\tau_s)) = 1$ whenever $Q(\tau_s) \neq 0$. We conclude

$$E[L'Q(\tau_{s+1})|Q(\tau_s)] = L'Q(\tau_s) + \rho - 1,$$

whenever $Q(\tau_s) \neq 0$. Assuming that (5.4) holds we see that $\Phi(x) = L'x$ is both a Lyapunov function with drift $1 - \rho$ and exception parameter $B = 0$, under any work conserving Markovian policy, and a lower Lyapunov function with the drift factor $\gamma = 1 - \rho$ under any Markovian policy. Note that the function Φ has a simple physical interpretation:

$$\Phi(Q(t)) = e'M^{-1}[I-P']^{-1}Q(t)$$

is the *workload* in the system - the total time required to process customers currently present in the system, assuming no future arrivals occur. We denote this quantity (which is time dependent) by $W(t)$. We now estimate the parameters $\nu^{\Phi}_{max}, \nu^{\Phi}_{min}, p^{\Phi}_{max}$ and p^{Φ}_{min} defined by (4.1)-(4.4). Note from (5.5) that $L \geq e'M^{-1} = (\mu_1^{-1}, \mu_2^{-1}, \ldots, \mu_n^{-1})' > 0$. The value of the Lyapunov function $L'Q(t)$ can change in two ways. It can increase or decrease by $L_i > 0$, if the transition corresponds to an external arrival to or departure from class i. It can also change by $L_i - L_j$ if the transition corresponds to a departure from class j becoming an internal arrival to class i. This can happen only if $p_{ji} > 0$. We conclude

$$\nu^{\Phi}_{max} = \max\{\max_{1\leq i\leq n}\{L_i\}, \max_{1\leq i,j\leq n}\{|L_i - L_j| : p_{ji} > 0\}\} = \max_{1\leq i\leq n}\{L_i\},$$

which we denote by L_{max} Similarly, the value of the lower Lyapunov function $L'Q(t)$ can increase in two ways. It can increase by $L_i > 0$, if the transition corresponds to an external arrival to class i which happens with probability α_i. It can also increase by $L_i - L_j$ if the transition corresponds to a departure from class j becoming an internal arrival to the class i. This can happen only if $L_i - L_j > 0$ and $p_{ji} > 0$. Then

$$p^{\Phi}_{max} = \sup_{Q\in Z^n_+}\Big\{\sum_{i=1}^n \alpha_i + \sum_{j=1}^n \mu_j w_j(Q) \sum_{i:L_i - L_j > 0}^n p_{ji}\Big\} \leq \sum_{i=1}^n \alpha_i + \sum_{i=1}^n \mu_j \sum_{i:L_i - L_j > 0}^n p_{ji},$$

$$\nu^{\Phi}_{min} = \min\{\min\{L_i : \alpha_i > 0\}, \min_{1\leq i,j\leq n}\{L_i - L_j : L_i - L_j > 0, p_{ji} > 0\}\},$$

which we denote by L_{min}, and

$$p^{\Phi}_{min} \geq \sum_{i=1}^n \alpha_i.$$

The following upper and lower bounds then hold on the steady-state workload in a multiclass single station queueing system.

Theorem 5.1 *Given a multiclass single station exponential type queueing system*

(α, μ, P), the steady-state probability distribution of the workload $W(t) = e'M^{-1}[I - P']^{-1}Q(t)$ satisfies the following lower bounds under any stable Markovian policy

$$\Pr_\pi\Big\{\frac{2W(t)}{L_{min}} \geq m\Big\} \geq \Big(\frac{(1/2)L_{min}\sum_{i=1}^n \alpha_i}{(1/2)L_{min}\sum_{i=1}^n \alpha_i + 1 - \rho}\Big)^m,$$

and the following upper bound under any Markovian work conserving policy

$$\Pr_\pi\Big\{\frac{W(t)}{2L_{max}} \geq m\Big\} \leq \Big(\frac{L_{max}(\sum_{i=1}^n \alpha_i + \sum_{i=1}^n \mu_j \sum_{i:L_i - L_j > 0}^n p_{ji})}{L_{max}(\sum_{i=1}^n \alpha_i + \sum_{i=1}^n \mu_j \sum_{i:L_i - L_j > 0}^n p_{ji}) + 1 - \rho}\Big)^m$$

for $m = 0, 1, 2, \ldots,$ where $L' = (L_1, \ldots, L_n)' = e'M^{-1}[I - P']^{-1}$ and

$$L_{max} = \max_{1 \leq i \leq n}\{L_i\}$$

$$L_{min} = \min\{\min\{L_i : \alpha_i > 0\}, \min_{1 \leq i,j \leq n}\{L_i - L_j : L_i - L_j > 0, p_{ji} > 0\}\}.$$

Also, the average workload in the system satisfies the following bounds under any Markovian work conserving policy

$$\frac{L_{min}^2 \sum_{i=1}^n \alpha_i}{4(1-\rho)} \leq E_\pi[W(t)] \leq \frac{2L_{max}^2(\sum_{i=1}^n \alpha_i + \sum_{i=1}^n \mu_j \sum_{i:L_i - L_j > 0}^n p_{ji})}{1-\rho}.$$

In particular, the expected workload is of the order $\Theta\big(\frac{1}{1-\rho}\big)$.

Proof:

The result is obtained directly by applying Theorem 4.1. Note that the bounds are the same whether we have rescaled the parameters so that $\sum_{i=1}^n \alpha_i + \sum_{i=1}^n \mu_i = 1$ or not. In fact, multiplying all the parameters α_i, μ_i by some constant does not change the value of ρ and of the products $L_{min}\alpha_i, L_{min}\mu_i, L_{max}\alpha_i, L_{max}\mu_i$. Therefore the bounds do not depend on scaling. □

5.2.2 Examples and Discussion

- **M/M/1 Queue.** Consider an M/M/1 queue with arrival rate α and service rate $\mu > \alpha$. Let us apply the lower bound on the queue length given by the

Theorem 5.2. We have $L = e'M^{-1}[I-P']^{-1} = \mu^{-1}$. Clearly, $L_{max} = L_{min} = \mu^{-1}$. Then from Theorem 5.1, we obtain

$$\Pr_\pi\{1/2Q(t) \geq m\} \leq \Big(\frac{\mu^{-1}\alpha}{\mu^{-1}\alpha + 1 - \rho}\Big)^m = \rho^m,$$

and

$$\Pr_\pi\{2Q(t) \geq m\} \geq \Big(\frac{(1/2)\mu^{-1}\alpha}{(1/2)\mu^{-1}\alpha + 1 - \rho}\Big)^m = (\frac{\rho}{2-\rho})^m,$$

or

$$\Big(\big(\frac{\rho}{2-\rho}\big)^2\Big)^m \leq \Pr_\pi\{Q(t) \geq m\} \leq (\rho^{1/2})^m.$$

Recall, that the correct value of $\Pr_\pi\{Q(t) \geq m\}$ is ρ^m. We see that the loss of accuracy is due to the factors 1/2 and 2 in the lower and upper bounds given by Theorem 5.2. For the expectations, we obtain

$$\frac{\rho}{4(1-\rho)} \leq E_\pi[Q(t)] \leq \frac{2\rho}{1-\rho}.$$

This agrees with the correct formula $E_\pi[Q(t)] = \frac{\rho}{1-\rho}$.

- **Re-entrant line single station queueing system** Consider a multiclass single station queueing system which is a re-entrant line. Thus, $\alpha = (\lambda, 0, 0, \ldots, 0)$, $p_{i,i+1} = 1$ for $i = 1, 2, \ldots, n-1$, and $p_{ij} = 0$ otherwise. Recall that in re-entrant line queueing systems $\rho_i = \lambda/\mu_i,\ \rho = \sum_{i=1}^n \rho_i$. Let $\rho_i^+ = \sum_{k=i}^n \rho_k$. The result of Theorem 5.1 can be reformulated as follows

Proposition 5.2 *In a re-entrant line single station queueing system with arrival rate λ and service rates $\mu_1, \mu_2, \ldots, \mu_n$ the following bounds hold on the steady-state number of customers. Under all stable Markovian policies*

$$\Pr_\pi\{\frac{2\sum_{i=1}^n \rho_i^+ Q_i(t)}{\rho} \geq m\} \geq \Big(\frac{(1/2)\rho}{1-(1/2)\rho}\Big)^m.$$

Under all stable Markovian work conserving policies

$$\Pr_\pi\{\frac{\sum_{i=1}^n \rho_i^+ Q_i(t)}{2\rho} \geq m\} \leq \rho^m,$$

for $m = 0, 1, 2, \ldots$.

Also

$$\frac{\rho^2}{4(1-\rho)} \leq E_\pi[\sum_{i=1}^n \rho_i^+ Q_i(t)] \leq \frac{2\rho^2}{1-\rho},$$

where the lower bounds hold for all stable Markovian and upper bounds hold for all stable Markovian work conserving policies.

Proof:

It is not hard to check that

$$L' = e'M^{-1}[I-P']^{-1} = (\sum_{i=1}^n \mu_i^{-1}, \sum_{i=2}^n \mu_i^{-1}, \ldots, \sum_{i=n}^n \mu_i^{-1}) = (1/\lambda)(\rho_1^+, \rho_2^+, \ldots, \rho_n^+).$$

Thus the workload of the system at time t is given by

$$W(t) = (1/\lambda)\sum_{i=1}^n \rho_i^+ Q_i(t).$$

We obtain $L_{max} = (1/\lambda)\rho_1^+ = (1/\lambda)\rho$. Note that $L_i - L_j > 0$ if and only if $i > j$. But $p_{ji} = 0$ for $i > j$. Also $\alpha_1 = \lambda, \alpha_i = 0, i > 1$. Therefore $L_{min} = (1/\lambda)\rho_1^+ = (1/\lambda)\rho$. Also $\sum_{i=1}^n \mu_j \sum_{i:L_i-L_j>0}^n p_{ji} = 0$. The result then follows immediately from Theorem 5.1 after cancelling $1/\lambda$. □

- **Multiclass single server queueing system without feedback**. Consider a multiclass single server queueing system (α, μ, P) in which there is no feedback. Namely, $P = 0$. Let $S_i = \mu_i^{-1}$, $i = 1, 2, \ldots, n$ and let $S_{max} = \max\{S_i\}, S_{min} = \min\{S_i\}$. Then $L' = e'M^{-1}[I-P']^{-1} = (S_1, \ldots, S_n)'$, and $L_{max} = S_{max}$. Assume that $\alpha_i > 0$ for all $i = 1, 2, \ldots, n$. Then $L_{min} = S_{min}$. From Theorem 5.1 we obtain

$$\Pr_\pi\{\frac{\sum_{i=1}^n S_i Q_i(t)}{2S_{max}} \geq m\} \leq \Big(\frac{S_{max}\sum_{i=1}^n \alpha_i}{S_{max}\sum_{i=1}^n \alpha_i + 1 - \rho}\Big)^m,$$

and

$$\Pr_\pi\{\frac{2\sum_{i=1}^n S_iQ_i(t)}{S_{min}} \geq m\} \geq \Big(\frac{(1/2)S_{min}\sum_{i=1}^n \alpha_i}{(1/2)S_{min}\sum_{i=1}^n \alpha_i + 1 - \rho}\Big)^m.$$

For the expectations

$$\frac{S_{min}^2 \sum_{i=1}^n \alpha_n}{4(1-\rho)} \leq E_\pi[\sum_{i=1}^n S_iQ_i(t)] \leq \frac{2S_{max}^2\sum_{i=1}^n \alpha_n}{1-\rho}.$$

The inequalities obtained in this section for a single station queueing system are not tight, as we have seen, even in the simplest case of M/M/1 system. In fact, the existing techniques can achieve better results. For example, one can analyze a single station exponential type multiclass queueing system with feedback as an M/G/1 system (see [56]). That approach gives an closed form solution to the steady-state workload in the system under any Markovian work conserving policy. Also, the quadratic Lyapunov function approach provides an exact description of the performance region for the expected queue lengths. In that sense, the results of this section are inferior when the *expectation* of queue lengths is concerned. However, unlike the methods mentioned above, the results of this section give a closed form and a simple upper and lower bound on the *distribution* of the queue lengths in such systems.

5.3 Lower Bounds in Networks via Linear lower Lyapunov Functions

In this and in the following sections we apply Theorem 4.1 to the performance analysis of multiclass exponential type queueing networks. We start by using linear lower Lyapunov functions to find closed form lower bounds on the distribution and expectation of steady-state queue lengths under any Markovian policy. We will show, in particular, that under any Markovian policy

$$E_\pi[\sum_{i=1}^n Q_i(t)] = \Omega\Big(\sum_{j=1}^J \frac{1}{1-\rho_{\sigma_j}}\Big) = \Omega\Big(\frac{1}{1-\rho}\Big),$$

where $\rho = \max_{1 \le j \le J}\{\rho_{\sigma_j}\}$ and ρ_{σ_j} is the traffic intensity of station σ_j. We will also apply these results to some examples.

5.3.1 Closed Form Lower Bounds on Expected Steady-State Number of Customers in Multiclass Queueing Networks

Suppose a multiclass exponential type queueing network (α, μ, P, C) is given. Recall that under any Markovian scheduling policy, the transitions of the uniformized embedded Markov chain are given by Proposition 5.1. In particular, for $\Phi(Q(\tau_s)) = \sum_{i=1}^n Q_i(\tau_s)$ we have $\Phi(Q(\tau_{s+1})) \ge \Phi(Q(\tau_s)) - 1$. Therefore the function Φ is a lower Lyapunov function with drift $\gamma = 1$ and $\nu_{min}^{\Phi} = 1$. Any external arrival corresponds to an increase of the value of Φ. The probability of an arrival is $\sum_{i=1}^n \alpha_i$. Therefore $p_{min}^{\Phi} \ge \sum_{i=1}^n \alpha_i$. In fact, since all departures decrease the value of Φ and all internal arrivals do not change the value of Φ, then p_{min}^{Φ} is exactly equal to $\sum_{i=1}^n \alpha_i$. We then have the following result

Proposition 5.3 *Given a multiclass exponential type queueing network* (α, μ, P, C) *operating under any stable Markovian policy, the following lower bound holds on steady-state probability distributions of the total number of customers in the network*

$$\Pr_\pi\{2\sum_{i=1}^n Q_i(t) \ge m\} \ge \Big(\frac{(1/2)\sum_{i=1}^n \alpha_i}{(3/2)\sum_{i=1}^n \alpha_i + \sum_{i=1}^n \mu_i}\Big)^m,$$

for m=0,1,2,.... Also

$$E_\pi[\sum_{i=1}^n Q_i(t)] \ge \frac{1}{4}\frac{\sum_{i=1}^n \alpha_i}{\sum_{i=1}^n \alpha_i + \sum_{i=1}^n \mu_i}.$$

Proof:

The result is obtained simply by rescaling α and μ so that $\sum_{i=1}^n \alpha_i + \sum_{i=1}^n \mu_i = 1$ and applying Theorem 4.1.□

The bound in Proposition 5.3 is obtained by using the linear lower Lyapunov function $L'Q(t)$, where $L = (1, 1, \dots, 1) = e'$. The bound, although very simple, is not the only possible one. We will use now a different linear lower Lyapunov function to obtain better lower bounds and in particular to show that $E_\pi[\sum_{i=1}^n Q_i(t)] = \Omega(\frac{1}{1-\rho})$ where $\rho \equiv \max_{1\leq j\leq J}\{\rho_{\sigma_j}\}$. In particular, for each station σ_j we consider workload of this station, and use it as a lower Lyapunov function. For each $j = 1, 2, \ldots, J$, let C_j denote the j-th column of the matrix C. Namely, C_j is an n-dimensional zero-one vector, with ones corresponding to coordinates $i \in \sigma_j$. Let $L'_j = (L^j_1, \ldots, L^j_n) = C'_j M^{-1}[I - P']^{-1} \geq 0$. Then $L'_j Q(\tau_s)$ is the workload of the station σ_j corresponding to a vector of queue lengths $Q(\tau_s)$. Let $\rho' = (\rho_1, \rho_2, \ldots, \rho_n)$. For each $j = 1, 2, \ldots, J$ we will show that

$$E_\pi[\sum_{i=1}^n Q_i(t)] = \Omega\Big(\frac{1}{1-\rho_j}\Big).$$

Proposition 5.4 $\Phi(Q(t)) = L'_j Q(t)$ *is a lower Lyapunov function with drift factor* $\gamma_j = 1 - \rho_{\sigma_j}$.

Proof:

Using Proposition 5.1, for any Markovian policy w, we have

$$E[L'_j Q(\tau_{s+1})|Q(\tau_s)] = L'_j Q(\tau_s) + L'_j[P' - I]Mw(Q(\tau_s)) = L'_j Q(\tau_s) + C'_j\rho - C'_j w(Q(\tau_s)).$$

Note $C'_j\rho = \rho_{\sigma_j}$ and $C'_j w(Q(\tau_s)) = \sum_{i\in\sigma_j} w_i(Q(\tau_s)) \leq 1$. Therefore

$$E[L'_j Q(\tau_{s+1})|Q(\tau_s)] \geq L'_j Q(\tau_s) - (1 - \rho_{\sigma_j}),$$

which completes the proof. $\Box$

We now estimate the parameters ν^Φ_{min} and p^Φ_{min} defined by (4.4) and (4.2) for $\Phi(Q(t)) = L'_j Q(t)$. Let

$$Cl(j) = \{i : L^j_i > 0\} \tag{5.6}$$

Intuitively, $Cl(j)$ is the set of all classes that have a positive probability of eventually becoming class j. Note that $\sigma_j \subset Cl(j)$ since $L_j \geq C'_j M^{-1}$ and for $i \in \sigma_j$ the i-th coordinate of $C'_j M^{-1}$ is $\mu_i^{-1} > 0$.

The increase of the value of Φ can happen in one of two cases:

1. An external customer arrives into a class $i \in Cl(j)$ Then the value of the increase is L_i^j. This happens with positive probability if $\alpha_i > 0$.

2. For some classes i_1 and i such that $1 \le i_1 \le n$, $i \in Cl(j)$ and $L_i^j - L_{i_1}^j > 0$, a customer departing from class i_1 becomes a customer of class $i \in Cl(j)$. The value of increase is then exactly $L_i^j - L_{i_1}^j$. Of course, this can happen only if $p_{i_1 i} > 0$.

We conclude

$$\nu_{min}^{\Phi} = \min\Big\{ \min\{L_i : \alpha_i > 0\}, \min_{i \in Cl(j), 1 \le i_1 \le n} \{L_i - L_{i_1} : L_i - L_{i_1} > 0, p_{i_1 i} > 0\}\Big\} \quad (5.7)$$

which we denote by c_{min}^j. We now argue that

$$p_{min}^{\Phi} \ge \sum_{i \in Cl(j)} \alpha_i \quad (5.8)$$

and show that without loss of generality we can assume

$$\sum_{i \in Cl(j)} \alpha_i > 0 \quad (5.9)$$

The first inequality follows simply from the definition of $Cl(j)$: whenever there is an external arrival to class $i \in Cl(j)$ the value of Φ increases by $L_i^j > 0$.

Now to prove (5.9), denote by $P^{(m)}$ the m-th product of the matrix P with itself. For any $k = 1, 2, \dots, n$ let $p_{k\sigma_j}^{(m)} = \sum_{i \in \sigma_j} p_{ki}^{(m)}$. It is not hard to check that

$$Cl(j) = \{k : \sum_{m=0}^{\infty} p_{k\sigma_j}^{(m)} > 0\}.$$

This means customers departing from classes in $\{1, 2, \dots, n\} - Cl(j)$ can never become customers of classes in $Cl(j)$. If $\sum_{i \in Cl(i)} \alpha_i = 0$ then there are no external arrivals to the classes of $Cl(j)$ either. Then after a certain transient period, the classes in $Cl(j)$ will be empty almost surely. Thus, we can consider just the reduced network of classes $\{1, 2, \ldots, n\} - Cl(j)$ without loss of generality.

We now are ready to state and prove the main result of this subsection.

Theorem 5.2 *Given a multiclass exponential type queueing network (α, μ, P, C), operating under any stable Markovian policy, there holds*

$$\Pr_\pi\{\frac{2C_j' M^{-1}[I - P']^{-1} Q(t)}{c_{min}^j} \geq m\} \geq \Big(\frac{(1/2) c_{min}^j \sum_{i \in Cl(j)} \alpha_i}{(1/2) c_{min}^j \sum_{i \in Cl(j)} \alpha_i + 1 - \rho_{\sigma_j}}\Big)^m,$$

for each $j = 1, 2, \ldots, J$ and $m = 0, 1, 2, \ldots$, where c_{min}^j and $Cl(j)$ are defined by (5.7) and (5.6) respectively.

Also

$$E_\pi[C_j' M^{-1}[I - P']^{-1} Q(t)] \geq \frac{(c_{min}^j)^2 \sum_{i \in Cl(j)} \alpha_i}{4(1 - \rho_{\sigma_j})},$$

and

$$E_\pi[e' M^{-1}[I - P']^{-1} Q(t)] \geq \sum_{j=1}^{J} \frac{(c_{min}^j)^2 \sum_{i \in Cl(j)} \alpha_i}{4(1 - \rho_{\sigma_j})}.$$

Proof:

The result follows simply by combining Proposition 5.4 with (5.7), and (5.8), and applying Theorem 4.1. □

Remarks:

1. The bounds hold whether we have rescaled the parameters to $\sum_{i=1}^n \alpha_i + \sum_{i=1}^n \mu_i = 1$ or not, since ρ_{σ_j}, the products $c_{min}^j \sum_{i \in Cl(j)} \alpha_i$ and the ratios

$$\frac{C_j' M^{-1}[I - P']}{c_{min}^j}$$

are insensitive to rescaling.

2. The bounds are nontrivial since $c_{min}^j > 0$ by (5.7) and $\sum_{i \in Cl(j)} \alpha_i > 0$ from (5.9).

3. The theorem essentially proves that for $\rho = \max\{\rho_{\sigma_1}, \ldots, \rho_{\sigma_J}\}$

$$E_\pi[e' M^{-1}[I - P']^{-1} Q(t)] \geq \Omega\Big(\frac{1}{1 - \rho}\Big),$$

in the following sense. Suppose we have fixed the parameters μ and P and only allow α to vary. Then as $\rho(\alpha) = C'M^{-1}[I - P']^{-1}\alpha$ approaches the J-dimensional unit vector $(1, 1, \ldots, 1)$, the value $E_\pi[e'M^{-1}[I - P']^{-1}Q(t)]$ goes to infinity at a rate of at least $\frac{G}{1-\rho(\alpha)}$, where the constant G depends on μ and P only, and $\rho(\alpha) = \max_{1 \leq j \leq J} \rho_{\sigma_j}(\alpha)$.

4. The quantity $E_\pi[e'M^{-1}[I - P']^{-1}Q(t)]$ has a certain physical meaning. It is a steady-state workload in the system - average time needed to process all the customers currently present in the system, assuming no future arrivals occur.

We now translate the bound given by the Theorem 5.2 to the lower bound on the total expected number of customers in the network $E_\pi[\sum_{i=1}^n Q_i(t)]$. Let c_{max} denote the maximal component of the vector $e'M^{-1}[I - P']^{-1}$. Clearly, $c_{max} > 0$.

Corollary 5.1 *Given a multiclass exponential type queueing network (α, μ, P, C), operating under any stable Markovian policy the following lower bound holds on the expected number of customers in the network*

$$E_\pi[\sum_{i=1}^n Q_i(t)] \geq \sum_{j=1}^J \frac{(c_{min}^j)^2 \sum_{i \in Cl(j)} \alpha_i}{4c_{max}(1 - \rho_{\sigma_j})}.$$

Proof:

The result follows easily from the last inequality of Theorem 5.2 and the definition of c_{max}.□

5.3.2 Re-entrant Line Queueing Networks

In this subsection we apply the bounds provided by Theorem 5.2 to re-entrant line exponential type queueing networks. The bounds will be expressed exclusively in terms of the load factors ρ_i.

Recall that in a re-entrant queueing network we have $\alpha = (\lambda, 0, 0, \ldots, 0)$, $p_{i,i+1} = 1$ for $i = 1, 2, \ldots, n-1$, and $p_{ij} = 0$ otherwise. As a result, $\rho_i = \lambda/\mu_i$, and for each station σ_j, $\rho_{\sigma_j} = \sum_{i \in \sigma_j} \lambda / \mu_i$. For each $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, J$, let

$$\rho_i^{\sigma_j +} \equiv \sum_{k \in \sigma_j, k \geq i} \rho_k \tag{5.10}$$

$$\rho_i^{+} \equiv \sum_{k \geq i} \rho_k \tag{5.11}$$

We now state and prove the main result of this subsection

Theorem 5.3 *Given a re-entrant line queueing network with arrival rate λ, service rates $\mu_1, \mu_2, \ldots, \mu_n$ and processing stations $\sigma_1, \ldots, \sigma_J$, the following lower bounds hold on the steady-state number of customers in the network, under any Markovian scheduling policy:*

$$\Pr_\pi\{\frac{2\sum_{i=1}^n \rho_i^{\sigma_j +} Q_i(t)}{\rho_{\sigma_j}} \geq m\} \geq \Big(\frac{\rho_{\sigma_j}}{2 - \rho_{\sigma_j}}\Big)^m,$$

for each $j = 1, 2, \ldots, J$, and $m = 0, 1, 2, \ldots$.

Also,

$$E_\pi[\sum_{i=1}^n \rho_i^{\sigma_j +} Q_i(t)] \geq \frac{1}{4} \frac{(\rho_{\sigma_j})^2}{1 - \rho_{\sigma_j}},$$

and

$$E_\pi[\sum_{i=1}^n \rho_i^{+} Q_i(t)] \geq \sum_{j=1}^J \frac{1}{4} \frac{(\rho_{\sigma_j})^2}{1 - \rho_{\sigma_j}}.$$

Proof:

It is not hard to check that

$$[I - P]^{-1} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix}.$$

Note that $C_j' M^{-1}$ is a vector in which all coordinates i with $i \in \sigma_j$ are equal to μ_i^{-1}

and all the other coordinates are equal to zero. Therefore

$$C_j' M^{-1}[I-P]^{-1} = \frac{1}{\lambda}(\rho_1^{\sigma_j+}, \rho_2^{\sigma_j+}, \ldots, \rho_n^{\sigma_j+})'.$$

Then from (5.7)

$$c_{min}^j = \frac{1}{\lambda}\min\Big\{\min\{\rho_i^{\sigma_j+} : \alpha_i > 0\}, \min_{i\in Cl(j), 1\le i_1\le n}\{\rho_i^{\sigma_j+} - \rho_{i_1}^{\sigma_j+} : \rho_i^{\sigma_j+} - \rho_{i_1}^{\sigma_j+} > 0, p_{i_1,i} > 0\}\Big\}.$$

But $\alpha_1 = \lambda$ and $\alpha_i = 0$ for $i > 1$. Also $p_{i_1,i} > 0$ only if $i_1 = i-1$. But $\rho_{i-1}^{\sigma_j+} \ge \rho_i^{\sigma_j+}$. Therefore

$$c_{min}^j = \frac{1}{\lambda}\rho_1^{\sigma_j} = \frac{1}{\lambda}\rho_{\sigma_j}.$$

Note also that class 1 belongs to $Cl(j)$ for all j, since $\rho_1^{\sigma_j+} = \rho_{\sigma_j} > 0$. Therefore, $\sum_{i\in Cl(j)}\alpha_i = \lambda$. We conclude $c_{min}^j \sum_{i\in Cl(j)}\alpha_i = \rho_{\sigma_j}$. The bounds follow then immediately from Theorem 5.2. □

5.3.3 Examples and Discussion

- **Lu-Kumar Network** Consider the network in Figure 5-1. The network is described by the following parameters

$$\alpha' = (\lambda, 0, 0, 0)', \mu' = (\mu_1, \mu_2, \mu_3, \mu_4)'$$

$$\rho_i = \lambda/\mu_i,\ i = 1,2,3,4,\ \ \rho_{\sigma_1} = \rho_1 + \rho_4, \rho_{\sigma_2} = \rho_2 + \rho_3.$$

We have $\rho_1^{\sigma_1+} = \rho_1 + \rho_4$ and $\rho_i^{\sigma_1+} = \rho_4$ for $i = 2,3,4$.

Similarly, $\rho_i^{\sigma_2+} = \rho_2 + \rho_3$ for $i = 1,2$, $\rho_3^{\sigma_2+} = \rho_3$ and $\rho_4^{\sigma_2+} = 0$.

Applying Theorem 5.3 we obtain

Proposition 5.5 *In a Lu-Kumar network the following lower bounds hold for the stationary probability distribution* π

$$\Pr_\pi\{2\frac{(\rho_1+\rho_4)Q_1(t) + \rho_4 Q_2(t) + \rho_4 Q_3(t) + \rho_4 Q_4(t)}{\rho_1+\rho_4} \ge m\} \ge \Big(\frac{\rho_1+\rho_4}{2-\rho_1-\rho_4}\Big)^m,$$

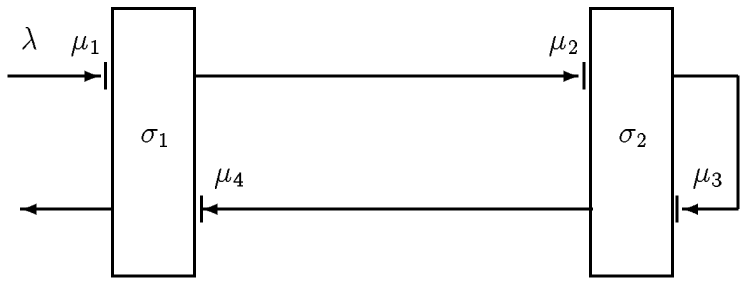

Figure 5-1: Lu-Kumar network.

and

$$\Pr_\pi\{2\frac{(\rho_2+\rho_3)Q_1(t)+(\rho_2+\rho_3)Q_2(t)+\rho_3Q_3(t)}{\rho_2+\rho_3}\geq m\}\geq\Big(\frac{\rho_2+\rho_3}{2-\rho_3-\rho_2}\Big)^m,$$

for all $m=0,1,2,\ldots,$. *Also*

$$E_\pi[(\rho_1+\rho_4)Q_1(t)+\rho_4Q_2(t)+\rho_4Q_2(t)+\rho_4Q_2(t)]\geq\frac{1}{4}\frac{(\rho_1+\rho_4)^2}{1-\rho_1-\rho_4}$$

$$E_\pi[(\rho_2+\rho_3)Q_1(t)+(\rho_2+\rho_3)Q_2(t)+\rho_3Q_2(t)]\geq\frac{1}{4}\frac{(\rho_2+\rho_3)^2}{1-\rho_2-\rho_3}.$$

- **3-class 2-station network** Consider a network described on the Figure 5-2. It is a re-entrant line consisting of two classes (with load factors ρ_1 and ρ_3) in the first station and one class (load factor ρ_2) in the second station. For this network we have $\rho_1^{\sigma_1+}=\rho_1+\rho_3, \rho_2^{\sigma_1+}=\rho_3^{\sigma_1+}=\rho_3$ and $\rho_1^{\sigma_2+}=\rho_2^{\sigma_2+}=\rho_2, \rho_3^{\sigma_2+}=0$.

  Applying Theorem 5.3 we obtain

Proposition 5.6 *In the network of Figure 5-2 the following lower bounds hold for the steady-state number of customers, under any Markovian scheduling policy*

$$\Pr_\pi\{2\frac{(\rho_1+\rho_3)Q_1(t)+\rho_3Q_2(t)+\rho_3Q_3(t)}{\rho_1+\rho_3}\geq m\}\geq\Big(\frac{\rho_1+\rho_3}{2-\rho_1-\rho_3}\Big)^m,$$

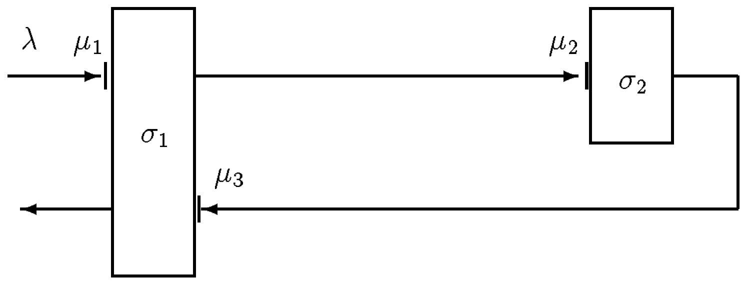

Figure 5-2: Three-Classes-Two-Station network.

and

$$\Pr_\pi\{2Q_1(t) + 2Q_2(t) \geq m\} \geq (\frac{\rho_2}{2-\rho_2})^m,$$

for all $m = 0, 1, 2, \ldots,$. *Also*

$$E_\pi[(\rho_1+\rho_3)Q_1(t) + \rho_3 Q_2(t) + \rho_3 Q_3(t)] \geq \frac{1}{4}\frac{(\rho_1+\rho_3)^2}{1-\rho_1-\rho_3}$$

$$E_\pi[Q_1(t) + Q_2(t)] \geq \frac{1}{4}\frac{\rho_2}{1-\rho_2}.$$

Several researchers have used quadratic Lyapunov function for the performance analysis of multiclass exponential type queueing networks (see [4],[36],[37], [39], [32], [33]). For specific cases, closed form bounds are obtained. For example, suppose we are given a re-entrant line queueing network with n classes, J stations $\sigma_1, \sigma_2, \ldots, \sigma_J$ such that the service rates for each class depend only on the station. Namely, there exist μ_{σ_j} for each $j = 1, 2, \ldots, J$ such that $\mu_i = \mu_{\sigma_j}$ for all $i \in \sigma_j$. The following lower bounds are obtained via quadratic Lyapunov function in [36], which hold for any stable Markovian policy w and the corresponding stationary distribution $\pi(w)$.

$$E_\pi[\sum_{k=1}^{n} kQ_k(t)] \geq \frac{\lambda n(n+1)}{2\sum_{j=1}^{J}\mu_{\sigma_j}(1-\rho_{\sigma_j})} \tag{5.12}$$

The bound (5.12) is inferior to the bounds provided by Theorem 5.3 in two respects. First, this bound assumes the same service rates for all classes within the same station. This can be relaxed, but at the price of getting only numerical bounds. More importantly, the bound above does not capture the right heavy traffic behavior of the network. Note that if the traffic intensity of one of the stations ρ_{σ_j} converges to one and the others remain fixed, then the bound above stays finite, even though, the total expected number of customers converges to infinity (as a function of ρ_{σ_j}). This is not the case, however, with the bound provided by Theorem 5.3. On the other hand for the case of balanced queueing networks (all traffic intensities ρ_{σ_j} are equal to each other) the bounds obtained using a quadratic Lyapunov function can lead to a stronger bound.

5.4 Upper Bounds via Piecewise Linear Lyapunov Function

The main focus of this section is in deriving upper bounds on the steady-state queue lengths in queueing networks by means of a piecewise linear Lyapunov function. We perform a certain "smoothing" of the piecewise linear function, so that the resulting function is Lyapunov for the embedded Markov chain $Q(\tau_s)$ of the queueing network. Then, finite bounds will be obtained on the tail distribution and the expectation of the queue lengths in the network, operating under any work conserving policy, whenever some piecewise linear Lyapunov function witnesses global stability. In that sense we approximate the *global performance region*. We will then use the notion of a *virtual traffic intensity* to analyze the performance of stable queueing networks with two stations in heavy-traffic. We will show that actual and virtual traffic intensities are fundamental performance characteristics.

5.4.1 Lyapunov Function in Multiclass Exponential Type Queueing Networks

Consider a two-station network. We have described in Chapter 3, Section 2.3, a certain linear program LP[md] which if feasible leads to a piecewise linear Lyapunov function for the fluid trajectory $Q(t)$ in the sense of Definition 2.3. Let us reformulate this linear program as follows

$$\begin{aligned}
\sum_{k=1}^{n} L_k\alpha_k + \sum_{k=1}^{n} L_k p_{ik}\mu_i - L_i\mu_i + V &\leq -\gamma \qquad i \in \sigma_1, &(5.13)\\
\sum_{k=1}^{n} L_k p_{jk}\mu_j - L_j\mu_j &\leq V \qquad j \in \sigma_2, &(5.14)\\
\sum_{k=1}^{n} Q_k\alpha_k + \sum_{k=1}^{n} Q_k p_{jk}\mu_j - Q_j\mu_j + W &\leq -\gamma \qquad j \in \sigma_2, &(5.15)\\
\sum_{k=1}^{n} Q_k p_{ik}\mu_i - Q_i\mu_i &\leq W \qquad i \in \sigma_1, &(5.16)\\
L_i &\geq Q_i \qquad i \in \sigma_1, &(5.17)\\
L_j &\leq Q_j \qquad j \in \sigma_2, &(5.18)\\
L,\ Q,\ V,\ W, \gamma &\geq 0. &(5.19)
\end{aligned}$$

This linear program is equivalent to LP[md] in the following sense. It has a feasible solution with positive value of γ if and only if LP[md] is feasible. We denote this linear program also by LP[md]. Thus, as we have shown in Chapter 2, LP[md] has a feasible solution (L_1, L_2, γ) with positive γ if and only if the fluid network is globally stable.

Down and Meyn [21] have constructed a more general linear program that yields sufficient conditions for global stability. Given a multiclass queueing network (α, μ, P, C) with J stations $\sigma_1, \sigma_2, \ldots, \sigma_J$ consider the following linear program in nonnegative variables $L_j = (L_1^j, L_2^j, \ldots, L_n^j)$, j=1,2,...,J:

GLP[md]:

$$\begin{aligned}
\sum_{k=1}^{n} L_k^j\alpha_k + \sum_{k=1}^{n} L_k^j p_{ik}\mu_i - L_i^j\mu_i + V_j &\leq -\gamma \qquad i \in \sigma_j,\ 1 \leq j \leq J, &(5.20)\\
\sum_{k=1}^{n} L_k^j p_{ik}\mu_i - L_i^j\mu_i &\leq V_j \qquad i \in \sigma_{j'},\ j' \neq j, &(5.21)\\
\frac{1}{J-1}\sum_{j'\neq j} L_i^{j'} &\geq L_i^j \qquad i \notin \sigma_j, &(5.22)\\
L_i^j,\ V_j,\ \gamma &\geq 0. &(5.23)
\end{aligned}$$

It was shown in [21] that if GLP[md] has a feasible solution with positive γ, then

$$\Phi(Q(t)) \equiv \max_{1\leq j\leq J}\{L_j' Q(t)\} \tag{5.24}$$

is a Lyapunov function in the sense of Definition 2.3 and the fluid (and stochastic) queueing network is globally stable. Note that this is exactly LP[md] for the case of networks with two stations. A variation of this linear program was derived by Dai and Vande-Vate in [20] and will be considered later in the section.

While the piecewise linear function of the form (5.24) is a Lyapunov function for the fluid model, it is not necessarily a Lyapunov function for the embedded Markov chain $Q(\tau_s)$ of the queueing network (α, μ, P, C) in the sense of Definition 4.1, of Chapter 4. In this subsection we build a certain variation of the function (5.24) which will be a Lyapunov function of the Markov chain $Q(\tau_s)$. It will be used in the following subsection for the analysis of the steady-state performance of exponential type queueing networks, operating under any Markovian work conserving policy.

We now show that if the linear program GLP[md] is feasible, then a certain modification of the function Φ is a Lyapunov function of the embedded Markov chain $Q(\tau_s)$ (in the sense of the Definition 4.1). This modification is a certain "smoothing" of the function (5.24) and is performed exactly as in [21], Theorem 3.7. But unlike [21], our derivation is explicit and a closed form estimate of the exception parameter B will be given. Again we rescale time so that

$$\sum_{i=1}^{n} \alpha_i + \sum_{i=1}^{n} \mu_i = 1.$$

Let $L_1, L_2, \ldots, L_J, \gamma$ be any feasible solution of GLP[md] with $\gamma > 0$. Let

$$L_{max} \equiv \max_{1\leq i\leq n, 1\leq j\leq J}\{L_i^j\}.$$

For all $j = 1, 2, \ldots, J$ we let

$$O_j = \Big\{z = (z_1, \ldots, z_n) \in R^n :$$

$$z_i \in \Big[L_i^j + \frac{J-1}{2J}\gamma, L_i^j + \frac{1}{2}\gamma\Big], \text{ for } i \in \sigma_j,\ z_i \in \Big[L_i^j, L_i^j + \frac{1}{2J}\gamma\Big], \text{ for } i \notin \sigma_j\Big\}. \quad (5.25)$$

Consider the uniform probability density function $p_j(z)$ on the set O_j. We will denote by Z_j a random variable chosen according to distribution $p_j(z)$ and denote by z_j a sample point from the set O_j.

For any $(z_1, z_2, \dots, z_J) \in O_1 \times \dots \times O_J$ and any $x \in R^n_+$ let

$$\Phi_0(z_1, z_2, \dots, z_J, x) = \max_{1 \le j \le J} \{z_j' x\} \quad (5.26)$$

and let

$$\Phi_s(x) = E_u[\Phi_0(Z_1, Z_2, \dots, Z_J, x)] =$$

$$\int \Phi_0(z_1, z_2, \dots, z_J, x) p_1(z_1) \dots p_J(z_J) dz_1 dz_2 \dots dz_J. \quad (5.27)$$

We put a subscript u to emphasize the uniform distribution u. We now prove that our modified function Φ_s is a Lyapunov function.

Proposition 5.7 *Let $L_1, L_2, \dots, L_J, \gamma$ be any feasible solution of GLP[md] with $\gamma > 0$. Then for any Markovian non-idling policy w, Φ_s is a Lyapunov function of the embedded Markov chain $Q(\tau_s)$ with drift equal to $\frac{1}{4}\gamma$ and exception parameter*

$$B = \frac{16nJ^2(J-1)(L_{max}+\gamma)^3}{\gamma^2}.$$

Proof:

We denote the vector

$$\sum_{i=1}^n \alpha_i e_i + \sum_{1 \le i \le n, 1 \le k \le n+1} \mu_i p_{ik} w_i(x)(e_k - e_i)$$

by $\triangle(x, w)$, where, by assumption $e_0 = 0$. Let us denote the vector

$$E[\Phi_s(Q(\tau_{s+1}))|Q(\tau_s) = x] - \Phi_s(x) =$$

$$\sum_{i=1}^n \alpha_i(\Phi_s(x+e_i) - \Phi_s(x)) + \sum_{1 \le i \le n, 1 \le k \le n+1} \mu_i p_{ik} w_i(x)(\Phi_s(x+e_k-e_i) - \Phi_s(x)),$$

by $\triangle\Phi_s(x,w)$. We need to show that the function Φ_s satisfies

$$\triangle\Phi_s(x,w) =$$

$$\sum_{i=1}^{n}\alpha_i(\Phi_s(x+e_i)-\Phi_s(x))+\sum_{1\le i\le n,1\le k\le n+1}\mu_i p_{ik} w_i(x)(\Phi_s(x+e_k-e_i)-\Phi_s(x)) \le -(1/4)\gamma, \tag{5.28}$$

for all Z_+^n such that $\Phi_s(x) \ge B$.

Lemma 5.1 *Suppose a nonzero vector $x \in R_+^n$ is such that for some station σ_{j_0} $\sum_{i\in\sigma_{j_0}} x_i = 0$. Then*

$$\Pr_{\mathrm{u}}\{\Phi_0(Z_1,\dots,Z_J,x) = Z'_{j_0}x\} = 0. \tag{5.29}$$

Proof:

Since $\sum_{i\in\sigma_{j_0}} x_i = 0$, then for any $z_{j_0} \in O_{j_0}$, (5.25) yields

$$z'_{j_0}x = \sum_{i\notin\sigma_{j_0}} z_i^{j_0}x_i \le \sum_{i\notin\sigma_{j_0}} L_i^{j_0}x_i + \frac{1}{2J}\gamma\sum_{i\notin\sigma_{j_0}} x_i \tag{5.30}$$

We have from the third constraint of GLP[md] that for all $i \notin \sigma_{j_0}$

$$L_i^{j_0}x_i \le \frac{1}{J-1}\sum_{j\ne j_0} L_i^j x_i. \tag{5.31}$$

Now consider any $i \notin \sigma_{j_0}$. Note from (5.25), that for j such that $i \notin \sigma_j$ and for any $z_j \in O_j$

$$L_i^j x_i \le z_i^j x_i,$$

and for j such that $i \in \sigma_j$ and for any $z_j \in O_j$

$$L_i^j x_i \le z_i^j x_i - \frac{J-1}{2J}\gamma x_i$$

Combining tow inequalities above we obtain that for every $i \notin \sigma_{j_0}$

$$\frac{1}{J-1}\sum_{j\ne j_0} L_i^j x_i \le \frac{1}{J-1}\Big(\sum_{j\ne j_0} z_i^j x_i\Big) - \frac{1}{J-1}\frac{J-1}{2J}\gamma x_i =$$

$$\frac{1}{J-1}\sum_{j\neq j_0} z_i^j x_i - \frac{1}{2J}\gamma x_i. \tag{5.32}$$

Then from (5.30), (5.31) and (5.32) we obtain

$$z_{j_0}' x \leq \frac{1}{J-1}\sum_{i\notin\sigma_{j_0}}\sum_{j\neq j_0} z_i^j x_i,$$

By assumption $\sum_{i\in\sigma_{j_0}} x_i = 0$, so

$$\sum_{i\notin\sigma_{j_0}}\sum_{j\neq j_0} z_i^j x = \sum_{j\neq j_0} z_j' x$$

We conclude

$$z_{j_0}' x \leq \frac{1}{J-1}\sum_{j\neq j_0} z_j' x$$

for all $z_j \in O_j, j = 1, 2, \ldots, J$. But the event

$$\Phi_0(Z_1, \ldots, Z_J, x) = Z_{j_0}' x$$

implies

$$Z_{j_0}' x \geq \frac{1}{J-1}\sum_{j\neq j_0} Z_j' x.$$

Therefore, the event $\Phi_0(Z_1, \ldots, Z_n, x) = Z_{j_0}' x$ implies

$$Z_{j_0}' x = \frac{1}{J-1}\sum_{j\neq j_0} Z_j' x.$$

The probability of this event is zero.□

Lemma 5.2 *If $z_j \in O_j$ and $x \in R_+^n$ is such that $\sum_{i\in\sigma_j} x_i > 0$ then for any Markovian non-idling policy w*

$$z_j' \triangle(x, w) \leq -\frac{1}{2}\gamma.$$

Proof:

We have

$$z_j' \triangle(x, w) = L_j' \triangle(x, w) + (z_j - L_j)' \triangle(x, w) =$$

$$L'_j\triangle(x,w)+\sum_{i=1}^{n}\alpha_i(z_i^j-L_i^j)+\sum_{1\le i\le n,1\le k\le n+1}\mu_i p_{ik} w_i(x)\Big((z_k^j-L_k^j)-(z_i^j-L_i^j)\Big).$$

From GLP[md] it can be shown that since $\sum_{i\in\sigma_j} x_i > 0$ then $L'_j\triangle(x,w)\le -\gamma$, (for details see [21]) Also since $L_i^j \le z_i^j \le L_i^j + 1/2\gamma$ and by assumption

$$\sum_{i=1}^{n}\alpha_i+\sum_{i=1}^{n}\mu_i=1,$$

then

$$\sum_{i=1}^{n}\alpha_i(z_i^j-L_i^j)+\sum_{1\le i\le n,1\le k\le n+1}\mu_i p_{ik} w_i(x)\Big((z_k^j-L_k^j)-(z_i^j-L_i^j)\Big)\le 1/2\gamma.$$

This completes the proof. □

We now return to the proof of the proposition. Recall our convention $e_0=e_{n+1}=0$. Fix $x\in Z_+^n$. Denote by $E_j,\ j=1,2,\ldots,J$ the event

$$\Phi_0(Z_1,\ldots,Z_J,x+e_k-e_i)=Z'_j(x+e_k-e_i),\ \text{for all } i,k=0,1,\ldots,n+1,$$

and by E_0 the remaining event that none of E_j occurs. Note that the probability that two of the events E_{j_1} and E_{j_2} occurring simultaneously is equal to zero.

We rewrite the left-hand side of (5.28) as

$$\triangle\Phi_s(x,w)=E_u\Big[\sum_{i=1}^{n}\alpha_i(\Phi_0(Z_1,\ldots,Z_J,x+e_i)-\Phi_0(Z_1,\ldots,Z_J,x))+$$

$$\sum_{1\le i\le n,1\le k\le n+1}\mu_i p_{ik} w_i(x)(\Phi_0(Z_1,\ldots,Z_J,x+e_k-e_i)-\Phi_0(Z_1,\ldots,Z_J,x))\Big],\qquad(5.33)$$

and condition consequently on $E_1,E_2,\ldots,E_J$ and E_0. We have

$$E_u\Big[\sum_{i=1}^{n}\alpha_i(\Phi_0(Z_1,\ldots,Z_J,x+e_i)-\Phi_0(Z_1,\ldots,Z_J,x))+$$

$$\sum_{1\le i\le n,1\le k\le n+1}\mu_i p_{ik} w_i(x)(\Phi_0(Z_1,\ldots,Z_J,x+e_k-e_i)-\Phi_0(Z_1,\ldots,Z_J,x))|E_j\Big]\mathrm{Pr_u}\{E_j\}=$$

$$E_u[Z'_j\triangle(x,w)|E_j]\mathrm{Pr_u}\{E_j\}.$$

From Lemma 5.1, if $\sum_{i\in\sigma_j} x_i = 0$ then $\Pr\{E_j\} = 0$. Otherwise, from Lemma 5.2,

$$E_u[Z_j' \triangle(x,w)|E_j]\mathrm{Pr}_{\mathrm{u}}\{E_j\} \leq -\frac{1}{2}\gamma \mathrm{Pr}_{\mathrm{u}}\{E_j\}.$$

Thus we have from (5.33)

$$\triangle\Phi_s(x,w) \leq -(1/2)\gamma \sum_{j=1}^{J} \mathrm{Pr}_{\mathrm{u}}\{E_j\} + \tag{5.34}$$

$$E_u\Big[\sum_{i=1}^{n} \alpha_i(\Phi_0(Z_1,\ldots,Z_J,x+e_i) - \Phi_0(Z_1,\ldots,Z_J,x)) +$$

$$\sum_{1\leq i\leq n, 1\leq k\leq n+1} \mu_i p_{ik} w_i(x)(\Phi_0(Z_1,\ldots,Z_J,x+e_k-e_i) - \Phi_0(Z_1,\ldots,Z_J,x))|E_0\Big]\mathrm{Pr}_{\mathrm{u}}\{E_0\}.$$

We now analyze the last term in the sum.

Lemma 5.3 *For any $i,k = 0,1,2,\ldots,n+1$, and $x \in R_+^n$ there holds*

$$\Big|\Phi_0(z_1,\ldots,z_J,x+e_k-e_i) - \Phi_0(z_1,\ldots,z_J,x)\Big| \leq L_{max} + (1/2)\gamma.$$

Proof:

Let $\Phi_0(z_1,\ldots,z_J,x) = z_{j_1}' x$ and $\Phi_0(z_1,\ldots,z_J,x+e_k-e_i) = z_{j_2}'(x+e_k-e_i)$. Then

$$\Phi_0(z_1,\ldots,z_J,x+e_k-e_i) - \Phi_0(z_1,\ldots,z_J,x) = z_{j_2}'(x+e_k-e_i) - z_{j_1}'x \leq$$

$$z_{j_2}'(x+e_k-e_i) - z_{j_2}'x \leq z_k^{j_2} \leq L_{max} + (1/2)\gamma.$$

Similarly we show that

$$\Phi_0(z_1,\ldots,z_J,x+e_k-e_i) - \Phi_0(z_1,\ldots,z_J,x) \geq -(L_{max} + (1/2)\gamma).$$

□

Applying Lemma 5.3 to (5.34), we obtain

$$\triangle\Phi_s(x,w) \leq -(1/2)\gamma \sum_{j=1}^{J} \mathrm{Pr}_{\mathrm{u}}\{E_j\} + (L_{max} + (1/2)\gamma)\mathrm{Pr}_{\mathrm{u}}\{E_0\} =$$

$$-(1/2)\gamma + (L_{max} + \gamma)\mathrm{Pr_u}\{E_0\}. \tag{5.35}$$

To complete the proof of the proposition we prove the following lemma:

Lemma 5.4 *For any x satisfying $\Phi_s(x) > B$ there holds*

$$\mathrm{Pr_u}\{E_0\} \leq (1/4)\frac{\gamma}{L_{max} + \gamma}.$$

Proof:

We have

$$\mathrm{Pr_u}\{E_0\} \leq \sum_{j_1 \neq j_2} \mathrm{Pr}\{E_{j_1 j_2}\}, \tag{5.36}$$

where $E_{j_1 j_2}$ denotes the event

$$\Phi_0(Z_1, \ldots, Z_J, x) = Z'_{j_1} x, \Phi_0(Z_1, \ldots, Z_J, x + e_k - e_i) = Z'_{j_2}(x + e_k - e_i), \text{ for some } i, k.$$

Now fix j_1, j_2 with $j_2 \neq j_1$. Note that the event $\Phi_0(Z_1, \ldots, Z_J, x) = Z'_{j_1} x,\ \Phi_0(Z_1, \ldots, Z_J, x + e_k - e_i) = Z'_{j_2}(x + e_k - e_i)$ implies

$$Z'_{j_2} x - Z'_{j_1} x \leq 0,$$

and

$$Z'_{j_2} x - Z'_{j_1} x = Z'_{j_2}(x + e_k - e_i) - Z'_{j_1} x + Z_i^{j_2} - Z_k^{j_2} \geq Z'_{j_1}(x + e_k - e_i) - Z'_{j_1} x + Z_i^{j_2} - Z_k^{j_2} =$$

$$Z_k^{j_1} - Z_i^{j_1} + Z_i^{j_2} - Z_k^{j_2} \geq -2(L_{max} + (1/2)\gamma).$$

Therefore the probability of the event E_{j_1, j_2} is not bigger than the probability of the event

$$-2(L_{max} + (1/2)\gamma) \leq Z'_{j_2} x - Z'_{j_1} x \leq 0.$$

Fix any i such that $x_i > 0$. Since $Z_i^{j_1}$ is chosen uniformly from $[L_i^{j_1}, L_i^{j_1} + \frac{1}{2J}\gamma]$ or $[L_i^{j_1} + \frac{J-1}{2J}\gamma, L_i^{j_1} + \frac{1}{2}\gamma]$ (depending on whether $i \in \sigma_{j_1}$ or not) and both intervals have the same length $\frac{\gamma}{2J}$ then by conditioning on Z_{j_2} and $Z_l^{j_1}$, $l \neq i$ we have

$$\Pr\{-2(L_{max}+(1/2)\gamma) \leq Z'_{j_2}x - Z'_{j_1}x \leq 0 | Z_{j_2}, Z_l^{j_1},\, l \neq i\} \leq$$

$$\frac{\frac{2(L_{max}+(1/2)\gamma)}{x_i}}{\frac{\gamma}{2J}} = 2J\frac{2(L_{max}+(1/2)\gamma)}{x_i\gamma}.$$

We conclude

$$x_i \Pr\{E_{j_1 j_2}\} \leq \frac{4J(L_{max}+(1/2)\gamma)}{\gamma},$$

for all i such that $x_i > 0$. Note that this inequality holds trivially if $x_i = 0$. Then for any $z_j = (z_1^j, \ldots, z_n^j) \in O_j$ there holds

$$z'_j x \Pr_{\mathrm{u}}\{E_{j_1 j_2}\} \leq \sum_{i=1}^{n} z_i^j \frac{4J(L_{max}+(1/2)\gamma)}{\gamma} \leq$$

$$n(L_{max}+(1/2)\gamma)\frac{4J(L_{max}+(1/2)\gamma)}{\gamma} = \frac{4nJ(L_{max}+(1/2)\gamma)^2}{\gamma}.$$

Then

$$\Phi_s(x)\Pr_{\mathrm{u}}\{E_{j_1 j_2}\} \leq \frac{4nJ(L_{max}+(1/2)\gamma)^2}{\gamma}.$$

From (5.36) we obtain

$$\Pr_{\mathrm{u}}\{E_0\} \leq J(J-1)\frac{4nJ(L_{max}+(1/2)\gamma)^2}{\gamma\Phi_s(x)}.$$

Thus whenever

$$\Phi_s(x) > \frac{16J^2(J-1)n(L_{max}+(1/2)\gamma)^2(L_{max}+\gamma)}{\gamma^2},$$

there holds

$$\Pr_{\mathrm{u}}\{E_0\} \leq (1/4)\frac{\gamma}{L_{max}+\gamma}.$$

Note that

$$\frac{16J^2(J-1)n(L_{max}+(1/2)\gamma)^2(L_{max}+\gamma)}{\gamma^2} \leq \frac{16nJ^2(J-1)(L_{max}+\gamma)^3}{\gamma^2}.$$

This completes the proof of the lemma. □

Applying Lemma 5.4 to (5.35) we complete the proof of the proposition. □

5.4.2 Upper Bounds on Steady-State Queue Lengths in Networks Operating under an Arbitrary Work Conserving Policy

In this subsection we derive upper bounds on the tail distribution and expectation of queue lengths in multiclass exponential type queueing networks. These bounds hold under any Markovian work conserving policy and are finite if and only if the linear program GLP[md] (or LP[md] for two-station networks) has a feasible solution with positive γ. We will then derive closed form bounds which involve actual and virtual (to be defined) traffic intensities for two-stations re-entrant type queueing networks. Since the linear program LP[md] provides an exact test for the global stability of two-station fluid networks (see Chapter 2), our bounds for two-station networks are finite if and only if the fluid network is globally stable.

We have shown in Proposition 5.7 that whenever the linear program GLP[md] has a solution $L_1, L_2, \ldots, L_J, \gamma$ with positive γ, then a certain Lyapunov function Φ_s can be constructed with drift $(1/4)\gamma$ and a certain exception parameter B. In order to fully use the theory of Chapter 4 we need to show that the parameter $\nu_{max}^{\Phi_s}$ defined by (4.1) is finite. But as a direct corollary of Lemma 5.3, we have

$$\nu_{max}^{\Phi_s} \le L_{max} + (1/2)\gamma.$$

We now apply Theorem 4.1 from Chapter 4 to obtain the following result.

Theorem 5.4 *Given a multiclass exponential type queueing network* (α, μ, P, C), *with parameters rescaled so that* $\sum_{i=1}^{n} \alpha_i + \sum_{i=1}^{n} \mu_i = 1$, *suppose that the corresponding linear program GLP[md] has a feasible solution* $L_1, L_2, \ldots, L_J, \gamma$ *with positive* γ. *Then the following upper bounds hold on the stationary distribution* $\pi = \pi(w)$ *corresponding* *to any work conserving Markovian policy w*

$$\Pr_\pi\{\frac{L_j'Q(t)-B}{2(L_{max}+(1/2)\gamma)} \geq m\} \leq \Big(\frac{L_{max}+(1/2)\gamma}{L_{max}+(3/4)\gamma}\Big)^m,$$

for all $m = 0, 1, 2, \ldots,$ *and all* $j = 1, 2, \ldots, J$, *where*

$$L_{max} = \max_{1\leq i\leq n, 1\leq j\leq J}\{L_i^j\},$$

and

$$B = \frac{16nJ^2(J-1)(L_{max}+\gamma)^3}{\gamma^2}.$$

Also

$$E_\pi[L_j'Q(t)] \leq \frac{16nJ^2(J-1)(L_{max}+\gamma)^3}{\gamma^2} + \frac{8(L_{max}+(1/2)\gamma)^2}{\gamma},$$

for all $j = 1, 2, \ldots, J,$. *For networks with two stations, both bounds are finite if and only if the corresponding fluid network is globally stable.*

Proof:

The bounds are a direct corollary of Proposition 5.7, Theorem 4.1, equation (5.26) and the fact

$$\Phi_0(z_1, z_2, \ldots, z_J, x) \geq z_j'x \geq L_j'x,$$

for all $z_j \in O_j$, $j = 1, 2, \ldots, J$. We also use $p^\Phi_{max} \leq 1$. It was shown in Chapter 2 that a fluid network is globally stable if and only if the linear program LP[md] has a feasible solution with positive γ. Therefore for networks with two stations, the bounds are finite if and only if the corresponding fluid network is globally stable. □

We now proceed to derive bounds, which explicitly involve the traffic intensity parameters ρ_{σ_j}. Following Dai and Vande-Vate [20], we introduce the notions of *virtual station* and *virtual traffic intensity.* Loosely speaking, a virtual station is a set of classes such that no two of them can be served simultaneously, when a particular priority policy is implemented. In that sense every "real" station is also a virtual station. The *virtual traffic intensity* is defined as the sum of the traffic intensities of the classes in the virtual station, rescaled in a certain way (called "push starting" in [20]). The goal is to derive closed form upper bounds on the steady-state queue lengths which involve the traffic intensities of the real and virtual stations.

Suppose we are given a re-entrant line exponential type queueing network with two stations σ_1, σ_2. The arrival rate and the vector of service rates are given by λ and $(\mu_1, \mu_2, \ldots, \mu_n)$, respectively. Following Dai and Vande-Vate [20], a sequence of classes $e = \{i, i+1, \ldots, k\}$ is called an excursion in station σ_1 if all of these classes are from station σ_1 and classes $i-1$ and $k+1$ are not from station σ_1, (i.e. they are from station σ_2, or $i = 1$, or $k = n$). We call the class k the *last* class and denote it by $l(e)$. The classes $i, i+1, \ldots, k-1$ are called the *first* classes of the excursion e. Similarly, we define excursions, last classes, and first classes in station σ_2. For example in the Lu-Kumar network of Figure 5-1 the set of excursions is $\{1\}, \{2,3\}, \{4\}$. In network of Figure 5-3, the set of excursions is $\{1\}, \{2,3,4\}, \{5\}$.

We denote the sequence of all excursions by $e_1, e_2, \ldots, e_K$. Note that excursions alternate between stations.

Definition 5.1 *(Def. 2.2 in [20]) A set S of excursions is called a separating set if $e_k \in S$ implies $e_{k-1}, e_{k+1} \notin S$. Namely, S does not contain consecutive excursions.*

If $e_1 \subset \sigma_1$ and a separating set S satisfies $e_1 \notin S$, then S is called σ_1 strictly separating. Similarly, if $e_1 \subset \sigma_2$ and a separating set S satisfies $e_1 \notin S$, then S is called σ_2 strictly separating.

Two separating sets consisting of all the excursions in σ_1 and all the excursions in σ_2 are called trivial separating sets.

Notice that from the re-entrant line structure we have either $e \subset \sigma_1$ or $e \subset \sigma_2$. From now on, for concreteness, we assume that $e \subset \sigma_1$. Thus every separating set S is also σ_2 strictly separating.

Definition 5.2 *(Def. 2.4 in [20]) Given a separating set S consider a set of classes $V(S)$ consisting of classes in excursions of S together with first classes of excursions,*

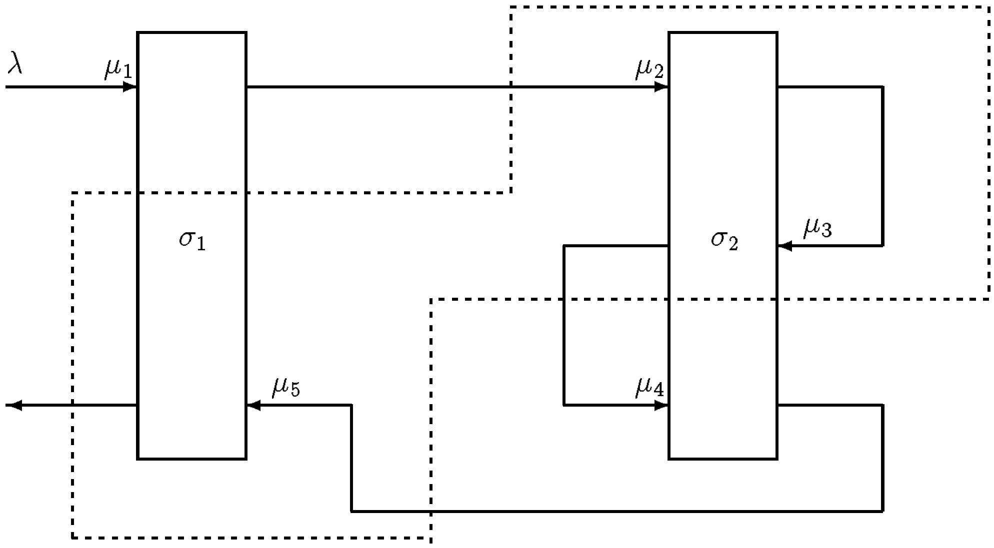

Figure 5-3: Classes 2,3 and 5 constitute a virtual station in a re-entrant line network

whose immediate predecessor is not in S :

$$V(S) = \cup_{k \in S} e_k \cup \Big(\cup_{k \notin S} \, e_{k+1} - \{l(e_k)\}\Big).$$

A set of classes $V \subset \{1, 2, \ldots, n\}$ *is called a virtual station if* $V = V(S)$ *for some separating set* S *and* V *does not contain classes from* e_1. *Thus* V *is a virtual station if*

1. *No classes of the first excursion are in* V: $e_1 \cap V = \emptyset$.

2. *If the last class of an excursion is in* V *then every class of that excursion is in* V, *and if a first class of an excursion is in* V *then every first class of that excursion is in* V.

3. *The last class of an excursion (except for the last excursion) is in* V *if and only if no classes of the next excursion is in* V.

For example, in the network of Figure 5-3 a set S consisting of a single excursion $e_3 = \{5\}$ is a separating set not containing the first excursion $e_1 = \{1\}$. Therefore the classes $V(S) = \{2, 3, 5\}$ constitute a virtual station. In the Lu-Kumar network of Figure 5-1 classes 2 and 4 constitute a virtual station.

Dai and Vande-Vate proved the following theorem (Theorem 2.1 in [20]) which provides explicit and exact conditions for the global stability of fluid networks with two stations.

Theorem 5.5 *A two station fluid re-entrant line queueing network is globally stable if and only if $\rho_{\sigma_1} < 1, \rho_{\sigma_2} < 1$, and for each excursion $e_k, 1 \le k \le K$ and separating set S*

$$\frac{\sum_{i\in\sigma_1\cap V(S), i>l(e_k)} \rho_i}{1-\sum_{i\in\sigma_1, i<l(e_k)} \rho_i} + \frac{\sum_{i\in\sigma_2\cap V(S), i>l(e_k)} \rho_i}{1-\sum_{i\in\sigma_2, i<l(e_k)} \rho_i} < 1. \tag{5.37}$$

The condition (5.37) will be referred to as a *virtual load* condition. Let

$$\rho(V(S),\sigma_1) \equiv \sum_{i\in\sigma_1\cap V(S)} \rho_i$$

$$\rho(V(S),e_k,\sigma_1) \equiv \sum_{i\in\sigma_1\cap V(S), i>l(e_k)} \rho_i \tag{5.38}$$

$$\rho(e_k,\sigma_1) \equiv \sum_{i\in\sigma_1, i<l(e_k)} \rho_i.$$

Similarly, we introduce $\rho(V(S),\sigma_2), \rho(V(S),e_k,\sigma_2)$ and $\rho(e_k,\sigma_2)$. Using these notations (5.37) becomes

$$\frac{\rho(V(S),e_k,\sigma_1)}{1-\rho(e_k,\sigma_1)} + \frac{\rho(V(S),e_k,\sigma_2)}{1-\rho(e_k,\sigma_2)} < 1$$

which is equivalent to

$$\rho(V(S),e_k) \equiv \rho(V(S),e_k,\sigma_1)(1-\rho(e_k,\sigma_2)) + \rho(V(S),e_k,\sigma_2)(1-\rho(e_k,\sigma_1)) +$$

$$\rho(e_k,\sigma_1) + \rho(e_k,\sigma_2) - \rho(e_k,\sigma_1)\rho(e_k,\sigma_2) < 1, \tag{5.39}$$

for all $k = 0,1,2\ldots,K$ and all separating sets S. We call this quantity *virtual traffic intensity* corresponding to separating set S and excursion e_k. It is not hard to see that the conditions of Theorem 5.5 can be restated as

$$\rho^* \equiv \max\Big\{ \max_{S,1\le k\le K} \{\rho(V(S),e_k)\}, \rho_{\sigma_1}, \rho_{\sigma_2} \Big\} < 1, \tag{5.40}$$

or in words "the maximum of all actual and virtual traffic intensities is less than one". Dai and Vande-Vate showed that this condition is "almost" necessary and sufficient for global stability of stochastic queueing networks as well. In particular, the following stochastic version of Theorem 5.5 was proven in [20].

Theorem 5.6 *A two station (stochastic) re-entrant line queueing network is globally stable if*

$$\rho^* \equiv \max_{S,e_k}\{\rho(V(S),e_k)\} < 1.$$

Moreover, if $\rho^* > 1$ then there exists an unstable priority policy.

We see that apart from a boundary case $\rho^* = 1$, the conditions of Theorem 5.5 are exact for global stability. For now, our goal is to derive closed form upper bounds on the steady-state number of customers in the network, in terms of the parameter ρ^*. We will show, in particular, that the expected number of customers in the network is of the order $O(\frac{1}{(1-\rho^*)^2})$.

The sufficiency part of Theorem 5.5 was proven by showing that if the conditions of the theorem hold, then a certain version of the linear program LP[md] has a feasible solution with positive γ. Our immediate goal is to prove that if the conditions of the Theorem 5.5 hold, namely, if $\rho^* < 1$, then there exists a feasible solution L, Q, γ of LP[md] such that $\gamma \geq \lambda(1-\rho^*)$.

The following linear program (parameterized by β) was considered in [20] (constraints (4.11)-(4.15), (5.1), (5.2)).

$$\begin{aligned}
\lambda\sum_{i\in\sigma_j} x_i - \mu_k/x_k + \lambda\epsilon &\leq 0 \qquad k\in\sigma_j,\ j=1,2, && (5.41)\\
\sum_{i\in\sigma_1, i>l(e)} x_i - \sum_{i\in\sigma_2, i\geq l(e)} x_i &\leq 0 \qquad \text{for any excursion } e\subset\sigma_2, && (5.42)\\
\sum_{i\in\sigma_2, i>l(e)} x_i - \sum_{i\in\sigma_1, i\geq l(e)} x_i &\leq 0 \qquad \text{for any excursion } e\subset\sigma_1, && (5.43)\\
\sum_{k\in\sigma_1} x_k + \epsilon &= 1, && (5.44)\\
\sum_{k\in\sigma_2} x_k + \epsilon &= \beta, && (5.45)\\
x, \epsilon &\geq 0. && (5.46)
\end{aligned}$$

We denote this linear program by LP[dv]. Linear programs LP[dv] and LP[md] are related to each other as the following proposition demonstrates.

Proposition 5.8 *Suppose LP[dv] has a feasible solution x, ϵ. Then LP[md] also has a feasible solution $L = (L_1, \ldots, L_n), Q = (Q_1, \ldots, Q_n), \gamma$ such that $\gamma = \lambda\epsilon$. The solution satisfies $L_1 \geq L_2 \geq \cdots \geq L_n$, $Q_1 \geq Q_2 \geq \cdots \geq Q_n$.*

Proof:

Given a solution x, γ to LP[dv] set $V = W = 0$

$$L_k = \sum_{i \in \sigma_1, i \geq k} x_i,$$

and

$$Q_k = \sum_{i \in \sigma_2, i \geq k} x_i.$$

This implies $L_k - L_{k+1} = x_k$, $Q_k - Q_{k+1} = 0$ if $k \in \sigma_1$, $k \leq n-1$ and implies $L_k - L_{k+1} = 0$, $Q_k - Q_{k+1} = x_k$ if $k \in \sigma_2$, $k \leq n-1$. In particular, $L_k \geq L_{k+1}$, $Q_k \geq Q_{k+1}$ for $1 \leq k \leq n-1$. Since $p_{ik} = 0$ when $k \neq i+1$, then the constraints (5.14), (5.16) are satisfied.

Fix any $k \in \sigma_2$, and let $e \subset \sigma_2$ be the excursion containing k. Then, we have

$$L_k - Q_k = \sum_{i \in \sigma_1, i > l(e)} x_i - \sum_{i \in \sigma_2, i \geq k} x_i.$$

Note that since $k \in \sigma_2$, we can write

$$L_k = \sum_{i \in \sigma_1, i > l(e)} x_i = \sum_{i \in \sigma_1, i > l(e)} x_i$$

But, since $k \leq l(e)$ then

$$Q_k = \sum_{i \in \sigma_2, i \geq k} x_i \geq \sum_{i \in \sigma_2, i \geq l(e)} x_i.$$

From the constraint (5.42), it follows $L_k \leq Q_k$ for all $k \in \sigma_2$. Similarly, we show that $L_k \geq Q_k$ for all $k \in \sigma_1$. This implies that constraints (5.17) and (5.18) are satisfied.

Finally, constraints (5.41) are equivalent to constraints (5.13) and (5.15) for $\gamma = \lambda\epsilon$. Thus $L = (L_1, \ldots, L_n)$, $Q = (Q_1, \ldots, Q_n), \gamma = \lambda\epsilon$ is a feasible solution to LP[md]. □

Note that constraints (5.41), (5.44), (5.45) and (5.46) of LP[dv] imply

$$x_i \geq \rho_i, \qquad i \in \sigma_1 \tag{5.47}$$

$$x_i \geq \beta\rho_i, \qquad i \in \sigma_2 \tag{5.48}$$

Recall that by assumption, $e_1 \subset \sigma_1$. By means of a network flow argument Dai and Vande-Vate proved the following result (see Sections 5, 6 in [20])

Proposition 5.9 *If the conditions of Theorem 5.5 hold (i.e., if $\rho^* < 1$), then there exists some β such that*

$$\frac{1 - \rho(V(S)\sigma_1)}{\rho(V(S), \sigma_2)} \geq \beta \geq \frac{\rho(V(S'), \sigma_1)}{1 - \rho(V(S'), \sigma_2)} \tag{5.49}$$

for all non-trivial σ_1 separating sets S' and all non-trivial separating sets S. Also for all such β and for $\epsilon(\beta)$ defined by

$$\epsilon(\beta) =$$

$$\min\left\{1-\rho_{\sigma_1}, 1-\rho_{\sigma_2}, \beta(1-\rho(V(S'), \sigma_2))-\rho(V(S'), \sigma_1), (1-\rho(V(S), \sigma_1))-\beta\rho(V(S), \sigma_2)\right\} \tag{5.50}$$

(where the minimum is over all σ_1 separating sets S' and all separating sets S) there exists a feasible solution x, ϵ, to LP[dv] with parameter β and with $\epsilon = \epsilon(\beta)$.

We use this to show that if the conditions of Theorem 5.5 hold, then there exists a feasible solution x, ϵ of LP[dv] satisfying $\epsilon \geq 1 - \rho^*$.

Lemma 5.5 *Suppose $\rho^* < 1$. Then, there exists β satisfying (5.49), such that $\epsilon(\beta) \geq 1 - \rho^*$. In particular, for this choice of β, there exists a feasible solution x, ϵ of LP[dv] such that $\epsilon \geq 1 - \rho^*$.*

Proof:

The proof is a slight modification of the argument in Section 6 of [20]. Choose any σ_1 separating set S' and any separating set S. Choose e_k as discussed in [20] (conditions a and b). We have from (5.39) and (5.40)

$$\frac{\rho(V(S'), e_k, \sigma_1)}{1-\rho(e_k, \sigma_1)} + \frac{\rho(V(S'), e_k, \sigma_2)}{1-\rho(e_k, \sigma_2)} + \frac{1-\rho^*}{(1-\rho(e_k, \sigma_1))(1-\rho(e_k, \sigma_2))} \leq 1 \tag{5.51}$$

By writing

$$1 = \frac{1-\rho(e_k, \sigma_2)}{1-\rho(e_k, \sigma_2)},$$

we obtain from above

$$\frac{\rho(V(S'), e_k, \sigma_1)}{1-\rho(e_k, \sigma_1)} + \frac{1-\rho^*}{(1-\rho(e_k, \sigma_1))(1-\rho(e_k, \sigma_2))} \leq \frac{1-\rho(V(S'), e_k, \sigma_2)-\rho(e_k, \sigma_2)}{1-\rho(e_k, \sigma_2)},$$

or

$$\frac{\rho(V(S'), e_k, \sigma_1)}{1-\rho(V(S'), e_k, \sigma_2)-\rho(e_k, \sigma_2)} + \frac{1-\rho^*}{(1-\rho(V(S'), e_k, \sigma_2)-\rho(e_k, \sigma_2))(1-\rho(e_k, \sigma_2))} \leq$$

$$\frac{1-\rho(e_k, \sigma_1)}{1-\rho(e_k, \sigma_2)}.$$

For this choice of e_k, it can be shown (see [20], Section 6) that

$$\rho(V(S'), e_k, \sigma_2) + \rho(e_k, \sigma_2) \geq \rho(V(S')\sigma_2),$$

and

$$\rho(V(S'), e_k, \sigma_1) = \rho(V(S'), \sigma_1).$$

Also $1-\rho(e_k, \sigma_2) \leq 1$. Therefore,

$$\frac{\rho(V(S'), \sigma_1)}{1-\rho(V(S'), \sigma_2)} + \frac{1-\rho^*}{1-\rho(V(S')\sigma_2)} \leq \frac{1-\rho(e_k, \sigma_1)}{1-\rho(e_k, \sigma_2)} \tag{5.52}$$

Writing (5.51) for the separating set S, we obtain

$$\frac{1-\rho(V(S), e_k, \sigma_1)-\rho(e_k, \sigma_1)}{1-\rho(e_k, \sigma_1)} \geq \frac{1-\rho^*}{(1-\rho(e_k, \sigma_1))(1-\rho(e_k, \sigma_2))} +$$

$$\frac{\rho(V(S), e_k, \sigma_2)}{1-\rho(e_k, \sigma_2)},$$

or

$$\frac{1-\rho(V(S), e_k, \sigma_1) - \rho(e_k, \sigma_1)}{\rho(V(S), e_k, \sigma_2)} \geq \frac{1-\rho^*}{\rho(V(S), e_k, \sigma_2)(1-\rho(e_k, \sigma_2))} + \frac{1-\rho(e_k, \sigma_1)}{1-\rho(e_k, \sigma_2)}.$$

Again, for this choice of e_k,

$$\rho(V(S), e_k, \sigma_1) + \rho(e_k, \sigma_1) \geq \rho(V(S)\sigma_1)$$

$$\rho(V(S), e_k, \sigma_2) = \rho(V(S), \sigma_2),$$

Also $1-\rho(e_k, \sigma_2) \leq 1$, then

$$\frac{1-\rho(V(S)\sigma_1)}{\rho(V(S), \sigma_2)} \geq \frac{1-\rho^*}{\rho(V(S), \sigma_2)} + \frac{1-\rho(e_k, \sigma_1)}{1-\rho(e_k, \sigma_2)}.$$

By combining this inequality with (5.52) we obtain

$$\frac{1-\rho(V(S)\sigma_1)}{\rho(V(S), \sigma_2)} - \frac{1-\rho^*}{\rho(V(S), \sigma_2)} \geq \frac{\rho(V(S'), \sigma_1)}{1-\rho(V(S'), \sigma_2)} + \frac{1-\rho^*}{1-\rho(V(S')\sigma_2)} \qquad (5.53)$$

Note that the left-hand side of this inequality depends only on the separating set S, and the right-hand side depends only on the separating set S'. Therefore there exists some β which is in between these two quantities for any S and S'. In particular, any such β satisfies (5.49). Also for any such β, from the inequality above, we obtain

$$1-\rho^* \leq \beta(1-\rho(V(S'), \sigma_2)) - \rho(V(S'), \sigma_1),$$

and

$$1-\rho^* \leq (1-\rho(V(S), \sigma_1)) - \beta\rho(V(S), \sigma_2).$$

By definition $1-\rho^* \leq 1-\rho_{\sigma_j},\ j=1,2$. Therefore, $1-\rho^* \leq \epsilon(\beta)$. This completes the proof of the lemma. $\square$

We now derive closed form upper bounds on the steady-state probability distribution and expectation of the number of customers in the network which hold under any Markovian work conserving scheduling policy. The bounds are expressed exclusively in terms of the traffic intensities ρ_i, ρ^* and $\rho_i^{\sigma_j +} = \sum_{k\in\sigma_j, k\geq i} \rho_k$.

Theorem 5.7 *We consider a reentrant line queueing network with n classes, two stations σ_1, σ_2, arrival rate λ, and service rates $\mu_1, \mu_2, \ldots, \mu_n$, with class 1 being from the station σ_1. If the conditions of Theorem 5.5 hold, namely, if $\rho^* < 1$, then, the following upper bounds hold on the steady-state number of customers in the network. Under any Markovian work conserving scheduling policy w, the corresponding stationary probability distribution $\pi = \pi(w)$ satisfies*

$$\Pr_\pi\Big\{\sum_{i=1}^{n} \rho_i^{\sigma_1 +} Q_i(t) - B \geq m\frac{1+\rho^*+2\rho^*\sum_{i=1}^n \rho^{-1}}{1+\sum_{i=1}^n \rho^{-1}}\Big\} \leq \Big(\frac{\frac{1}{2}+\frac{1}{2}\rho^*+\rho^*\sum_{i=1}^n \rho_i^{-1}}{\frac{3}{4}+\frac{1}{4}\rho^*+\rho^*\sum_{i=1}^n \rho_i^{-1}}\Big)^m \tag{5.54}$$

and

$$\Pr_\pi\Big\{\sum_{i=1}^{n} \rho_{l(e_2)+1}^{\sigma_1 +} \rho_i^{\sigma_2 +} Q_i(t) - B \geq m\frac{1+\rho^*+2\rho^*\sum_{i=1}^n \rho^{-1}}{1+\sum_{i=1}^n \rho^{-1}}\Big\} \leq \Big(\frac{\frac{1}{2}+\frac{1}{2}\rho^*+\rho^*\sum_{i=1}^n \rho_i^{-1}}{\frac{3}{4}+\frac{1}{4}\rho^*+\rho^*\sum_{i=1}^n \rho_i^{-1}}\Big)^m \tag{5.55}$$

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

$$E_\pi[\sum_{i=1}^{n} \rho_i^{\sigma_1 +} Q_i(t)] \leq \frac{64n(\rho^*\sum_{i=1}^n \rho_i^{-1})^3}{(1+\sum_{i=1}^n \rho_i^{-1})(1-\rho^*)^2} + \frac{2(1+\rho^*+2\rho^*\sum_{i=1}^n \rho_i^{-1})^2}{(1+\sum_{i=1}^n \rho_i^{-1})(1-\rho^*)}, \tag{5.56}$$

and

$$E_\pi[\sum_{i=1}^{n} \rho_{l(e_2)+1}^{\sigma_1 +} \rho_i^{\sigma_1 +} Q_i(t)] \leq \frac{64n(\rho^*\sum_{i=1}^n \rho_i^{-1})^3}{(1+\sum_{i=1}^n \rho_i^{-1})(1-\rho^*)^2} + \frac{2(1+\rho^*+2\rho^*\sum_{i=1}^n \rho_i^{-1})^2}{(1+\sum_{i=1}^n \rho_i^{-1})(1-\rho^*)}. \tag{5.57}$$

In particular,

$$E_\pi[\sum_{i=1}^{n} Q_i(t)] = O\Big(\frac{1}{(1-\rho^*)^2}\Big). \tag{5.58}$$

The bounds are finite if and only if the fluid queueing network is globally stable.

Remark. The bounds are asymmetric with respect to the order of the stations. If class 1 belongs to station σ_2, the corresponding bounds are obtained trivially by exchanging σ_1 and σ_2.

Proof:

We rescale the system parameters, without loss of generality, in the following way:

$$\bar{\lambda} = \frac{\lambda}{\lambda + \sum_{i=1}^{n} \mu_i}, \qquad \bar{\mu}_i = \frac{\mu_i}{\lambda + \sum_{i=1}^{n} \mu_i}, \; i = 1, 2, \ldots, n. \tag{5.59}$$

For the new set of parameters, there holds $\bar{\lambda} + \sum_{i=1}^{n} \bar{\mu}_i = 1$. Combining Proposition 5.8 with Lemma 5.5, we conclude: if $\rho^* < 1$, then there exists a feasible solution of the form $L, Q, \gamma = \bar{\lambda}\epsilon$ to LP[md] such that $\epsilon \geq 1 - \rho^*$. We now apply Theorem 5.4. Since, by assumption, class 1 is from station σ_1, then from the constraint (5.17) of LP[md], $L_1 \geq Q_1$. Using Proposition 5.8, we obtain for this solution $L_{max} = \max_{1 \leq i,k \leq n}\{L_i, Q_k\} = L_1$. From constraint (5.44) of LP[dv], $L_1 = \sum_{i \in \sigma_1} x_i = 1 - \epsilon$. Since $\bar{\lambda} \leq 1$ then

$$L_{max} + \frac{1}{2}\gamma = 1 - \epsilon + \frac{1}{2}\bar{\lambda}\epsilon \leq \rho^* + \frac{1}{2}\bar{\lambda}(1-\rho^*),$$

and

$$L_{max} + \gamma = 1 - \epsilon + \bar{\lambda}\epsilon \leq \rho^* + \bar{\lambda}(1-\rho^*).$$

Therefore,

$$\frac{(L_{max}+\gamma)^3}{\gamma^2} \leq \frac{(\rho^* + \bar{\lambda}(1-\rho^*))^3}{\bar{\lambda}^2(1-\rho^*)^2}.$$

It is also easy to check that from $\epsilon \geq 1 - \rho^*$ and $\bar{\lambda} \leq 1$, it follows that

$$\frac{L_{max} + (1/2)\gamma}{L_{max} + (3/4)\gamma} \leq \frac{\rho^* + (1/2)\bar{\lambda}(1-\rho^*)}{\rho^* + (3/4)\bar{\lambda}(1-\rho^*)}.$$

Also, for the solution $L = (L_1, L_2, \ldots, L_n), Q = (Q_1, Q_2, \ldots, Q_n)$, we obtain from (5.47) and (5.48) that

$$L_i = \sum_{k \in \sigma_1, k \geq i} x_i \geq \sum_{k \in \sigma_1, k \geq i} \rho_k = \rho_i^{\sigma_1 +}$$

and similarly

$$Q_i \geq \beta \rho_i^{\sigma_2 +}$$

Applying Theorem 5.4 to $L_1 = L$, we obtain

$$\Pr_\pi \Big\{ \frac{\sum_{i=1}^n \rho_i^{\sigma_1 +} Q_i(t) - B_0}{2(\rho^* + (1/2)\bar{\lambda}(1-\rho^*))} \geq m \Big\} \leq \Big(\frac{\rho^* + (1/2)\bar{\lambda}(1-\rho^*)}{\rho^* + (3/4)\bar{\lambda}(1-\rho^*)} \Big)^m,$$

where

$$B_0 = \frac{64n(\rho^* + \bar{\lambda}(1-\rho^*))^3}{\bar{\lambda}^2(1-\rho^*)^2},$$

and

$$E_\pi[\sum_{i=1}^n \rho_i^{\sigma_1 +} Q_i(t)] \leq \frac{64n(\rho^* + \bar{\lambda}(1-\rho^*))^3}{\bar{\lambda}^2(1-\rho^*)^2} + \frac{8(\rho^* + \frac{1}{2}\bar{\lambda}(1-\rho^*))^2}{\bar{\lambda}(1-\rho^*)}.$$

Observe that from (5.59)

$$\bar{\lambda} = \frac{1}{1 + \sum_{i=1}^n \rho_i^{-1}}.$$

Substituting, we obtain the bounds (5.54) and (5.56).

Applying Theorem 5.4 to $L_2 = Q$ we obtain

$$\Pr_\pi \Big\{ \frac{\sum_{i=1}^n \beta \rho_i^{\sigma_2 +} Q_i(t) - B_0}{2(\rho^* + (1/2)\bar{\lambda}(1-\rho^*))} \geq m \Big\} \leq \Big(\frac{\rho^* + (1/2)\bar{\lambda}(1-\rho^*)}{\rho^* + (3/4)\bar{\lambda}(1-\rho^*)} \Big)^m,$$

where again

$$B_0 = \frac{64n(\rho^* + \bar{\lambda}(1-\rho^*))^3}{\bar{\lambda}^2(1-\rho^*)^2},$$

and

$$E_\pi[\sum_{i=1}^n \beta \rho_i^{\sigma_2 +} Q_i(t)] \leq \frac{64n(\rho^* + \bar{\lambda}(1-\rho^*))^3}{\bar{\lambda}^2(1-\rho^*)^2} + \frac{8(\rho^* + \frac{1}{2}\bar{\lambda}(1-\rho^*))^2}{\bar{\lambda}(1-\rho^*)}.$$

But from the constraint (5) of LP[dv] $\beta > \sum_{i \in \sigma_2} x_i$. Recall that by assumption $e_1 \subset \sigma_1$

and $e_2 \subset \sigma_2$. Therefore from the constraint (2) of LP[dv]

$$\sum_{i \in \sigma_2, i \geq l(e_2)} x_i \geq \sum_{i \in \sigma_1, i > l(e_2)} x_i.$$

But then from (5.47)

$$\sum_{i \in \sigma_1, i > l(e_2)} x_i \geq \sum_{i \in \sigma_1, i > l(e_2)} \rho_i = \rho^{\sigma_1 +}_{l(e_2)+1}.$$

As a result $\beta > \rho^{\sigma_1 +}_{l(e_2)+1}$. By substituting β with $\rho^{\sigma_1 +}_{l(e_2)+1}$ in the bounds above and using (5.59), we obtain (5.55) and (5.57).

This completes the proof of the theorem. □

5.4.3 Examples

We now demonstrate the power of Theorems 5.4 and 5.7 on some examples.

- **Feedforward Networks.** Recall from Subsection 2.5.2 of Chapter 2 that a multiclass queueing network (α, μ, P, C) is feedforward (acyclic) if for any two stations σ_i, σ_j, with $i < j$, and any $k \in \sigma_i$, $l \in \sigma_j$ we have $p_{lk} = 0$. The stability of these networks under the usual load conditions $\rho_{\sigma_j} < 1$ was proven in [14] and [21]. We have also proven the stability of feedforward networks in Subsection 2.5.2 of Chapter 2, using the trajectory decomposition method.

  Since the stability conditions for feedforward networks are given explicitly as $\rho_{\sigma_j} < 1$, then it is natural to expect that performance bounds can be constructed, which are finite whenever the load condition $\rho_{\sigma_j} < 1$ holds. In the next theorem we will show that this is the case.

  Theorem 5.8 *Suppose we are given a feedforward queueing network (α, μ, P, C) with n classes and J stations $\sigma_1, \sigma_2, \ldots, \sigma_J$. Under any work conserving policy w, the following upper bounds hold on the steady-state number of customers in* *the network:*

$$\Pr_\pi\Big\{\frac{\bar{L}_j'Q(t)-B}{2(L_{max}^J+(1/2)\bar{\gamma})}\geq m\Big\}\leq\Big(\frac{L_{max}^J+(1/2)\bar{\gamma}}{L_{max}^J+(3/4)\bar{\gamma}}\Big)^m,$$

for all $m = 0,1,2,\ldots,$ *and all* $j = 1,2,\ldots,J$*, where* L_{max}^j *and* L_{min}^j *are the maximal and minimal positive components of* $C_j'M^{-1}[I-P']^{-1}$ *respectively,*

$$\bar{L}_j' = (\bar{L}_1^j,\bar{L}_2^j,\ldots,\bar{L}_n^j)' = \frac{1}{(J-1)^{J-j}}\frac{\prod_{l=j+1}^J L_{min}^l}{\prod_{l=j}^{J-1} L_{max}^l}C_j'M^{-1}[I-P']^{-1}$$

$$\bar{\gamma} = \min_{1\leq j\leq J}\frac{1}{(J-1)^{J-j}}\frac{\prod_{l=j+1}^J L_{min}^l}{\prod_{l=j}^{J-1} L_{max}^l}(1-\rho_{\sigma_j}),$$

and

$$B = \frac{16nJ^2(J-1)(L_{max}^J+\bar{\gamma})^3}{\bar{\gamma}^2}.$$

Also, there holds

$$E_\pi[\bar{L}_j'Q(t)]\leq\frac{16nJ^2(J-1)(L_{max}^J+\bar{\gamma})^3}{\bar{\gamma}^2}+\frac{8(L_{max}^J+(1/2)\bar{\gamma})^2}{\bar{\gamma}},$$

for all $j = 1,2,\ldots,J$*. In particular,*

$$E_\pi[\sum_{i=1}^n Q_i(t)] = O\Big(\frac{1}{(1-\rho)^2}\Big),$$

where $\rho = \max_{1\leq j\leq J}\rho_{\sigma_j}$*.*

Proof:

For each station σ_j consider the workloads introduced by customers in σ_j

$$L_j' = C_j'M^{-1}[I-P']^{-1},$$

where, as before, $M = diag\{\mu_1,\mu_2,\ldots,\mu_n\}$ and C_j is the j-th column of the matrix C. We have used $L_j'Q$ as lower Lyapunov functions in Section 5.3 to get lower bounds on steady-state number of customers in the network. We now consider a certain modification of the piecewise linear function

$$\Phi(x) = \max_{1 \le j \le J} \{L_j' x\},$$

which will be a Lyapunov function. This will allow us to use Theorem 5.4 to get upper bounds.

It is not hard to see that from the feedforward structure, for any $j_1 < j_2$ and any $k \in \sigma_{j_1}$, there holds

$$L_k^{j_2} = 0. \tag{5.60}$$

In words, customers in the station σ_{j_2} do not introduce any workload for station σ_{j_1}. For each j, let L_{max}^j and L_{min}^j be (as defined above) the maximal and minimal positive components of L_j. Namely,

$$L_{max}^j \equiv \max_{1 \le i \le n} \{L_i^j\}, \qquad L_{min}^j \equiv \min_{1 \le i \le n} \{L_i^j : L_i^j > 0\}. \tag{5.61}$$

Now, we rescale the vectors $L_1, L_2, \ldots, L_J$ as follows

$$\bar{L}_j = \frac{1}{(J-1)^{J-j}} \frac{\prod_{l=j+1}^{J} L_{min}^l}{\prod_{l=j}^{J-1} L_{max}^l} L_j, \tag{5.62}$$

for all $j = 1, 2, \ldots, J-1$.

Lemma 5.6 *$\bar{L}_j, \bar{\gamma}, V_j = 0$, $1 \le j \le J$ is a feasible solution to GLP[md].*

Proof:

First let us show that for this solution, the constraints (5.21) are satisfied with equality for $V_j = 0$, $j = 1, 2, \ldots, J$. Note that for each j and each $i \notin \sigma_j$

$$\sum_{k=1}^{n} \bar{L}_k^j \mu_i p_{ik} - \bar{L}_i^j \mu_i = \bar{L}_j' [P' - I] M e_i.$$

But from (5.62) and

$$\bar{L}_j'[P'-I]Me_i = \frac{1}{(J-1)^{J-j}} \frac{\prod_{l=j+1}^{J} L_{min}^l}{\prod_{l=j}^{J-1} L_{max}^l} L_j'[P'-I]Me_i$$

$$= \frac{1}{(J-1)^{J-j}} \frac{\prod_{l=j+1}^{J} L_{min}^l}{\prod_{l=j}^{J-1} L_{max}^l} C_j' M^{-1}[I-P']^{-1}[P'-I]Me_i$$

$$= \frac{1}{(J-1)^{J-j}} \frac{\prod_{l=j+1}^{J} L_{min}^l}{\prod_{l=j}^{J-1} L_{max}^l} C_j' e_i = 0.$$

The last equality follows since $i \notin \sigma_j$. Thus constraint (5.21) is satisfied with equality for $V_j = 0$. To prove the constraint (5.20) note that for each j and each $i \in \sigma_j$, by a similar argument

$$\sum_{k=1}^{n} \bar{L}_k^j \alpha_k + \sum_{k=1}^{n} \bar{L}_k^j \mu_i p_{ik} - \bar{L}_i^j \mu_i = \bar{L}_j' \alpha + \bar{L}_j'[P'-I]Me_i$$

$$= \frac{1}{(J-1)^{J-j}} \frac{\prod_{l=j+1}^{J} L_{min}^l}{\prod_{l=j}^{J-1} L_{max}^l} (C_j' M^{-1}[I-P']^{-1}\alpha + C_j' e_i)$$

$$= \frac{1}{(J-1)^{J-j}} \frac{\prod_{l=j+1}^{J} L_{min}^l}{\prod_{l=j}^{J-1} L_{max}^l} (\rho_{\sigma_j} - 1) \le -\bar{\gamma}.$$

Thus constraints (5.20) are satisfied.

We now prove the constraints (5.22). Select j_0 and $i \notin \sigma_{j_0}$. Let $i \in \sigma_j$. If $j < j_0$ then from (5.60) $\bar{L}_i^{j_0} = 0$ and (5.22) is satisfied automatically by non-negativity of the variables.

Suppose now $j > j_0$. We have from $L_{min}^j \le L_{max}^j$ that

$$\bar{L}_i^{j_0} = \frac{1}{(J-1)^{J-j_0}} \frac{\prod_{l=j_0+1}^{J} L_{min}^l}{\prod_{l=j_0}^{J-1} L_{max}^l} L_i^{j_0} \le \frac{1}{(J-1)^{J-j_0}} \frac{L_{min}^j}{L_{max}^{j_0}} \frac{\prod_{l=j+1}^{J} L_{min}^l}{\prod_{l=j}^{J-1} L_{max}^l} L_i^{j_0}.$$

But

$$\frac{L_{min}^j}{L_{max}^{j_0}} L_i^{j_0} \le L_i^j, \tag{5.63}$$

by the definition of L^l_{min}, L^l_{max}. Therefore

$$\bar{L}^{j_0}_i \leq \frac{1}{(J-1)^{J-j_0}} L^j_i \frac{\prod_{l=j+1}^J L^l_{min}}{\prod_{l=j}^{J-1} L^l_{max}} = \frac{1}{(J-1)^{j_0-j}} \bar{L}^j_i \leq \frac{1}{J-1} \bar{L}^j_i.$$

The constraint (5.22) follow from this inequality.

The last constraints are satisfied trivially since all the variables are nonnegative. □

We now apply Theorem 5.4. By definition $\bar{\gamma}$ is positive. Note, that Theorem 5.7 is not applicable here, since the number of stations is, generally, more than two. Observe, from (5.63), that for L_{max} defined by

$$L_{max} \equiv \max_{1 \leq k \leq n, 1 \leq j \leq J} \{\bar{L}^j_k\},$$

there holds

$$L_{max} \leq L^J_{max}.$$

Then from Theorem 5.4 we obtain the result. □

- **The Lu-Kumar Network.** Consider the network in Figure 5-1. The network is described by the following parameters

$$\alpha' = (\lambda, 0, 0, 0)', \ \mu' = (\mu_1, \mu_2, \mu_3, \mu_4)',$$

$$\rho_i = \lambda/\mu_i, \ i = 1, 2, 3, 4, \ \rho_{\sigma_1} = \rho_1 + \rho_4, \rho_{\sigma_2} = \rho_2 + \rho_3.$$

We have $\rho^{\sigma_1+}_1 = \rho_1 + \rho_4$ and $\rho^{\sigma_1+}_i = \rho_4$ for $i = 2, 3, 4$.

The set of excursions in this network is given as $e_1 = \{1\}, e_2 = \{2, 3\}, e_3 = \{4\}$. Then $\rho^{\sigma_1+}_{l(e_2)+1} = \rho_4$. The only two nontrivial separating sets in this network are composed of single excursions $e_1 = \{1\}$ and $e_3 = \{4\}$. The separating set e_1 with its set of classes $V(\{e_1\}) = \{1\}$ has $\rho(V(S), e_k) = 0$ for all $k = 1, 2, 3$. For the separating set e_3 with its set of classes $V(4) = \{2, 4\}$ we have

$$\rho(V(4), e_1) = \rho_2 + \rho_4, \qquad \rho(V(4), e_2) = \rho_1 + \rho_4.$$

But the second term is equal to $\rho_{\sigma_2} < 1$. We conclude that

$$\rho^* = \max\{\rho_2 + \rho_4, \rho_{\sigma_1}, \rho_{\sigma_2}\}.$$

Assume now, in addition, that $\rho_2 \geq \rho_3$ and $\rho_4 \geq \rho_1$. Then $\rho^* = \rho_2 + \rho_4$. Applying now Theorem 5.7 we obtain the following result

Proposition 5.10 *In a Lu-Kumar network satisfying $\rho_2 \geq \rho_3$, $\rho_4 \geq \rho_1$ and $\rho_2 + \rho_4 < 1$ the following upper bounds hold for any work conserving policy w and the corresponding stationary probability distribution π*

$$\Pr{}_\pi\Big\{(\rho_1 + \rho_4)Q_1(t) + \rho_4 Q_2(t) + \rho_4 Q_3(t) + \rho_4 Q_4(t) - B$$

$$\geq m \frac{1 + \rho_2 + \rho_4 + 2(\rho_2 + \rho_4)\sum_{i=1}^4 \rho_i^{-1}}{1 + \sum_{i=1}^4 \rho_i^{-1}}\Big\} \leq$$

$$\Big(\frac{\frac{1}{2} + \frac{1}{2}(\rho_2 + \rho_4) + (\rho_2 + \rho_4)\sum_{i=1}^4 \rho_i^{-1}}{\frac{3}{4} + \frac{1}{4}(\rho_2 + \rho_4) + (\rho_2 + \rho_4)\sum_{i=1}^4 \rho_i^{-1}}\Big)^m,$$

and

$$\Pr{}_\pi\Big\{\rho_4(\rho_2 + \rho_3)Q_1(t) + \rho_4(\rho_2 + \rho_3)Q_2(t) + \rho_4\rho_3 Q_2(t) - B \geq$$

$$m \frac{1 + \rho_2 + \rho_4 + (\rho_2 + \rho_4)\sum_{i=1}^4 \rho_i^{-1}}{1 + \sum_{i=1}^4 \rho_i^{-1}}\Big\} \leq$$

$$\Big(\frac{\frac{1}{2} + \frac{1}{2}(\rho_2 + \rho_4) + (\rho_2 + \rho_4)\sum_{i=1}^4 \rho_i^{-1}}{\frac{3}{4} + \frac{1}{4}(\rho_2 + \rho_4) + (\rho_2 + \rho_4)\sum_{i=1}^4 \rho_i^{-1}}\Big)^m,$$

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

$$B = \frac{256(\rho_2 + \rho_4)^3(\sum_{i=1}^4 \rho_i^{-1})^3}{(1 + \sum_{i=1}^4 \rho_i^{-1})(1 - \rho_2 - \rho_4)^2}$$

Also

$$E_\pi[(\rho_1+\rho_4)Q_1(t)+\rho_4Q_2(t)+\rho_4Q_3(t)+\rho_4Q_4(t)]$$
$$\le \frac{256(\rho_2+\rho_4)^3(\sum_{i=1}^4\rho_i^{-1})^3}{(1+\sum_{i=1}^4\rho_i^{-1})(1-\rho_2-\rho_4)^2}+\frac{2(1+\rho_2+\rho_4+2(\rho_2+\rho_4)\sum_{i=1}^4\rho_i^{-1})^2}{(1+\sum_{i=1}^4\rho_i^{-1})(1-\rho_2-\rho_4)}$$

and

$$E_\pi[\rho_4(\rho_2+\rho_3)Q_1(t)+\rho_4(\rho_2+\rho_3)Q_2(t)+\rho_4\rho_3Q_3(t)]$$
$$\le \frac{256(\rho_2+\rho_4)^3(\sum_{i=1}^4\rho_i^{-1})^3}{(1+\sum_{i=1}^4\rho_i^{-1})(1-\rho_2-\rho_4)^2}+\frac{2(1+\rho_2+\rho_4+2(\rho_2+\rho_4)\sum_{i=1}^4\rho_i^{-1})^2}{(1+\sum_{i=1}^4\rho_i^{-1})(1-\rho_2-\rho_4)}.$$

Similar bounds can be obtained for the cases $\rho_1>\rho_4$ or $\rho_3>\rho_2$.

- **3-classes 2-station network.** Consider a network described in Figure 5-2. The network is described by the parameters

$$\alpha'=(\lambda,0,0),\ \ \mu'=(\mu_1,\mu_2,\mu_3)',$$

$$\rho_i=\frac{\lambda}{\mu_i},\ \ i=1,2,3,\ \ \rho_{\sigma_1}=\rho_1+\rho_3,\ \ \rho_{\sigma_2}=\rho_2.$$

We also have $\rho_1^{\sigma_1+}=\rho_1+\rho_3, \rho_2^{\sigma_1+}=\rho_3^{\sigma_1+}=\rho_3\ \rho_1^{\sigma_2+}=\rho_2^{\sigma_2+}=\rho_2, \rho_3^{\sigma_2+}=0$.

The usual stability conditions are $\rho_1+\rho_3<1$ and $\rho_2<1$. The set of excursions for this network is given as $e_1=\{1\}, e_2=\{2\}, e_3=\{3\}$ and a simple check shows that it does not generate any extra conditions for stability. In fact, this also follows from Theorem 2.8 of Chapter 2. We conclude

$$\rho^*=\max\{\rho_1+\rho_3,\rho_2\}.$$

Assume now in addition that $\rho_1+\rho_3\ge\rho_2$. Then $\rho^*=\rho_1+\rho_3$. We also have $\rho_{l(e_2)+1}^{\sigma_1+}=\rho_3$. Applying Theorem 5.7 we obtain the following result.

Proposition 5.11 *If the network described in Figure 5-2 satisfies $\rho_1+\rho_3<1$ and $\rho_2\le\rho_1+\rho_3$, then the following upper bounds hold for any work conserving* *policy w and the corresponding stationary probability distribution π*

$$\Pr_\pi\Big\{(\rho_1+\rho_3)Q_1(t)+\rho_3Q_2(t)+\rho_3Q_3(t)-B \geq m\frac{1+\rho_1+\rho_3+2(\rho_1+\rho_3)\sum_{i=1}^{3}\rho^{-1}}{1+\sum_{i=1}^{3}\rho_i^{-1}}\Big\}$$

$$\leq \Big(\frac{\frac{1}{2}+\frac{1}{2}(\rho_1+\rho_3)+(\rho_1+\rho_3)\sum_{i=1}^{3}\rho_i^{-1}}{\frac{3}{4}+\frac{1}{4}(\rho_1+\rho_3)+(\rho_1+\rho_3)\sum_{i=1}^{n}\rho_i^{-1}}\Big)^m,$$

and

$$\Pr_\pi\Big\{\rho_3\rho_2Q_1(t)+\rho_3\rho_2Q_2(t)-B \geq m\frac{1+\rho_1+\rho_3+2(\rho_1+\rho_3)\sum_{i=1}^{3}\rho^{-1}}{1+\sum_{i=1}^{3}\rho_i^{-1}}\Big\}$$

$$\leq \Big(\frac{\frac{1}{2}+\frac{1}{2}(\rho_1+\rho_3)+(\rho_1+\rho_3)\sum_{i=1}^{3}\rho_i^{-1}}{\frac{3}{4}+\frac{1}{4}(\rho_1+\rho_3)+(\rho_1+\rho_3)\sum_{i=1}^{3}\rho_i^{-1}}\Big)^m,$$

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

$$B = \frac{192(\rho_1+\rho_3)^3(\sum_{i=1}^{3}\rho_i^{-1})^3}{(1+\sum_{i=1}^{3}\rho_i^{-1})(1-\rho_1-\rho_3)^2}$$

Also

$$E_\pi[(\rho_1+\rho_3)Q_1(t)+\rho_3Q_2(t)+\rho_3Q_3(t)] \leq$$

$$\frac{192(\rho_1+\rho_3)^3(\sum_{i=1}^{3}\rho_i^{-1})^3}{(1+\sum_{i=1}^{3}\rho_i^{-1})(1-\rho_1-\rho_3)^2}+\frac{2(1+\rho_1+\rho_3+2(\rho_1+\rho_3)(\sum_{i=1}^{3}\rho_i^{-1}))^2}{(1+\sum_{i=1}^{3}\rho_i^{-1})(1-\rho_1-\rho_3)}$$

and

$$E_\pi[\rho_3\rho_2Q_1(t)+\rho_3\rho_2Q_2(t)] \leq$$

$$\frac{192(\rho_1+\rho_3)^3(\sum_{i=1}^{3}\rho_i^{-1})^3}{(1+\sum_{i=1}^{3}\rho_i^{-1})(1-\rho_1-\rho_3)^2}+\frac{2(1+\rho_1+\rho_3+2(\rho_1+\rho_3)(\sum_{i=1}^{3}\rho_i^{-1}))^2}{(1+\sum_{i=1}^{3}\rho_i^{-1})(1-\rho_1-\rho_3)}.$$

5.4.4 Discussion

We have constructed in this section upper bounds on the steady-state number of customers in exponential type multiclass queueing networks, operating under an arbitrary Markovian work conserving policy. The bounds are finite whenever some piecewise linear Lyapunov function witnesses the global stability of the network. These results are the first that use piecewise linear Lyapunov function for performance analysis and thereby establish an interesting connection between the power of

Lyapunov functions to establish stability and to provide performance bounds.

One of the implications of the bounds above is that if the linear program GLP[md] for finding piecewise linear Lyapunov functions has a feasible solution with positive γ, (or for networks with two stations, if $\rho^* < 1$), then there exists a positive constant $c < 1$ such that for any work conserving policy w

$$\Pr_{\pi(w)}\{\text{total number of customers} \geq m\} \leq c^m.$$

Namely, the tail probabilities decay exponentially fast. This is the first result proving exponential decay of tail probabilities in multiclass queueing networks. In the following chapter, we extend these results for priority policies.

Some earlier results on performance of multiclass exponential type queueing networks involve quadratic Lyapunov functions. In particular, Kumar and Meyn in [38] have constructed a certain linear program, which, if feasible, establishes the global stability of the network. The data of this linear program involve the primary parameters of the network - arrival and service rates (routing probabilities are 0-1 since the authors considered a re-entrant line type model). It was shown that the dual of this linear program provides upper and lower bounds on expected queue lengths in steady-state. The upper bounds are finite if and only if the primal linear program is feasible and thereby establishes global stability.

The limitation of the quadratic Lyapunov function approach is two-fold. First, it is proven not to have an exact power of checking the global stability of queueing networks with two stations. In particular, a certain two-station queueing network was constructed in [21] for which there exists a piecewise linear Lyapunov function proving global stability, but the linear program corresponding to quadratic Lyapunov functions is infeasible. As a result, the dual of this linear program leads to infinite (and therefore inconclusive) upper bounds on the expected queue lengths. On the other hand, the upper bounds constructed in this section, as indicated in Theorem 5.4, are finite if and only if the maximal (actual or virtual) traffic intensity ρ^* is less than one, which was shown in [20] to be an exact stability condition.

The second limitation of the quadratic Lyapunov function approach is that, apart from special cases like feedforward networks in [33] (see also discussion in subsection 5.3.3), it only provides numerical bounds and only on the expected queue lengths. On the other hand, we have constructed in this section closed form bounds on tail probabilities of queue lengths. In particular, we have proved that tail probabilities decay exponentially fast and provided lower and upper bounds on the rates of decay.

A different approach to performance of multiclass queueing networks was proposed by Bertsimas and Nino-Mora in [3]. Using conservation laws in queueing networks they have constructed bounds equivalent to the ones provided by the quadratic Lyapunov function, but the approach is simpler and more intuitive. This raises the question whether there exists an analogous method to bypass the piecewise linear Lyapunov function approach and to derive the bounds constructed in this chapter directly.

An important limitation of the upper bounds constructed in this chapter seems to be the square exponent in the bound

$$E_\pi[\sum_{i=1}^{n} Q_i(t)] = O\Big(\frac{1}{(1-\rho^*)^2}\Big).$$

We conjecture that a stronger bound

$$E_\pi[\sum_{i=1}^{n} Q_i(t)] = O\Big(\frac{1}{1-\rho^*}\Big)$$

holds, which seems very likely to be the tight upper bound. In the following chapter we will see that this bound is achieved by some priority policy in some examples of multiclass queueing networks.

Notice that the bounds on expected steady-state queue lengths constructed in this chapter are of the polyhedral type. That is, they show that the set of vectors of steady-state expected queue lengths

$$P \equiv \Big\{(E_\pi[Q_1(t)], E_\pi[Q_2(t)], \ldots, E_\pi[Q_n(t)])\Big\}$$

corresponding to the family of work conserving policies, is a subset of a certain polyhedron P_0. We have proven that for networks with two stations the polyhedron P_0 is bounded if and only if the fluid network is globally stable. The situation is presented schematically on Figures 5-4 and 5-5 for two dimensional vectors of queue lengths $(E_\pi[Q_1], E_\pi[Q_2])$.

An important implication of this result is that for networks with two stations, global stability is equivalent to bounded performance region. For the special case of re-entrant line queueing networks these conditions are equivalent to an explicit condition $\rho^* < 1$.

It seems to be an intractable problem to calculate the performance region P exactly. It is not even clear whether it is polyhedron. A more tractable, but still very challenging problem, is getting exact global stability conditions for networks with more than two stations and checking whether global stability is also equivalent to finiteness of the performance region P. The finiteness of the performance region will be proven using the bounds of this chapter, if piecewise linear functions do have the power of checking exactly the global stability of networks with more than two stations.

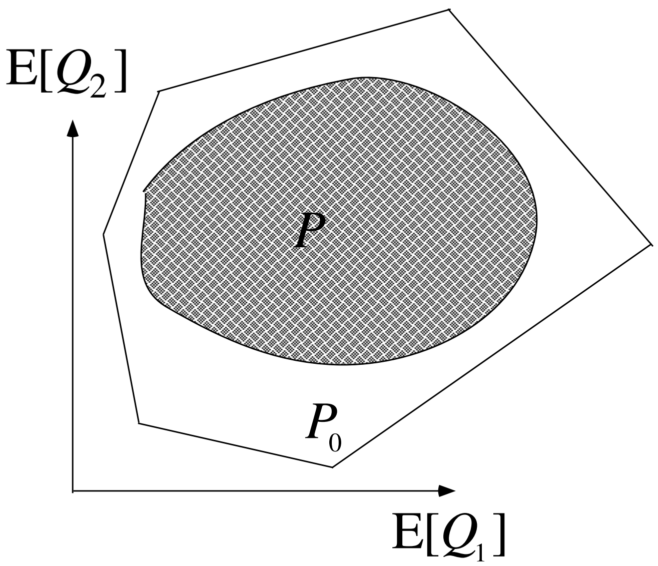

Figure 5-4: Bounded performance region. $\rho^* < 1$. Shaded area represents actual performance region. Polyhedron represents estimated performance region.

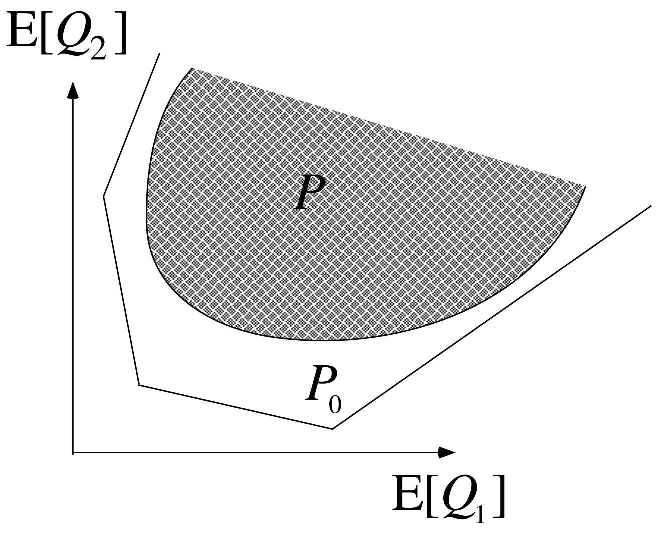

Figure 5-5: Unbounded performance region. $\rho^* > 1$. Shaded area represents actual performance region. Polyhedron represents estimated performance region.

Chapter 6

Performance of Priority Policies in Multiclass Exponential Type Queueing Networks

In this chapter, we develop bounds on the steady-state number of customers in queueing networks operating under priority policies. We will apply the techniques developed in Chapter 4. Linear and piece-wise linear Lyapunov functions will be used to develop lower and upper bounds on the tail probabilities and expectation of the number of customers in the network when a particular priority scheduling policy is implemented. In that sense, we refine the results of the previous chapter specifically for priority policies. We will introduce the notion of a K-virtual station, which generalizes the corresponding definition from the previous chapter. We will show that for policies that give priority to classes from some K-virtual station V, the expected number of customers in the network satisfies

$$E[\sum_{i=1}^{n} Q_i(t)] = \Omega(\frac{1}{K - 1 - \sum_{i \in V} \rho_i}).$$

The structure of this chapter is as follows. In the following section we describe preemptive priority policies in multiclass exponential type queueing networks. Priority policies lead to some special structure of the underlying uniformized Markov chain of the queue lengths $Q(\tau_s)$ in Z_+^n. In the subsequent two sections we will derive upper and lower bounds on steady-state queue lengths by means of piece-wise linear Lyapunov and linear Lower Lyapunov functions respectively. The lower bounds are closed form and are analogous to bounds provided by Theorem 5.3, adopted to virtual stations. The upper bounds are only numerical and are finite if and only if some piece-wise linear function witnesses stability of the priority policy.

We will conclude the chapter with some examples and discussion.

6.1 Priority Policies in Multiclass Exponential Type Queueing Networks

The model under consideration is a multiclass exponential type queueing network with n classes and J service stations $\sigma_1, \sigma_2, \ldots, \sigma_J$. The interarrival and service times are exponentially distributed. The vectors of arrival rates and service rates are given as $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ and $\mu = (\mu_1, \mu_2, \ldots, \mu_n)$ respectively. We again rescale the parameters so that $\sum_{i=1}^n \alpha_i + \sum_{i=1}^n \mu_i = 1$ and resort to uniformization as in the previous chapter. We consider the corresponding embedded Markov chain $Q(\tau_s)$ with its state space Z_+^n. In addition we assume that a certain priority policy w_θ corresponding to some priority rule θ is implemented. A priority rule θ can be described as a permutation of the set of classes

$$\theta : \{1, 2, \ldots, n\} \rightarrow \{1, 2, \ldots, n\}.$$

For two classes i, k which are from the same station σ_j we say that class i has a higher priority than class k if $\theta(i) < \theta(k)$. A corresponding priority policy w_θ can be described as follows: for each state $x \in Z_+^n$, $w_\theta(x)$ is a unique n-dimensional $0-1$ vector $(w_1^\theta(x), w_2^\theta(x), \ldots, w_n^\theta(x))$ satisfying

$$w_i^\theta(x) = 1$$

if and only if $x_i > 0$ and $x_k = 0$ for all $k \in \sigma(i)$ such that $\theta(k) < \theta(i)$. In words the policy w_θ at each transition epoch τ_s selects within each station the highest priority class with positive number of customers and works on customers from this class. The coordinates of the vector $w_\theta(x)$ do not change during the transition interval (τ_s, τ_{s+1}). They do change at epochs τ_s, if epochs correspond to an (external or internal) arrival or a departure of a customer. We assume that θ is a preemptive resume priority policy. Clearly policy w_θ is a Markovian policy. Formally, if $w_i^\theta(Q(\tau_s)) = 1$ and for some class $k \in \sigma(i)$ with $\theta(k) < \theta(i)$ $Q_k(\tau_{s+1}) = 1$ (an arrival occurred to class k) then $w_i^\theta(Q(\tau_{s+1}) = 1$.

6.2 Lower Bounds Based on a Linear Lower Lyapunov Function

In this section, we derive lower bounds on tail probabilities and expected number of customers in re-entrant line type queueing network, when the network operates under a certain priority policy w_θ. Much of the work is based on the concept of a virtual station and virtual traffic intensity introduced by Dai and Vande-Vate in [20] and described in detail in the previous chapter. However, the discussion in the previous chapter did not particularly illuminate the characteristic property of virtual stations - sharing of service resources among classes in *different* stations. Loosely speaking, virtual station is a collection of classes such that no two classes can receive service simultaneously when a particular priority policy is implemented.

We extend the definition of excursions from the previous chapter to networks with an arbitrary number of stations. A collection of classes $e = \{i, i+1, \ldots, k\}$ is defined to be an excursion if all these classes are from some station σ_j, but classes $i-1$ and $k+1$ are not from the station σ_j. This includes the possibility $i = 1$ or $k = n$. We denote the sequence of all excursions by $e_1, e_2, \ldots, e_L$. The definition of an K-virtual station is introduced as follows.

Definition 6.1 *Consider a re-entrant line type queueing network operating under* *a priority policy w_θ. Suppose a collection of stations $\Lambda \subset \{\sigma_1, \sigma_2, \ldots, \sigma_J\}$ with size $|\Lambda| = K$, and non-empty collections of classes $V_j \subset \sigma_j$, $j \in \Lambda$ are selected, so that classes of V_j have priority over classes in $\sigma_j - V_j$. Namely, for any $j \in \Lambda$ and any $i \in V_j, k \in \sigma_j - V_j$ there holds $\theta(i) < \theta(k)$. The set of classes $V = \cup_{j\in S} V_j$ is defined to be K-virtual station if the following conditions hold:*

1. *No classes of the first excursion are in V: $e_1 \cap V = \emptyset$.*

2. *If $k \in V$ and $\sigma(k-1) = \sigma(k)$ then $k - 1 \in V$.*

3. *If $k \in V$ and $\sigma(k-1) \neq \sigma(k)$ then $k - 1 \notin V$.*

We see that the definition of a virtual station is tied to a particular priority policy under consideration. Thus the set of classes $V = \cup_{j\in S} V_j$ satisfying conditions 1-3 can be a K-virtual station for one priority policy and not a virtual station for some other priority policy.

Notice that any virtual station $V(S)$, corresponding to some non-trivial separating set S, as introduced in the previous chapter for networks with two stations, satisfies the definition above for any priority policy that gives priority to classes in $V(S)$ over classes not in $V(S)$. However the definition above is less restrictive, since it allows that only a portion of the set of "first" classes $\{i, i+1, \ldots, k-1\}$ of some excursion $e = \{i, i+1, \ldots, k\}$ belongs to V.

The following theorem was proven in [7] and extends the result obtained by Dai and Vande-Vate in [20].

Theorem 6.1 *Given a re-entrant line type queueing network operating under a priority policy w_θ, suppose some collection of classes $V = \cup_{j\in\Lambda} V_j$ forms a K-virtual station. Then there exists a time, after which, with probability one not more than $K-1$ out of K sets V_j, $j \in \Lambda$ have positive amount of customers. Formally, there exists a (random) time $t_0 = t(Q(0))$, which depends on the initial state $Q(0)$, such that for all $t \geq t_0$*

$$\Pr\Big\{ \sum_{i\in V_j} Q_i(t) > 0, \, \forall j \in \Lambda \Big\} = 0,$$

where the probability is taken with respect to the sample paths of processes $Q(t)$ with initial state $Q(0)$. The result holds even if interarrival and service are generally distributed (with mild restriction on distribution functions).

We see that, in particular, for networks with two stations, if V is a 2 virtual station for a priority policy w_θ, then, after a certain transient period t_0, only one of the classes of V can receive service st a time. Thus V acts as a station sharing its resources among its classes. This justifies the name *virtual station.* The main idea of this section is to use this notion to build lower bounds on queue lengths of the type given by Theorem 5.3, with actual stations substituted with virtual ones.

One implication of Theorem 6.1 is that the sharing of resources among the classes in virtual station is preserved in steady state.

Corollary 6.1 *Suppose w_θ is a stable priority policy in a re-entrant line queueing network, and set of classes $V = \cup_{j\in\Lambda}V_j$ forms a K-virtual station. Then for the corresponding stationary distribution π_θ there holds*

$$\Pr_{\pi_\theta}\Big\{\sum_{i\in V} w_i^\theta(Q(t)) \le K-1\Big\} = 1.$$

Namely, in steady state at most $K-1$ of classes in V can receive service simultaneously.

Proof:

We consider our process $Q(t)$ which is in steady-state at time t. From Theorem 6.1 there exists a (random) time $t_0 \ge t$ after which, with probability one, at most $K-1$ out of J sets of classes $V_j, j\in\Lambda$ can have customers present. Since we consider the process in steady-state, the same is true at time t. Thus

$$\Pr_{\pi_\theta}\Big\{\sum_{j\in\Lambda} 1\{\sum_{i\in V_j} Q_i(t) > 0\} \le K-1\Big\} = 1,$$

where $1\{A\}$ is an indicator function of the event A. Since within each station σ_j we have the usual constraints

$$\sum_{i \in V_j} w_i^\theta(Q(t)) \le 1$$

and

$$\sum_{i \in V_j} w_i^\theta(Q(t)) = 0,$$

if $\sum_{i \in V_j} Q_i(t) = 0$, then we obtain by combining

$$\Pr_{\pi_\theta}\Big\{\sum_{i \in V} w_i^\theta(Q(t)) \le K-1\Big\} = 1.$$

This completes the proof. □

We now construct a certain linear function which is a Lower Lyapunov function of the embedded Markov chain $Q(\tau_s)$, whenever some priority policy is implemented and some $(K-1)$-virtual station is found. The methodology is very similar to the one in Chapter 5, Subsection 5.3.2. We consider the total workload brought by the customers in the classes of the virtual station V as a candidate for a Lower Lyapunov function. Thus, let $V = \cup_{j \in \Lambda} V_j$ be a $K-1$ virtual station in a re-entrant line queueing network, corresponding to some priority policy w_θ. For each $i = 1, 2, \ldots, n$ let

$$\rho(V) \equiv \sum_{k \in V} \rho_k, \tag{6.1}$$

$$\rho_i^{V+} \equiv \sum_{k \in V, k \ge i} \rho_k. \tag{6.2}$$

We now show that for $L = \frac{1}{\lambda}(\rho_1^{V+}, \rho_2^{V+}, \ldots, \rho_n^{V+})$ the linear function $\Phi(x) = L'x$ is a lower Lyapunov function.

Proposition 6.1 *Suppose w_θ is a stable priority policy in a re-entrant line exponential type queueing network. Suppose also that the set of classes $V = \cup_{j \in \Lambda} V_j$ forms a $(K-1)$-virtual station. Then for the corresponding embedded Markov chain $Q(\tau_s)$ and for*

$$L = \frac{1}{\lambda}(\rho_1^{V+}, \rho_2^{V+}, \ldots, \rho_n^{V+})$$

there holds

$$E[L'Q(\tau_{s+1})|Q(\tau_s)] - L'Q(\tau_s) \geq \rho(V) - (K-1).$$

for all $x \in Z_+^n$. In particular, $\Phi(x) = L'x$ is a Lower Lyapunov function with drift $K - 1 - \rho(V)$.

Proof:

From Proposition 5.1, Chapter 5, we have

$$E[L'Q(\tau_{s+1})|Q(\tau_s)] - L'Q(\tau_s) = \lambda\frac{1}{\lambda}\rho_1^{V+} + \sum_{i=1}^{n} \mu_i w_i^\theta(Q(\tau_s))\frac{1}{\lambda}(\rho_{i+1}^{V+} - \rho_i^{V+}),$$

where we assume for the convenience that $\rho_{n+1}^{V+} = 0$.

Note that $\rho_{i+1}^{V+} - \rho_i^{V+} = -\rho_i$ if $i \in V$ and $\rho_{i+1}^{V+} - \rho_i^{V+} = 0$ if $i \notin V$. Therefore

$$E[L'Q(\tau_{s+1})|Q(\tau_s)] - L'Q(\tau_s) = \rho_1^{V+} - \sum_{i \in V} w_i^\theta(Q(\tau_s)).$$

From Corollary 6.1

$$\Pr_{\pi_\theta}\Big\{\sum_{i \in V} w_i^\theta(Q(\tau_s)) \leq K - 1\Big\} = 1.$$

The result follows from this inequality. □

An interesting implication of the result above is that the queue lengths process is transient if the workload of the virtual station is not smaller than one. We formalize this in the following corollary.

Corollary 6.2 *Suppose w_θ is a priority policy in a re-entrant line exponential type queueing network. Suppose also that the set of classes V forms a $(K-1)$-virtual station such that $\rho(V) \geq K - 1$. Then the Markov chain $Q(\tau_s)$ is transient and the priority policy is unstable.*

Proof:

Consider the same as above linear Lower Lyapunov function $\Phi(x) = L'x$. For this

Lyapunov function we have shown in Proposition 6.1 that

$$E[L'Q(\tau_{s+1})|Q(\tau_s)] - L'Q(\tau_s) = \rho_1^{V+} - \sum_{i\in V} w_i^\theta(Q(\tau_s)).$$

In particular, if $\sum_{i\in V} Q_i(\tau_s) > 0$, then, from Corollary 6.1, it follows $\sum_{i\in V} w_i^\theta(Q(\tau_s)) \leq K-1$ and

$$E[L'Q(\tau_{s+1})|Q(\tau_s)] - L'Q(\tau_s) = \rho(V) - (K-1) \geq 0.$$

If however, $\sum_{i\in V} Q_i(\tau_s) = 0$ then $\sum_{i\in V} w_i^\theta(Q(\tau_s)) = 0$ and

$$E[L'Q(\tau_{s+1})|Q(\tau_s)] - L'Q(\tau_s) = \rho(V) > 0.$$

We see that at each transition the expected value of the function $\Phi(Q(\tau_s))$ is not decreasing and is increasing whenever states $\sum_{i\in V} Q_i(\tau_s) = 0$ are reached. Since these states are visited infinitely often, then $\Phi(Q(\tau_s)) \to \infty$ as $s \to \infty$.□

The result above supplements Theorem 5.9, proven by Dai and Vande-Vate in [20] for networks with two stations and generally distributed interarrival and service times.

We now turn our attention to deriving lower bounds on steady-state queue lengths, when traffic intensity of $(K-1)$-virtual station is less than $K-1$. We start from estimating the parameters ν_{min}^Φ and p_{min}^Φ, given by (4.2) and (4.4) and corresponding to our Lower Lyapunov function $\Phi(x) = L'x = \frac{1}{\lambda}(\rho_1^{V+}, \rho_2^{V+}, \dots, \rho_n^{V+})$.

Note that any transition of a Markov chain $Q(\tau_s)$ corresponding to a service completion does not increase the value of Φ since $\rho_1^{V+} \geq \rho_2^{V+} \geq \cdots \geq \rho_n^{V+}$. Also external arrivals can occur only to the first class. We conclude

$$\nu_{min}^\Phi = \frac{1}{\lambda}\rho_1^{V+} = \rho(V), \tag{6.3}$$

and

$$p_{min}^\Phi = \lambda. \tag{6.4}$$

We now state and prove the main result of this section.

Theorem 6.2 *Suppose we are given a J-station re-entrant line queueing network with parameters $\lambda, \mu_1, \ldots, \mu_n$, operating under some stable priority policy w_θ. If a set of classes V forms a $K-1$ virtual station, then the following lower bounds hold on distribution and expectation of the number of customers in the network*

$$\Pr_\pi\Big\{\frac{2\sum_{i=1}^n \rho_i^{V+} Q_i(t)}{\rho(V)} \geq m\Big\} \geq \Big(\frac{\rho(V)}{2(K-1)-\rho(V)}\Big)^m,$$

for each $j = 1, 2, \ldots, J$, and $m = 0, 1, 2, \ldots$.

Also,

$$E_\pi[\sum_{i=1}^n \rho_i^{V+} Q_i(t)] \geq \frac{1}{4}\frac{\rho^2(V)}{K-1-\rho(V)}.$$

Remarks.

1. Note that the bounds are insensitive to whether we have scaled parameters to have $\lambda + \sum_{i=1}^n \mu_i = 1$ or not, since the bounds are expressed exclusively in terms of load factors $\rho(V), \rho_i^{V+}$.

2. As a simple corollary of this theorem using $\rho_i^{V+} \leq \rho(V)$ we obtain

$$E_\pi[\sum_{i=1}^n Q_i(t)] \geq \frac{1}{4}\frac{\rho(V)}{K-1-\rho(V)}.$$

   It follows then that

$$E_\pi[\sum_{i=1}^n Q_i(t)] = \Omega\Big(\frac{1}{K-1-\rho(V)}\Big).$$

   The term $K-1-\rho(V)$ describes a new heavy-traffic behavior of the system, which has not been described before in the literature. It is not clear whether it exactly describes a heavy-traffic behavior, since exact conditions for stability of priority policies are not known in general.

Proof:

The result follows immediately from the second part of Theorem 4.1, Proposition 6.1 and observations (6.3), (6.4). □

6.3 Piece-Wise Linear Lyapunov Functions and Upper Bounds on Steady-State Queue Lengths

In this section we derive upper bounds on distribution and expectation of queue lengths in multiclass queueing networks operating under some priority policy. Our bounds will be exclusively numerical. They will be derived from solutions to some linear program which leads to a piece-wise linear Lyapunov function proving stability of priority policy. In particular, Meyn and Down in [21] have proposed a linear program as a test for stability of priority policy w_θ. We modify this linear program for networks with Bernoulli routing and require in addition all the variables $L_j = (L_1^j, L_2^j, \ldots, L_n^j), \gamma$ to be nonnegative. For each $i = 1, 2, \ldots, n$ let

$$P_i = \{k : k \in \sigma(i), \theta(k) < \theta(i)\}.$$

Consider the following linear program, which we denote by $LP_\theta[md]$.

$$\begin{array}{lcrr}
\sum_{k=1}^n L_k^i \alpha_k + U_i + \sum_{j:\sigma_j \neq \sigma(i)} V_{\sigma_j,i} & \leq & -\gamma & 1 \leq i \leq n \\
\sum_{k=1}^n L_k^i p_{lk}\mu_l - L_l^i \mu_l & \leq & U_i & 1 \leq i \leq n \; l \in P_i \\
\sum_{k=1}^n L_k^i p_{lk}\mu_l - L_l^i \mu_l & \leq & V_{\sigma_j,i} & 1 \leq i \leq n, \; l \in \sigma_j, \sigma_j \neq \sigma(i) \\
V_{\sigma_j,i} & \geq & 0 & i = 1, 2, \ldots, n, \; j = 1, 2, \ldots, J \\
\frac{1}{n-|P_i|} \sum_{k \notin P_i} L_l^k & \geq & L_l^s & 1 \leq i \leq n \; s \in P_i, \; l \notin P_i \\
L_i^l, V_{\sigma_j,i} & \geq & 0 & 1 \leq i, l \leq n, 1 \leq j \leq j
\end{array}.$$

This linear program takes into account a particular priority rule θ and is a direct analog of linear program GLP[md], described in the previous chapter. It was shown in [21] that if this linear program has a feasible solution with positive γ, then the following piece-wise linear function with n pieces

$$\Phi(x) = \max_{1 \leq i \leq n} L_i' x \tag{6.5}$$

is a Lyapunov function (in a fluid sense of Definition 2.3) proving stability of the priority policy w_θ. We will show that whenever $LP_\theta[md]$ has a feasible solution with positive γ, implying stability, certain finite upper bounds hold on steady-state tail probabilities and expectation of the queue lengths.

Again we start from "smoothing" the function (6.5) to make it a Lyapunov function for the Markov chain $Q(\tau_s)$. This is done exactly as in the previous chapter for general work conserving policies. Let $L_1, \ldots, L_n, \gamma$ be a feasible solution to $LP_\theta[md]$ with positive γ. Let also

$$L_{max} \equiv \max_{1 \le i,k \le n} \{L_i^k\}.$$

For all $i = 1, 2, \ldots, n$ set

$$O_i = \{z = (z_1, \ldots, z_n) \in R^n :$$

$$z_k \in [L_k^i + \frac{n-1}{2n}\gamma, L_k^i + 1/2\gamma], \text{ for } k \in P_i,\ z_k \in [L_k^i, L_k^i + \frac{1}{2n}\gamma], \text{ for } k \notin P_i\}. \quad (6.6)$$

Choose the uniform probability density function $p_i(z)$ on the set O_i. We will denote by Z_i a random variable chosen according to distribution $p_i(z)$ and denote by z_i a sample point from the set O_i.

For any $(z_1, z_2, \ldots, z_n) \in O_1 \times \ldots \times O_n$ and any $x \in R_+^n$ let

$$\Phi_0^\theta(z_1, z_2, \ldots, z_n, x) = \max_{1 \le i \le n} \{z_i' x\}, \quad (6.7)$$

and let

$$\Phi_s^\theta(x) = E_u[\Phi_0^\theta(Z_1, Z_2, \ldots, Z_n, x)] = \int \Phi_0^\theta(z_1, z_2, \ldots, z_n, x) p_1(z_1) \ldots p_n(z_n) dz_1 dz_2 \ldots dz_n. \quad (6.8)$$

The next proposition shows that our modified function Φ_s is a Lyapunov function. The proof is almost identical to the one of Proposition 5.7 and is omitted.

Proposition 6.2 *Let $L_1, L_2, \ldots, L_n, \gamma$ be any feasible solution of $LP_\theta[md]$ with $\gamma > 0$. Then Φ_s^θ is a Lyapunov function of the embedded Markov chain $Q(\tau_s)$, corresponding to a priority policy w_θ. The drift of the Lyapunov function is equal to $1/4\gamma$ and* *exception parameter is*

$$B = \frac{16n^3(n-1)(L_{max}+\gamma)^3}{\gamma^2}.$$

The proposition above shows that whenever the linear program $LP_\theta[md]$ has a solution $L_1, L_2, \ldots, L_n, \gamma$ with $\gamma > 0$, then a certain Lyapunov function Φ_s^θ can be constructed with drift $(1/4)\gamma$ and a certain exception parameter B. From (5.3), adopted to function Φ_s^θ we have

$$\nu_{max}^{\Phi_s^\theta} \le L_{max} + (1/2)\gamma.$$

We now apply Theorem 4.1 from Chapter 4 to obtain the following result.

Theorem 6.3 *Suppose we are given a multiclass exponential type queueing network (α, μ, P, C), operating under stable priority policy w_θ, and with parameters rescaled so that $\sum_{i=1}^n \alpha_i + \sum_{i=1}^n \mu_i = 1$. If the corresponding linear program $LP_\theta[md]$ has a feasible solution $L_1, L_2, \ldots, L_n, \gamma$ with positive γ, then the following upper bounds hold on the stationary distribution $\pi = \pi_\theta$ corresponding to priority policy w_θ*

$$\Pr_\pi\Big\{\frac{L_k' Q(t) - B}{2(L_{max}+(1/2)\gamma)} \ge m\Big\} \le \Big(\frac{L_{max}+(1/2)\gamma}{L_{max}+(3/4)\gamma}\Big)^m$$

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

$$L_{max} = \max_{1\le i,k\le n}\{L_i^k\}$$

and

$$B = \frac{16n^3(n-1)(L_{max}+\gamma)^3}{\gamma^2}.$$

Also

$$E_\pi[L_k' Q(t)] \le \frac{16nJ^2(J-1)(L_{max}+\gamma)^3}{\gamma^2} + \frac{8(L_{max}+(1/2)\gamma)^2}{\gamma},$$

for all $k = 1, 2, \ldots, n$.

Proof:

The bounds are a direct corollary of Proposition 6.2, Theorem 4.1 and the fact

$$\Phi_0^\theta(z_1, z_2, \ldots, z_n, x) \ge z_k' x \ge L_k' x,$$

for all $z_k \in O_k,\ k = 1, 2, \dots, n$. We also use $p^{\Phi}_{max} \leq 1$. $\square$

6.4 Examples

We now apply Theorems 6.2 and 6.3 to some examples of queueing networks and priority policies.

- **Lu-Kumar Network.** Consider again the network on Figure 5-1. Let priority rule θ be given as $\theta(4) < \theta(1), \theta(2) < \theta(3)$. Then according to Definition 6.1, $V = \{2, 4\}$ is a 2 virtual station corresponding to this priority policy. Applying Theorem 6.2 we obtain the following result.

  Theorem 6.4 *Consider a Lu-Kumar network on Figure 5-1 which operates under priority policy w_θ with priority rule $\theta(4) < \theta(1), \theta(2) < \theta(3)$. Under usual stability conditions $\rho_{\sigma_1}, \rho_{\sigma_2} < 1$ and additional condition $\rho_2 + \rho_4 < 1$ this policy is stable. Moreover, for the corresponding stationary probability distribution π_θ there holds for each $m = 0, 1, 2, \dots$.*

  $$\Pr_{\pi_\theta}\Big\{2\frac{(\rho_2 + \rho_4)Q_1(t) + (\rho_2 + \rho_4)Q_2(t) + \rho_4 Q_3(t) + \rho_4 Q_4(t)}{\rho_2 + \rho_4} \geq m\Big\} \geq$$

  $$\Big(\frac{\rho_2 + \rho_4}{2 - \rho_2 - \rho_4}\Big)^m,$$

  Also,

  $$E_{\pi_\theta}[(\rho_2 + \rho_4)Q_1(t) + (\rho_2 + \rho_4)Q_2(t) + \rho_4 Q_3(t) + \rho_4 Q_4(t)] \geq \frac{1}{4}\frac{(\rho_2 + \rho_4)^2}{1 - \rho_2 - \rho_4}.$$

  Remark. A simple corollary of this result is

  $$E_{\pi_\theta}[Q_1(t) + Q_2(t) + Q_3(t) + Q_4(t)] \geq \frac{1}{4}\frac{\rho_2 + \rho_4}{1 - \rho_2 - \rho_4},$$

  and thus supplements the upper bound result presented in the previous chapter for Lu-Kumar network operating under any work conserving policy. It also supplements the lower bound results obtained in Theorem 5.3 in the previous chapter for all stable Markovian policies.

Recall from the previous chapter, that for Lu-Kumar network $\rho^* = \max\{\rho_1 + \rho_4, \rho_2 + \rho_3, \rho_2 + \rho_4\}$. Therefore, for this priority policy

$$E_{\pi_\theta}[Q_1(t) + Q_2(t) + Q_3(t) + Q_4(t)] = \Omega\Big(\frac{1}{1-\rho^*}\Big).$$

This extends the result

$$E_{\pi_\theta}[Q_1(t) + Q_2(t) + Q_3(t) + Q_4(t)] = O\Big(\frac{1}{(1-\rho^*)^2}\Big),$$

which we proved holds under any work conserving policy.

Proof:

The stability of this policy follows from stability of all work conserving policies proved in the previous chapter in Section 5.4.3. The bounds are a direct corollary of Theorem 6.2. □

- Consider a re-entrant line queueing network with six classes and three stations described on Figure 1-1. Introduce priority rule θ with $\theta(4) < \theta(1)$, $\theta(2) < \theta(5)$, $\theta(6) < \theta(3)$. Then $V = \{2, 4, 6\}$ forms a 2-virtual station. Applying Theorem 6.2 we obtain the following result.

Theorem 6.5 *Consider a network on Figure 1-1 which operates under priority policy w_θ with priority rule $\theta(4) < \theta(1), \theta(2) < \theta(5), \theta(6) < \theta(3)$. Suppose in addition that the policy is stable. Then for the corresponding stationary probability distribution π_θ there holds*

$$\Pr_{\pi_\theta}\Big\{2\frac{(\rho_2+\rho_4+\rho_6)Q_1(t) + (\rho_2+\rho_4+\rho_6)Q_2(t) + (\rho_4+\rho_6)Q_3(t)}{\rho_2+\rho_4+\rho_6} +$$

$$\frac{(\rho_4+\rho_6)Q_4(t) + \rho_6 Q_5(t) + \rho_6 Q_6(t)}{\rho_2+\rho_4+\rho_6} \geq m\Big\} \geq$$

$$\left(\frac{\rho_2+\rho_4+\rho_6}{4-\rho_2-\rho_4-\rho_6}\right)^m,$$

for each $m = 0, 1, 2, \ldots$.

Also,

$$E_{\pi_\theta}[(\rho_2+\rho_4+\rho_6)Q_1(t)+(\rho_2+\rho_4+\rho_6)Q_2(t)+$$
$$(\rho_4+\rho_6)Q_3(t)+(\rho_4+\rho_6)Q_4(t)+\rho_6Q_5(t)+\rho_6Q_6(t)] \geq$$
$$\frac{1}{4}\,\frac{(\rho_2+\rho_4+\rho_6)^2}{2-\rho_2-\rho_4-\rho_6}.$$

One of the implications of Theorem 6.3 is that the tail probabilities of queue lengths have geometric bounds. Combining this with lower bounds given for general policies by Theorem 5.2 and Theorem 6.2 this allows us to conclude that given a priority policy w_θ, if the conditions of Theorem 6.3 hold, then for certain positive constants $c_1 < c_2 < 1$ there holds

$$c_1^m \leq \Pr_{\pi_\theta}\Big\{\text{total number of customers} \geq r\Big\} \leq c_2^m.$$

Namely, the steady-state probability that the number of customers is bigger than m decays exponentially fast in m.

6.5 Discussion

In this chapter we have constructed upper and lower bounds on distribution and expectation of queue lengths in queueing networks, operating under a specific priority policy. The lower bounds are closed form and are expressed in terms of traffic intensities of actual and virtual stations. The upper bounds provided by Theorem 6.3 are, on the other hand, only numerical. Partly, the reason for that is the absence of closed form and exact conditions for stability of priority policies. Such conditions (of the form $\rho^* < 1$) are only proven for general policies in networks with two stations. It seems that virtual stations is a very promising approach in establishing such conditions. The existing examples support the conjecture that priority policies are stable if and only if the corresponding virtual stations have traffic intensities less than one. If this is the case then the heavy traffic behavior of priority policies is probably described by

$$E_{\pi_\theta}[\sum_{i=1}^{n} Q_i(t)] = \max_V \Theta\Big(\frac{1}{K-1-\rho(V)}\Big).$$

One part of it is proven already by Theorem 6.2.

The existing results in the performance analysis of priority policies in multiclass queueing networks involve quadratic Lyapunov functions or conservation laws of queueing networks (see [37], [38], [33],[3]). The technique is very similar to the one used for the performance analysis of general work conserving policies. In most of the cases upper and lower bounds obtained are only numerical. In comparison, the lower bounds constructed in this chapter are closed form and capture a non-trivial interaction between the stations - the virtual station phenomenon. Also Theorem 6.5 provides a completely new heavy-traffic behavior, which has not been described before in the literature.

Chapter 7

Conclusions

The theory of multiclass queueing networks has attracted numerous researchers during the last decade. The research has been focused on the problems of stability, performance and optimization of multiclass queueing networks. Despite numerous efforts, the theory is far from being complete and the problems generally remain unsolved.

In this thesis, we have concentrated on two such problems - stability and performance. In the first part of the thesis, we have addressed the question of global stability of multiclass queueing networks - convergence to steady-state of the underlying stochastic process when an arbitrary work conserving policy is implemented. Using fluid models of multiclass queueing networks, we have proposed a trajectory decomposition approach for stability analysis. A certain linear program was constructed that provides an upper bound on the duration of the time until the fluid network empties. We have shown that linear programs with finite objective values correspond to globally stable networks. We have proven that this criterion for stability is both necessary and sufficient for networks with two stations. This is the first sharp stability result in such generality. This approach also provides sufficient conditions for global stability of networks with more than two processing stations. These sufficient conditions were shown to capture the global stability of feedforward (acyclic) networks.

We have modified the trajectory decomposition approach specifically for priority policies in multiclass fluid queueing networks. Again linear programming sufficient conditions for stability have been derived. We have proven that linear programs with finite objective values correspond to stable priority rules. Some early known and new stability results are proven using our trajectory decomposition approach.

In the second part of the thesis we have concentrated on the performance analysis of multiclass queueing networks. We have proposed a general methodology of using 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.

We have proven that whenever some piece-wise linear Lyapunov function witnesses global stability of the network, certain finite upper bounds hold on the probability distribution and expectation of queue lengths. Lower bounds are also constructed by means of a linear Lyapunov functions. Specifically we have constructed bounds of the form

$$c_1^m \leq \Pr\{\text{total number ofcustomers} \geq m\} \leq c_2^m$$

for some computable constants $0 < c_1 < c_2 < 1$. The bounds hold uniformly under any work conserving policy. These are the first results on steady-state *distribution* of queue lengths in networks of such generality, which prove exponential decay of the tail probabilities. Since piece-wise linear Lyapunov functions provide an exact test for stability of fluid networks with two stations then our bounds for networks with two stations are finite if and only if the fluid network is globally stable. An important corollary is that global stability is equivalent to bounded performance region (set of possible mean queue lengths, under arbitrary work-conserving policies) for networks with two stations.

Specifically, for re-entrant line type queueing networks with two processing stations the constants c_1 and c_2 can be expressed explicitly in terms of traffic intensities (actual and virtual) of the network. Closed form bounds were also constructed on the total expected number of customers in the network. In particular, we have proven that

$$E[\text{total number of customers}] = O\Big(\frac{1}{(1-\rho^*)^2}\Big)$$

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

We then concentrated on the performance analysis of queueing networks operating under some priority policy. Piece-wise linear and linear functions were used again to construct upper and lower bounds on steady-state queue lengths in such networks. The proposed lower bounds are in closed form and are expressed exclusively in terms of traffic intensities (actual and virtual) of the network. The upper bounds are numerical and are constructed using the solution of a linear program which establishes the existence of a piece-wise linear Lyapunov function.

The results proposed 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.

Bibliography

[1] D. Bertsimas and G. Mortzinou. Mutliclass queuing systems in heavy traffic: An asymptotic approach based on distributional and conservation laws. *Oper. Res.*, 45, 470-487, 1997.

[2] D. Bertsimas and J. Nino-Mora. Conservation laws, extended polymatroids and the multiarmed-bandit problem: a polyhedral approach to indexable systems. *Math. Oper. Res.*, 21, 257-306, 1996.

[3] 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.*, to appear.

[4] D. Bertsimas, I. Paschalidis and J. Tsitsiklis. Branching Bandits and Klimov's problem: achievable region and side constraints. *IEEE Trans. Automat. Control*, 40, 12, 2063-2075, 1995.

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

[6] D. Bertsimas and J. Tsitsiklis. *Introduction to Linear Optimization*. Belmont: Athena Scientific, 1997.

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

[8] P.P Bhattacharaya, L. Georgiadis and P. Tsoucas. Extended polymatroids: Properties and optimization. Research Report, IBM Research Division, T.J. Watson Research Center, Yorktown Heights, NY, 1992.

[9] A.A. Borovkov. Limit theorems for queueing networks. *J. Theory Probab. Appl.*,, 31 (3), 413-427, 1986.

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

[11] D. D. Botvich and A. A. Zamyatin. Ergodicity of conservative communication networks. Rapports de Recherche, no. 1772, INRIA, France, 1992.

[12] H. Chen. Fluid approximations and stability of multiclass queuing networks I: Work conserving disciplines. *Ann. Appl. Probab.*, 5, 637-666, 1995.

[13] H. Chen and H. Zhang. Stability of multiclass queueing networks under FIFO service discipline. Technical report, University of British Columbia, 1994.

[14] 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.

[15] J.G. Dai. A fluid-limit model criterion for instability of multiclass queueing networks. *Ann. Appl. Probab.*, 6, 751-757, 1996.

[16] J.G. Dai and T. Kurtz. A multiclass station with Markov feedback in heavy traffic. *Math. Oper. Res.* 20, 721-742, 1995.

[17] J.G. Dai and S.P. Meyn. Stability and Convergence of Moments for Multiclass Queueing Networks via Fluid Limit Models. *IEEE Transcation on Automatic Control*, 40 , 1889-1904, 1995.

[18] J.D. Dai and Y. Wang. Nonexistence of Brownian models of certain multiclass queueing networks. *Queueing Systems*, 13, 41-46, 1993.

[19] J. G. Dai and G. Weiss. Stability and instability of fluid models for certain reentrant lines. Technical report, Georgia Institute of Technology, 1994.

[20] J. G. Dai and J. H. Vande Vate. The Stability of Two-Station Fluid Networks. 1997. Preprint.

[21] D. D. Down and S. P. Meyn. Piecewise linear test functions for stability and instability of queueing networks. Technical report, University of Illinois at Urbana Champaigne, 1994.

[22] D.D. Down and S.P. Meyn. Stability of acyclic multiclass queueing networks. *IEEE Trans. Automat. Control*, 40, 5, 916-920, 1995.

[23] G. Fayolle. On random walks arising in queueing systems: Ergodicity and transience via quadratic forms as Lyapunov functions. Part I. *Queueing Systems*, 5, 167-184, 1989.

[24] G. Fayolle, V.A. Malyshev, M.V. Menshikov and A.F. Sidorenko. Lyapunov Functions for Jackson Networks. *Math. Oper. Res.*, 18, 916-927, 1993.

[25] A. Federgruen and H. Groenvelt. Characterization and optimization of achievable performance in queueing systems. *Oper. Res.*, 36, 733-741, 1988.

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

[27] J.C. Gittins. *Bandit Processes and Dynamic Allocation Indices* Wiley, New York, 1989.

[28] B. Hajek. Hitting-time and occupation-time bounds implied by drift analysis with applications. *Ann. Appl. Prob.*, 14, 502-525, 1982.

[29] J.M. Harrison and V. Nguen. Some badly behaved closed queueing networks. *Stochastic Networks*, IMA Volumes in Mathematics and Its Applications, F. Kelly and R. Williams (eds.), 71 Springer-Verlag, New York, 117-124, 1995.

[30] J.M. Harrison and L. Wein. Scheduling networks of queues: Heavy traffic analysis of a two-station closed network. *Oper. Res.* , 38, 1052-1064, 1990.

[31] J.M. Harrison and R. Williams. A multiclass closed queueing network with unconventional heavy traffic behavior. *Ann. Appl. Probab.*, 6, 1-47, 1996.

[32] H. Jin, J. Ou and P. R. Kumar. The throughput of irreducible closed Markovian queueing networks: Functional bounds, asymptotic loss, efficiency, and the Harrison-Wein conjectures. To appear in *Math. Oper. Res.*

[33] H. Jin, J. Ou and P. R. Kumar. The delay of open Markovian queueing networks: Uniform functional bounds, heavy traffic pole multiplicities, and stability. To appear in *Math. Oper. Res.*

[34] H. Kaspi and A. Mandelbaum. Regenerative closed queueing networks. *Stochastics*, 39, 239-258, 1992.

[35] G.P. Klimov. Time-sharing service systems I. *Theory Probab. Appl.*, 19, 1974.

[36] P.R. Kumar. Re-entrant lines. *Queueing Systems: Theory and Applications*, 13, 87-110, 1993.

[37] 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.

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

[39] P. R. Kumar and S. P. Meyn. Duality and linear programs for stability and performance analysis of queueing networks and scheduling policies. To appear in *IEEE Transactions on Automatic Control.*

[40] P.R. Kumar and T.I. Seidman. Dynamic instabilities and stabilization methods in distributed realtime scheduling of manufacturing stems. *IEEE Trans. Autom. Control*, 35, 3, 289-298, 1990.

[41] S. Lippman. Applying a new device in the optimization of exponential queueing systems. *Oper. Res*, 21, pp. 652-666, 1983.

[42] 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.

[43] V.A. Malyshev. Classification of two-dimensional positive random walks and almost linear semimartingales. *Dokl. Akad. Nauk SSSR*, 202, 526-528, 1972.

[44] V.A. Malyshev and M.V. Menshikov. Ergodicity, continuity and analyticity of countable Markov chains. *Trans. Moscow Math. Soc.*, 7, 1-48, 1981

[45] M.V. Menshikov. Ergodicity and transience conditions for random walks in the positive octant of space. *Soviet.Math.Dokl.*, 15, 1979.

[46] S.P. Meyn. Transience of queueing networks via fluid limit models. Technical report, University of Illinois at Urbana Champaigne, 1994.

[47] S. P. Meyn and D. Down. Stability of generalized Jackson networks. *Ann. Appl. Probab.*, 4, 124-148, 1994.

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

[49] I. Paschalidis. *Large Deviations in High Speed Communication Networks.* MIT Thesis, 1996.

[50] 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.

[51] A. Schwartz and A. Weiss. *Large Deviations for Performance Analysis*, Chapman and Hall, New York, 1995.

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

[53] J.G. Shanthikumar and D.D. Yao. Multiclass queueing systems: Polymatroid structure and optimal scheduling control. *Oper. Res.* 40, 293-299, 1992.

[54] K. Sigman. The stability of open queueing networks. *Stochastic Processes and their Applications*, 34, 11-25, 1990.

[55] P. Tsoucas. The region of achievable performance in a model of Klimov. Research Report, IBM Research Division, T.J. Watson Research Center, Yorktown Heights, NY, 1991.

[56] J. Walrand. *An introduction to queueing networks*, Prentice Hall, 1988.

[57] L. Wein. Optimal control of a two-station Brownian network. *Math. Oper. Res.* 15, 215-242, 1990.

[58] G. Weiss. Branching bandit processes. *Pro. Eng. Inform. Sci.*, 2, 269-278, 1988.