December 1992

LIDS-TH-2156

Research Supported By:

National Science Foundation
Grant ECS-8552419

PYI Award DDM-9158118
with matching funds from
Draper Laboratory

Army Research Office
ARO DAAL03-92-G-0309

Scheduling of Multiclass Queueing Networks: Bounds on Achievable Performance

Ioannis Ch. Paschalidis

December 1992

LIDS-TH-2156

Scheduling of Multiclasss Queueing Networks: Bounds on Achievable Performance

Ioannis Ch. Paschalidis

This report is based on the unaltered thesis of Ioannis Ch. Paschalidis submitted to the Department of Electrical Engineering and Computer Science in partial fulfillment of the requirements for the degree of Master of Science in Electrical Engineering and Computer Science at the Massachusetts Institute of Technology in December 1992 . This research was carried out at the Laboratory for Information and Decision Systems and supported in part by the National Science Foundation under Grant ECS-8552419, by a PYI award DDM-9158118 with matching funds from Draper Laboratory, andby the Army Research Office under grant DAAL03-92-G-0309.

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, MA 02139

Scheduling of Multiclass Queueing Networks : Bounds on Achievable Performance

by

Ioannis Ch. Paschalidis

Professional Diploma in Electrical and Computer Engineering
National Technical University of Athens (1991)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

December 1992

Author .....
Department of Electrical Engineering and Computer Science
December , 1992

Certified by .....
Dimitris Bertsimas, Associate Professor
Thesis Supervisor

Certified by .....
John N. Tsitsiklis, Associate Professor
Thesis Supervisor

Accepted by .....
Campbell L. Searle
Chairman, Departmental Committee on Graduate Students

Scheduling of Multiclass Queueing Networks : Bounds on Achievable Performance

by

Ioannis Ch. Paschalidis

Submitted to the Department of Electrical Engineering and Computer Science
on December , 1992, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering and Computer Science

Abstract

We consider a general model of an open multiclass queueing network with Poisson arrivals and exponentially distributed service times. The routing may be class dependent and can have deterministic or probabilistic character. Different classes can have different service requirements at each node of the network. The performance objective is to minimize a weighted sum of the expected response times of different classes. We propose a new method for finding a lower bound on achievable performance. Based mainly on conservation laws ideas, we derive a polyhedral space which includes the whole set of points with achievable response times (i.e. the achievable region) by stable and preemptive scheduling policies. Optimizing over this polyhedron, we derive a lower bound on achievable performance. We check its tightness by simulating heuristic scheduling policies. We prove that for the special case of single-station networks (multiclass queues and Klimov's model), this polyhedron is the achievable region. Moreover, the proposed method can be viewed as an extension of conservation laws to a general model of an open multiclass queueing network. In terms of computational complexity and in contrast to simulation-based existing methods, the calculation of the lower bound consists of solving an LP with both the number of variables and constraints being polynomial to the number of classes. In terms of the tightness of the bound, our method is at least as good as existing methods.

Thesis Supervisor: Dimitris Bertsimas
Title: Associate Professor

Thesis Supervisor: John N. Tsitsiklis
Title: Associate Professor

Acknowledgments

I wish to express my deepest appreciation to both my research advisors Prof. Dimitris Bertsimas and Prof. John Tsitsiklis. Their contribution throughout the development of this thesis, with ideas, support and constant encouragement was indispensable. They gave me the chance to work in the stimulating environment of MIT and they created a very friendly and supportive atmosphere. Few things can match John's sense of humor and Dimitris' enthusiasm. I definitely enjoyed our meetings, which revealed to me not only two comprehensive advisors but also two exceptional persons. For all that I am deeply indebted to them.

Special thanks to my graduate counsellor Prof. Dimitri Bertsekas for his encouragement and advice.

Lots of thanks to all the members of LIDS and of OR Center for their help and for the nice environment they created for me. Special thanks to Rodrigo who helped me with the simulation package.

On a personal note, I would like to thank my friends: Dimitris and Georgia, Angela, Helen and John Halaris for being my "family" here, Nicole for sharing the first day of her life with me, Lakis for being such a nice guy and Srimathy for helping me and teasing me when I was not in the mood. I cannot resist mentioning a group of very special people, my best friends. Thank you: Alina, Eleni, Regina, Apostole, Kosta, Niko, Viviana, Antoni, Spyro and Nano.

Last, but definitely not least, I wish to express my love to my parents Charalambos and Vasso. I owe everything to them and nothing can describe my feelings. They made my journey into life a wonderful experience. My love to "my Gina" for being the closest to me and for being always there for me.

This research was supported by the National Science Foundation under Grant ECS-8552419, by a PYI award DDM-9158118 with matching funds from Draper Laboratory, and by the ARO under grant DAAL03-92-G0309.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A *multiclass queueing network* is one that services multiple types of customers which may differ in their arrival processes, service requirements, route through the network as well as cost per unit of waiting time. A very important problem that often arises in such a system is to find out the optimal way to slip customers by the network while minimizing the expected sojourn time of each customer type. There are two kinds of decisions involved in this optimization problem: *sequencing* and *routing* decisions. A *sequencing policy* determines which type of customer to serve at each station of the network every moment in time; a *routing policy* determines which route each type of customer should follow to get through the network.

Numerous applications point out the importance of the problem described above. Packet-switching communication networks with different types of packets and different priorities between those packet-types as well as job shop manufacturing systems are among them. The scheduling control of a CPU in a multi-level programming computer system constitutes another application. In all these systems it is desirable to specify, if possible, or just to characterize an optimal strategy that minimizes a linear cost function of the expected waiting times in every queue of the network.

The control of multiclass queueing networks is a mathematically challenging problem. In order to achieve optimality, stations have to decide which of the competing

customer types to serve at each point in time, based on information about the load conditions of various other stations. Additionally, customers can choose their route through the network taking into account the current state of various queues. Those interactions between various queues create serious dependencies among them and prevent exact performance analysis which, whenever it can be done, is generally achieved through approximations. Moreover, optimizing a multiclass queueing network is an even harder problem. Thus, not surprisingly, simulation is the most common practice among researchers as a tool of evaluating heuristic policies.

The research community has not developed until now analytical tools to evaluate the closeness to optimality of proposed heuristic policies. The derivation of lower and upper bounds on achievable performance is the aim of this thesis. Tight lower bounds provide a good estimation of the proximity to optimality of the proposed policies. Upper bounds, other than the straw policies like first-come first-serve (FCFS), which come from heuristic scheduling policies restrict the feasible space and can probably give a relatively good intuition on the character of the optimal policy. Taking into consideration that a lower bound on achievable performance is not necessarily achievable by a specific policy, a policy within, say, 10 % of the lower bound is considered a very good policy.

1.1 Literature Review

A large variety of multiclass networks is presented in a survey paper by Kelly and Laws [KeLa]. Results from many researchers are provided when the network is in heavy traffic. In particular, extensive use is made of the so-called *Brownian network models* developed by Harrison and his co-authors Reiman and Wein [Harr, HaDa, Reim, HaWe, Wei1, Wei2, Wei3].

Perhaps, this heavy traffic scheduling approach, is one of the most successful approaches for controlling multiclass queueing networks. It proposes heuristic policies

which typically outperform more traditional policies. It has been more successful in closed networks and networks with controllable input, but has not been particularly successful in scheduling open networks which is the focus of this thesis. In the only study that concerns lower bounds, Ou and Wein in [OuWe] derive "pathwise" lower bounds for general open queueing networks with deterministic routing. By *pathwise* we mean that a bound on a measure of the "work to be done in the network" is derived for every sample path. They also calculate steady-state bounds by basically averaging over all sample paths. But, since their bounds are "pathwise", *simulation* is needed for their derivation.

Harrison and Wein in [HaWe] include some heavy-traffic scheduling results for a simple two-station network that we are also considering in this thesis (Chapter 2). Wein in [Wei1, Wei2] also includes heavy traffic-scheduling results in a networks where admission control is applied.

Kumar, in [Kuma] treats a category of manufacturing systems which he calls *re-entrant lines.* Some simple bounds are proposed, stability is proven for some policies and a number of policies are compared via simulation.

One very interesting approach, that stimulates the work in this thesis is to try to characterize the whole region of achievable performance in the problem of scheduling multiclass queueing networks. This region, which is constrained in the positive orthant, is defined such that every point in it corresponds to the performance of a valid policy. In addition, there is no valid policy with performance outside of the region. Having determined the achievable region, the scheduling problem reduces to a mathematical programming problem. More precisely, optimization of the objective function over this space yields the optimal performance. In [ShYa], Shantikumar and Yao follow this approach and study several variations of a multiclass queue. They are able to exactly characterize the achievable space and prove that, in the cases they studied, this space has a very special combinatorial structure; it is a *polymatroid*

polytope (see Chapter 4 for a definition). They also prove that the optimal policy is a strict priority rule. Their results partially extend to some queueing networks also, but under stringent restrictions. Namely, they assume that different types of customers are identically treated in the network, by having the same routing probabilities and the same service requirements at each station of the network.

In a recent work ([Tsou]) Tsoucas, derives the achievable region for a scheduling problem introduced by Klimov ([Klim]). This model is basically a multiclass queue with feedback. It turns out that in this problem, the optimal policy is again a strict priority policy that can be found by an algorithm initially given by Klimov. This algorithm is also derived in [Tsou] by different means.

1.2 Background Material; Conservation Laws

Since this thesis extends the notion of conservation laws in single-server systems to open multiclass queueing networks, in this section, we are discussing about the conservation laws (equalities and inequalities) in a single server multiclass queue.

The following discussion is mainly based in [GeMi, chap. 6]. Consider a multiclass single-server queue with N classes, where the server has unit speed. Interarrival and service times can be arbitrary. Let us define as *virtual load*, denoted by $V_S(t)$, the total amount of work in the system, working under scheduling strategy S, at time t. Let us also define a policy to be *work-conserving* if does not allow the server to be idle when there are jobs in the system and does not cause jobs to depart before they are finished. We are making the following assumptions:

- The stochastic process $V_S(t)$ has an equilibrium distribution with finite steady-state mean V_S.
- Service times are independent and identically distributed.
- The scheduling strategies we are considering are *non-anticipative* (i.e. only

information about the present and the past of the queueing process is used in making scheduling decisions).

The third assumption is necessary in order that the service time distribution of a job from a particular class, given that the job is in the system, is the same as the unconditional distribution. The second assumption asserts that the server cannot obtain information for future service times from past service times.

It can be seen from the definition of $V_S(t)$ that it is independent of the scheduling strategy S as long as we consider work-conserving policies. If by $V_{S,r}$ we denote the steady-state average virtual load of class r (i.e. the sum of the average remaining service times of the class r jobs present in the system) then this observation is equivalent to the equation:

$$\sum_{r=1}^{N} V_{S,r} = V \tag{1.1}$$

where V is a constant independent from the scheduling strategy. Equation (1.1) is the so-called conservation law and is true for every scheduling strategy S.

Let us now denote by g any non-empty subset of classes. If by V_S^g we denote the steady-state average virtual load of jobs whose class belongs to the set g and if we allow preemptive policies then there exists a scheduling strategy that minimizes V_S^g. Namely, it holds that:

$$V_S^g \geq V_{S^*}^g \tag{1.2}$$

where S^* is the strategy that assigns preemptive priority to the classes in g over the classes not in g. Equation (1.2) can be rewritten as:

$$\sum_{r \in g} V_{S,r} \geq V^g \tag{1.3}$$

where V^g is the steady-state average virtual load in a system where only classes in the set g arrive. Thus, (1.3) is a conservation inequality and it basically states that the best we can do for jobs whose class belongs in g is to give them preemptive priority

over the remaining jobs.

The space defined by (1.1) and (1.3) is a polytope that includes the achievable region. Moreover, since at each vertex of the polytope some of the inequalities (1.3) are satisfied with equality and V^g is achievable by a policy, it can be shown that every vertex corresponds to a policy. Since every point in the polytope can be written as a convex combination of the vertices, every point in the above polytope is achieved by a randomized policy. This argument proves that the space defined by (1.1) and (1.3) is the achievable region for work-conserving and preemptive policies that satisfy the assumptions introduced. Using these formulas one can derive the achievable region in an M/M/1 multiclass queue. In this case, it is also possible to calculate the constants V and V^g and thus to obtain the achievable region explicitly.

1.3 Problem Formulation

In this section we define the most general queueing network model that we are considering in this thesis. All the other models that we are using can be easily transformed to it.

Consider an open multiclass queueing network with N single server stations and R different job types. Jobs may change type as they move from one node to another. In particular, a job of type r, when completing service at node i goes to node j as a type s job with probability $p_{i,r;j,s}$ and leaves the network with probability $p_{i,r;0} = 1 - \sum_{j,s} p_{i,r;j,s}$. There are r independent Poisson streams of arrivals to the network, one for each type of customers. The Poisson arrival process for customers of type r has rate $\lambda_{0,r}$ and these customers join the i station with probability $q_{i,r}$. The pair (i,r) is called class and the class (i,r) requires an exponentially distributed service with rate $\mu_{i,r}$. Let $n_{(i,r)}(t)$ be the number of class (i,r) customers, present in the network at time t. The optimization problem is to determine a global scheduling policy that minimizes a linear cost function of the form $\sum_{(i,r)} c_{(i,r)} x_{(i,r)}$, $x_{(i,r)}$ being

the expected response time (waiting + service time) of class (i,r) and $c_{(i,r)}$ being given finite weights. Note that only sequencing decisions are involved. The routing probabilities, which are class dependent, are given and are not subject to optimization. Our objective is to derive a region that includes the region of achievable performance for this network model.

1.4 Assumptions

In this section we define the class of policies that we are considering valid. Throughout the thesis, unless explicitly stated otherwise, we are considering policies such that:

Assumption A *The stochastic process $n_{i,r}(t)$ has a unique invariant distribution with steady-state mean $n_{i,r}$, for every class (i,r) of customers.*

Assumption B *For every class (i,r) of customers, $E[n_{i,r}^2(t)] < \infty$, where the expectation is taken with respect to the invariant distribution.*

Assumption C *The scheduling strategies we are considering are* non-anticipative *(i.e. only information about the present and the past of the queueing process is used in making scheduling decisions).*

Assumption D *Preemption is allowed.*

Notice that we are not restricting ourselves to work-conserving policies.

1.5 Contributions of the Thesis

The main contribution of the thesis is that it proposes a new method, based mainly on conservation laws ideas (see [GeMi, ShYa]) and on ideas in [Kuma], to derive linear inequalities and equalities involving the mean response time of the different classes of customers in the network. These expressions define a polyhedron that contains the achievable region. Thus, optimization over this approximate region provides a lower bound on the achievable performance. We prove that our characterization is

exact for the well known case of the multiclass queue [GeMi, Klv2] and for the case of the Klimov's problem (see [Klim, Tsou]) [1]. As a result, our approach can be seen as a natural extension of conservation laws to multiclass queueing networks. Our technique includes the case of both deterministic and probabilistic routing. To the best of our knowledge, this is the first attempt to apply conservation ideas to an open multiclass queueing network with a general structure.

In the examples that we studied, we found that the tightness of our lower bound on achievable performance is, approximately, in the the same order of magnitude as "pathwise" bounds derived in [OuWe] with a technique that needs a simulation experiment for the calculation of the bound. Moreover, our lower bound can be computed in a number of steps which is a polynomial function of the number of classes in the network. Namely, the calculation of the lower bound consists of solving a linear programming problem with $O(n^2)$ variables and $O(n^2)$ constraints, n being the number of classes in the network.

1.6 Detailed Summary of Results and Organization of the Thesis

The rest of the thesis is organized as follows:

In Chapter 2, in order to illustrate our approach, we start with a well-studied, simple, open network. We derive the conservation equalities and inequalities mentioned in the previous section. These equations, along with some more based on different ideas, define an approximate performance region for the problem. Optimization over this region yields the lower bound on achievable performance. We also discuss some heuristic policies whose performance yields an upper bound. At the end of the chapter, we consider some simple limiting cases for the traffic parameters of the network

[1] Since in the literature these problems are studied within the class of work-conserving policies, in order to establish the equivalence we restrict ourselves to work-conserving policies.

and show that for policies which are "intuitively" optimal we can prove asymptotic optimality via our lower bounds.

In Chapter 3, we apply the bounding method illustrated in Chapter 2 to a general open multiclass network with Poisson arrivals and exponentially distributed service times, where different classes have the same service requirement at each node of the network. The method basically consists of writing $2^n - 1$ inequalities, n being the total number of classes in the network. We also slightly modify this method to include networks with probabilistic routing and different service requirements among classes at each node of the network.

In Chapter 4, we prove that we can get the exact characterization for an M/M/1 multiclass queue and for Klimov's problem [2] with Poisson arrivals and exponentially distributed service times. Thus, we justify our claim that the proposed bounding method can be viewed as an extension of conservation laws to networks. Moreover, it provides an alternative method for deriving the achievable spaces for the multiclass queue and Klimov's model. In particular, we derive the conservation equalities and inequalities for both models and prove that the region they define is the achievable region. For the multiclass queue the result is also proven in [GeMi] using a different approach. For Klimov's problem, we prove that the achievable region we get has the same structure as the one described in [Tsou]. This is a new result since Tsoucas considers general interarrival and services times along with non-preemptive policies. He provides only the structure of the achievable region by deriving it in terms of an arbitrary function, while we are deriving it explicitly.

In Chapter 5, we propose a refined bounding method that yields our tightest lower bound. The method consists of writing only conservation equalities. In addition, it makes the numerical calculation of the lower bound less burdensome, computationally. Moreover, we provide a proof, for the case of the multiclass queue, that the

[2] in this chapter we restrict ourselves to work-conserving policies

refined method is consistent with the earlier one. Namely, we prove that the polyhedral space derived in Chapter 4 for the multiclass queue is a projection of the polyhedral space derived via the refined method. This is an interesting result from a purely combinatorial point of view.

In Chapter 6, both methods are applied to three specific network examples considered in the literature and numerical results are obtained for various traffic conditions. We believe that these experiments demonstrate that the methods proposed in this thesis yield in most of the cases adequately tight lower bounds.

Finally, in Chapter 7, we conclude and we mention some open problems.

Chapter 2

A Simple Two-Station Network

Our goal in this chapter is to illustrate our methodology in a relatively simple example and to gain some useful insights.

Consider the network, with two types of customers, depicted in Figure 2-1. Type 1 customers visit stations 1 and 2, in that order, before exiting the network and type 2 customers visit only station 1 before exiting the network. Both arrival processes are Poisson with rates λ_1, λ_2 for customers of type 1,2 respectively. Service times at stations 1 and 2 are exponentially distributed with parameters μ_1, μ_2 respectively. In order to ensure that at least one stable policy exists, we assume that $\lambda_1 + \lambda_2 < \mu_1$ and $\lambda_1 < \mu_2$. We define *class 1 customers* to be type 1 customers at station 1, *class 2 customers* to be type 2 customers at station 1 and *class 3 customers* to be type 1 customers at station 2. Let x_i, $i = 1, 2, 3$, be the expected response time (waiting + service time) of class i customers. Note that the existence and finiteness of these expected response times is guaranteed by Assumption A in Sec. 1.4. The optimization problem is to determine a scheduling policy at station 1 that minimizes a linear cost function of the form $\sum_{i=1}^{3} c_i x_i$ where c_i are given finite weights.

In [HaWe], Harrison and Wein examine this particular network within the Brownian motion model framework when interarrival and service times have a general distribution. Their objective function being the total expected number of customers

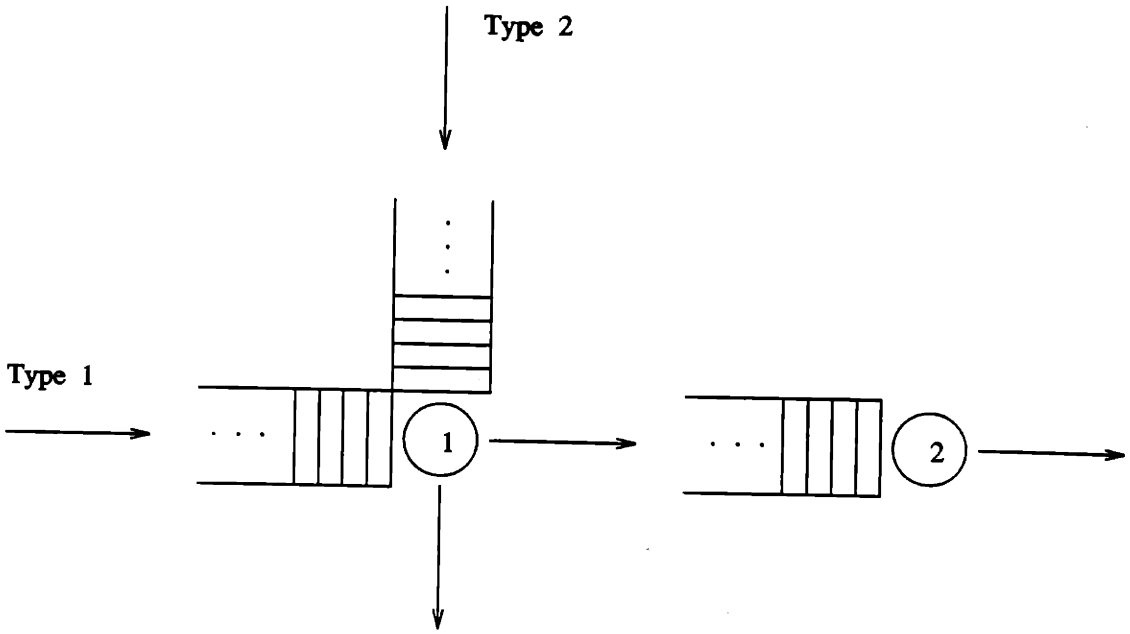

Figure 2-1: A simple two-station network

in the network, they derive a pathwise lower bound on achievable performance and propose a *threshold policy*. They then show via simulation that the relative difference between the performance of the proposed policy and the lower bound becomes small as the load is increased towards the heavy traffic limit. In [ChYY], Chen et al. follow a *stochastic intensity control* approach. They model the arrival and service processes as counting processes with controllable stochastic intensities and they try to schedule the server in station 1 in order to minimize a discounted cost function over an infinite time horizon. They establish a switching curve structure and they prove the optimality of simple policies in some specific cases. They also propose heuristic policies for other more difficult cases.

In this chapter we will at first derive lower bounds by :

1. A unified approach which yields $2^3 - 1$ *conservation inequalities* (that is, one inequality for each non-empty subset of the three classes). We refer to them as *conservation inequalities* because they are based on the conservation of the work to be done in the network.

2. Using some more ideas, such as stability conditions, which properly exploit the special structure of this network but can also be extended to the more general setting of open multiclass queueing networks.

We will then derive several upper bounds and prove, for some special load conditions, asymptotic optimality of certain policies, using the lower bounds.

2.1 Lower Bounds

2.1.1 Conservation Inequalities

Let $n_i(t), i = 1, 2, 3$ be the number of customers of class i present in the network at time t. Consider the following potential function :

$$R^S(t) = \sum_{i \in S} f^S(i) n_i(t) \tag{2.1}$$

where S is a subset of classes and $f^S(i)$ are positive multiplying constants, depending on the class and on the subset S, which we hereafter call *f-parameters*. This function can be thought as a measure of the work to be done in the network at time t. Using a probabilistic argument we obtain a lower bound on the mean "amount of work to be done". The results are summarized in the following theorem:

Theorem 2.1 *For the specific network we are considering in this chapter and for every policy satisfying the assumptions introduced in Section 1.4 the following inequalities hold:*

$$\lambda_1 x_1 + \lambda_2 x_2 \geq \frac{\lambda_1 + \lambda_2}{\mu_1 - \lambda_1 - \lambda_2} \tag{2.2}$$

$$x_1 \geq \frac{1}{\mu_1 - \lambda_1} \tag{2.3}$$

$$x_2 \geq \frac{1}{\mu_1 - \lambda_2} \tag{2.4}$$

$$x_3 \geq \frac{1}{\mu_2} \tag{2.5}$$

$$x_1 + x_3 \geq \frac{1}{\mu_2 - \lambda_1} \tag{2.6}$$

$$\lambda_2 x_2 + \lambda_1 x_3 \geq \frac{\lambda_1 + \lambda_2}{\mu_1 + \mu_2 - \lambda_2} \tag{2.7}$$

$$2\lambda_1 x_1 + \lambda_2 x_2 + \lambda_1 x_3 \geq \frac{3\lambda_1 + \lambda_2}{\mu_1 + \mu_2 - 2\lambda_1 - \lambda_2} \tag{2.8}$$

Proof : We shall apply a common technique and force servers to work continuously even when the station is empty. Thus, at service completions customers depart only if they are present. This modification does not alter the behavior of the system because of the memoryless property of the exponential distribution. In particular, when an arriving customer finds the server working on a fictitious customer, his service time is the residual life of the service time for the fictitious customer which is still exponentially distributed. Let τ_n be the sequence of times immediately after an arrival or a service completion (fictitious or real). Let also denote by $1\{\cdot\}$ the indicator function; that is, $1\{A\} = 1$ if event A occurs and zero otherwise. In addition, by σ_{τ_n}, we denote the σ-field generated by events up to and including time τ_n or, intuitively, the previous history. Finally, without loss of generality, let

$$\lambda_1 + \lambda_2 + \mu_1 + \mu_2 = 1$$

in order to make notation easier. Note that what we are doing is that we uniformize the underlying Markov chain corresponding to the network. Events occur according to a common "Poisson clock" of rate 1 and the self-transitions correspond to departures of fictitious customers. We are going to apply this uniformization again in Chapter 3 where we extend this method to open multiclass networks where classes may have different service requirements at each node. Next, we demonstrate the derivation of the bounds given in the theorem's statement, for each specific choice of the subset S. Therefore, dropping S from $R^S(t)$ and $f^S(i)$, we obtain:

- for $S = \{1, 2\}$:

$$E[R^2(\tau_{n+1}) \mid \sigma_{\tau_n}] \;=\; \lambda_1(R(\tau_n) + f(1))^2 + \lambda_2(R(\tau_n) + f(2))^2 +$$

$$\mu_1 1\{\text{server 1 busy from class 1 at } \tau_n\}(R(\tau_n) - f(1))^2 +$$
$$\mu_1 1\{\text{server 1 busy from class 2 at } \tau_n\}(R(\tau_n) - f(2))^2 +$$
$$\mu_1 1\{\text{server 1 idle at } \tau_n\}R^2(\tau_n) + \mu_2 R^2(\tau_n)$$

We expand the squared terms and observe that if we set $f(1) = f(2) = f_a$ the term:

$$2\mu_1 1\{\text{server 1 busy from class 1 at } \tau_n\}R(\tau_n)f(1) +$$
$$2\mu_1 1\{\text{server 1 busy from class 2 at } \tau_n\}R(\tau_n)f(2)$$

can be written as:

$$2\mu_1 1\{\text{server 1 busy at } \tau_n\}R(\tau_n)f_a$$

Using also that:

$$1\{\text{server 1 busy at } \tau_n\} \leq 1 \tag{2.9}$$

we get:

$$\begin{aligned} E[R^2(\tau_{n+1}) \mid \sigma_{\tau_n}] \;\geq\; & R^2(\tau_n) + \lambda_1 f^2(1) + \lambda_2 f^2(2) + \\ & \mu_1 1\{\text{server 1 busy from class 1 at } \tau_n\}f^2(1) + \\ & \mu_1 1\{\text{server 1 busy from class 2 at } \tau_n\}f^2(2) - \\ & 2\mu_1 R(\tau_n)f_a + \\ & (2\lambda_1 f(1) + 2\lambda_2 f(2))R(\tau_n) \end{aligned} \tag{2.10}$$

From (2.9) it is apparent that the bound will be tighter for heavy traffic conditions because then the probability that server 1 is busy is closer to one. Note also that we did a proper matching between the f-parameters ($f(1) = f(2) = f_a$) in order to obtain tighter bounds. According to Assumption (B) in Sec. 1.4, $E[R^2(\tau_n)]$ is finite. In addition, under the invariant distribution of $n_i(t)$ $i = 1, 2, 3$, considered in Assumption (A), we have:

$$E[1\{\text{server i busy from class j at } \tau_n\}] = E[1\{\text{server i busy from class j at t}\}] \;\; \forall t, n$$

and

$$E[R(\tau_{n+1})] = E[R(\tau_n)] = E[R(t)] \quad \forall t, n$$

because the events τ_n are triggered by a "Poisson clock" of rate 1, and it is a fact that Poisson arrivals see time averages (PASTA property). Thus, by taking expectations with respect to the invariant distribution of $n_i(t)$ $i = 1, 2, 3$, in equation (2.10) we obtain:

$$E[R(\tau_n)] \geq \frac{\lambda_1 f^2(1) + \lambda_2 f^2(2)}{\mu_1 f_a - \lambda_1 f(1) - \lambda_2 f(2)}$$

using :

$$E[1\{\text{server i busy from class j at } \tau_n\}] = \frac{\lambda_j}{\mu_i} \text{ for } i, j = 1, 2 \tag{2.11}$$

In addition, by Little's law we have :

$$E[R(\tau_n)] = \lambda_1 f(1) x_1 + \lambda_2 f(2) x_2 \tag{2.12}$$

Therefore since $f(1) = f(2) = f_a$ we finally obtain (2.2).

Similarly, for the other possible choices of S, we get:

- for $S = \{1\}$:

$$\begin{aligned} E[R^2(\tau_{n+1}) \mid \sigma_{\tau_n}] \;=\; & \lambda_1 (R(\tau_n) + f(1))^2 + \lambda_2 R^2(\tau_n) + \\ & \mu_1 1\{\text{server 1 busy from class 1 at } \tau_n\}(R(\tau_n) - f(1))^2 + \\ & \mu_1 1\{\text{server 1 idle from class 1 at } \tau_n\} R^2(\tau_n) + \mu_2 R^2(\tau_n) \end{aligned}$$

Observing that:

$$1\{\text{server 1 busy from class 1 at } \tau_n\} \leq 1, \tag{2.13}$$

by using (2.11) and by taking expectations we get:

$$E[R(\tau_n)] \geq \frac{\lambda_1 f^2(1)}{\mu_1 f(1) - \lambda_1 f(1)}$$

By Little's law we obtain (2.3).

- for $S = \{2\}$, similarly, we obtain (2.4).

• for $S = \{3\}$:

$$\begin{aligned} E[R^2(\tau_{n+1}) \mid \sigma_{\tau_n}] &= \lambda_1 R^2(\tau_n) + \lambda_2 R^2(\tau_n) + \\ &\quad \mu_1 1\{\text{server 1 busy from class 1 at } \tau_n\}(R(\tau_n) + f(3))^2 + \\ &\quad \mu_1 1\{\text{server 1 idle from class 1 at } \tau_n\} R^2(\tau_n) + \\ &\quad \mu_2 1\{\text{server 2 busy from class 3 at } \tau_n\}(R(\tau_n) - f(3))^2 + \\ &\quad \mu_2 1\{\text{server 2 idle at } \tau_n\} R^2(\tau_n) \end{aligned}$$

Observing that:

$$1\{\text{server 1 busy from class 1 at } \tau_n\} \geq 0$$

$$1\{\text{server 2 busy from class 3 at } \tau_n\} \leq 1,$$

by using (2.11) and the following :

$$E[1\{\text{server 2 busy from class 3 at } \tau_n\}] = \frac{\lambda_1}{\mu_2} \tag{2.14}$$

and by taking expectations we get :

$$E[R(\tau_n)] \geq \frac{\lambda_1 f^2(3)}{\mu_2 f(3)}$$

By Little's law we get (2.5).

• for $S = \{1, 3\}$:

$$\begin{aligned} &E[R^2(\tau_{n+1}) \mid \sigma_{\tau_n}] = \\ &\quad \lambda_1 (R(\tau_n) + f(1))^2 + \lambda_2 R^2(\tau_n) + \\ &\quad \mu_1 1\{\text{server 1 busy from class 1 at } \tau_n\}(R(\tau_n) - f(1) + f(3))^2 + \\ &\quad \mu_1 1\{\text{server 1 idle from class 1 at } \tau_n\} R^2(\tau_n) + \\ &\quad \mu_2 1\{\text{server 2 busy from class 3 at } \tau_n\}(R(\tau_n) - f(3))^2 + \\ &\quad \mu_2 1\{\text{server 2 idle at } \tau_n\} R^2(\tau_n) \end{aligned}$$

When

$$f(1) \geq f(3)$$

using the inequalities:

$$1\{\text{server 1 busy from class 1 at } \tau_n\} \leq 1$$

$$1\{\text{server 2 busy from class 3 at } \tau_n\} \leq 1,$$

by using (2.11), (2.14) and by taking expectations we get:

$$E[R(\tau_n)] \geq \frac{\lambda_1 f^2(1) + \lambda_1 (f(1) - f(3))^2 + \lambda_1 f^2(3)}{2[\mu_1(f(1) - f(3)) + \mu_2 f(3) - \lambda_1 f(1)]}$$

The above equation can be written as:

$$x_1 + f_{3,1} x_3 \geq \frac{\lambda_1 + \lambda_1 (1 - f_{3,1})^2 + \lambda_1 f_{3,1}^2}{2[\mu_1(1 - f_{3,1}) + \mu_2 f_{3,1} - \lambda_1]}$$

where $f_{3,1} = f(3)/f(1) \leq 1$. One can now, intuitively, argue, that since the values $f_{3,1} = 0$ and $f_{3,1} = 1$ leave in the denominator of the previous equation the "heavy traffic" terms $\mu_1 - \lambda_1$ and $\mu_2 - \lambda_1$, respectively, they yield tighter bounds. A plot in the space of x_1-x_3, for different values of $f_{3,1}$ has the form depicted in Fig. 2-2. The bound corresponding to the choice $f_{3,1} = 0$ is the same as (2.3). Therefore we set $f(1) = f(3)$ in order to get the bound corresponding to the choice $f_{3,1} = 0$ and by using Little's law we get (2.6). In this example we were able to argue that a specific choice of the f-parameters yields tighter bounds. Intuitively, we chose the f-parameters such that the denominator of the rhs [1] of the bounding equation takes the form $1 - \sum_{i \in S} \rho_i$ where ρ_i is the traffic intensity of class i. However, in the general case of a multiclass queueing network it is difficult to argue similarly. Therefore, we are just going to select the f-parameters, in the general case developed in Chapter 3, following the intuition developed here. In Chapter 5, we are going to propose a refined method that yields bounds independent of the choice of the f-parameters.

[1] right hand side

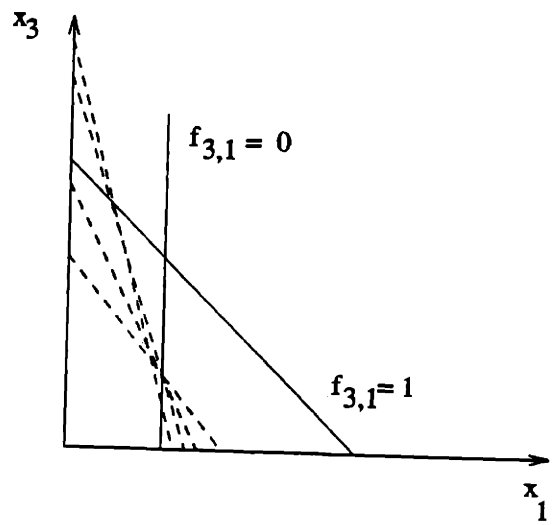

Figure 2-2: A plot for different values of $f_{3,1}$.

Continuing the proof of the theorem,

• for $S = \{2,3\}$:

$$\begin{aligned} E[R^2(\tau_{n+1}) \mid \sigma_{\tau_n}] \;=\; & \lambda_1 R^2(\tau_n) + \lambda_2 (R(\tau_n) + f(2))^2 + \\ & \mu_1 1\{\text{server 1 busy from class 1 at } \tau_n\}(R(\tau_n) + f(3))^2 + \\ & \mu_1 1\{\text{server 1 busy from class 2 at } \tau_n\}(R(\tau_n) - f(2))^2 + \\ & \mu_1 1\{\text{server 1 idle at } \tau_n\} R^2(\tau_n) + \\ & \mu_2 1\{\text{server 2 busy from class 3 at } \tau_n\}(R(\tau_n) - f(3))^2 + \\ & \mu_2 1\{\text{server 2 idle at } \tau_n\} R^2(\tau_n) \end{aligned}$$

Observing that:

$$1\{\text{server 1 busy from class 1 at } \tau_n\} \geq 0$$

$$1\{\text{server 1 busy from class 2 at } \tau_n\} \leq 1$$

$$1\{\text{server 2 busy from class 3 at } \tau_n\} \leq 1,$$

by using (2.11), (2.14) and by taking expectations we get :

$$E[R(\tau_n)] \geq \frac{2\lambda_2 f^2(2) + 2\lambda_1 f^2(3)}{2(\mu_1 f(2) + \mu_2 f(3) - \lambda_2 f(2))}$$

Setting $f(2) = f(3)$ in order to get a tighter bound (in the sense that we discussed before) and by using Little's law we have (2.7).

- Finally for for $S = \{1, 2, 3\}$:

$$
\begin{aligned}
&E[R^2(\tau_{n+1}) \mid \sigma_{\tau_n}] = \\
&\quad \lambda_1(R(\tau_n) + f(1))^2 + \lambda_2(R(\tau_n) + f(2))^2 + \\
&\quad \mu_1 1\{\text{server 1 busy from class 1 at } \tau_n\}(R(\tau_n) - f(1) + f(3))^2 + \\
&\quad \mu_1 1\{\text{server 1 busy from class 2 at } \tau_n\}(R(\tau_n) - f(2))^2 + \\
&\quad \mu_1 1\{\text{server 1 idle at } \tau_n\}R^2(\tau_n) + \\
&\quad \mu_2 1\{\text{server 2 busy from class 3 at } \tau_n\}(R(\tau_n) - f(3))^2 + \\
&\quad \mu_2 1\{\text{server 2 idle at } \tau_n\}R^2(\tau_n)
\end{aligned}
$$

When

$$f(1) \geq f(3)$$

setting:

$$f_c = f(1) - f(3) = f(2),$$

observing that:

$$1\{\text{server 1 busy at } \tau_n\} \leq 1$$

$$1\{\text{server 2 busy from class 3 at } \tau_n\} \leq 1,$$

by using (2.11), (2.14) and by taking expectations we get :

$$E[R(\tau_n)] \geq \frac{\lambda_1 f^2(1) + \lambda_2 f^2(2) + \lambda_1 (f(1) - f(3))^2 + \lambda_2 f^2(2) + \lambda_1 f^2(3)}{2(\mu_1 f_c + \mu_2 f(3) - \lambda_1 f(1) - \lambda_2 f(2))}$$

Setting $f(1) = 2, f(2) = f(3) = 1$ in order to get a tighter bound and by using Little's law we have (2.8). □

Discussion : Note that equation (2.2) is the same as the conservation law for the multiclass M/M/1 queue (see [GeMi, Chap. 6]), with an inequality sign instead of

an equality. Within the class of policies we are considering the conservation law does not hold since we allow idling. If, however, we restrict ourselves to work-conserving policies then it is possible to derive the conservation law via this approach. We are going to provide the proof in Chapter 4 where we will address the general case of the multiclass queue.

Note also, that equations (2.3) and (2.4) have a very intuitive explanation; they are the two inequalities that with the conservation law define the achievable region for the multiclass queue at station 1. In Chapter 4, also, we are going to prove for the general case of the multiclass queue that for work conserving policies equations (2.3) [(2.4)] hold with equality if we give preemptive priority to customers of class 1 [class 2], respectively.

A final note is that (2.5) says nothing more than the fact that the mean response time of class 3 is at least its mean service time. Actually, this conservation inequality suggests that there exists a scheduling policy which makes zero the waiting time of customers of class 3. Indeed, a policy which serves class 1 customers only if server 2 is idle, is such a policy.

2.1.2 A Bound Based on Stability

Since we want to derive a lower bound we can eliminate type 2 customers. The intuitive idea behind the bound derived in this subsection is that if we wanted customers of class 3 to have zero waiting time we would have to make server 1 idle most of the time and work only when station 3 becomes empty. If this was the case, we could make station 1 unstable under heavy traffic conditions. A simple probabilistic argument, yields the following theorem :

Theorem 2.2 *(Stability bound) In the network of Figure 2-1, the following bound holds for any policy satisfying the assumptions given in Section 1.4:*

$$x_3 \geq \hat{p}\frac{\mu_1}{\mu_2(\mu_1+\mu_2)} + \frac{1}{\mu_2} \tag{2.15}$$

where :

$$\hat{p} = \max(0, \frac{2\lambda_1 - \max(\mu_1, \mu_2)}{\mu_1 + \mu_2 - \max(\mu_1, \mu_2)}) \tag{2.16}$$

Proof : Let p denote the steady state probability of both servers being busy with type 1 customers. Then since:

$$1 - p = \mathrm{E}[1\{\text{only one server is busy}\}] + \mathrm{E}[1\{\text{both servers are idle}\}]$$

it holds that:

$$1 - p \geq \mathrm{E}[1\{\text{only one server is busy}\}]$$

According to assumption (A) in Sec. 1.4 the system is stable and therefore, the expected queue lengths are finite. Customers are entering both servers with rate $2\lambda_1$ (since every customer has to to visit both servers). They are departing with rate $\mu_1 + \mu_2$ if both servers are busy and with rate μ_i if only server $i, i = 1, 2$ is busy. Due to stability these arrival and service rates are equal. Writing this argument down we have:

"departure rate from both stations" = "arrival rate to both stations"

But it holds that:

$$p(\mu_1 + \mu_2) + (1 - p)\max(\mu_1, \mu_2) \geq \text{"departure rate from both stations"}$$

Therefore we get :

$$p(\mu_1 + \mu_2) + (1 - p)\max(\mu_1, \mu_2) \geq 2\lambda_1 \Rightarrow$$

$$p \geq \frac{2\lambda_1 - \max(\mu_1, \mu_2)}{\mu_1 + \mu_2 - \max(\mu_1, \mu_2)}$$

Because p is a probability we finally have

$$p \geq \hat{p} \tag{2.17}$$

where $\hat{p}$ is given by (2.16). Since the arrivals are Poisson and the service times exponentially distributed there exist an underlying continuous-time Markov chain describing the network. With probability $\frac{\mu_1}{\mu_1+\mu_2}$ the chain leaves the set of states where both servers are busy, due to a server 1 completion. Thus, $p\frac{\mu_1}{\mu_1+\mu_2}$ is the steady-state number of transitions in the Markov chain, from the set of states where both servers are busy, corresponding to arrivals at station 2. Each of these arrivals has to wait, in queue, at least $1/\mu_2$ on the average. Therefore if by w_3 we denote the expected waiting time of class 3 customers we have just proved that:

$$w_3 \geq p\frac{\mu_1}{\mu_2(\mu_1+\mu_2)} \tag{2.18}$$

The inequality holds because there may be arrivals to station 2 even when both stations are not busy. But since $x_3 = w_3 + 1/\mu_2$ from (2.17) and (2.18) it is seen that (2.15) holds. □

2.1.3 A Bound Based on A Tandem Queue Argument

Eliminating type 2 customers and setting $c_1 = c_3 = 1$ we get a tandem queue with equal costs at both servers. Since we can't do better than FCFS, we get the bound :

$$x_1 + x_3 \geq \frac{1}{\mu_1 - \lambda_1} + \frac{1}{\mu_2 - \lambda_1} \tag{2.19}$$

Note that $\frac{1}{\mu-\lambda}$ is the mean response time of an M/M/1 queue with arrival and service rate λ, μ respectively.

2.1.4 A Bound Based on the c-μ Rule

It is known that for a multiclass queue the c-μ rule (see [Klv2]) is optimal. Therefore the performance of the c-μ rule at station 1 bounds the weighted sum of x_1, x_2. Let denote by $\rho_1 = \frac{\lambda_1}{\mu_1}$ and by $\rho_2 = \frac{\lambda_2}{\mu_1}$ the traffic intensities of class 1 and class 2 at station 1, respectively. The bound is given by :

- if $c_1/\rho_1 \geq c_2/\rho_2$

$$c_1 x_1 + c_2 x_2 \geq c_1 x_1^A + c_2 x_2^A \tag{2.20}$$

where x_1^A, x_2^A are the response times of class 1 and 2,respectively, when preemptive priority is given to class 1, and can be easily calculated (see [Klv2]). In particular, they are given by:

$$x_1^A = \frac{1}{\mu_1 - \lambda_1}$$

$$x_2^A = \frac{1}{\mu_1(1-\rho_1)} + \frac{\rho_1/\mu_1 + \rho_2/\mu_1}{(1-\rho_1)(1-\rho_1-\rho_2)}$$

- if $c_2/\rho_2 \geq c_1/\rho_1$

$$c_1 x_1 + c_2 x_2 \geq c_1 x_1^B + c_2 x_2^B \tag{2.21}$$

where x_1^B, x_2^B are the response times of class 1 and 2,respectively, when preemptive priority is given to class 2. In particular, they are given by:

$$x_2^B = \frac{1}{\mu_1 - \lambda_2}$$

$$x_1^B = \frac{1}{\mu_1(1-\rho_2)} + \frac{\rho_2/\mu_1 + \rho_1/\mu_1}{(1-\rho_2)(1-\rho_1-\rho_2)}$$

Note : If for some reason we want to restrict ourselves to non-preemptive policies the bound (2.20) holds with x_1^A, x_2^A the response times of class 1 and 2, respectively, when non-preemptive priority is given to class 1. The same comment applies to (2.21).

In fact, it will be seen later (Chapter 4) that this bound, based on the optimality of the c-μ rule is redundant. Namely, we are going to prove that the "conservation inequalities" for a multiclass queue fully characterize the achievable region and from that the optimality of the c-μ rule. Thus, for the specific network under consideration in this chapter, the conservation inequalities for subsets of classes involving only classes 1 and 2 (in particular, (2.2), (2.3), (2.4)) fully characterize the achievable

space when $c_3 = 0$. Therefore, they capture the information provided by the c-μ rule bound.

2.1.5 Lower Bound on achievable performance

It is now seen that the space of the expected response times for the network under consideration is constrained by the bounds derived so far. Thus, the derivation of the lower bound on the attainable performance consists of solving the following linear programming (LP) where LB stands for lower bound :

$$Z_{LB} = \min c_1 x_1 + c_2 x_2 + c_3 x_3$$

$$\text{subject to :} \tag{2.22}$$

$$\lambda_1 x_1 + \lambda_2 x_2 \geq \frac{\lambda_1 + \lambda_2}{\mu_1 - \lambda_1 - \lambda_2}$$

$$x_1 \geq \frac{1}{\mu_1 - \lambda_1}$$

$$x_2 \geq \frac{1}{\mu_1 - \lambda_2}$$

$$x_3 \geq \frac{1}{\mu_2}$$

$$x_1 + x_3 \geq \frac{1}{\mu_2 - \lambda_1}$$

$$\lambda_2 x_2 + \lambda_1 x_3 \geq \frac{\lambda_1 + \lambda_2}{\mu_1 + \mu_2 - \lambda_2}$$

$$2\lambda_1 x_1 + \lambda_2 x_2 + \lambda_1 x_3 \geq \frac{3\lambda_1 + \lambda_2}{\mu_1 + \mu_2 - 2\lambda_1 - \lambda_2}$$

$$x_3 \geq \hat{p}\frac{\mu_1}{\mu_2(\mu_1 + \mu_2)} + \frac{1}{\mu_2}$$

$$x_1 + x_3 \geq \frac{1}{\mu_1 - \lambda_1} + \frac{1}{\mu_2 - \lambda_1}$$

$$x_1, x_2, x_3 \geq 0$$

where $\hat{p}$ is given by (2.16).

2.2 Upper Bounds

One obvious candidate for an upper bound is the FCFS policy which is easily analyzable. Other candidates are strict priority policies which can be simulated. Finally, heuristic policies, based mainly on intuitive grounds can also provide upper bounds. In this section we will first analyze the FCFS policy and then discuss a heuristic *threshold policy.*

2.2.1 Analysis of the FCFS Policy

We will first analyze the FCFS policy from first principles. We will also analyze it using the BCMP network notation (see [GeMi] and Sec. 3.2 where a brief summary of this notation is given) because this is the easiest way to analyze simple policies (like FCFS and LCFS) in complex networks where a first principles analysis is not available.

Analysis from First Principles

Let us denote by $\rho_A = \frac{\lambda_1+\lambda_2}{\mu_1}$, $\rho_B = \frac{\lambda_1}{\mu_2}$ the total traffic intensities at stations 1,2 respectively. Recall that by $\rho_i = \frac{\lambda_i}{\mu_1}, i = 1,2$ we denoted the traffic intensities of class i.

The total input at station 1 is an aggregate Poisson process with rate $\lambda_1 + \lambda_2$. Therefore, the steady state probability distribution of the total number of customers at station 1 is (M/M/1 result):

$$p_A(n_A) = (1 - \rho_A)\rho_A^{n_A} \tag{2.23}$$

Also at station 2 the total number of customers has probability distribution in steady-state:

$$p_B(n_3) = (1 - \rho_B)\rho_B^{n_3} \tag{2.24}$$

Moreover, since the scheduling policy at station 1 is independent of the class of each

customer (FCFS), the number of class 1 customers versus the number of class 2 customers should depend only on their arrival rates. Therefore, using that $\rho/(1-\rho)$ is the expected number of customers in an M/M/1 queue we get:

$$E[n_1] = \frac{\lambda_1}{\lambda_1+\lambda_2}\frac{\rho_A}{1-\rho_A} \Rightarrow x_1 = \frac{1}{\mu_1-\lambda_1-\lambda_2} \tag{2.25}$$

$$E[n_2] = \frac{\lambda_2}{\lambda_1+\lambda_2}\frac{\rho_A}{1-\rho_A} \Rightarrow x_2 = \frac{1}{\mu_1-\lambda_1-\lambda_2} \tag{2.26}$$

$$E[n_3] = \frac{\rho_B}{1-\rho_B} \Rightarrow x_1 = \frac{1}{\mu_2-\lambda_1} \tag{2.27}$$

Thus, the performance of the FCFS policy is given by:

$$Z_{FCFS} = \frac{c_1+c_2}{\mu_1-\lambda_1-\lambda_2} + \frac{c_3}{\mu_2-\lambda_1} \tag{2.28}$$

Analysis Using BCMP Notation

The state of the system is $\vec{n} = (\vec{n}_A, \vec{n}_B)$ where $\vec{n}_A = (n_1, n_2)$ is the vector of the number of customers of class 1, 2 at station 1 and $\vec{n}_B = n_3$ is the number of customers at station 3. If we denote by $g_A(\vec{n}_A)$, $g_B(\vec{n}_B)$ the terms corresponding to stations 1, 2 respectively, we get:

$$g_A(\vec{n}_A) = \frac{1}{G_A}\frac{(n_1+n_2)!}{n_1!n_2!}\frac{1}{\mu_1^{n_1+n_2}}$$

$$g_B(\vec{n}_B) = \frac{1}{G_B}\frac{1}{\mu_2^{n_3}}$$

$$d(\vec{n}) = \lambda_1^{n_1+n_3}\lambda_2^{n_2}$$

$$p(\vec{n}) = d(\vec{n})g_A(\vec{n}_A)g_B(\vec{n}_B) \tag{2.29}$$

$p(\vec{n})$ being the steady state probability distribution of the number of customers in the network.

The constant G_A can be calculated easily as follows:

$$G_A = \sum_{n_1,n_2} \binom{n_1+n_2}{n_1,n_2} \left(\frac{\lambda_1}{\mu_1}\right)^{n_1} \left(\frac{\lambda_2}{\mu_1}\right)^{n_2} =$$

$$\sum_k \sum_{k=n_1+n_2} \binom{k}{n_1,n_2} \left(\frac{\lambda_1}{\mu_1}\right)^{n_1} \left(\frac{\lambda_2}{\mu_1}\right)^{n_2} =$$

$$\sum_k \left(\frac{\lambda_1+\lambda_2}{\mu_1}\right)^k =$$

$$\frac{1}{1-\rho_A}$$

Also, G_B is given by:

$$G_B = \frac{1}{1-\rho_B}$$

Thus, we are able finally to calculate the mean number of customers of each class in the system :

$$E[n_1] = \sum_{n_1,n_2} n_1 G_A^{-1} \binom{n_1+n_2}{n_1,n_2} \left(\frac{\lambda_1}{\mu_1}\right)^{n_1} \left(\frac{\lambda_2}{\mu_1}\right)^{n_2} =$$

$$\left(\frac{\lambda_1}{\mu_1}\right) G_A^{-1} \frac{\vartheta}{\vartheta w} \left[\sum_{n_1,n_2} \binom{n_1+n_2}{n_1,n_2} (w)^{n_1} \left(\frac{\lambda_2}{\mu_1}\right)^{n_2}\right]_{w=\lambda_1/\mu_1} =$$

$$= \frac{\lambda_1}{\mu_1-\lambda_1-\lambda_2}$$

which is the same as (2.25). Using similar reasoning, we can also derive (2.26) and (2.27). Therefore, (2.28) can be derived via this alternate route.

2.2.2 Heuristic Threshold Policies

Intuition suggests that a *threshold policy* would perform well under heavy traffic conditions. What this policy has to avoid is to leave station 2 idle (by making the server at station 1 serve customers of type 2) when there are type 1 customers available at station 1. Depending on whether idling is desirable the following two

policies achieve this goal.

> **Policy 1** : Give priority to type 1 customers at station 1 when there are B or fewer customers at station 2. Otherwise give priority to type 2 customers. Never idle.

and

> **Policy 2** : Give priority to type 1 customers at station 1 when there are B or fewer customers at station 2. Otherwise give priority to type 2 customers. Idle at station 1 when there are B or more customers at station 2 and no type 2 customer is present at station 1.

Both policies can be preemptive or not. Here, B is a constant threshold and its optimal value can be calculated via simulation. Policy 1 was proposed in [HaWe] where the Brownian network model approach was used. Intuition seems to suggest that when c_1 and c_3 are comparable, policy 1 which is work-conserving is preferable. But when $c_3 \gg c_1$ then policy 2 should be closer to optimal.

We were not able to analyze these policies. Thus, the only available tool for performance analysis is simulation.

2.3 Some Special Cases

In this section we discuss optimal policies for some special heavy traffic cases. We were able to prove optimality using our lower bounds. We observed that the LP (2.22) yields equal cost with the cost of some specific policy. Thus, this specific policy is in a sense optimal. The optimal policies for the following special cases are also intuitively obvious. What is really important is that our lower bounds have the proper limiting behavior. That is, they approach optimality as certain parameters tend to various limits. In the proofs of the following theorems we are using the notation:

$$\text{"as } x \to y,\ A(x) \approx B(x)\text{"} \equiv \lim_{x \to y} \frac{A(x)}{B(x)} = 1$$

$$\text{"as } x \to y,\ o(A(x)) = B(x)\text{"} \equiv \lim_{x \to y} \frac{A(x)}{B(x)} = 0$$

Theorem 2.3 *Fix μ_1 and μ_2. For $\lambda_1 \to \mu_1$, $\lambda_2 \to 0$ and $\mu_2 > \mu_1$, non-idling FCFS is optimal, in the sense that:*

$$\lim_{\lambda_1 \to \mu_1,\, \lambda_2 \to 0} \frac{Z_{L.B}}{Z_{FCFS}} = 1 \tag{2.30}$$

where Z_{FCFS} is the performance of the FCFS and $Z_{L.B}$ the lower bound on achievable performance in (2.22).

Proof : Note that only station 1 is in heavy-traffic. We are only going to use the bounds (2.3), (2.15), (2.19). Rewriting them we have :

$$x_1 \geq \frac{1}{\mu_1 - \lambda_1} = r_1 \tag{2.31}$$

$$x_1 + x_3 \geq \frac{1}{\mu_1 - \lambda_1} + \frac{1}{\mu_2 - \lambda_1} = r_2 \tag{2.32}$$

$$x_3 \geq \hat{p}\frac{\mu_1}{\mu_2(\mu_1 + \mu_2)} + \frac{1}{\mu_2} = r_3 \tag{2.33}$$

where $\hat{p}$ is given by (2.16). We defined r_1, r_2 and r_3 to be equal with the rhs of (2.31), (2.32) and (2.33) respectively. Figure 2-3 illustrates these inequalities in the space of x_1, x_3.

Note that under the heavy traffic condition $\lambda_1 \to \mu_1$ both r_1 and r_2 tend to infinity. In, contrast r_3 remains finite. Let the corners z_1 and z_2 in the Figure 2-3 have the following coordinates in the x_1–x_3 space.

$$z_1 = (x_{11}, x_{13}), \quad z_2 = (x_{21}, x_{23}).$$

Therefore, as $\lambda_1 \to \mu_1$ it is seen that:

$$x_{11} = x_{21} \approx \frac{1}{\mu_1 - \lambda_1}$$

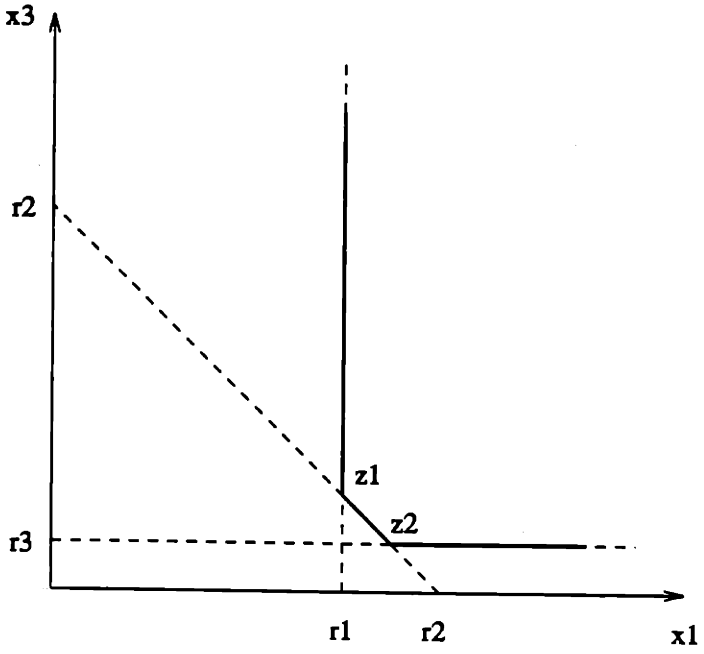

Figure 2-3: Bounds in the x_1, x_3 space for 2.4 theorem.

Instead

$$x_{13} = x_{23} = o(1)$$

In addition, when $\lambda_1 \to \mu_1$,

$$Z_{FCFS} \approx \frac{c_1}{\mu_1 - \lambda_1}$$

Thus, (2.30) is proven. □

Theorem 2.4 *For $\lambda_1 + \lambda_2 \to \mu_1$, and within the class of non-preemptive work-conserving policies satisfying the assumptions A, B and C of Sec. 1.4: the non-preemptive c-μ rule is optimal in the sense that*

$$\lim_{\lambda_1+\lambda_2 \to \mu_1} \frac{Z_{L.B}}{Z_{c-\mu}} = 1 \tag{2.34}$$

where $Z_{c-\mu}$ is the performance of the c-μ rule and $Z_{L.B}$ the lower bound on achievable performance in (2.22).

Proof : Since we are considering work-conserving policies, the conservation law at station 1 asserts that:

$$\lambda_1 x_1 + \lambda_2 x_2 = \frac{\lambda_1 + \lambda_2}{\mu_1 - \lambda_1 - \lambda_2} = g_4 \tag{2.35}$$

We also have the bounds (2.3), (2.4):

$$x_1 \geq \frac{1}{\mu_1 - \lambda_1} = g_1 \tag{2.36}$$

$$x_2 \geq \frac{1}{\mu_1 - \lambda_2} = g_2 \tag{2.37}$$

the bound (2.19):

$$x_1 + x_3 \geq \frac{1}{\mu_1 - \lambda_1} + \frac{1}{\mu_2 - \lambda_1} = g_3 \tag{2.38}$$

and the c-μ rule bound (2.20) where, without loss of generality, we are making the assumption $c_1/\lambda_1 \geq c_2/\lambda_2$:

$$c_1 x_1 + c_2 x_2 \geq c_1 g_5 + c_2 g_6 \tag{2.39}$$

g_5, g_6 being the mean response times when non-preemptive priority is given to class 1 at station 1. We defined g_4, g_1, g_2 and g_3 to be equal with the rhs of the equations (2.35), (2.36), (2.37) and (2.38), respectively. Finally, we also have the stability bound (2.15):

$$x_3 \geq \hat{p}\frac{\mu_1}{\mu_2(\mu_1 + \mu_2)} + \frac{1}{\mu_2} = g_7 \tag{2.40}$$

where g_7 is defined to be equal with the rhs of (2.40). Solving (2.35) for x_2, by substituting into (2.37) we get:

$$x_1 \leq \frac{g_4 - \lambda_2 g_2}{\lambda_1} = x_{1_u} \tag{2.41}$$

and by substituting into (2.39) we get:

$$x_1 \geq \frac{\lambda_2 + \mu_1}{\mu_1(\mu_1 - \lambda_1)} = x_{1_d} \tag{2.42}$$

where we defined x_{1_u} and x_{1_d} to be equal with the rhs of the equations (2.41) and (2.42), respectively. It can be shown that the rhs of (2.42) is greater than g_1 therefore (2.41) along with (2.42) define the feasible [2] region for x_1. Additionally, since we are considering heavy traffic conditions ($\lambda_1 + \lambda_2 \rightarrow \mu_1$) in station 1, $g_4 \rightarrow \infty$ and as a consequence g_3 is smaller than the rhs of (2.41). Thus, the feasible region has the shape depicted in Figure (2-4).

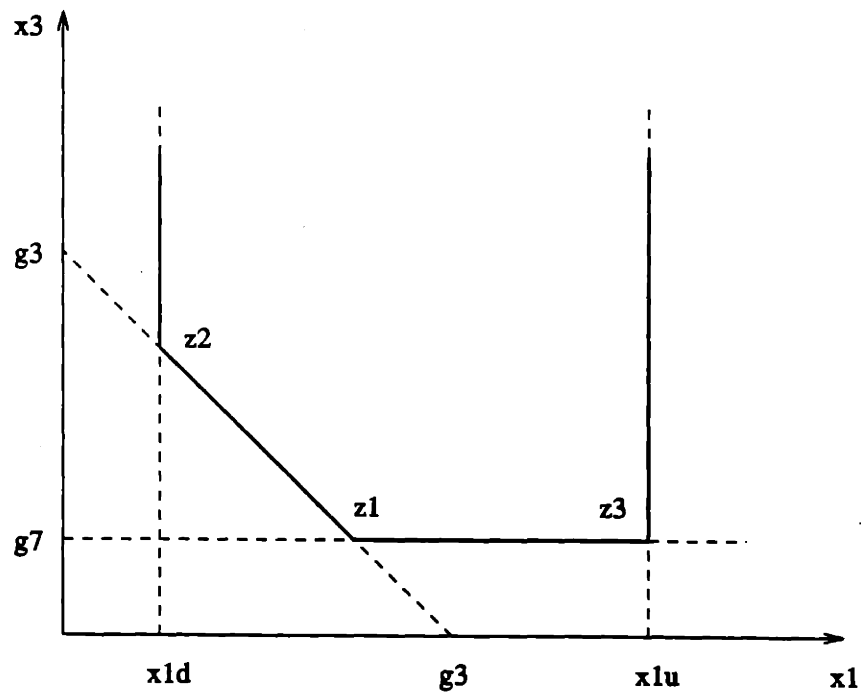

Figure 2-4: Bounds in the x_1, x_3 space for Theorem 2.5.

The lower bound is achieved at one of the corners z_1, z_2, z_3. Using mathematica [3] we found that as $\lambda_1 + \lambda_2 \rightarrow \mu_1$:

$$Z_{z_1} \approx \frac{c_2}{\lambda_2} \frac{\lambda_1 + \lambda_2}{\mu_1 - \lambda_1 - \lambda_2} \tag{2.43}$$

$$Z_{z_2} \approx \frac{c_2}{\mu_1 - \lambda_1} \frac{\mu_1}{\mu_1 - \lambda_1 - \lambda_2} \tag{2.44}$$

$$Z_{z_3} \approx \frac{c_1}{\lambda_1} \frac{\lambda_1 + \lambda_2}{\mu_1 - \lambda_1 - \lambda_2} \tag{2.45}$$

[2]The term feasible here has the linear programming meaning; it doesn't necessarily mean that this is the achievable region.

[3]mathematica is a software package for symbolic algebraic calculations

$$Z_{c-\mu} \approx \frac{c_2}{\mu_1 - \lambda_1} \frac{\lambda_1 + \lambda_2}{\mu_1 - \lambda_1 - \lambda_2} \tag{2.46}$$

where Z_{z_1}, Z_{z_2} and Z_{z_3} is the performance at corners z_1, z_2 and z_3 respectively. In the limit $\lambda_1 + \lambda_2 \to \mu_1$, and since $c_1/\lambda_1 \geq c_2/\lambda_2$, it is seen from (2.43), (2.44) and (2.45) that the minimum is achieved at either z_1 or z_2. Using also (2.46), it is seen that as $\lambda_1 + \lambda_2 \to \mu_1$ we have:

$$Z_{z_1} \approx Z_{z_2} \approx Z_{c-\mu}.$$

□

For $\lambda_1 \to \mu_2$ and $c_3 > c_1$ we were not able to find the optimal policy. We believe that a threshold-type policy should be quite close to the optimal. However, we proved that an insertion of a buffer with a deterministic server in the stream of type 1 customers from station 1 to station 2, considered as a policy, outperforms FCFS because it smoothes the input traffic to station 2. Note that this is not an admissible policy since it alters the network configuration. Our proof, assumes $\lambda_2 \to 0$ because otherwise it would be very difficult to characterize the output of station 1. However, the smoothing behaviour of the buffer is present even when class 2 customers are not eliminated and therefore it is reasonable to believe that the qualitative result holds for this case also. Summarizing we have the following theorem:

Theorem 2.5 *Let $\lambda_1 \to \mu_2$ such that $\lambda_1 = \mu_2 - \epsilon$, where $\epsilon \to 0$. Let also $c_3 > c_1$, and $\lambda_2 \to 0$. Then the insertion of an infinite buffer with a deterministic server, with rate $\mu = \mu_2 - \alpha\epsilon$, where $\alpha \in [0,1]$ and $\epsilon \in (0, \mu_2)$ in the stream of type 1 customers from station 1 to station 2, considered as a policy, outperforms FCFS.*

Proof : Under the conditions given in the theorem a FCFS analysis (M/M/1 queues in tandem) yields:

$$x_1^{FCFS} = \frac{1}{\mu_1 - \lambda_1}, x_3^{FCFS} = \frac{1}{\mu_2 - \lambda_1}$$

where x_1^{FCFS}, x_3^{FCFS} are the mean response times of classes 1,3 under FCFS, respectively. If by Z_{FCFS} we denote the performance of the FCFS then as $\lambda_1 \rightarrow \mu_2$

$$Z_{FCFS} \approx \frac{c_3}{\mu_2 - \lambda_1} \tag{2.47}$$

Let us now analyze the network after the insertion of the deterministic device. The modified network is depicted in Figure 2-5.

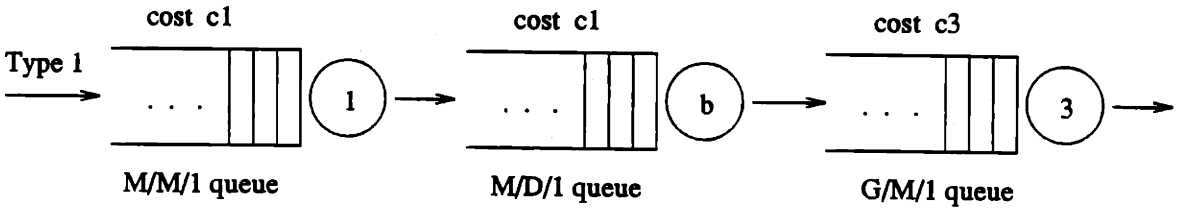

Figure 2-5: The modified network with the deterministic device.

The mean response time at station 1 (M/M/1 queue) and the mean response time at the buffer (M/D/1 queue) are:

$$x_1^D = \frac{1}{\mu_1 - \lambda_1}, \quad x_b^D = \frac{\lambda_1}{2\mu(\mu - \lambda_1)} + \frac{1}{\mu} \tag{2.48}$$

where x_1^D, x_b^D are the mean response times at stations 1,b, respectively, in the network depicted in Figure 2-5. Since we want the performance of the modified network to be better than the performance of FCFS we have to drive station b also in heavy traffic. We choose $\mu \rightarrow \mu_2$ (that is, the M/D/1 queue is in heavy traffic), such that:

$$\mu = \mu_2 - \alpha\epsilon \tag{2.49}$$

where $\epsilon \rightarrow 0$ and α is a multiplying constant. Then the contribution of both station 1 and the deterministic device to the objective value function is dominated by:

$$c_1(x_1^D + x_b^D) \approx \frac{c_1\lambda_1}{2\mu(\mu - \lambda_1)} \tag{2.50}$$

The probability density function of the service times in the M/D/1 queue is $\delta(t - \frac{1}{\mu})$

and therefore the Laplace transform for the interdeparture time distribution is (see [BeNa, pg. 14])

$$\frac{\lambda_1}{\mu}\frac{s+\mu}{s+\lambda_1}e^{-s/\mu}$$

Interdeparture times are not independent but in [BeNa] is shown that in the heavy traffic limit ($\rho \to 1$) the covariance between different interdeparture times goes to zero very fast, as the distance in time between these interdeparture times increases. Therefore, making an approximation, we are going to ignore them and use the results for a G/M/1 queue with independent and identically distributed interarrival times. Thus, the parameter σ for the G/M/1 queue discussed in [Klv1] is given by:

$$\sigma = \frac{\lambda_1}{\mu}\frac{\mu_2 - \mu_2\sigma + \mu}{\mu_2 - \mu_2\sigma + \lambda_1}e^{-(\mu_2-\mu_2\sigma)/\mu} \tag{2.51}$$

From the theorem's statement we have:

$$\lambda_1 = \mu_2 - \epsilon \tag{2.52}$$

where ϵ is a small number. Since in heavy traffic $\sigma \to 1$, we expand (2.51) in a neighborhood of 1 and by keeping second order terms, after some algebra, we get:

$$\sigma = 1 + \frac{2\lambda_1\mu^2}{(2\mu^2 - \lambda_1^2)\mu_2}(\frac{\lambda_1}{\mu_2} - 1) \leq 1 \tag{2.53}$$

Now, since the mean waiting time in a G/M/1 queue with parameter σ and rate μ is given by $\sigma/[\mu(1-\sigma)]$ (see [Klv1]), the cost in the G/M/1 queue is dominated in heavy traffic (as $\sigma \to 1$) by $c_3x_3 \approx c_3\sigma/[\mu_2(1-\sigma)]$ and therefore the total performance, in heavy traffic, of the system (taking into consideration 2.50) is given by:

$$Z_D \approx \frac{c_1\lambda_1}{2\mu(\mu-\lambda_1)} + \frac{c_3\sigma}{\mu_2(1-\sigma)} \tag{2.54}$$

where Z_D denotes the performance of the network depicted in Fig. 2-5. Plugging (2.52) and (2.49) into (2.53) and the result into (2.54), by expanding with respect to ϵ and by minimizing the result over α we finally get for the performance of the system

with the deterministic device:

$$Z_D = \frac{c_1 + c_3}{2} + o(\epsilon) \tag{2.55}$$

From (2.47) and (2.52) it is clear (since $c_3 \geq c_1$) that the insertion of the deterministic device, if considered as a policy outperforms FCFS by $(c_3 - c_1)/2$. □

Remarks : A simple analysis similar to the one followed in the two previous theorems yields that the lower bound on attainable performance is in this case:

$$Z_{L.B} = \frac{c_1}{\epsilon} + o(\epsilon) \tag{2.56}$$

Chapter 3

A General Open Multiclass Queueing Network: Approximate Characterization

Our goal in this chapter, is to derive lower and upper bounds on the achievable performance for a general multiclass queueing network. We are considering two network models; one with deterministic routing where different customer classes have the same service requirements at a specific station and another with probabilistic routing where customer classes may have different service requirements at a specific station. The second model is the most general model that we consider in this thesis and is defined in Sec. 1.3. Next we define both models for the chapter to be self-contained.

Network model with Deterministic Routing: Consider an open multiclass queueing network with J single server stations and I different customer types. Customers of type i $(i = 1, 2, \ldots, I)$ enter the system as a Poisson stream at rate λ_i and pass through a sequence of stations

$$r(i,1), r(i,2), \ldots, r(i, M(i))$$

before leaving the system. $M(i)$ is the number of stages that an arriving customer of type i has to complete before exiting the network. The station which a type i customer visits at stage m ($m = 1, 2, \ldots, M(i)$) of his route is station $r(i,m)$. Each visit of a customer type at a station defines a different class. So, (i,m) is the class formed by type i customers at the m^{th} stage of their route. Service times are exponentially distributed and let $\mu_{r(i,m)}$ be the service rate of station $r(i,m)$. We assume that the total input rate at a station is less than $\mu_{r(i,m)}$ to ensure that at least one stable policy exists. Note here that we let service rates depend only on the station. Let finally $n_{(i,m)}(t)$ be the number of customers of class (i,m) present in the network at time t. The optimization problem is to determine a scheduling policy that minimizes a linear cost function of the form $\sum_{(i,m)} c_{(i,m)} x_{(i,m)}$, $x_{(i,m)}$ being the expected response time (waiting + service time) of class (i,m) and $c_{(i,m)}$ being given finite weights.

Network model with Probabilistic Routing: Consider an open multiclass queueing network with N single server stations and R different job types. Jobs may change type as they move from one node to another. In particular, a job of type r, when completing service at node i goes to node j as a type s job with probability $p_{i,r;j,s}$ and leaves the network with probability $p_{i,r;0} = 1 - \sum_{j,s} p_{i,r;j,s}$. There are r independent Poisson streams of arrivals to the network, one for each type of customers. The Poisson arrival process, for customers of type r, has rate $\lambda_{0,r}$ and these customers join the i station with probability $q_{i,r}$. The pair (i,r) is called class and the class (i,r) requires an exponentially distributed service with rate $\mu_{i,r}$. Let $n_{(i,r)}(t)$ be the number of class (i,r) customers, present in the network at time t. The optimization problem is to determine a global scheduling policy that minimizes a linear cost function of the form $\sum_{(i,r)} c_{(i,r)} x_{(i,r)}$, $x_{(i,r)}$ being the expected response time (waiting + service time) of class (i,r) and $c_{(i,r)}$ being given finite weights.

Note that the probabilistic network model, includes the deterministic one. To bound the performance, we will mainly use the ideas illustrated in the previous chapter for the specific two-station network mentioned there. In Section 3.1 we derive

lower bounds that include the achievable region for the open multiclass queueing network. Then in Section 3.2 we describe the BCMP networks notation because it is a concrete method to analyze simple policies that provide upper bounds to the performance in a general multiclass network.

3.1 Lower Bounds

3.1.1 Deterministic Routing; Conservation Inequalities

Consider a subset S of the set of classes. We define a measure of the work to be done in the network as follows:

$$R^S(t) = \sum_{(i,m)\in S} f^S(i,m) n_{(i,m)}(t) \tag{3.1}$$

where $f^S(i,m)$ are the multiplying constants which we called f-parameters, in Chapter 2.

The following conditions on the f-parameters emerge from the proof of Theorem 3.1. Although they may appear unmotivated at this point, the proof suggests that they lead to tighter bounds. So, for each S we have the following conditions holding (note that we drop S from the notation):

For all classes $(i,m) \in S$ queued in the same station $r(i,m)$:

1.

$$f(i,m) = f_{r(i,m)} \text{ when } (i,m+1) \notin S \tag{3.2}$$

and

2.

$$f_{r(i,m)} = f(i,m) - f(i,m+1) \text{ when } (i,m+1) \in S \tag{3.3}$$

where $f_{r(i,m)}$ is a non-negative constant depending only on the station. We then have the following theorem:

Theorem 3.1 *For all f-parameters satisfying conditions (3.2), (3.3), and within the class of policies described in Sec. 1.4, the following inequality (lower bound) holds for each subset S of the set of classes:*

$$\sum_{(i,m)\in S} \lambda_i f(i,m) x_{(i,m)} \geq \frac{N(S)}{D(S)} \tag{3.4}$$

where:

$$\begin{aligned} N(S) = & \sum_{\{(i,m)\in S | m=1\}} \lambda_i f^2(i,m) + \sum_{\{(i,m)\in S | (i,m+1)\notin S\}} \lambda_i f^2(i,m) + \\ & \sum_{\{(i,m)\in S | (i,m+1)\in S\}} \lambda_i (f(i,m) - f(i,m+1))^2 + \\ & \sum_{\{(i,m)\notin S | (i,m+1)\in S\}} \lambda_i f^2(i,m+1) \end{aligned}$$

$$D(S) = 2\left[\sum_{\{r(i,m) | (i,m)\in S\}} f_{r(i,m)} \mu_{r(i,m)} - \sum_{\{(i,m)\in S | m=1\}} \lambda_i f(i,m)\right]$$

$x_{(i,m)}$ *being the mean response time of class (i,m).*

Proof : In order to prove the theorem we use the same procedure we used for the derivation of the bounds named conservation inequalities in Chapter 2. So, we are again applying the same trick and force the single servers at each station to work continuously even when the station is empty. As we explained in Chapter 2 this modification does not alter the behaviour of the system. Let again τ_n be the sequence of times immediately after an arrival or service completion (fictitious or real). Let also denote by $1\{\cdot\}$ the indicator function and by σ_{τ_n} the σ-field generated by events up to time τ_n or intuitively the previous history. Let us finally, without loss of generality scale time such that:

$$\sum_i \lambda_i + \sum_{r(i,m)} \mu_{r(i,m)} = 1$$

for ease of notation. We use the notation $\{(i,m) \in S \mid m = 1\}$ to represent all the classes in S formed by type i customers at the first stage of their route through the network. In addition, in the set $\{(i,m) \notin S\}$ we also include the external world of

the network. Dropping S from $R^S(t)$ and $f^S(i)$, we get:

$$E[R^2(\tau_{n+1}) \mid \sigma_{\tau_n}] =$$

$$\sum_{\{(i,m)\in S|m=1\}} \lambda_i (f(i,m)+R(\tau_n))^2 + \sum_{\{(i,m)\notin S|m=1\}} \lambda_i R^2(\tau_n) +$$

$$\sum_{\{(i,m)\in S|(i,m+1)\notin S\}} \mu_{r(i,m)} 1\{r(i,m) \text{ busy from } (i,m) \text{ at } \tau_n\}(R(\tau_n)-f(i,m))^2 +$$

$$\sum_{\{(i,m)\in S|(i,m+1)\in S\}} \mu_{r(i,m)} 1\{r(i,m) \text{ busy from } (i,m) \text{ at } \tau_n\}$$

$$(R(\tau_n)-f(i,m)+f(i,m+1))^2 +$$

$$\sum_{\{(i,m)\in S\}} \mu_{r(i,m)} 1\{r(i,m) \text{ idle from } (i,m) \text{ at } \tau_n\} R^2(\tau_n) +$$

$$\sum_{\{(i,m)\notin S|(i,m+1)\in S\}} \mu_{r(i,m)} 1\{r(i,m) \text{ busy from } (i,m) \text{ at } \tau_n\}(R(\tau_n)+f(i,m+1))^2 +$$

$$\sum_{\{(i,m)\notin S|(i,m+1)\in S\}} \mu_{r(i,m)} 1\{r(i,m) \text{ idle from } (i,m) \text{ at } \tau_n\} R^2(\tau_n) +$$

$$\sum_{\{(i,m)\notin S|(i,m+1)\notin S\}} \mu_{r(i,m)} R^2(\tau_n)$$

In order to derive a tighter bound we are going to make a proper matching of the f-parameters at each station. Thus, we observe that if we set (3.2) and (3.3) in the term:

$$\sum_{\{(i,m)\in S|(i,m+1)\notin S\}} 2\mu_{r(i,m)} 1\{r(i,m) \text{ busy from } (i,m) \text{ at } \tau_n\} R(\tau_n) f(i,m) +$$

$$\sum_{\{(i,m)\in S|(i,m+1)\in S\}} 2\mu_{r(i,m)} 1\{r(i,m) \text{ busy from } (i,m) \text{ at } \tau_n\}$$

$$R(\tau_n)(f(i,m)-f(i,m+1))$$

it can be written as:

$$\sum_{\{r(i,m)|(i,m)\in S\}} 2 f_{r(i,m)} \mu_{r(i,m)} R(\tau_n) 1\{r(i,m) \text{ busy from } (i,m)\in S \text{ at } \tau_n\}$$

To bound this term we are going to use the fact that:

$$1\{r(i,m) \text{ busy from } (i,m)\in S \text{ at } \tau_n\} \leq 1$$

Moreover we bound the term:

$$2 \sum_{\{(i,m)\notin S | (i,m+1)\in S\}} \mu_{r(i,m)} 1\{r(i,m) \text{ busy from } (i,m) \text{ at } \tau_n\} R(\tau_n) f(i,m+1)$$

by using that:

$$1\{r(i,m) \text{ busy at } \tau_n \text{ from } (i,m) \notin S \mid (i,m+1) \in S\} \geq 0$$

According to Assumption (B) in Sec. 1.4, $E[R^2(\tau_n)]$ is finite. In addition, under the invariant distribution of $n_{(i,m)}(t)$, considered in Assumption (A), we have:

$$E[1\{r(i,m) \text{ busy from } (i,m) \text{ at } \tau_n\}] = E[1\{r(i,m) \text{ busy from } (i,m) \text{ at } t\}] \quad \forall t,n$$

and

$$E[R(\tau_{n+1})] = E[R(\tau_n)] = E[R(t)] \quad \forall t,n$$

because the events τ_n are triggered by a "Poisson clock" of rate 1, and it is a fact that Poisson arrivals see time averages (PASTA property). Now, using:

$$E[1\{r(i,m) \text{ busy from } (i,m) \text{ at } \tau_n\}] = \frac{\lambda_i}{\mu_{r(i,m)}}$$

we finally obtain, after some algebra:

$$E[R(\tau_n)] \geq \frac{N(S)}{D(S)}$$

where $N(S)$ and $D(S)$ are given in the statement of the theorem. At this point it is a simple matter to see that the application of Little's law yields (3.4). □

Remarks : The procedure to obtain the bound was very similar to the one described in Chapter 2. The last task is to set the f-parameters in such a way to make the bound even tighter. As we noted in Chapter 2 we try to form a term $1-\rho$ in the denominator $D(S)$ under restrictions (3.2),(3.3). That is, we first do

the proper matchings of the f-parameters described by equations (3.2) and (3.3) and we then determine the remaining f-parameters in such a way that the denominator $D(S)$ takes the heavy-traffic form. An other choice for the f-parameters is to select $f^S(i,m)$ to be the expected remaining service time of class (i,m) within the subset S. This choice satisfies (3.2) and (3.3) yielding $f_{r(i,m)} = 1/\mu_{r(i,m)}$ which is positive and depends only on the station. In general, it is not known if there exists a selection of the f-parameters that provides dominant bounds. But, even if this is the case, it is difficult to determine these "best" f-parameters. In Chapter 5, we will prove that it is not important to do so, because a refined bounding method that we propose there yields better bounds independent of the choice of the f-parameters. Let us now denote by $\overline{S}$ the cardinality of the set S. This bounding technique summarized in the above theorem yields $2^{\overline{S}} - 1$ linear inequalities. The initial idea was found in [Kuma] where the f-parameters were chosen to be the "remaining number of stages", that is, the number of service completions a customer has in front of him ($f(i,m) = M(i) - m + 1$). Our generalization yields much tighter bounds.

3.1.2 Probabilistic Routing; Conservation Inequalities

The traffic equations for the probabilistic network model take the form:

$$\lambda_{i,r} = \lambda_{0,r} q_{i,r} + \sum_{j=1}^{N} \sum_{r'=1}^{R} \lambda_{j,r'} p_{j,r';i,r} \tag{3.5}$$

We assume that the inequality:

$$\sum_{(j,r')|j=i} \frac{\lambda_{j,r'}}{\mu_{j,r'}} < 1$$

holds at each station i, in order to ensure that at least one stable policy exists. As we did in the previous subsection we consider a subset S of the set of classes and we define a measure of the work to be done in the network as follows:

$$R^S(t) = \sum_{(i,r)\in S} f^S(i,r) n_{(i,m)}(t) \tag{3.6}$$

where $f^S(i,m)$ are the multiplying constants which we called f-parameters, in Chapter 2.

The following conditions on the f-parameters emerge from the proof of Theorem 3.2. Although they may appear unmotivated at this point, the proof suggests that they lead to tighter bounds. So, if for each S we set (note that we drop S from the notation):

For all classes $(i,r) \in S$ queued in the same station i:

$$f_i = \mu_{i,r} \left[\sum_{(j,r') \in S} p_{i,r;j,r'}(f(i,r) - f(j,r')) + \sum_{(j,r') \notin S} p_{i,r;j,r'} f(i,r) \right] \quad (3.7)$$

where f_i is a non-negative constant depending only on the station.

we have the following theorem:

Theorem 3.2 *For all f-parameters satisfying the restriction (3.7) and within the class of policies described in Sec. 1.4, the following inequality (lower bound) holds for each S:*

$$\sum_{(i,r) \in S} \lambda_{i,r} f(i,r) x_{(i,r)} \geq \frac{N'(S)}{D'(S)} \quad (3.8)$$

where :

$$N'(S) = \sum_{(i,r) \in S} \lambda_{0,r} q_{i,r} f^2(i,r) + \sum_{(i,r) \notin S} \lambda_{i,r} \sum_{(j,r') \in S} p_{i,r;j,r'} f^2(j,r') +$$
$$\sum_{(i,r) \in S} \lambda_{i,r} \left[\sum_{(j,r') \in S} p_{i,r;j,r'}(f(i,r) - f(j,r'))^2 + \sum_{(j,r') \notin S} p_{i,r;j,r'} f^2(i,r) \right]$$

$$D'(S) = 2 \left[\sum_{\{i | \exists r \text{ with } (i,r) \in S\}} f_i - \sum_{(i,r) \in S} \lambda_{0,r} q_{i,r} f(i,r) \right]$$

S being a subset of the set of classes and $x_{(i,r)}$ the mean response time of class (i,r).

Proof : The proof is quite similar to the proof of Theorem 3.1. We shall apply uniformization to the continuous-time Markov-chain corresponding to the network. So let the uniform rate be:

$$\nu = \sum_{r} \lambda_{0,r} + \sum_{i,r} \mu_{i,r}$$

Without loss of generality we scale time such that $\nu = 1$. Let τ_n be the sequence of times immediately after a transition and let σ_{τ_n} be the σ-field generated by events up to time τ_n or more intuitively the previous history. Let finally $1\{\cdot\}$ be an indicator function. With the uniformization we are inserting a common "Poisson clock" for every transition in the Markov-chain. Being in a specific state of the chain, only some of the generally possible events can occur. For example, if the chain is in a state where no class k customer is present in the network then a departure of a class k customer cannot occur. Thus, the actual transition rate (ν_1) from a specific state, say u, is less than ν. In the uniformized chain however, from state u transitions occur with rate ν. That is, actual transitions occur with rate ν_1 and self-transitions with rate $\nu - \nu_1$. Thus, the recursive equation for $R(t)$ takes the form:

$$\begin{aligned}
&E[R^2(\tau_{n+1}) \mid \sigma_{\tau_n}] = \\
&\quad \sum_{(i,r)\in S} \lambda_{0,r} q_{i,r} (R(\tau_n) + f(i,r))^2 + \sum_{(i,r)\notin S} \lambda_{0,r} q_{i,r} R^2(\tau_n) + \\
&\quad \sum_{(i,r)\in S} \mu_{i,r} 1\{\text{server } i \text{ busy from type } r \text{ at } \tau_n\} \\
&\quad \left[\sum_{(j,r')\in S} p_{i,r;j,r'} (R(\tau_n) - f(i,r) + f(j,r'))^2 + \sum_{(j,r')\notin S} p_{i,r;j,r'} (R(\tau_n) - f(i,r))^2 \right] + \\
&\quad \sum_{(i,r)\in S} \mu_{i,r} 1\{\text{server } i \text{ idle from type } r \text{ at } \tau_n\} R^2(\tau_n) + \\
&\quad \sum_{(i,r)\notin S} \mu_{i,r} 1\{\text{server } i \text{ busy from type } r \text{ at } \tau_n\} \\
&\quad \left[\sum_{(j,r')\in S} p_{i,r;j,r'} (R(\tau_n) + f(j,r'))^2 + \sum_{(j,r')\notin S} p_{i,r;j,r'} R^2(\tau_n) \right] + \\
&\quad \sum_{(i,r)\notin S} \mu_{i,r} 1\{\text{server } i \text{ idle from type } r \text{ at } \tau_n\} R^2(\tau_n)
\end{aligned}$$

Note here that the set of classes $(j,r') \notin S$ includes the external world of the network. As we did before we are going to make a proper matching of the f-parameters in order to get tighter bounds. Using (3.7) the term:

$$2 \sum_{(i,r)\in S} \mu_{i,r} 1\{\text{server } i \text{ busy from type } r \text{ at } \tau_n\}$$
$$\left[\sum_{(j,r')\in S} p_{i,r;j,r'} R(\tau_n)(f(i,r) - f(j,r')) + \sum_{(j,r')\notin S} p_{i,r;j,r'} R(\tau_n) f(i,r) \right]$$

can be written as follows:

$$\sum_{\{i|\exists r \text{with}(i,r)\in S\}} f_i R(\tau_n) 1\{\text{server i busy from classes } (i,r) \in S \text{ at } \tau_n\}$$

Now, to bound the above term we are using the fact that:

$$1\{\text{server i busy from classes } (i,r) \in S \text{ at } \tau_n\} \leq 1$$

In addition, to bound the term:

$$\sum_{(i,r)\notin S} 2\mu_{i,r} 1\{\text{server } i \text{ busy from type } r \text{ at } \tau_n\} \sum_{(j,r')\in S} p_{i,r;j,r'} R(\tau_n) f(j,r')$$

we are using that:

$$1\{\text{server } i \text{ busy from type } r \text{ at } \tau_n\} \geq 0$$

From here the procedure to get the bound is identical to the one described in Theorem 3.1. Thus, having (3.7) holding and using that:

$$E[1\{\text{server } i \text{ busy from type } r \text{ at } \tau_n\}] = \frac{\lambda_{i,r}}{\mu_{i,r}}$$

where $\lambda_{i,r}$ is the solution of the traffic equations (3.5), we finally get (3.8). □

Remarks : The last task is again to determine the f-parameters, satisfying (3.7) in order to get the dominant bounds from the class of bounds defined by (3.8). One

choice for those parameters that we believe gives these dominant bounds is to set:

$$f^S(i,r) = \text{"expected remaining service time for jobs } (i,r) \text{ within the subset } S\text{"}$$

This suggests:

$$f^S(i,r) = \frac{1}{\mu_{i,r}} + \sum_{(j,r')\in S} p_{i,r;j,r'} f^S(j,r') \tag{3.9}$$

This choice satisfies (3.7) because:

$$f_i = \mu_{i,r} f(i,r) \sum_{(j,r')\in S} p_{i,r;j,r'} - \mu_{i,r} \sum_{(j,r')\in S} p_{i,r;j,r'} f(j,r')) + \mu_{i,r} f(i,r) \sum_{(j,r')\notin S} p_{i,r;j,r'} \Rightarrow$$

$$f_i = \mu_{i,r} f(i,r) - \mu_{i,r} f(i,r) + 1 = 1$$

where we used (3.9). Moreover, this choice of the f-parameters would make the denominator of (3.8) in the form $1 - \sum_{(i,r)\in S} \rho_{(i,r)}$ and this why is going to yield tighter bounds. This claim is also justified by the fact that in the Klimov's problem (see Chap. 4) this choice of the f-parameters yields the tightest bounds. But, we were not able to prove it in the general case. Instead, in Chapter 5 we modify the method that we are using to derive the bounds and we derive stronger bounds independent of the choice of the f-parameters.

3.1.3 A Bound Based on Stability

For the deterministic network model a generalization of the results of Section 2.1.2 yields :

Theorem 3.3 *For every pattern of the form depicted in Figure 3-1, where customers enter (from the outside world or another station) with rate λ and stations $i, i+1$ have exponentially distributed service times with rates rates μ_i, μ_{i+1} respectively, and within the class of policies satisfying the assumptions in Sec. 1.4, the following inequality holds:*

$$x_k \geq \hat{p} \frac{\mu_i}{\mu_i + \mu_{i+1}} \frac{1}{\mu_{i+1}} + \frac{1}{\mu_{i+1}} \tag{3.10}$$

where:

$$\hat{p} = \max\left(0, \frac{2\lambda - \max(\mu_i, \mu_{i+1})}{\mu_i + \mu_{i+1} - \max(\mu_i, \mu_{i+1})}\right)$$

x_k being the mean response time of the class of customers at station i+1.

Proof : If in the proof of Theorem 2.2 we replace λ_1 by λ, μ_1 by μ_i, μ_2 by μ_{i+1} and x_3 by x_k then the proof carries through. □

Figure 3-1: The pattern considered in the stability bound

3.1.4 Lower Bound on Achievable Performance

It is now seen that the space of the expected response times for the network under consideration is constrained by the bounds derived so far. Note, that we didn't write down for this network model a bound analogous to the c-μ rule bound of Chapter 2 because, as we mentioned there and we will prove in Chapter 4, such a bound is captured by the "conservation inequalities". We denote by E the entire set of classes in the network. The derivation of the lower bound on the attainable performance consists of solving the following linear programming (LP) where LB stands for lower bound:

Deterministic Routing:

$$Z_{LB} = \min \sum_{\forall \text{ class } k} c_k x_k$$

subject to: (3.11)

$$\sum_{k \in S} \lambda_i f(k) x_k \geq \frac{N(S)}{D(S)} \quad \forall S \subseteq E$$

$$x_k \geq \hat{p}\frac{\mu_i}{\mu_i + \mu_{i+1}}\frac{1}{\mu_{i+1}} + \frac{1}{\mu_{i+1}}$$

where $N(S), D(S)$ are defined in the statement of Theorem 3.1 and $\hat{p}, \mu_i, \mu_{i+1}$ are defined in the statement of Theorem 3.3.

Probabilistic Routing

$$Z_{LB} = \min \sum_{\forall \text{ class } k} c_k x_k$$

$$\text{subject to:} \tag{3.12}$$

$$\sum_{k \in S} \lambda_k f(k) x_k \geq \frac{N'(S)}{D'(S)} \qquad \forall S \subseteq E$$

where N', D' are defined in the statement of Theorem 3.3.

3.2 Upper Bounds

Upper bounds are obtained by using simple policies which can be analyzed in closed form since the network models we considered can be easily transformed to the model of BCMP and Kelly networks. We will also, simulate strict priority policies and other heuristics that exploit the special structure of each specific topology. We will, next, briefly discuss the BCMP networks notation (see [GeMi, chap. 3]). The BCMP network model is the most general model known to have a closed-form solution.

The network topology is represented by an arbitrary graph with N nodes (excluding the "outside world" node). There are R *job types* and jobs may change type as they move from one node to another. In particular, a job of type r, after completing service at node i, goes to node j as a job of type s with probability $p_{i,r;j,s}$ and leaves the network with probability $p_{i,r;0}$. The pair (i, r) is called *class*. The set of classes is split into one or more non-intersecting subsets, called *subchains* according to the following rule: two classes belong to the same subchain if there is a non-zero probability that a job will be in both classes during its sojourn through the network. We denote these subchains by $E_1, E_2, \ldots, E_m$ $(m \geq 1)$. Let $\vec{S}$ be the state of the network, $M(\vec{S})$ the total number of jobs in the network in state $\vec{S}$ and $M(\vec{S}, E_k)$ the number of jobs

in subchain E_k when the network is in state $\vec{S}$.

The external arrivals are generated by one independent non-homogeneous Poisson process for each subchain with instantaneous rate $\lambda_k(M(\vec{S}, E_k))$ corresponding to the k^{th} process. Note that the arrival process may depend on the state of the system. An arrival from the k^{th} process goes to node i as a type r job with probability $p_{0;i,r}$.

There are the following four possibilities for the stations (node types) of the network:

1. The service requirements for each class depend on the station and not on the specific class. More precisely, all classes queued in the same node i have exponentially distributed service times with rate μ_i. The policy is FCFS. The speed of the single server $C_i(n_i)$ depends on the number of jobs in the node n_i.

2. The service requirements for type r jobs can have Coxian distributions but we are only considering the case of the exponential distribution. The difference from the type 1 node is that each class can have distinct service requirements. Thus, the service requirements for class (i,r) are exponentially distributed with rate $\mu_{i,r}$. The scheduling strategy is processor-sharing and the speed of the single server may also depend on n_i as for type 1 nodes.

3. The assumptions for the service requirements are the same as for type 2 nodes and the scheduling strategy is server-per-job (that is there are infinite number of servers and a job is assigned to one as soon as it enters the node).

4. The assumptions for the service requirements are the same as for type 2 nodes and the scheduling strategy is LCFS preemptive-resume. The speed of the single server may also depend on n_i as for type 1 nodes.

The traffic equations for the network take the form:

$$e_{j,s} = p_{0;j,s} + \sum_{i,r} e_{i,r} p_{i,r;j,s} \quad i,j = 1,2,\ldots,N;\ r,s = 1,2,\ldots,R$$

We are looking at a node state which is $\vec{n}_i = (n_{i,1}, n_{i,2}, \ldots, n_{i,R})$ where $n_{i,r}$ is the number of type r customers at node i. Let also n_i be the total number of customers at node i. The aggregate network state is $\vec{S} = (\vec{n}_1, \vec{n}_2, \ldots, \vec{n}_N)$. The BCMP theorem asserts that the steady state distribution is:

$$p(\vec{S}) = \frac{1}{G} d(\vec{S}) g_1(\vec{n}_1) g_2(\vec{n}_2) \ldots g_N(\vec{n}_N)$$

where:

- G is a normalizing constant;
- $d(\vec{S}) = \prod_{k=1}^{m} \left[\prod_{n=0}^{M(\vec{S}, E_k)-1} \lambda_k(n) \right]$;
- the factor $g_i(\vec{n}_i)$ depends on the type of node i as follows:
  - if node i is of type 1 then

$$g_i(\vec{n}_i) = \frac{n_i! \prod_{r=1}^{R} \frac{e_{i,r}^{n_{i,r}}}{n_{i,r}!}}{\prod_{j=1}^{n_i} \mu_i C_i(j)}$$

  - if node i is of type 2 or 4 then

$$g_i(\vec{n}_i) = \frac{n_i! \prod_{r=1}^{R} \frac{(e_{i,r}/\mu_{i,r})^{n_{i,r}}}{n_{i,r}!}}{\prod_{j=1}^{n_i} C_i(j)}$$

  - if node i is of type 3 then

$$g_i(\vec{n}_i) = \prod_{r=1}^{R} \frac{(e_{i,r}/\mu_{i,r})^{n_{i,r}}}{n_{i,r}!}.$$

For an open network it is easy to calculate the expectation of the queue lengths for each class of customers by using some known formulas as we did in Chapter 2 for the specific network topology we were discussing there. For the deterministic routing network model we can analyze the FCFS strategy as an upper bound. Note that the nodes of the network can be modeled according to the BCMP notation as type 1 nodes. For the probabilistic routing network model (where different classes have

different service requirements) we can no longer analyze the FCFS policy. But we can analyze the LCFS [1] preemptive-resume and the processor-sharing policy. More precisely we can model the nodes of the network as nodes of type 2 or type 4.

[1] last come first serve

Chapter 4

The Single Station Case: Complete Characterization

In this chapter we will prove that our bounds fully characterize the achievable region for single station networks. In particular, we examine an M/M/1 multiclass queue where each class can have distinct service requirements and a model introduced by Klimov [Klim] which is a multiclass M/M/1 queue with Bernoulli feedback. Again, in Klimov's model, we allow different classes to have distinct service requirements. In this chapter we are considering work-conserving policies satisfying the assumptions (A), (B), (C) and (D) in Sec. 1.4.

4.1 Multiclass Queue

In this section we prove that the polytope which our lower bounds define is a *polymatroid polytope* with the associated function being *supermodular*. We then prove that this space is the achievable region for this problem. This is a special case of the result in [ShYa] where it is proven that the achievable region of any system that satisfies *strong conservation laws* (see [ShYa] for a definition) is a polymatroid polytope. Also, the achievable region of the multiclass queue is given in [GeMi]. As a consequence of the conservation laws, the c-μ rule is the optimal. The optimality of the c-μ rule can also be proved based on different arguments ([Klv2]). The proof we are giving for the

optimality of the c-μ rule is not new. But we do provide it for the purpose of completeness and because it illustrates in a simple example the power of our approach. Next come the definitions of polymatroids and supermodularity:

Definition 4.1 *Let N be a finite set and let y be a real-valued function on the subsets of N. Then y is supermodular if:*

$$y(S) + y(T) \leq y(S \cup T) + y(S \cap T) \quad \text{for } S, T \subseteq N$$

Definition 4.2 *Given a finite set N and a nondecreasing supermodular function y on N with $y(\emptyset) = 0$, the polytope*

$$P(y) = \left\{ x \in \Re_+^n \mid \sum_{j \in S} x_j \geq y(S) \quad \text{for } S \subseteq N \right\}$$

is the polymatroid polytope *associated with (N,y).*

Consider now, a multiclass queue with n classes of customers, as depicted in Figure 4-1. Customers of type i enter the station in a Poisson stream of rate λ_i and form class i. The station has a single server and each class of customers requires service time exponentially distributed with rate μ_i. Let x_i be the expected response time for customers of class i and let $n_i(t)$ be the number of customers of class i present in the system at time t.

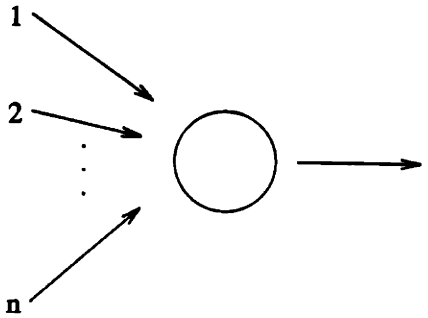

Figure 4-1: A multiclass queue

Our goal is to prove that the lower bounds in Chapter 3 define, for this case, a polyhedron which is the achievable region for the problem. In order to prove this, we

first prove some lemmas.

We will first derive the lower bounds for the above described system of the multiclass queue. We use the method described in the chapter 3. The next lemma summarizes the result:

Lemma 4.1 *For the M/M/1 multiclass queue and within the class of policies satisfying the assumptions of Sec. 1.4. the following inequalities hold for every subset S of the set of classes:*

$$\sum_{i \in S} \rho_i x_i \geq \frac{\sum_{i \in S}(\rho_i/\mu_i)}{1 - \sum_{i \in S} \rho_i} \tag{4.1}$$

where $\rho_i = \lambda_i/\mu_i$ is the traffic intensity of class i customers.

Proof : We again need to uniformize and we define the uniform rate to be:

$$\nu = \sum_{i=1}^{n}(\lambda_i + \mu_i)$$

Without loss of generality we scale time in order to have $\nu = 1$. So following the steps of our method for a subset S of $N = \{1, 2, \ldots, n\}$ we have:

$$R^S(t) = \sum_{i \in S} f^S(i) n_i(t)$$

Dropping S from $R^S(t)$ and $f^S(i)$ we get:

$$\begin{aligned} E[R^2(\tau_{n+1}) \mid \sigma_{\tau_n}] \;=\; & \sum_{i \in S} \lambda_i (R(\tau_n) + f(i))^2 + \sum_{i \notin S} \lambda_i R^2(\tau_n) + \\ & \sum_{i \in S} \mu_i 1\{\text{server busy from class } i \text{ at } \tau_n\}(R(\tau_n) - f(i))^2 + \\ & \sum_{i \in S} \mu_i 1\{\text{server idle from class } i \text{ at } \tau_n\} R^2(\tau_n) + \\ & \sum_{i \notin S} \mu_i R^2(\tau_n) \end{aligned}$$

To get tighter bounds we now choose (following the remarks in Chap.3):

$$f(i) = \frac{1}{\mu_i}, \ \forall i \in S$$

Therefore using:

$$1\{\text{server busy from classes } i \in S \text{ at } \tau_n\} \leq 1 \tag{4.2}$$

we get (4.1). □

The following lemma proves that the rhs of the above bounds is an achievable performance when preemptive priority is given to the classes in S.

Lemma 4.2 *Inequality (4.1) holds with equality for work-conserving policies satisfying the assumptions in Sec. 1.4, when preemptive priority is given to the classes in the subset S.*

Proof : Observe that in the derivation of the bound we used (4.2) in order to bound the term:

$$2f_i\mu_i 1\{\text{server busy from classes } i \in S \text{ at } \tau_n\}R(\tau_n)$$

If preemptive priority is given to the classes in S, however, we have:

$$R(\tau_n)1\{\text{server busy from classes } i \in S \text{ at } \tau_n\} = R(\tau_n)$$

because when $R(\tau_n) \neq 0$ (that is a customer of classes $i \in S$ is present) and preemptive priority is given to the classes $\in S$ then the server should definitely be working on a customer of classes $i \in S$. Otherwise, when $R(\tau_n) = 0$ the above equation holds trivially. □

Corollary 4.3 *When $S \equiv N$, then equation (4.1) holds with equality for work-conserving policies satisfying the assumptions in Sec. 1.4.*

Lemma 4.4 *The optimal solution of the LP:*

$$\min \sum_{i=1}^{n} c_i x_i$$

subject to:

$$\sum_{i \in S} \rho_i x_i \geq \frac{\sum_{i \in S}(\rho_i/\mu_i)}{1 - \sum_{i \in S} \rho_i}$$

$$x_i \in \Re_+$$

is the performance vector of the c-μ rule.

Proof : We are first going to establish the polymatroid structure of the polytope defined by (4.1). If for $A \subseteq N$ we define:

$$\rho(A) = \sum_{i \in A} \rho_i \ , \quad \rho'(A) = \sum_{i \in A}(\rho_i/\mu_i) \tag{4.3}$$

and

$$y(A) = \frac{\rho'(A)}{1 - \rho(A)} \tag{4.4}$$

then the LP takes the form:

$$\min \sum_{i=1}^{n} c_i x_i$$

subject to:

$$\sum_{i \in S} \rho_i x_i \geq y(S) \ \text{ for } S \subseteq N$$

$$x_i \in \Re_+$$

The function y on subsets of N is obviously non-decreasing. We are going to need the following lemma which is proven in Appendix A.

Lemma 4.5 *The real-valued function y on subsets of N is supermodular.*

Continuing the proof of the lemma 4.4 the above LP has as dual the following:

$$\max \sum_{S \subseteq N} \pi_S \, y(S)$$

subject to:

$$\sum_{S | j \in S} \pi_S \leq \frac{c_j}{\rho_j}$$

$$\pi_S \in \Re_+$$

Now, assuming that the costs are positive, suppose that:

$$\frac{c_1}{\rho_1} \geq \frac{c_2}{\rho_2} \geq \ldots \geq \frac{c_n}{\rho_n} \geq 0$$

and let $S^j = \{1, 2, \ldots, j\}$ for $j \in N$ and $S^0 = \emptyset$. Then the solution:

$$x_j = \frac{y(S^j) - y(S^{j-1})}{\rho_j} \tag{4.5}$$

is primal feasible because $\forall \, T \subseteq N$:

$$\begin{aligned} \sum_{j \in T} \rho_j x_j &= \sum_{j \in T} [y(S^j) - y(S^{j-1})] \\ &\geq \sum_{j \in T} [y(S^j \cap T) - y(S^{j-1} \cap T)] \quad \text{(supermodularity)} \\ &= y(N \cap T) - y(\emptyset) = y(T) \end{aligned}$$

The primal objective for this solution is:

$$\sum_{j=1}^{n} \frac{c_j}{\rho_j} (y(S^j) - y(S^{j-1})) \tag{4.6}$$

Additionally, the solution:

$$\pi_S = \begin{cases} \frac{c_j}{\rho_j} - \frac{c_{j+1}}{\rho_{j+1}} & \text{if } S \equiv S^j \text{ and } 1 \leq j < n \\ \frac{c_n}{\rho_n} & \text{if } S \equiv S^j \text{ and } j = n \\ 0 & \text{otherwise} \end{cases}$$

is dual feasible because $\pi_S \geq 0$ and:

$$\sum_{S|j\in S} \pi_S = \pi_{Sj} + \pi_{Sj+1} + \ldots + \pi_{SN} = \frac{c_j}{\rho_j}$$

The dual objective function is:

$$\sum_{j=1}^{n-1} \left[\frac{c_j}{\rho_j} - \frac{c_{j+1}}{\rho_{j+1}}\right] y(S^j) + c_n y(N) = \sum_{j=1}^{n} \frac{c_j}{\rho_j}(y(S^j) - y(S^{j-1})) \tag{4.7}$$

Therefore comparing (4.6), (4.7) we see that the proposed solutions of the primal and of the dual are also optimal since they are feasible and achieve the same objective value. What is now left to do is to prove that (4.5) gives the expected response time x_j for class j when preemptive priority is give in ascending order of the index j and highest priority is given to class 1. Indeed, as the following lemma (proved in Appendix B) asserts this is the case.

Lemma 4.6 *Equation (4.5) gives the expected response times when preemptive priority is given to classes* $1, 2, \ldots, n$ *in that order.*

Thus, the c-μ rule is proved to be optimal. □

Now, we have all the required tools to prove that the polyhedron defined by our bounds is the achievable region for the multiclass queue. We have the following theorem:

Theorem 4.7 *(Multiclass queue) The polyhedron:*

$$\sum_{i\in S} \rho_i x_i \geq \frac{\sum_{i\in S}(\rho_i/\mu_i)}{1 - \sum_{i\in S} \rho_i} \quad \forall S \subset N$$

$$\sum_{i\in N} \rho_i x_i = \frac{\sum_{i\in N}(\rho_i/\mu_i)}{1 - \sum_{i\in N} \rho_i}$$

$$x_i \in \Re_+$$

is the achievable space for the multiclass queue.

Proof : We have already proved in lemma 4.1 that this polyhedron includes the achievable region. In addition, lemma 4.4 shows that the performance at every extreme point is achieved via a preemptive priority rule (the c-μ rule) since every extreme point can become optimal by a proper selection of the costs. Thus, since every point in the polyhedron can be written as a convex combination of the extreme $(z_1, z_2, \ldots)$ points with coefficients $a_1, a_2, \ldots$, there exists a randomized policy that achieves the performance at this point. Namely, the randomized policy uses the priority rule corresponding to z_i with probability a_i. □

4.2 Klimov's Problem

In this section we prove that for this problem, also, our bounds fully characterize the achievable region. More precisely we prove that the polytope defined by the bounds described in Chapter 3 for the M/M/1 case of the Klimov's problem is the achievable region. The derived polytope has exactly the same structure as the polytope derived in [Tsou] for the M/G/1 case, under non-preemptive policies. In fact, the explicit form of the polytope is not given in [Tsou]; the rhs of the inequalities that define the polytope is an unknown function satisfying some properties. In contrast, we will explicitly define the polytope, in the M/M/1 case, and prove that it is the achievable region via a different technique. We are again, as in the previous section, considering work-conserving policies satisfying the assumptions in Sec. 1.4.

Let us first define the problem. Consider the single-server station of Fig. 4-2. Customers of type $i = 1, 2, \ldots, n$ enter the station in a Poisson stream of rate λ_i and form class i. Each class i of customers requires an exponentially distributed service time with rate μ_i. Upon service completion, customers of class i, with probability p_{ij} are fed back as customers of type j and with probability p_{i0} leave the system. Let $n_i(t)$ be the number of customers of class i at time t. The objective function has the form $\sum_{i=1}^{n} c_i x_i$ where x_i is the mean response time of class i.

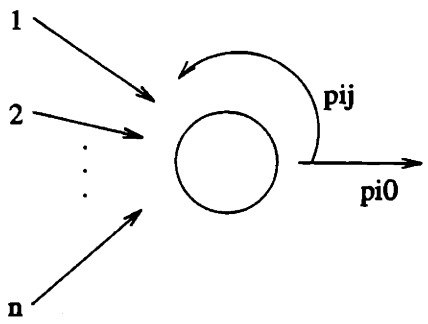

Figure 4-2: Klimov's problem.

The traffic equations for the above system are:

$$\hat{\lambda}_i = \lambda_i + \sum_{j=1}^{n} \hat{\lambda}_j p_{ji} \tag{4.8}$$

We are at first going to derive the lower bounds in the following lemma:

Lemma 4.8 *For every policy in the class of policies satisfying the assumptions in Sec. 1.4 the following inequality holds for every subset S of the set of classes:*

$$\sum_{i \in S} \hat{\lambda}_i f_S(i) x_i \geq \frac{N(S)}{D(S)} \tag{4.9}$$

where:

$$N(S) = \sum_{i \in S} \lambda_i f_S^2(i) + \sum_{i \in S} \hat{\lambda}_i \left[\sum_{j \in S} p_{ij} (f_S(i) - f_S(j))^2 + \sum_{j \notin S} p_{ij} f_S^2(i) \right] + \sum_{i \notin S} \hat{\lambda}_i \sum_{j \in S} p_{ij} f_S^2(j) \, ,$$

$$D(S) = 2 \left[1 - \sum_{i \in S} \lambda_i f_S(i) \right]$$

and

$$f_S(i) = \frac{1}{\mu_i} + \sum_{j \in S} p_{ij} f_S(j) \tag{4.10}$$

Proof : We uniformize by setting a uniform rate:

$$\nu = \sum_{k=1}^{n} \lambda_k + \sum_{k=1}^{n} \mu_k$$

and we scale time so that $\nu = 1$. So, following the steps of our method for a subset S of $\{1, 2, \ldots, n\}$, we have:

$$R^S(t) = \sum_{i \in S} f_S(i) n_i(t)$$

Dropping S from $R^S(t)$ and $f_S(i)$ we get:

$$\begin{aligned}
&E[R^2(\tau_{n+1}) \mid \sigma_{\tau_n}] = \\
&\sum_{i \in S} \lambda_i (R(\tau_n) + f(i))^2 + \sum_{i \notin S} \lambda_i R^2(\tau_n) + \\
&\sum_{i \in S} \mu_i 1\{\text{server busy from class } i \text{ at } \tau_n\} \left[\sum_{j \in S} p_{ij} (R(\tau_n) - f(i) + f(j))^2 + \right. \\
&\qquad \left. \sum_{j \notin S} p_{ij} (R(\tau_n) - f(i))^2 \right] + \\
&\sum_{i \in S} \mu_i 1\{\text{server idle from class } i \text{ at } \tau_n\} R^2(\tau_n) + \\
&\sum_{i \notin S} \mu_i 1\{\text{server busy from class } i \text{ at } \tau_n\} \left[\sum_{j \in S} p_{ij} (R(\tau_n) + f(j))^2 + \right. \\
&\qquad \left. \sum_{j \notin S} p_{ij} R^2(\tau_n) \right] + \\
&\sum_{i \notin S} \mu_i 1\{\text{server idle from class } i \text{ at } \tau_n\} R^2(\tau_n)
\end{aligned}$$

Note here that the subset $\{j \mid j \notin S\}$ includes also node 0 which is the external world of the network. As we did before we have to make the proper matchings in terms of the f-parameters in order to get tighter bounds. Therefore we set:

$$\mu_i \left[\sum_{j \in S} p_{ij} (f(i) - f(j)) + \sum_{j \notin S} p_{ij} f(i) \right] = f \quad \forall i \in S \tag{4.11}$$

Then by using the fact that:

$$1\{\text{server busy from classes } i \in S \text{ at } \tau_n\} \leq 1 \tag{4.12}$$

we are able to bound the term:

$$2f1\{\text{server busy from classes } i \in S \text{ at } \tau_n\}R(\tau_n)$$

On the other hand we use:

$$1\{\text{server busy from classes } i \notin S \text{ at } \tau_n\} \geq 0 \tag{4.13}$$

to bound the term:

$$\sum_{i \notin S} \mu_i 1\{\text{server busy from class } i \text{ at } \tau_n\} 2 \sum_{j \in S} p_{ij} R(\tau_n) f(j)$$

Thus, by using Little's law and the solution of the traffic equations (4.8), we finally derive the following bound:

$$\sum_{i \in S} \hat{\lambda}_i f(i) x_i \geq \frac{N(S)}{D(S)}$$

where:

$$N(S) = \sum_{i \in S} \lambda_i f^2(i) + \sum_{i \in S} \hat{\lambda}_i \left[\sum_{j \in S} p_{ij} (f(i) - f(j))^2 + \sum_{j \notin S} p_{ij} f^2(i) \right] + \sum_{i \notin S} \hat{\lambda}_i \sum_{j \in S} p_{ij} f^2(j)$$

and

$$D(S) = 2 \left[f - \sum_{i \in S} \lambda_i f(i) \right]$$

Now, the only task left is to determine the value of the f-parameters that yields the tightest bounds satisfying the conditions that were imposed to them in the proof. According to the remarks of theorem 3.3, we choose $f(i)$ to be the expected remaining service time within the class S, that is we choose $f(i)$ satisfying (4.10). Let us check if this choice satisfies (4.11). Plugging it to (4.11) we get:

$$\begin{aligned} f &= \mu_i f(i) \sum_{j \in S} p_{ij} - \mu_i \sum_{j \in S} p_{ij} f(j) + \mu_i f(i) \sum_{j \notin S} p_{ij} \\ &= f(i)\mu_i - \mu_i \sum_{j \in S} p_{ij} f(j) \\ &= 1 \end{aligned}$$

Note from (4.9) that we are not interested in the absolute values of the f-parameters but in their ratio $f(i)/f$. Thus plugging $f = 1$ into the expression for $D(S)$ we proved the lemma. □

The following lemma proves that the rhs of the above bounds is an achievable performance when preemptive priority is given to the classes in S.

Lemma 4.9 *Inequality (4.9) holds with equality, for work-conserving policies satisfying the assumptions of Sec. 1.4 when preemptive priority is given to the classes in the subset S of the set of classes.*

Proof : The proof is similar to the proof of lemma 4.2. In the derivation of (4.9) we used (4.12) in order to bound the term

$$2f1\{\text{server busy from classes } i \in S \text{ at } \tau_n\}R(\tau_n)$$

Moreover, it holds that:

$$R(\tau_n)1\{\text{server busy from classes } i \in S \text{ at } \tau_n\} = R(\tau_n)$$

because when $R(\tau_n) \neq 0$ (that is a customer of classes $i \in S$ is present) and preemptive priority is given to the classes $\in S$ then the server should definitely be working on a customer of classes $i \in S$. Otherwise, when $R(\tau_n) = 0$ the above equation holds trivially. In addition to that, we also used (4.13) to bound the term

$$\sum_{i \notin S} \mu_i 1\{\text{server busy from class } i \text{ at } \tau_n\} \sum_{j \in S} p_{ij} 2R(\tau_n)f(j)$$

Moreover, it holds that:

$$R(\tau_n)1\{\text{server busy from classes } i \notin S \text{ at } \tau_n\} = 0$$

because when $R(\tau_n) \neq 0$ (that is a customer of classes $i \in S$ is present) and preemptive priority is given to the classes $\in S$ then the server should definitely not be working

on a customer of classes $i \notin S$. Otherwise, when $R(\tau_n) = 0$ the above equation holds trivially. □

Corollary 4.10 *When $S = \{1, 2, \ldots, n\}$ then (4.9) holds with equality within the class of policies considered in lemma 4.9.*

The next lemma characterizes the extreme points of the polytope defined by (4.9) and proves that Klimov's algorithm [Klim] is the optimal priority rule. Let us introduce some notation. By $E = \{1, 2, \ldots, n\}$ we denote the entire set of classes. Since $N(S), D(S)$ defined in lemma 4.8 are functions of the subset S of E we define a real-valued function $G(S)$ on the subsets of E. Let also denote by $\{\pi_1, \pi_2, \ldots, \pi_n\}$ an ordering of the set of classes in E. Let finally $c_i' = \frac{c_i}{\lambda_i}$. We have the following lemma:

Lemma 4.11 *The solution of the LP:*

$$\min \sum_{i \in E} c_i' n_i$$

$$\textit{subject to:} \tag{4.14}$$

$$\sum_{i \in S} f_S(i) n_i \geq G(S) \quad S \subset E$$

$$\sum_{i \in E} f_E(i) n_i = G(E)$$

$$n_i \in \Re^+$$

is the solution of the system of equations:

$$\sum_{i=1}^{k} f_{\{\pi_1, \pi_2, \ldots, \pi_i\}}(\pi_i) n_{\pi_i} = G(\{\pi_1, \pi_2, \ldots, \pi_k\}) \quad k = 1, 2, \ldots, n \tag{4.15}$$

where the optimal ordering $\pi_1, \pi_2, \ldots, \pi_n$ is given by the following adaptive algorithm:

Step 1:

$$E^0 \leftarrow E$$

$$y_{E^0} = \min_{i \in E^0} \left\{ \frac{c_i'}{f_{E^0}(i)} \right\}$$

$$\pi_n = \arg\min\left\{\frac{c_i'}{f_{E^0}(i)}\right\}$$

Step 2: *For* $k = 1, 2, \ldots, n-1$

$$E^k \leftarrow N^{k-1} \setminus \{\pi_{n-k+1}\}$$

$$y_{E^k} = \min_{i \in E^k}\left\{\frac{c_i' - \sum_{j=0}^{k-1} f_{Ej}(i) y_{Ej}}{f_{E^k}(i)}\right\}$$

$$\pi_{n-k} = \arg\min\left\{\frac{c_i' - \sum_{j=0}^{k-1} f_{Ej}(i) y_{Ej}}{f_{E^k}(i)}\right\}$$

Proof : Note at first that in the statement of the lemma, without loss of generality, we have written the polytope defined in lemma 4.8 in the space of the mean number of customers in the system instead of mean response times. We are going to give a duality proof. The dual of (4.14) is:

$$\max \sum_S y_S G(S)$$

$$\text{subject to:} \tag{4.16}$$

$$\sum_{S|i \in S} y_S f_S(i) \leq c_i'$$

$$y_S \geq 0$$

$$y_E \quad \text{unconstrained}$$

Let a proposed primal solution be the solution of the system of equations (4.15). This solution is feasible. To see that consider the following subsets of E:

$$S^k = \{\pi_1, \pi_2, \ldots, \pi_k\} \quad k = 1, 2, \ldots, n$$

For each of these subsets lemma 4.9 asserts that the bound (4.9) is satisfied with equality if preemptive priority is given to classes in S^k. But the union of those rules is just the policy "assign preemptive priorities according to the ordering $\pi_1, \pi_2, \ldots, \pi_n$".

This is a valid policy and therefore according to lemma 4.8 has performance in the feasible space of the LP (4.14). Thus, primal feasibility is proven. Now, as dual solution consider:

$$y_S = \begin{cases} y_{E^k} & \text{if } S \equiv E^k \; k = 0, 1, \ldots, n-1 \\ 0 & \text{otherwise} \end{cases} \tag{4.17}$$

We are going to prove dual feasibility using induction.

At the first step of the induction consider the solution:

$$y_S = \begin{cases} y_{E^0} & \text{if } S \equiv E^0 \\ 0 & \text{otherwise} \end{cases} \tag{4.18}$$

At each step we will update the solution. We have:

$$\sum_{S|i \in S} y_S f_S(i) = y_{E^0} f_{E^0}(i) \leq c'_i$$

which verifies that (4.18) is feasible from the definition of y_{E^0}.

For the the second step of induction consider the solution:

$$y_S = \begin{cases} y_{E^0} & \text{if } S \equiv E^0 \\ y_{E^1} \geq 0 & \text{if } S \equiv E^1 \\ 0 & \text{otherwise} \end{cases} \tag{4.19}$$

To have dual feasibility:

$$\sum_{S|i \in S} y_S f_S(i) = y_{E^0} f_{E^0}(i) + y_{E^1} f_{E^1}(i) \leq c'_i$$

$$\Rightarrow y_{E^1} \leq \frac{c'_i - y_{E^0} f_{E^0}(i)}{f_{E^1}(i)}$$

So selecting the minimum according to the adaptive algorithm we satisfy dual feasibility. Thus, inductively we prove that (4.17) is dual feasible. Moreover from (4.15) note that for the non-zero dual variables the primal constraints are satisfied with

equality. Thus, complementary-slackness is available. □

Remarks : The algorithm to determine the optimal ordering is Klimov's algorithm. The real-valued function $G(S)$ on the subsets of E is not supermodular. It has the property that the system of equations (4.15) has a solution in the polytope. The polytope defined in the statement of this lemma with such a G function is named extended-polymatroid in [BGeT].

Now, we have all the required tools to prove that the polyhedron defined by our bounds is the achievable region for the Klimov's problem. We have the following theorem:

Theorem 4.12 *(Klimov's problem) The polyhedron:*

$$\sum_{i \in S} f_S(i) n_i \geq G(S) \quad S \subset E$$

$$\sum_{i \in E} f_E(i) n_i = G(E)$$

$$n_i \in \Re^+$$

with $f_S(i)$ given by (4.10) and $G(S) = N(S)/D(S)$, is the achievable space for the Klimov's problem.

Proof : We have already proved in lemma 4.8 that this polyhedron includes the achievable region. In addition, lemma 4.11 shows that the performance at every extreme point is achieved via a preemptive priority rule since every extreme point can become optimal via by a proper selection of the costs. Thus, since every point in the polyhedron can be written as a convex combination of the extreme $(z_1, z_2, \ldots)$ points with coefficients $a_1, a_2, \ldots$, there exists a randomized policy that achieves the performance at this point. Namely, the randomized policy uses the priority rule corresponding to z_i with probability a_i. □

As a final remark, our bounding technique provides the exact characterization of

the achievable region in the multiclass scheduling problem with and without feedback. An additional advantage of our method is that it provides explicit formulae for the constants involved in the description of the performance polytope. As a result, by solving the system of equations (4.15), one can get the performance of the optimal policy. To the best of our knowledge the optimal policy has not been analyzed in the Klimov's problem. The reason is that it is very difficult to characterize the feedback process.

Chapter 5

A Refined Bounding Technique

In this chapter we are going to present a more refined method based on the conservation inequalities discussed in the previous chapters. Among the advantages of this methods are:

- It yields tighter bounds.
- It yields a more appealing, in terms of computation, LP than the LP (3.11) or (3.12) proposed in Section 3.1.
- It yields bounds independent of the choice of the f-parameters which was based only on intuitive grounds in Chapter 3.

We are going to derive these bounds for the most general network model we have considered so far, the probabilistic routing model defined in Section 1.3 and in Chapter 3. This model as we have mentioned there includes the deterministic routing model presented in Chapter 3. Consider, therefore, the probabilistic routing network model and let $n_{(i,r)}$ be the steady-state number of customers of class (i,r) in the system. The traffic equations of this network are given in (3.5).

5.1 Lower Bounds; Conservation Equalities

We denote by $E = \{(i,r) \mid \ i = 1,2,\ldots,N,\ r = 1,2,\ldots,R\}$ the entire set of classes and by $\bar{E}$ its cardinality. We define again a measure of the work to be done in the

network as follows:

$$R(t) = \sum_{(i,r)\in E} f(i,r) n_{(i,r)}(t) \tag{5.1}$$

where $f(i,r)$ are the multiplying constants which we named f-parameters. Letting τ_n be the sequence of transition times in the uniformized Markov chain corresponding to the network we also define the following variables:

$$I_{i,r;j,r'} = E[1\{\text{server } i \text{ busy from type } r \text{ at } \tau_n\} n_{(j,r')}(\tau_n)] \tag{5.2}$$

$$N_{i;j,r'} = E[1\{\text{server } i \text{ idle at } \tau_n\} n_{(j,r')}(\tau_n)] \tag{5.3}$$

where $1\{\cdot\}$ is the indicator function and the expectations are taken with respect to the invariant distribution mentioned in Assumption (A). We finally define an ordering between different classes:

$$(i_1, r_1) \ \ldots \ (i_k, r_k) \ \ldots \ (i_E, r_E)$$

where (i_k, r_k) is the k^{th} class and also define a function that gives the order of a class by:

$$\text{ind}(i,r) = k$$

when (i,r) is the k^{th} class. We then have the following theorem for scheduling strategies satisfying the assumptions imposed in the introduction of the thesis:

Theorem 5.1 *For the probabilistic routing network model the following equalities hold independent of the scheduling strategy, satisfying the assumptions in Sec. 1.4:*

$$2\mu_{i,r} I_{i,r;i,r} - 2 \sum_{(j,r')\in E} \mu_{j,r'} p_{j,r';i,r} I_{j,r';i,r} - 2\lambda_{0,r} q_{i,r} \lambda_{i,r} x_{i,r} =$$
$$\lambda_{0,r} q_{i,r} + \lambda_{i,r}(1 - p_{i,r;i,r}) + \sum_{\{(j,r')\in E | j\neq i, r'\neq r\}} \lambda_{j,r'} p_{j,r';i,r} \qquad \forall (i,r) \in E \tag{5.4}$$

$$\mu_{i,r} I_{i,r;j,r'} + \mu_{j,r'} I_{j,r';i,r} - \sum_{(k,w)\in E} \mu_{k,w} p_{k,w;i,r} I_{k,w;j,r'} - \sum_{(k,w)\in E} \mu_{k,w} p_{k,w;j,r'} I_{k,w;i,r} -$$
$$\lambda_{0,r} q_{i,r} \lambda_{j,r'} x_{j,r'} - \lambda_{0,r'} q_{j,r'} \lambda_{i,r} x_{i,r} =$$
$$-\lambda_{i,r} p_{i,r;j,r'} - \lambda_{j,r'} p_{j,r';i,r} \qquad \forall (i,r),(j,r') \in E \mid ind(i,r) > ind(j,r') \quad (5.5)$$

Proof : We are again applying uniformization as in Theorem 3.2 and writing the recursion for $R(t)$. Let the uniform rate be:

$$\nu = \sum_{r} \lambda_{0,r} + \sum_{i,r} \mu_{i,r}$$

Without loss of generality, we scale time such that $\nu = 1$. Let σ_{τ_n} be the σ-field generated by events up to time τ_n or more intuitively the previous history. We have:

$$E[R^2(\tau_{n+1}) \mid \sigma_{\tau_n}] =$$
$$\sum_{(i,r)\in E} \lambda_{0,r} q_{i,r} (R(\tau_n) + f(i,r))^2 +$$
$$\sum_{(i,r)\in E} \mu_{i,r} 1\{\text{server } i \text{ busy from type } r \text{ at } \tau_n\}$$
$$\left[\sum_{(j,r')\in E} p_{i,r;j,r'} (R(\tau_n) - f(i,r) + f(j,r'))^2 + p_{i,r;0} (R(\tau_n) - f(i,r))^2 \right] +$$
$$\sum_{(i,r)\in E} \mu_{i,r} 1\{\text{server } i \text{ idle from type } r \text{ at } \tau_n\} R^2(\tau_n)$$

Rearranging terms, taking expectations and using that:

$$E[1\{\text{server } i \text{ busy from type } r \text{ at } \tau_n\}] = \frac{\lambda_{i,r}}{\mu_{i,r}}$$

where $\lambda_{i,r}$ is the solution of the traffic equations (3.5) we obtain:

$$2\sum_{(i,r)\in E}\mu_{i,r}\left[\sum_{(j,r')\in E}p_{i,r;j,r'}(f(i,r)-f(j,r'))+p_{i,r;0}f(i,r)\right]$$

$$\lim_{n\to\infty}E[1\{\text{server } i \text{ busy from type } r \text{ at } \tau_n\}R(\tau_n)]-$$

$$2\sum_{(i,r)\in E}\lambda_{0,r}q_{i,r}f(i,r)E[R(\tau_n)]=$$

$$\sum_{(i,r)\in E}\lambda_{0,r}q_{i,r}f^2(i,r)+\sum_{(i,r)\in E}\lambda_{i,r}\left[\sum_{(j,r')\in E}p_{i,r;j,r'}(f(i,r)-f(j,r'))^2+p_{i,r;0}f^2(i,r)\right] \quad (5.6)$$

Moreover, it is seen from (5.1) and (5.2) that:

$$E[1\{\text{server } i \text{ busy from type } r \text{ at } \tau_n\}R(\tau_n)]=\sum_{(j,r')\in E}f(j,r')I_{i,r;j,r'}$$

Therefore if we enumerate the different classes by letting (i_k,r_k) be the k^{th} class and define a vector:

$$\vec{f}=[f(i_1,r_1)\ \ldots\ f(i_k,r_k)\ \ldots\ f(i_E,r_E)]^T$$

then (5.6) which is a symmetric quadratic form in terms of the f-parameters can be written in the form:

$$\vec{f}^TQ\vec{f}=\vec{f}^TQ_0\vec{f} \quad (5.7)$$

for some symmetric matrices Q,Q_0. Since (5.7) is valid for every choice of the f-parameters, $Q=Q_0$. Let us write down explicitly the equations in (5.7). From (5.6) the diagonal terms have the form:

$$2\mu_{i,r}\left[p_{i,r;0}+\sum_{\{(j,r')\in E|j\neq i,r'\neq r\}}p_{i,r;j,r'}\right]I_{i,r;i,r}-2\sum_{\{(j,r')\in E|j\neq i,r'\neq r\}}\mu_{j,r'}p_{j,r';i,r}I_{j,r';i,r}-$$

$$2\lambda_{0,r}q_{i,r}\lambda_{i,r}x_{i,r}=$$

$$\lambda_{0,r}q_{i,r}+\lambda_{i,r}\left[p_{i,r;0}+\sum_{\{(j,r')\in E|j\neq i,r'\neq r\}}p_{i,r;j,r'}\right]+\sum_{\{(j,r')\in E|j\neq i,r'\neq r\}}\lambda_{j,r'}p_{j,r';i,r}$$

from which we easily obtain (5.4) since the transition probabilities add up to one. For the off-diagonal terms we consider only terms above the diagonal in the matrix

equation (5.7). We then have:

$$\mu_{i,r}\left[p_{i,r;0}+\sum_{\{(j,r')\in E|j\neq i,r'\neq r\}}p_{i,r;j,r'}\right]I_{i,r;j,r'}-\mu_{i,r}p_{i,r;j,r'}I_{i,r;i,r}+$$

$$\mu_{j,r'}\left[p_{j,r';0}+\sum_{\{(k,w)\in E|k\neq j,w\neq r'\}}p_{j,r';k,w}\right]I_{j,r';i,r}-\mu_{j,r'}p_{j,r';i,r}I_{j,r';j,r'}-$$

$$\sum_{\{(k,w)\in E|k\neq i,j,w\neq r,r'\}}\mu_{k,w}p_{k,w;i,r}I_{k,w;j,r'}-\sum_{\{(k,w)\in E|k\neq i,j,w\neq r,r'\}}\mu_{k,w}p_{k,w;j,r'}I_{k,w;i,r}-$$

$$\lambda_{0,r}q_{i,r}\lambda_{j,r'}x_{j,r'}-\lambda_{0,r'}q_{j,r'}\lambda_{i,r}x_{i,r}=$$

$$-\lambda_{i,r}p_{i,r;j,r'}-\lambda_{j,r'}p_{j,r';i,r}$$

which implies (5.5). □

Remarks : Note that in the proof of theorem 3.2 we were making different approximations for every non-empty subset of E (namely, we were using the fact that certain indicator functions were less than 1) and thus all the inequalities (one for each non-empty subset of E) were useful. On the contrary here, since we are proceeding with equalities (we do not bound indicator functions with 1) we need only consider the entire set E. For every other non-empty subset S of E we can get the corresponding to $Q = Q_0$ matrix equation, by setting equal to zero the f-parameters corresponding to classes out of S. But since (5.7) holds for every choice of the f-parameters we are not going to get more information.

In addition to (5.4) and (5.5) we can also derive some more equalities by taking the product of all the possible events at each node i of the network with $n_{(j,r')}$ for every class (j,r'). Namely, we have the following theorem:

Theorem 5.2 *For each node i of the network and each class $(j,r') \in E$ and within the considered class of policies, the following equality holds:*

$$\sum_{r|(i,r)\in E}I_{i,r;j,r'}+N_{i;j,r'}=\lambda_{j,r'}x_{j,r'} \tag{5.8}$$

Proof : Note that at node i of the network the events:

$$B_{i,r}(\tau_n) = \text{"server } i \text{ busy from type } r \text{ at } \tau_n\text{"}$$

for every class (i,r) along with the event:

$$BN_i(\tau_n) = \text{"server } i \text{ idle at } \tau_n\text{"}$$

are mutually exclusive and exhaustive. Thus:

$$E\left[n_{(j,r')}(\tau_n)\left(\sum_{\{r|(i,r)\in E\}} 1\{BI_{i,r}(\tau_n)\} + 1\{BN_i(\tau_n)\}\right)\right] = n_{(j,r')} = \lambda_{j,r'}x_{j,r'}$$

It is not hard to see considering the definitions (5.2), (5.3) that we get (5.8). □

The polyhedron in standard form defined by (5.4), (5.5) and (5.8) is an approximation of the achievable for the network region at least as good as the polyhedron obtained by the approach in Chapter 3. This is due to the fact that they are derived using the same recursive equation. The following theorem proves this claim. In particular, we prove that the polytope defined by (5.4), (5.5) and (5.8) is a subset of the union of polytopes defined by (3.8) for all the choices of the f-parameters. Therefore the polytope derived via the refined method is a tighter approximation of the performance space than the polytope derived in Chapter 3. Let us denote by $R1$ the polyhedron defined by (3.8) and the positivity constraints of the $x_i's$. Let also denote by $R2$ the polyhedron defined by (5.4), (5.5) and (5.8) along with the positivity constraints of the variables $x_{(i,r)}, I_{i,r;j,r'}$ and $N_{i;j,r'}$.

Theorem 5.3 *If* $\{x_{(i,r)}\ I_{i,r;j,r'}\ N_{i;j,r'},\ (i,r),(j,r') \in E\}$ *is a feasible solution of* $R2$ *then* $\{x_{(i,r)},\ (i,r) \in E\}$ *is a feasible solution of* $R1$.

Proof : Consider a feasible solution of $R2$. Since equation (5.7) holds for every choice of the f-parameters, it is seen that we can write down an equality for every non-empty $S \subseteq E$, if we set equal to zero the f-parameters corresponding to classes

outside S. For any such S, it is apparent from (5.8) that:

$$\sum_{\{r|(i,r)\in S\}} I_{i,r;j,r'} + \sum_{\{r|(i,r)\notin S\}} I_{i,r;j,r'} + N_{i;j,r'} = \lambda_{j,r'} x_{j,r'}$$

which implies that

$$\sum_{\{r|(i,r)\in S\}} I_{i,r;j,r'} \leq n_{(j,r')}$$

and

$$E[1\{\text{server i busy from classes } (i,r) \in S \text{ at } \tau_n\} n_{(j,r')}(\tau_n)] \leq n_{(j,r')} \qquad (5.9)$$

Now recall that in the proof of Theorem 3.2 we used that:

$$1\{\text{server i busy from classes } (i,r) \in S \text{ at } \tau_n\} \leq 1 \qquad (5.10)$$

and

$$1\{\text{server } i \text{ busy from type } r \text{ at } \tau_n\} \geq 0 \qquad (5.11)$$

in order to get the bound (3.8). That is, we first wrote down the recursive equation, we then applied (5.10) and (5.11) and we finally took expectations to get (3.8). It can be seen that exactly the same bound is obtained by first writing down the recursive equation, then taking expectations and finally using (5.9) along with the positivity constraint for the variables $I_{i,r;j,r'}$. Thus, from the equality in (5.7) corresponding to the subset S, the inequality (3.8) is derived by using (5.8). In other words we proved that if $\{x_{(i,r)}\ I_{i,r;j,r'}\ N_{i;j,r'},\ (i,r),(j,r') \in E\}$ is a feasible solution of $R2$ then $\{x_{(i,r)},\ (i,r) \in E\}$ is a feasible solution of $R1$. □

Remarks : We can intuitively argue that (5.8) contains more information than the somewhat "crude" (5.9). Thus, we strongly believe that there are, in general, feasible solutions $\{x_{(i,r)},\ (i,r) \in E\}$ of $R1$ such that $\{x_{(i,r)}\ I_{i,r;j,r'}\ N_{i;j,r'},\ (i,r),(j,r') \in E\}$ is not a feasible solution of $R2$.

The bound on achievable performance (LB) can now be obtained from the follow-

ing LP:

$$Z_{LB} = \min \sum_{\forall \text{ class } (i,r)} c_{(i,r)} x_{(i,r)}$$

$$\text{subject to:} \tag{5.12}$$

$$2\mu_{i,r} I_{i,r;i,r} - 2 \sum_{(j,r') \in E} \mu_{j,r'} p_{j,r';i,r} I_{j,r';i,r} - 2\lambda_{0,r} q_{i,r} \lambda_{i,r} x_{i,r} =$$
$$\lambda_{0,r} q_{i,r} + \lambda_{i,r}(1 - p_{i,r;i,r}) + \sum_{\{(j,r') \in E | j \neq i, r' \neq r\}} \lambda_{j,r'} p_{j,r';i,r} \qquad \forall (i,r) \in E$$

$$\mu_{i,r} I_{i,r;j,r'} + \mu_{j,r'} I_{j,r';i,r} - \sum_{(k,w) \in E} \mu_{k,w} p_{k,w;i,r} I_{k,w;j,r'} - \sum_{(k,w) \in E} \mu_{k,w} p_{k,w;j,r'} I_{k,w;i,r} -$$
$$\lambda_{0,r} q_{i,r} \lambda_{j,r'} x_{j,r'} - \lambda_{0,r'} q_{j,r'} \lambda_{i,r} x_{i,r} =$$
$$-\lambda_{i,r} p_{i,r;j,r'} - \lambda_{j,r'} p_{j,r';i,r} \qquad \forall (i,r),(j,r') \in E \mid \text{ind}(i,r) > \text{ind}(j,r')$$

$$\sum_{\{r | (i,r) \in E\}} I_{i,r;j,r'} + N_{i;j,r'} = \lambda_{j,r'} x_{j,r'}$$

$$x_{(i,r)}, I_{i,r;j,r'}, N_{i;j,r'} \in \Re_+ \quad \forall i, r, j, r'$$

According to Theorem 5.3 the lower bound on achievable performance calculated with the LP (5.12) is an upper bound to the one calculated with the LP (3.12). Moreover, note that the LP (5.12) is more tractable computationally than the LP (3.12). The former has $O(\bar{E}^2)$ variables and $O(\bar{E}^2)$ constraints while the latter had $\bar{E}$ variables and $O(2^{\bar{E}})$ constraints.

5.2 Consistency of the Refined Method with the Earlier Approach

To check the consistency of the two methods, we will consider the case of the multiclass queue where we are able to exactly characterize the achievable region. Thus, consider a multiclass queue with n classes of customers. By $N = \{1, 2, \ldots, n\}$ we denote the entire set of classes. Customers of type i enter the station in a Poisson

stream of rate λ_i and form class i. The station has a single server and each class of customers requires service time exponentially distributed with rate μ_i. Let $n_i(t)$ be the number of customers of class i present in the system at time t, and let n_i be the steady state quantity. Let also $\rho_i = \lambda_i/\mu_i$ be the traffic intensity of class i customers.

As we have shown in Chapter 4, the performance space of the multiclass queue is described by the following polyhedron $P1$:

$$\textbf{P1:} \qquad \sum_{i \in S} \frac{n_i}{\mu_i} \geq \frac{\sum_{i \in S}(\rho_i/\mu_i)}{1 - \sum_{i \in S} \rho_i} \quad \forall S \subset N$$

$$\sum_{i \in N} \frac{n_i}{\mu_i} = \frac{\sum_{i \in N}(\rho_i/\mu_i)}{1 - \sum_{i \in N} \rho_i}$$

$$n_i \in \Re^+$$

The polyhedron $(P2)$ derived via the refined method is given in the following lemma. Let us first define in analogy with (5.2) and (5.3):

$$I_{ij} = E[1\{\text{server busy from class } i \text{ at } \tau_n\} n_j(\tau_n)] \tag{5.13}$$

$$N_j = E[1\{\text{server idle at } \tau_n\} n_j(\tau_n)] \tag{5.14}$$

Note that $N_j = 0 \ \forall j$, because when $n_j(\tau_n) \neq 0$ the server should be definitely working [1].

Lemma 5.4 *For the multiclass queue and for work-conserving policies satisfying the assumptions in Sec. 1.4 the following polyhedron P2 includes the performance space:*

$$\textbf{P2:} \qquad \mu_i I_{ii} - \lambda_i n_i = \lambda_i \quad \forall i \tag{5.15}$$

$$\mu_i I_{ij} + \mu_j I_{ji} - \lambda_j n_i - \lambda_i n_j = 0 \quad \forall i, j \mid j > i \tag{5.16}$$

[1] Recall that throughout the analysis of this problem in Chapter 4 we were considering work-conserving policies. Therefore to demonstrate the equivalence we retain this assumption now.

$$\sum_{i \in N} I_{ij} = n_j \quad \forall j \tag{5.17}$$

$$n_i, I_{ij} \in \Re^+ \quad \forall i, j$$

Proof : The recursive equation for $R(t)$ takes the form:

$$\begin{aligned} E[R^2(\tau_{n+1}) \mid \sigma_{\tau_n}] \;=\; & \sum_{i \in N} \lambda_i (R(\tau_n) + f(i))^2 + \\ & \sum_{i \in N} \mu_i 1\{\text{server busy from class } i \text{ at } \tau_n\}(R(\tau_n) - f(i))^2 + \\ & \sum_{i \in N} \mu_i 1\{\text{server idle from class } i \text{ at } \tau_n\} R^2(\tau_n) \end{aligned}$$

Following the steps illustrated in Theorem 5.1 we get:

$$\sum_{i \in N} \mu_i f(i) \sum_{j \in N} f(j) I_{ij} - \sum_{i,j \in N} \lambda_i f(i) f(j) n_j = \sum_{i \in N} \lambda_i f^2(i) \tag{5.18}$$

From the last equation one can derive the equations that define $P2$. Equation (5.15) corresponds to (5.4) and equation (5.16) to (5.5). We also have to consider the equations corresponding to (5.8) which take the form (5.17). Therefore the refined method yields the polyhedron $P2$, defined by (5.15), (5.16), (5.17) along with the positivity constraints of all the variables. □

Next we will show that the polyhedron derived by the elimination of the I_{ij} variables in $P2$ is the polyhedron $P1$. In other words $P1$ is the projection of $P2$ in the space of n_i's.

Theorem 5.5 *The polyhedron $P2$, derived via the refined method for the multiclass queue, projected in the $n_i, i = 1, 2, \ldots, n$, space yields $P1$.*

Proof : Let $P2'$ the projection of $P2$ in the space of n_i's. We want to show that $P2' \equiv P1$. We first show that $P2' \subset P1$.

Consider $P2$. We eliminate the variables I_{ij}. Namely, dividing (5.16) with $\mu_i \mu_j$

and adding for all i, j such that $j > i$ we get:

$$\sum_{i,j \in N | j>i} \left[\frac{I_{ij}}{\mu_j} + \frac{I_{ji}}{\mu_i} - \rho_j \frac{n_i}{\mu_i} - \rho_i \frac{n_j}{\mu_j} \right] = 0 \tag{5.19}$$

which implies

$$\sum_{j \in N} \frac{1}{\mu_j} \sum_{i \in N | i<j} I_{ij} + \sum_{i \in N} \frac{1}{\mu_i} \sum_{j \in N | j>i} I_{ji} - \sum_{j \in N} \rho_j \sum_{i \in N | i<j} \frac{n_i}{\mu_i} - \sum_{i \in N} \rho_i \sum_{j \in N | j>i} \frac{n_j}{\mu_j} = 0 \Rightarrow$$

$$\sum_{j \in N} \frac{1}{\mu_j} \sum_{i \in N | i<j} I_{ij} + \sum_{j \in N} \frac{1}{\mu_j} \sum_{i \in N | i>j} I_{ij} - \sum_{j \in N} \rho_j \sum_{i \in N | i<j} \frac{n_i}{\mu_i} - \sum_{j \in N} \rho_j \sum_{i \in N | i>j} \frac{n_i}{\mu_i} = 0 \tag{5.20}$$

Observing that the jj term is missing from the above summation we use (5.15) and (5.17) to get:

$$\sum_{j \in N} \frac{1}{\mu_j} n_j - \sum_{j \in N} \frac{1}{\mu_j} (\rho_j + \rho_j n_j) - \sum_{j \in N} \rho_j \sum_{i \in N | i \neq j} \frac{n_i}{\mu_i} = 0 \tag{5.21}$$

or equivalently,

$$\sum_{i \in N} \frac{n_i}{\mu_i} = \frac{\sum_{i \in N} (\rho_i / \mu_i)}{1 - \sum_{i \in N} \rho_i}$$

which is the equality of $P1$.

Next, to get the inequalities we divide (5.16) with $\mu_i \mu_j$ and adding for all $i, j \in S$ such that $j > i$ we get:

$$\sum_{i,j \in S | j>i} \left[\frac{I_{ij}}{\mu_j} + \frac{I_{ji}}{\mu_i} - \rho_j \frac{n_i}{\mu_i} - \rho_i \frac{n_j}{\mu_j} \right] = 0 \tag{5.22}$$

from which we get:

$$\sum_{j \in S} \frac{1}{\mu_j} \sum_{i \in S | i<j} I_{ij} + \sum_{i \in S} \frac{1}{\mu_i} \sum_{j \in S | j>i} I_{ji} - \sum_{j \in S} \rho_j \sum_{i \in S | i<j} \frac{n_i}{\mu_i} - \sum_{i \in S} \rho_i \sum_{j \in S | j>i} \frac{n_j}{\mu_j} = 0 \Rightarrow$$

$$\sum_{j \in S} \frac{1}{\mu_j} \sum_{i \in S | i<j} I_{ij} + \sum_{j \in S} \frac{1}{\mu_j} \sum_{i \in S | i>j} I_{ij} - \sum_{j \in S} \rho_j \sum_{i \in S | i<j} \frac{n_i}{\mu_i} - \sum_{j \in S} \rho_j \sum_{i \in S | i>j} \frac{n_i}{\mu_i} = 0 \tag{5.23}$$

Observing that the jj term is missing from the above summation we use (5.15) and

(5.17) to get:

$$\sum_{j\in S}\frac{n_j}{\mu_j}-\sum_{j\in S}\frac{1}{\mu_j}(\rho_j+\rho_j n_j)-\sum_{j\in S}\rho_j\sum_{i\in S|i\neq j}\frac{n_i}{\mu_i}-\sum_{j\in S}\frac{1}{\mu_j}\sum_{i\notin S}I_{ij}=0 \qquad (5.24)$$

which yields

$$\sum_{i\in S}\frac{n_i}{\mu_i}\geq\frac{\sum_{i\in S}(\rho_i/\mu_i)}{1-\sum_{i\in S}\rho_i}$$

since $I_{ij}\geq 0$. These are the inequalities of $P1$. Therefore we have proven that a feasible solution of $P2'$ is also a feasible solution of $P1$.

We now show the converse. Consider a feasible solution of $P1$. It then holds that:

$$\sum_{i\in S}\frac{n_i}{\mu_i}\geq\frac{\sum_{i\in S}(\rho_i/\mu_i)}{1-\sum_{i\in S}\rho_i}\Rightarrow$$

$$\sum_{j\in S}\frac{n_j}{\mu_j}-\sum_{j\in S}\frac{1}{\mu_j}(\rho_j+\rho_j n_j)-\sum_{j\in S}\rho_j\sum_{i\in S|i\neq j}\frac{n_i}{\mu_i}\geq 0 \qquad (5.25)$$

We now choose $I_{i,j}$ satisfying (5.15) and (5.17) which also implies that:

$$\sum_{i\in N|i\neq j}I_{i,j}=n_j-\rho_j(1+n_j) \qquad (5.26)$$

Note that there exist positive $I_{i,j}$ variables satisfying (5.26). This is due to the fact that $n_i, i=1,2,\ldots,n$, is a feasible solution of $P1$. Namely for $S\equiv j$ the inequality of $P1$ implies that $n_j-\rho_j(1+n_j)\geq 0$. Writing now (5.25) as an equality by introducing the $I_{i,j}$ variables we get (5.24). Using (5.15) and (5.17) we get (5.23) and from that (5.22).

Moreover from the equation of $P1$ we get (5.21) and by choosing $I_{i,j}$ satisfying (5.15) and (5.17) we get (5.20) and from that (5.19).

Summarizing we started from a feasible solution of $P1$ and by choosing $I_{i,j}$ satisfying (5.15) and (5.17) we got (5.22) and (5.19). From the set of $I_{i,j}$ satisfying (5.22) and (5.19) we select a solution such that:

$$\frac{I_{ij}}{\mu_j}+\frac{I_{ji}}{\mu_i}-\rho_j\frac{n_i}{\mu_i}-\rho_i\frac{n_j}{\mu_j}=0 \quad \forall i,j\in N\mid j>i$$

This is (5.16). Thus we have constructed a feasible solution of $P2$. $\square$

Discussion : The result just proven is quite interesting even from a combinatorial optimization point of view. It states that a polymatroid polytope which is defined by $2^n - 1$ constraints can be transformed to a polytope defined in a different space of dimension $O(n^2)$ that has $O(n^2)$ constraints. This theorem suggests that such a relation may also exist between the approximate polytopes derived for the open multiclass queueing network. In other words, it may provide a way to obtain the "optimal" f-parameters in the sense that they yield the dominant and tightest bounds.

Finally, to conclude this chapter, we would like to point out that since there is nothing subject to optimization in the way we are deriving the bounds, in this chapter, and since no approximation is made throughout the derivation we believe that the proposed bounds are the tightest one can get using this approach.

Chapter 6

Numerical Results

In this chapter we provide some numerical results in order to evaluate the performance of our bounding techniques. In particular, we provide three network examples and for each of these examples and for various traffic conditions we will evaluate:

- The lower bound on achievable performance according to the approach developed in Chapter 3.
- The lower bound on achievable performance according to the approach developed in Chapter 5.
- The performance of the FCFS policy.
- The performance of the best policy we were able to found which serves as an upper bound.

Thus, we are able to evaluate the tightness of our lower bound. In fact, since the optimal is not known for each case, we cannot calculate the closeness of our lower bound to the optimal policy. Instead, we will calculate its closeness to the upper bound which of course is an overestimate. In particular, we will calculate the *efficiency* of the bound which we define as:

$$\text{efficiency} = \frac{\text{Best Lower Bound}}{\text{Best Upper Bound}} 100\%$$

6.1 A Simple Two-Station Network; Revisited

Consider the two-station network example studied in Chapter 2 and depicted in Figure 2-1. Table 6.1 compares our lower bounds on attainable performance with FCFS and the threshold policy 1 (see Chapter 2) for various load conditions and provides the efficiency of the bound. "Lower Bnd. 1" and "Lower Bnd. 2" in the table correspond to the bound developed in Chapter 3 and Chapter 5, respectively. Costs were chosen in order to have as objective function the total expected number of customers in the network. Actually, this is the reason that the threshold policy 1 was simulated and not the threshold policy 2. As we mentioned in Chapter 2 we expect policy 2 to be better only when $c_3 \gg c_1$. Note that the performance reported in the table for the threshold policy corresponds to the optimal value of the threshold B which was found for each case by doing several simulation runs. Table 6.2 contains the data used for each case reported in Table 6.1. Finally, recall that by ρ_A, ρ_B we denote the total traffic intensities at station 1 and station 2, respectively.

Load Node 1-Node 2	Lower Bnd. 1	Lower Bnd. 2	FCFS	Thresh. Policy	Effic.
HEAVY-HEAVY	14.15	14.15	19.43	16.98	83%
HEAVIER-HEAVIER	19.9	19.9	28	23.76	84%
VERY HEAVY-VERY HEAVY	49.96	49.96	73	57.38	87%
MEDIUM-HEAVY	9.18	9.18	10.5	10.44	88%
LIGHT-MEDIUM	1.61	1.61	2.17	2.16	75%
HEAVY-MEDIUM	9.6	9.6	10.5	9.98	96%
MEDIUM-LIGHT	1.9	1.9	2.17	2.14	89%

Table 6.1: Numerical results for the network of Figure 2-1.

It is interesting that the efficiency of our lower bound is of approximately the same order of magnitude as the efficiency of the "pathwise bound" derived in [OuWe], which is based on simulation. Our bound, however, does not need a simulation experiment. Note also that the threshold policy clearly outperforms FCFS. From Table 6.1 it is apparent that as $\rho \rightarrow 1$ the efficiency of the bound increases. This

Load	ρ_A	ρ_B	λ_1	λ_2	μ_1	μ_2
HEAVY-HEAVY	0.93	0.86	0.86	1	2	1
HEAVIER-HEAVIER	0.95	0.90	0.90	1	2	1
VERY HEAVY-VERY HEAVY	0.98	0.96	0.96	1	2	1
MEDIUM-HEAVY	0.6	0.9	0.9	0.3	2	1
LIGHT-MEDIUM	0.4	0.6	0.6	0.2	2	1
HEAVY-MEDIUM	0.9	0.6	0.6	1.2	2	1
MEDIUM-LIGHT	0.6	0.4	0.4	0.8	2	1

Table 6.2: Data for the experiments of Table 6.1.

is true for both balanced and imbalanced traffic conditions. In particular, the efficiency increases as we go from HEAVY-HEAVY to HEAVIER-HEAVIER and to VERY HEAVY-VERY HEAVY conditions. It also increases as we go from LIGHT-MEDIUM to MEDIUM-HEAVY and from MEDIUM-LIGHT to HEAVY-MEDIUM conditions. This behaviour is mainly due to the fact that the threshold policy behaves better as the traffic gets heavier (see [HaWe]). One final observation is that the efficiency of the bounds is better in imbalanced traffic conditions. An intuitive explanation is that when one station is not so loaded as the other the scheduling problem is somewhat "easier" in the sense that the scheduler should focus on the heavy loaded station. Since our bounds explicitly characterize the performance space for the single-station case they approximate better the achievable region in imbalanced traffic conditions.

6.2 A Four-Class Network Example

Consider the network of Figure 6-1. Customers enter the network in a Poisson stream of rate λ and they visit stations 1,2,1,2, in that order before exiting the network, forming classes 1,2,3,4 respectively. The single servers at stations 1,2 has service times exponentially distributed with rates μ_1,μ_2 respectively.

Table 6.3 compares our lower bounds on attainable performance with FCFS and

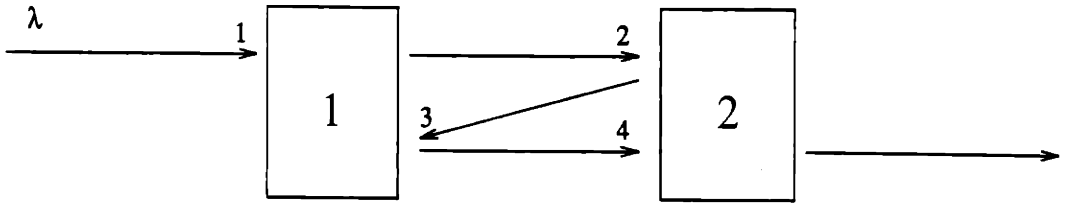

Figure 6-1: A Four-Class Network Example.

the best found policy [1] for various load conditions, providing also the efficiency of the bound. "Lower Bnd. 1" and "Lower Bnd. 2" in the table correspond to the bound developed in Chapter 3 and Chapter 5, respectively. Costs throughout the experiments reported in the table were chosen to be:

$$c_1 = 1.5,\ c_2 = 1.3,\ c_3 = 1.2,\ c_4 = 1.$$

In this specific example the best policy we were able to find, for each load condition we considered, happens to be a strict priority one. Note that we only considered non-preemptive policies. It is interesting that not a single policy was optimal for every case we considered. More precisely the following two policies were competing:

Policy 1: Give at station 1 highest priority to class 3 and lowest to class 1 ($3 \rightarrow 1$) and give at station 2 highest priority to class 4 and lowest to class 2 ($4 \rightarrow 2$).

Policy 2: Give at station 1 highest priority to class 3 and lowest to class 1 ($3 \rightarrow 1$) and give at station 2 highest priority to class 2 and lowest to class 4 ($2 \rightarrow 4$).

with the one outperforming the other in some cases and vice versa. In the table, next to the performance of the best policy for each case, we are giving in parenthesis the policy identifier, denoting by p1 and p2, policy 1 and policy 2, respectively. Table 6.4 contains the data used for each case reported in Table 6.3. Note that by ρ_A, ρ_B we denote the total traffic intensities at station 1 and station 2, respectively.

[1] we only considered non preemptive policies

Load Node 1-Node 2	Lower Bnd. 1	Lower Bnd. 2	FCFS	Best Policy	Effic.
HEAVY-HEAVY	42.24	45.36	70.55	65.58 (p2)	69%
MEDIUM-MEDIUM	16.07	20.07	28.83	27.88 (p1)	72%
MEDIUM-HEAVY	17.06	17.35	23.2	20.55 (p1)	85%
LIGHT-MEDIUM	3.44	3.69	5.23	5.00 (p1)	74%
HEAVY-MEDIUM	20.08	20.55	25.93	22.00 (p2)	94%
MEDIUM-LIGHT	4.25	4.56	5.56	5.29 (p1)	86%

Table 6.3: Numerical results for the network of Figure 6-1.

Load	ρ_A	ρ_B	λ	μ_1	μ_2
HEAVY-HEAVY	0.85	0.80	0.17	0.40	0.43
MEDIUM-MEDIUM	0.57	0.63	0.13	0.46	0.41
MEDIUM-HEAVY	0.6	0.9	0.5	1.67	1.12
LIGHT-MEDIUM	0.4	0.6	0.5	2.5	1.67
HEAVY-MEDIUM	0.9	0.6	0.5	1.12	1.67
MEDIUM-LIGHT	0.6	0.4	0.5	1.67	2.5

Table 6.4: Data for the experiments of Table 6.3.

The efficiency of our lower bound is again of approximately the same order of magnitude as the efficiency of the "pathwise bound" derived in [OuWe]. As we argued in the beginning of this Chapter the efficiency of the bounds depends both on the their closeness to optimality and on the suboptimality of the upper bound. In order to understand which factor is more important we calculated the performance of the optimal policy for one specific case via dynamic programming. In particular, we applied the value iteration algorithm given in [Bert] to the corresponding to the network Markov chain for the MEDIUM-MEDIUM traffic case. The dynamic programming algorithm yielded an optimal for the objective function of 27.7 proving policy p1 almost optimal. But it is not apparent that this is the case for all the other traffic conditions.

One different aspect of this example, compared to the example of Section 6.1,

is that in the balanced traffic case the efficiency of the bound deteriorates as the traffic gets heavier. Unfortunately, the traffic intensities are too large in this case to make the problem solvable by dynamic programming. Therefore we cannot specify if the deterioration of the efficiency is due to the deterioration of the policy. On the contrary, in the imbalanced case this example has the same behaviour as the previous one. In particular, we observe again that as the traffic goes from LIGHT-MEDIUM to MEDIUM-HEAVY the efficiency increases. The same is true as the traffic goes from MEDIUM-LIGHT to HEAVY-MEDIUM. A last note on this example is that the efficiency becomes better when station 1 is more loaded than station 2.

6.3 A Six-Class Network Example

Consider the network depicted in figure 6-2. Customers of type 1 enter the network in a Poisson stream of rate λ_1 and they visit stations 1,2,1,2, in that order, before exiting the network, forming classes 1,2,3,4 respectively. Customers of type 2 enter the network in a Poisson stream of rate λ_2 and they visit stations 1,2 before exiting the network, forming classes 5,6 respectively. The single servers at stations 1,2 have service times exponentially distributed with rates μ_1,μ_2 respectively.

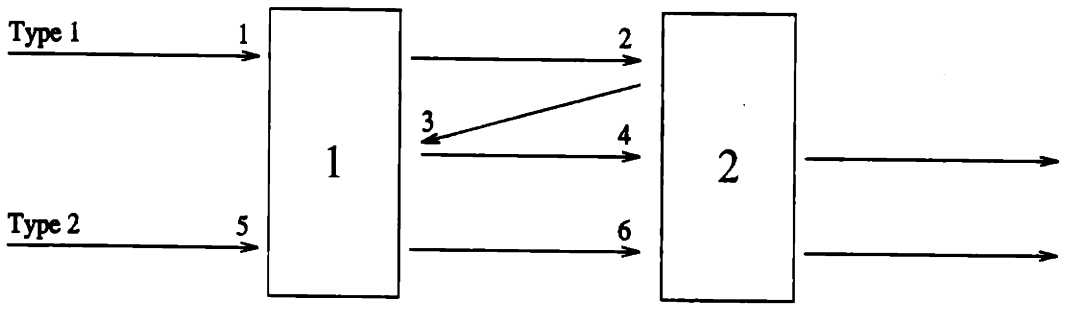

Figure 6-2: A Six-Class Network Example.

Table 6.5 compares our lower bounds on attainable performance with FCFS and the best found policy [2] for various load conditions, providing also the efficiency of

[2] we only considered non preemptive policies

the bound. "Lower Bnd. 1" and "Lower Bnd. 2" in the table correspond to the bound developed in Chapter 3 and Chapter 5, respectively. Costs throughout the experiments reported in the table were chosen to be:

$$c_1 = 1.5,\ c_2 = 1.3,\ c_3 = 1.2,\ c_4 = 1\ c_5 = 1.1,\ c_6 = 1.1.$$

In this specific example, also, the best policy we were able to find, for each load condition we considered, happens to be a strict priority one. Note that we only considered non-preemptive policies. It is interesting that not a single policy was optimal for every case we considered. More precisely the following two policies were competing:

Policy 1: Give at station 1 highest priority to class 3 and lowest to class 5 ($3 \rightarrow 1 \rightarrow 5$) and give at station 2 highest priority to class 6 and lowest to class 2 ($6 \rightarrow 4 \rightarrow 2$).

Policy 2: Give at station 1 highest priority to class 1 and lowest to class 5 ($1 \rightarrow 3 \rightarrow 5$) and give at station 2 highest priority to class 2 and lowest to class 6 ($2 \rightarrow 4 \rightarrow 6$).

with the one outperforming the other in some cases and vice versa. In the table, next to the performance of the best policy for each case, we are giving in parenthesis the policy identifier, denoting by p1 and p2, policy 1 and policy 2, respectively. Table 6.6 contains the data used for each case reported in Table 6.5. Recall that by ρ_A, ρ_B we denote the total traffic intensities at station 1 and station 2, respectively.

The efficiency of our lower bound is again of approximately the same order of magnitude as the efficiency of the "pathwise bound" derived in [OuWe].

In this example, as in the example of the previous section, we observe that in the balanced traffic case the efficiency of the bound deteriorates as the traffic gets heavier. We also observe that, as in the example of the previous section, the efficiency becomes better when station 1 is more loaded than station 2.

Load Node 1-Node 2	Lower Bnd. 1	Lower Bnd. 2	FCFS	Best Policy	Effic.
HEAVY-HEAVY	15.72	16.67	30.56	26.89 (p2)	62%
MEDIUM-MEDIUM	5.83	6.17	9.86	9.25 (p2)	67%
MEDIUM-HEAVY	15.77	15.85	21.26	18.20 (p1)	87%
HEAVY-MEDIUM	18.77	18.79	23.00	19.80 (p1)	95%

Table 6.5: Numerical results for the network of Figure 6-2.

Load	ρ_A	ρ_B	λ_1	λ_2	μ_1	μ_2
HEAVY-HEAVY	0.85	0.90	0.5	0.7	2	1.89
MEDIUM-MEDIUM	0.7	0.7	0.5	0.7	2.43	2.43
MEDIUM-HEAVY	0.6	0.9	0.5	0.7	2.83	1.89
HEAVY-MEDIUM	0.9	0.6	0.5	0.7	1.89	2.83

Table 6.6: Data for the experiments of Table 6.3.

A conclusion that can be drawn from all the cases studied numerically is that our lower bounds are very efficient in imbalanced traffic conditions. In these conditions the efficiency of the bounds increases with the traffic intensity. In balanced traffic conditions, they also behave well especially when the traffic intensity is not very close to one. But, even in these heavy-balanced traffic conditions, in the examples that we studied the efficiency does not get worse than 62%.

Chapter 7

Conclusions and Open Problems

In this thesis we considered the problem of scheduling an open multiclass queueing network, with Poisson arrivals and exponentially distributed service times, when only sequencing decisions are involved. The objective was to minimize a weighted sum of the expected response times of different classes in the network. We proposed a new method to derive a lower bound on the achievable performance based mainly on ideas of conservation laws. Our method consists of deriving a polyhedral space which includes the achievable region of the network. Thus, optimization of the objective function over this polyhedral space yields a lower bound on the achievable performance. We were able to prove that in single-station network models, namely in a multiclass queue with and without feedback, the above mentioned polyhedron exactly characterizes the achievable region. Thus, our method can be viewed as a natural extension of known results for single-station networks to the general setting of an open multiclass queueing network.

More precisely, we proposed two variations of the same method that define two different polyhedral spaces. In the first variation, we define a class of polyhedral spaces, by changing the values of a set of parameters, which we named f-parameters. In order to get the tightest of these polyhedra, one has to find the values of the f-parameters corresponding to this polyhedron. This is not an easy task to do, in general, and thus we chose the "optimal" values of the f-parameters based on intuitive grounds.

The second variation is more efficient in terms of computational effort needed for the calculation of the lower bound and yields only one polyhedron, explicitly defined by the data (arrival rates, service rates and routing matrix). In the single-station case we were able to find the "optimal" value of the f-parameters such that the polyhedra defined by both variations exactly characterize the achievable region.

Comparing with the existing literature on lower bounds, our method is an analytical one that calculates the lower bound on achievable performance in a number of steps which is a polynomial function of the number of classes in the network. On the contrary, existing methods are simulation-based. Moreover, in terms of tightness of the bound, a numerical study in various network topologies that we presented, suggests that our method is at least as good as existing methods.

A way to find or at least to characterize the optimal policy still remains an open problem. Among more tractable open problems we would like to point out:

- A way to find a relation between the two polyhedral spaces defined by the two variations of our method. This was done in the single-station case. However, we were not able to generalize in the general case.
- Ways to improve the efficiency of our bounds, especially in the case of networks with nodes in balanced-heavy traffic.
- A way to generalize the proposed method in order to include closed networks.
- A way to generalize the proposed method in order to include networks with general arrivals and general service times.

Appendix A

Proof of Lemma 4.5

The real-valued function y on the subsets of $N = \{1, 2, \ldots, n\}$ was defined in chapter 4 to be:

$$y(S) = \frac{\rho'(S)}{1 - \rho(S)}$$

where $\rho'(S) = \sum_{i \in S}(\rho_i/\mu_i)$ and $\rho(S) = \sum_{i \in S} \rho_i$. In order to prove the submodularity of the above function we are using the following proposition from [NeWo].

Proposition A.1 *A real-valued function f on the subsets of a set N is supermodular if and only if*

$$f(S \cup \{j\}) - f(S) \geq f(S \cup \{j,k\}) - f(S \cup \{k\}) \text{ for } j,k \in N,\ j \neq k \text{ and } S \subseteq N\{j,k\}.$$

We therefore have:

$$\begin{aligned} & f(S \cup \{j,k\}) - f(S \cup \{k\}) + f(S) - f(S \cup \{j\}) = \\ & \quad \frac{\rho'(S) + (\rho_j/\mu_j) + (\rho_k/\mu_k)}{1 - \rho(S) - \rho_j - \rho_k} - \frac{\rho'(S) + (\rho_k/\mu_k)}{1 - \rho(S) - \rho_k} + \frac{\rho'(S)}{1 - \rho(S)} - \frac{\rho'(S) + (\rho_j/\mu_j)}{1 - \rho(S) - \rho_j} \\ = & \quad \frac{\rho_j}{\mu_j}\rho_k(1 - \rho(S))(1 - \rho(S) - \rho_k) + \frac{\rho_k}{\mu_k}\rho_j(1 - \rho(S))(1 - \rho(S) - \rho_j) + \\ & \quad \rho'(S)\rho_j[(1 - \rho(S) - \rho_j)(1 - \rho(S)) - (1 - \rho(S) - \rho_k - \rho_j)(1 - \rho(S) - \rho_k)] \end{aligned}$$

The first two terms of the last equivalence are positive so it suffices to show the positivity of the last term. Thus,

$$
\begin{aligned}
& f(S \cup \{j,k\}) - f(S \cup \{k\}) + f(S) - f(S \cup \{j\}) \geq \\
& \qquad \rho'(S)\rho_j[(1-\rho(S)-\rho_j)(1-\rho(S)) - (1-\rho(S)-\rho_k-\rho_j)(1-\rho(S)-\rho_k)] \\
&= \rho'(S)\rho_j\rho_k(2-2\rho(S)-\rho_k-\rho_j) \\
&\geq 2\rho'(S)\rho_j\rho_k(1-\rho(S)-\rho_k-\rho_j) \\
&\geq 0 \quad \square
\end{aligned}
$$

Appendix B

Proof of Lemma 4.6

From (4.3), (4.4) and (4.5) we have that:

$$\begin{aligned}
\rho_j x_j &= \frac{\sum_{i=1}^{j}(\rho_i/\mu_i)}{1-\sum_{i=1}^{j}\rho_i} - \frac{\sum_{i=1}^{j-1}(\rho_i/\mu_i)}{1-\sum_{i=1}^{j-1}\rho_i} \\
&= \frac{\left(1-\sum_{i=1}^{j-1}\rho_i\right)\sum_{i=1}^{j}\frac{\rho_i}{\mu_i} - \left(1-\sum_{i=1}^{j}\rho_i\right)\sum_{i=1}^{j-1}\frac{\rho_i}{\mu_i}}{\left(1-\sum_{i=1}^{j-1}\rho_i\right)\left(1-\sum_{i=1}^{j}\rho_i\right)} \\
&= \frac{\sum_{i=1}^{j}\frac{\rho_i}{\mu_i} - \sum_{k=1}^{j-1}\rho_k\sum_{i=1}^{j}\frac{\rho_i}{\mu_i} - \sum_{i=1}^{j-1}\frac{\rho_i}{\mu_i} + \sum_{k=1}^{j}\rho_k\sum_{i=1}^{j-1}\frac{\rho_i}{\mu_i}}{\left(1-\sum_{i=1}^{j-1}\rho_i\right)\left(1-\sum_{i=1}^{j}\rho_i\right)} \\
&= \frac{\frac{\rho_j}{\mu_j} - \frac{\rho_j}{\mu_j}\sum_{k=1}^{j-1}\rho_k + \rho_j\sum_{k=1}^{j-1}\frac{\rho_k}{\mu_k}}{\left(1-\sum_{i=1}^{j-1}\rho_i\right)\left(1-\sum_{i=1}^{j}\rho_i\right)} \Rightarrow \\
x_j &= \frac{\frac{1}{\mu_j}\left(1-\sum_{i=1}^{j-1}\rho_i\right) + \sum_{i=1}^{j-1}\frac{\rho_i}{\mu_i} + \frac{\rho_j}{\mu_j} - \frac{\rho_j}{\mu_j}}{\left(1-\sum_{i=1}^{j-1}\rho_i\right)\left(1-\sum_{i=1}^{j}\rho_i\right)} \\
&= \frac{1}{\mu_j\left(1-\sum_{i=1}^{j-1}\rho_i\right)} + \frac{\sum_{i=1}^{j}(\rho_i/\mu_i)}{\left(1-\sum_{i=1}^{j-1}\rho_i\right)\left(1-\sum_{i=1}^{j}\rho_i\right)}
\end{aligned}$$

But this is exactly the formula from priority queueing (see [GeMi, eq. (1.81)]) of the response time x_j for class j customers when preemptive priority is given to classes $1, 2, \ldots, n$ in that order. □

Bibliography

[Bert] Bertsekas, D.P., (1976), *Dynamic Programming and Stochastic Control*, Academic Press, New York.

[BeNa] Bertsimas, D. and Nakazato, D., (1990), "The Departure Process from a GI/G/1 Queue and its Applications to the Analysis of Tandem Queues", *Operations Research Center*, Massachusetts Institute of Technology, Working paper, OR 245-91.

[BGeT] Bhattacharya, P.P, Georgiadis, L., Tsoucas, P. (1992), "Extended Polymatroids: Properties and Optimization", *IBM Research Division*, Research Report, T.J Watson Research Center, Yorktown Heights, New York.

[ChYY] Chen, H., Yang, P. and Yao, D.D., (1991), "Control and Scheduling in a Two-Station Queueing Network: Optimal Policies and Heuristics", Preprint.

[Flor] Flores, C., (1985), "Diffusion Approximations for Computer Communications Networks", *Proceedings of Symposia in Applied Mathematics*, Vol. 31.

[Fosc] Foschini, G.J., (1982), "Equilibria for Diffusion Models of Pairs of Communicating Computers— Symmetric Case", *IEEE Transactions on Information Theory*, 28, 273–284.

[GeMi] Gelenbe, E. and Mitrani, I., (1980), *Analysis and Synthesis of Computer Systems*, Academic Press, London.

[HaDa] Harrison, M. and Dai, J., (1991), "Solving Stationary Densities of Reflected Brownian Motion", *The Annals of Applied Probability*, Vol. 1, No. 1, 16–35.

[Harr] Harrison, J.M., (1978), "The Diffusion Approximation for Tandem Queues in Heavy Traffic", *Adv. Appl. Prob.*, 10, 886–905.

[HaWe] Harrison, J.M. and Wein, L.M., (1989), "Scheduling Networks of Queues: Heavy Traffic Analysis of a simple Open Network", *Queueing Systems Theory and Applications*, 5, 265–280.

[KeLa] Kelly, F.P. and Laws, C.N., "Dynamic Routing in Open Queueing Networks", Preprint.

[Klv1] Kleinrock, L., (1975), *Queueing Systems, Vol. 1: Theory*, Wiley, New York.

[Klv2] Kleinrock, L., (1976), *Queueing Systems, Vol. 2: Computer Applications*, Wiley, New York.

[Klim] Klimov, G.P., (1974), "Time-Sharing Service Systems. I", *Theory of Probability and its Applications, Vol. XIX*, No 3.

[Koba] Kobayashi, H., (1974), "Application of the Diffusion Approximation to Queueing Networks I: Equilibrium Queue Distributions", *Journal of the Association for Computing Machinery*, Vol. 21, No. 2, 316–328.

[Kuma] Kumar, P.R., "Re-Entrant Lines", Preprint.

[Lemo] Lemoine, A.J., (1978), "Networks of Queues— A Survey of Weak Convergence Results", *Management Science*, Vol. 24, No. 11, 1175–1193.

[NeWo] Nemhauser, G.L. and Wolsey, L.A. (1988), *Integer and Combinatorial Optimization*, Wiley, New York.

[OuWe] Ou, J. and Wein, L.M. (1992), "Performance Bounds for Scheduling Queueing Networks", *The Annals of Applied Probability*, Vol. 2, No. 2, 460–480.

[Reim] Reiman, M.I., (1984), "Open Queueing Networks in Heavy Traffic", *Mathematics of Operations Research*, Vol. 9, No. 3, 441–458.

[RVWa] Rosberg, Z., Varaiya, P.P. and Walrand, J.C., (1982), "Optimal Control of Service in Tandem Queues", *IEEE Transactions on Automatic Control*, Vol. 27, No. 3, 600–610.

[ShYa] Shanthikumar, J.G. and Yao, D.D., (1992), "Multiclass Queueing Systems: Polymatroid Structure and Optimal Scheduling Control", *Operations Research*, Vol. 40, No. 2, 293–299.

[Tsou] Tsoucas P., (1991), "The Region of Achievable Performance in a Model of Klimov", *IBM Research Division*, Research Report, T.J Watson Research Center, Yorktown Heights, New York.

[WeSt] Weber, R.R. and Stidham, S., (1987), "Optimal Control of Service Rates in Networks of Queues", *Adv. Appl. Prob.*, 19, 202–218.

[Wei1] Wein, L.M. (1988), "Optimal Control of a Two-Station Brownian Network", *Operations Research Center*, Massachusetts Institute of Technology, Working paper, OR 2015-88.

[Wei2] Wein, L.M. (1988), "Scheduling a Two-Station Multiclass Queueing Network in Heavy Traffic", *Operations Research Center*, Massachusetts Institute of Technology, Working paper, OR 2016-88.

[Wei3] Wein, L.M. (1990), "Dynamic Scheduling of a Multiclass Make-to-Stock Queue", *Operations Research Center*, Massachusetts Institute of Technology, Working paper, OR 3113-90-MSA.

[Whit] Whitt, W., (1982), "Refining Diffusion Approximations for Queues", *Operations Research Letters*, Vol. 1, No. 5, 165–169.