Lecture Notes in Mathematics

Maury Bramson

Stability of Queueing Networks

1950

École d'Été de Probabilités de Saint-Flour XXXVI-2006

Springer

Lecture Notes in Mathematics 1950

Saint-Flour Probability Summer School

The Saint-Flour volumes are reflections of the courses given at the Saint-Flour Probability Summer School. Founded in 1971, this school is organised every year by the Laboratoire de Mathématiques (CNRS and Université Blaise Pascal, Clermont-Ferrand, France). It is intended for PhD students, teachers and researchers who are interested in probability theory, statistics, and in their applications.

The duration of each school is 13 days (it was 17 days up to 2005), and up to 70 participants can attend it. The aim is to provide, in three high-level courses, a comprehensive study of some fields in probability theory or Statistics. The lecturers are chosen by an international scientific board. The participants themselves also have the opportunity to give short lectures about their research work.

Participants are lodged and work in the same building, a former seminary built in the 18th century in the city of Saint-Flour, at an altitude of 900 m. The pleasant surroundings facilitate scientific discussion and exchange.

The Saint-Flour Probability Summer School is supported by:

– Université Blaise Pascal
– Centre National de la Recherche Scientifique (C.N.R.S.)
– Ministère délégué à l'Enseignement supérieur et à la Recherche

For more information, see back pages of the book and
http://math.univ-bpclermont.fr/stflour/

Jean Picard
Summer School Chairman
Laboratoire de Mathématiques
Université Blaise Pascal
63177 Aubière Cedex
France

Maury Bramson

Stability of Queueing Networks

École d'Été de Probabilités
de Saint-Flour XXXVI - 2006

Springer

Maury Bramson
University of Minnesota
School of Mathematics
Minneapolis, MN 55455
USA
bramson@math.umn.edu

Library of Congress Control Number: 2008928773

Mathematics Subject Classification (2000): 60K25, 68M20, 90B22

ISSN print edition: 0075-8434
ISSN electronic edition: 1617-9692
ISSN Ecole d'Eté de Probabilités de St. Flour, print edition: 0721-5363
ISBN 978-3-540-68895-2 Springer Berlin Heidelberg New York
DOI 10.1007/978-3-540-68896-9

Typesetting by the author and SPi using a Springer LaTeX macro package

Cover art: Tree of Life, with kind permission of David M. Hillis, Derrick Zwickl, and Robin Gutell, University of Texas.

Cover design: SPI, Heidelberg

Printed on acid-free paper SPIN: 12114894 VA41/3100/SPi 5 4 3 2 1 0

Preface

These notes include the material from a series of nine lectures given at the Saint-Flour Probability Summer School, July 2 - July 15, 2006. Lectures were also given by Alice Guionnet and Steffen Lauritzen. It was my first visit to Saint-Flour, which I enjoyed considerably.

The topic of these notes, the stability of queueing networks, has been of interest to the queueing community since the early 1990s. At that point, little was known about this and other aspects of queueing networks in general settings. There is now a theory (albeit incomplete), with positive criteria for stability as well as examples where such stability fails. I felt that the time was right to combine in one place a summary of such work.

I wish to thank Jean Picard for his work in organizing the Saint-Flour Summer School and the other participants of the School. I thank the following individuals with whom I consulted at various points: John Baxter, Jim Dai, Michael Harrison, John Hasenbein, Haya Kaspi, Thomas Kurtz, Sean Meyn, and Ruth Williams. I also thank John Baxter for his technical help in preparing the manuscript.

The research that led to these notes was supported in part by U.S. National Science Foundation grants DMS-0226245 and CCF-0729537.

Minneapolis, Minnesota, U.S.A. *Maury Bramson*
March 2008

Contents

1

Introduction

Queueing networks constitute a large family of models in a variety of settings, involving "jobs" or "customers" that wait in queues until being served. Once its service is completed, a job moves to the next prescribed queue, where it remains until being served. This procedure continues until the job leaves the network; jobs also enter the network according to some assigned rule.

In these lectures, we will study the evolution of such networks and address the question: When is a network *stable*? That is, when is the underlying Markov process of the queueing network positive Harris recurrent? When the state space is countable and all states communicate, this is equivalent to the Markov process being positive recurrent. An important theme, in these lectures, is the application of fluid models, which may be thought of as being, in a general sense, dynamical systems that are associated with the networks.

The goal of this chapter is to provide a quick introduction to queueing networks. We will provide basic vocabulary and attempt to explain some of the concepts that will motivate later chapters. The chapter is organized as follows. In Section 1.1, we discuss the $M/M/1$ queue, which is the "simplest" queueing network. It consists of a single queue, where jobs enter according to a Poisson process and have exponentially distributed service times. The problem of stability is not difficult to resolve in this setting.

Using $M/M/1$ queues as motivation, we proceed to more general queueing networks in Section 1.2. We introduce many of the basic concepts of queueing networks, such as the discipline (or policy) of a network determining which jobs are served first, and the traffic intensity ρ of a network, which provides a natural condition for deciding its stability. In Section 1.3, we provide a preliminary description of fluid models, and how they can be applied to provide conditions for the stability of queueing networks.

In Section 1.4, we summarize the topics we will cover in the remaining chapters. These include the product representation of the stationary distributions of certain classical queueing networks in Chapter 2, and examples of unstable queueing networks in Chapter 3. Chapters 4 and 5 introduce fluid

M. Bramson, *Stability of Queueing Networks*. Lecture Notes in Mathematics 1950.

models and apply them to obtain criteria for the stability of queueing networks.

1.1 The $M/M/1$ Queue

The $M/M/1$ *queue*, or *simple queue*, is the most basic example of a queueing network. It is familiar to most probabilists and is simple to analyze. We therefore begin with a summary of some of its basic properties to motivate more general networks.

The setup consists of a server at a workstation, and "jobs" (or "customers") who line up at the server until they are served, one by one. After service of a job is completed, it leaves the system. The jobs are assumed to arrive at the station according to a Poisson process with intensity 1; equivalently, the interarrival times of succeeding jobs are given by independent rate -1 exponentially distributed random variables. The service times of jobs are given by independent rate-μ exponentially distributed random variables, with $\mu > 0$; the mean service time of jobs is therefore $m = 1/\mu$. We are interested here in the behavior of $Z(t)$, the number of jobs in the queue at time t, including the job currently being served (see Figure 1.1).

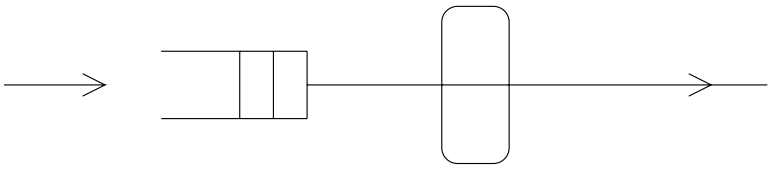

Fig. 1.1. Jobs enter the system at rate 1 and depart at rate μ. There are currently 2 jobs in the queue.

The process $Z(\cdot)$ can be interpreted in several ways. Because of the independent exponentially distributed interarrival and service times, $Z(\cdot)$ defines a Markov process, with states $0, 1, 2, \ldots$. ($M/M/1$ stands for Markov input and Markov output, with one server.) It is also a birth and death process on $\{0, 1, 2, \ldots\}$, with birth rate 1 and death rate μ. Because of the latter interpretation, it is easy to compute the stationary (or invariant) probability measure π_m of $Z(\cdot)$ when it exists, since the process will be reversible. Such a measure satisfies

$$\pi_m(n+1) = m\pi_m(n) \quad \text{for } n = 0, 1, 2, \ldots,$$

since it is constant, over time, on the intervals $[0, n]$ and $[n+1, \infty)$. It follows that when $m < 1$, π_m is geometrically distributed, with

$$\pi_m(n) = (1-m)m^n, \quad n = 0, 1, 2, \ldots. \tag{1.1}$$

All states clearly communicate with one another, and the process $Z(\cdot)$ is positive recurrent. The mean of π_m is $m(1-m)^{-1}$, which blows up as $m \uparrow 1$. When $m \geq 1$, no stationary probability measure exists for $Z(\cdot)$. Using standard reasoning, one can show that $Z(\cdot)$ is null recurrent when $m = 1$ and is transient when $m > 1$.

The behavior of $Z(\cdot)$ that was observed in the last paragraph provides the basic motivation for these lectures, in the context of the more general queueing networks which will be introduced in the next section. We will investigate when the Markov process corresponding to a queueing network is stable, i.e., is positive Harris recurrent. As mentioned earlier, this is equivalent to positive recurrence when the state space is countable and all states communicate.

For $M/M/1$ queues, we explicitly constructed a stationary probability measure to demonstrate positive recurrence of the Markov process. Typically, however, such a measure will not be explicitly computable, since it will not be reversible. This, in particular, necessitates a new, more qualitative, approach for showing positive recurrence. We will present such an approach in Chapter 4.

1.2 Basic Concepts of Queueing Networks

The $M/M/1$ queue admits natural generalizations in a number of directions. It is unnecessary to assume that the interarrival and service distributions are exponential. For general distributions, one employs the notation $G/G/1$; or $M/G/1$ or $G/M/1$, if one of the distributions is exponential. (To emphasize the independence of the corresponding random variables, one often uses the notation GI instead of G.)

The single queue can be extended to a finite system of queues, or a *queueing network* (for short, *network*), where jobs, upon leaving a queue, line up at another queue, or *station*, or leave the system. The queueing network in Figure 1.2 is also a *reentrant line*, since all jobs follow a fixed route.

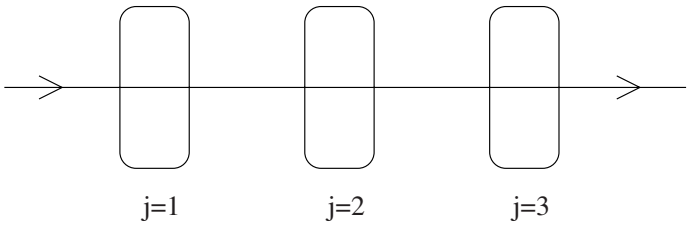

Fig. 1.2. A reentrant line with 3 stations.

Depending on a job's previous history, one may wish to prescribe different service distributions at its current station or different routing to the next

station. This is done by assigning one or more *classes*, or *buffers*, to each station. Except when stated otherwise, we label stations by $j = 1, \ldots, J$ and classes by $k = 1, \ldots, K$; we use $\mathcal{C}(j)$ to denote the set of classes belonging to station j and $s(k)$ to denote the station to which class k belongs. In Figure 1.3, there are 3 stations and 5 classes. Classes are labelled here in the order they occur along the route, with $\mathcal{C}(1) = \{1, 5\}$, $\mathcal{C}(2) = \{2, 4\}$, and $\mathcal{C}(3) = \{3\}$.

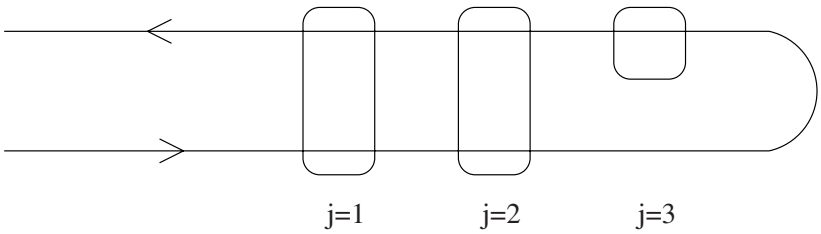

Fig. 1.3. A reentrant line with 3 stations and 5 classes. The stations are labelled by $j = 1, 2, 3$; the classes are labelled by $k = 1, \ldots 5$, in the order they occur along the route.

Other examples of queueing networks are given in Figures 1.4 and 1.5. Figure 1.4 depicts a network with 2 stations, each possessing 2 classes. The network is not a reentrant line but still exhibits *deterministic routing*, since each job entering the network at a given class follows a fixed route. When the individual routes are longer, it is sometimes more convenient to replace the above labelling of classes by (i, k), where i gives the route that is followed and k the order of the class along the route.

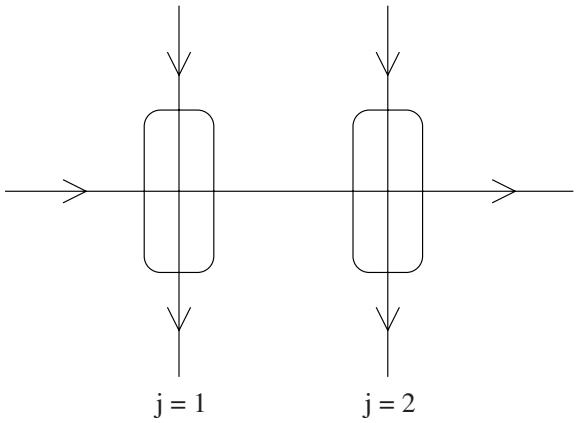

Fig. 1.4. A queueing network having 3 routes, with 2 stations and 4 classes.

Figure 1.5 depicts a network with 2 stations and 3 classes. The routing at class 2 is random, with each job served there having probability 0.4 of being sent to class 1 and probability 0.6 of being set to class 3. For queueing networks in general, we will assume that the interarrival times, service times, and routing of jobs at each class are given by sequences of i.i.d. random variables that are mutually independent (but whose distributions may depend on the class).

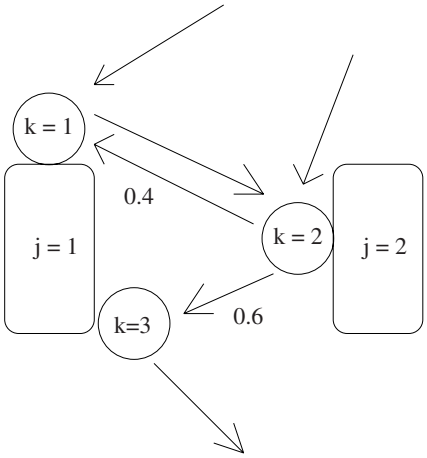

Fig. 1.5. A queueing network with 2 stations and 3 classes. The random routing at class 2 is labelled with the probability each route is taken.

Queueing networks occur naturally in a wide variety of settings. "Jobs" can be interpreted as products of some sort of complex manufacturing process with multiple steps, as tasks that need to be performed by a computer or communication system, or as people moving about through a bureaucratic maze. Such networks can be quite complicated. A portion of the procedure in the manufacture of a semiconductor wafer is depicted by the reentrant line in Figure 1.6. Another simplified example, with classes emphasized, is given in Figure 1.7. Typically, such procedures can require hundreds of steps.

We will say that a queueing network is *multiclass* when at least one station has more than one class; otherwise, the network is *single class*. The term *Jackson network* is frequently used for a single class network with exponentially distributed interarrival and service times, and *generalized Jackson network* is used for a single class network with arbitrary distributions. (The networks in Figures 1.3-1.7 are all multiclass; the networks in Figures 1.1–1.2 are single class.) Unless otherwise specified, it will be assumed that there is a single server at each station j. This server will be assumed to be *non-idling* (or *work conserving*), that is, it remains busy as long as there are jobs present at any

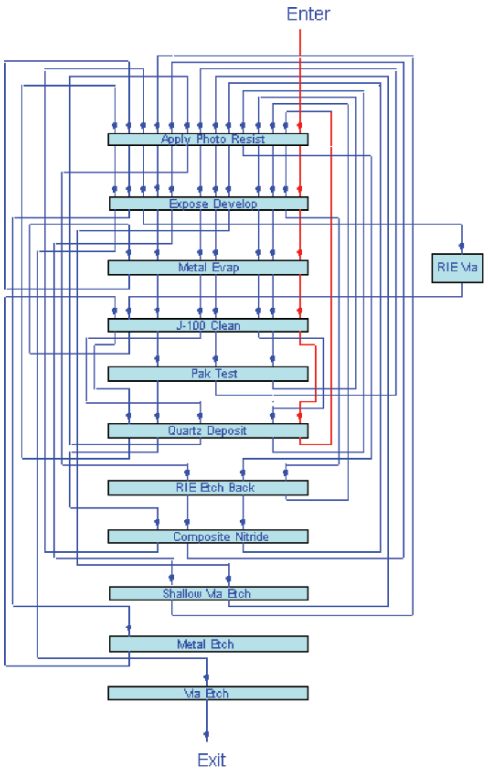

Fig. 1.6. Part of the procedure in the manufacture of a semiconductor wafer. The procedure starts at the top and ends at the bottom. (Example courtesy of P. R. Kumar.)

$k \in \mathcal{C}(j)$. Stations are assumed to have infinite capacity, with jobs never being turned away. (That is, there is no limit to the allowed length of queues.)

As already indicated, the interarrival times, service times, and routing of jobs are given by sequences of independent random variables. Although their specific distributions will be relevant for some matters, their means will be of much greater importance. We therefore introduce here the following notation and terminology for the means; systematic notation for the random variables will be introduced in Chapter 4.

We denote by α_k the rate at which jobs arrive at a class k from outside the network. When $k \in \mathcal{A}$, the subset of classes where external arrivals are allowed, α_k is the reciprocal of the corresponding mean interarrival time; when $k \notin \mathcal{A}$, we set $\alpha_k = 0$. The vector $\alpha = \{\alpha_k,\, k = 1, \ldots, K\}$ is referred to as the *external arrival rate*. (Throughout the lectures, we interpret vectors as column vectors unless stated otherwise.) We denote by m_k, $m_k > 0$, the *mean service time* at class k, and by M the diagonal matrix with m_k, $k = 1, \ldots, K$, as its entries; we set $\mu_k = 1/m_k$, which is the *service rate*. We also denote

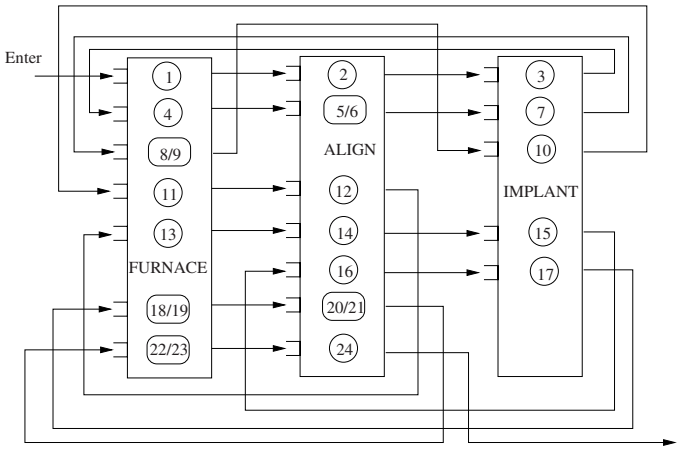

Fig. 1.7. Part of the procedure in the manufacture of another semiconductor wafer. (Example courtesy of J. G. Dai.)

by $P = \{P_{k,\ell},\, k, \ell = 1, \ldots, K\}$ the *mean transition matrix* (or *mean routing matrix*), where $P_{k,\ell}$ is the probability a job departing from class k goes to class ℓ. (In Figure 1.5, $P_{2,1} = 0.4$ and $P_{2,3} = 0.6$.)

In these lectures, we will be interested in *open queueing networks*, that is, those networks for which the matrix

$$Q \stackrel{\text{def}}{=} (I - P^T)^{-1} = I + P^T + (P^T)^2 + \ldots \tag{1.2}$$

is finite. (" ^{T} " denotes the transpose.) This means that jobs at any class are capable of ultimately leaving the network. *Closed queueing networks*, where neither arrivals to, nor departures from, the network are permitted, are not discussed here, although there are certain similarities between the two cases.

Disciplines for multiclass queueing networks

We have so far neglected an important aspect of multiclass queueing networks. Namely, when more than one job is present at a station, what determines the order in which jobs are served? There are numerous possibilities. As we will see in these lectures, the choice of the service rule, or *discipline* (also known as *policy*), can have a major impact on the evolution of the queueing network.

Perhaps the most natural discipline is *first-in, first-out* (FIFO), where the first job to arrive at a station (or "oldest" job) receives all of the service, irrespective of its class. (If the job later returns to the station, it starts over again as the "youngest" job. The jobs originally at a class are assigned some arbitrary order.) Another widely used discipline is *processor sharing* (PS), where the service at a station is simultaneously equally divided between all jobs presently there. PS can be thought of as a limiting *round robin* discipline,

where a server alternates among the jobs currently present at the station, dispensing a fixed small amount of service until it moves on to the next job.

The FIFO and PS disciplines are egalitarian in the sense that no job, because of its class, is favored over another. The opposite is the case for *static buffer priority* (SBP) disciplines, where classes are assigned a strict ranking, and jobs of higher ranked (or priority) classes are always served before jobs of lower ranked classes, irrespective of when they arrived at the station. Within a class, the jobs are served in the order of their arrival there. The disciplines are called *preemptive resume*, or more simply, *preemptive*, if arriving higher ranked jobs interrupt lower ranked jobs currently in service; when the service of such higher ranked jobs has been completed, service of lower ranked jobs continues where it left off. If the service of lower ranked jobs is not interrupted, then the discipline is *nonpreemptive*.

For reentrant lines, two natural SBP disciplines are *first-buffer-first-served* (FBFS) and *last-buffer-first-served* (LBFS). For FBFS, jobs at earlier classes along the route have priority over later classes. For example, jobs in class 1, in the network given in Figure 1.3, have priority over those in class 5, and jobs in class 2 have priority over those in class 4. For LBFS, the priority is reversed, with jobs at later classes along the route having priority over earlier classes.

In these lectures, we will concentrate on *head-of-the-line* (HL) disciplines, where only the first job in each class of a station may receive service. (This property is frequently referred to as FIFO within a class.) FIFO and SBP disciplines are HL, but PS is not. The single class networks we consider will always be assumed to be HL unless explicitly stated otherwise. A major simplifying feature of HL disciplines is that the total amount of "work" (or "effort") already devoted to partially served jobs does not build up. Since large numbers of jobs will therefore tend not to complete service at a station at close times, this avoids "bursts" of jobs from being suddenly routed elsewhere in the network.

Another non-HL discipline is *last-in, first-out* (LIFO), where the last job to arrive at a station is served first. Other plausible non-HL disciplines from a somewhat different setting are *first-in-system, first-out* (FISFO) and *last-in-system, first-out* (LISFO). FISFO is the same as FIFO, except that the first job to enter the queueing network (rather than the station) has priority over other jobs; the discipline may be preemptive or nonpreemptive. LISFO gives priority to the last job to enter the network.

Throughout these lectures, the term queueing network will indicate that a discipline, such as FIFO or FBFS, has already been assigned to the system. In much of the literature, the discipline is specified afterwards. This linguistic difference does not affect the theory, of course.

For the $M/M/1$ queue in the previous section, the interarrival and service times were assumed to be exponentially distributed. As a consequence, the process $Z(\cdot)$ counting the number of jobs is Markov. For queueing networks with exponentially distributed interarrival and service times, the situation

is often analogous. For instance, for preemptive SBP disciplines, the vector valued process $Z(t) = \{Z_k(t),\ k = 1, \ldots, K\}$ counting the number of jobs at each class is Markov; the same is true for PS. For the FIFO and LIFO disciplines, the process will be Markov if one appends additional information giving the order in which each job entered the class. In all of these cases, the state space is countable, so one can apply standard Markov chain theory. We will denote by $X(\cdot)$ the corresponding Markov processes.

The situation becomes more complicated when the interarrival and service times are not exponentially distributed, since the residual interarrival and service times need to be appended to the state in order for the process to be Markov. The resulting state space is uncountable, and so a more general theory, involving Harris recurrence, is required for the corresponding Markov process. We will give a careful construction of such processes and will summarize the needed results on positive Harris recurrence at the beginning of Chapter 4. We avoid this issue until then, since the material in Chapters 2 and 3 typically does not involve Markov processes on an uncountable state space. (The sole exception is the uncountable state space extension for symmetric processes at the end of Section 2.4.) In a first reading of the lectures, not too much will be lost by assuming the interarrival and service times are exponentially distributed.

Traffic intensity, criticality, and stability

The mean quantities α, M, and P introduced earlier perform an important role in determining the long-time behavior of a queueing network. To provide motivation, we first consider reentrant lines.

The (long term) rate at which jobs enter a reentrant line is α_1, and the mean time required to serve a job over all of its visits to a station j is $\sum_{k \in \mathcal{C}(j)} m_k$. So, the rate at which future "work" for the station j enters the network is

$$\rho_j = \alpha_1 \sum_{k \in \mathcal{C}(j)} m_k. \tag{1.3}$$

We recall that a queueing network is defined to be stable when its underlying Markov process is positive Harris recurrent. When the state space is countable and all states communicate, this is equivalent to the Markov process being positive recurrent. It is intuitively clear that, in order for a reentrant line to be stable, $\rho_j \leq 1$, for all j, is necessary. Otherwise, by the law of large numbers, the work in the system, corresponding to j, will increase linearly to infinity as $t \to \infty$, and so the same should be true for the total number of jobs in the system. (When the state space is countable, it follows from standard Markov chain theory that a stable network must be empty a fixed fraction of the time, as $t \to \infty$, from which it follows that, in fact, $\rho_j < 1$ must hold for all j.)

A natural condition for stability (assuming all states communicate), is that $\rho_j < 1$ for all j. This turns out, in fact, not to be sufficient for multiclass

queueing networks, as we will see in Chapter 3. Not surprisingly, the much stronger condition

$$\sum_j \rho_j < 1 \tag{1.4}$$

suffices for stability, since whenever the network is not empty, the total work in the network tends to decrease faster than it increases. (This will follow from Proposition 4.5.)

The situation is analogous for queueing networks with general routing, after the correct quantities have been introduced. We set

$$\lambda = Q\alpha, \tag{1.5}$$

and refer to the vector λ as the *total arrival rate*. (Recall that vectors are to be interpreted as column vectors.) Its components $\lambda_k,\ k = 1, \ldots, K$, are the rates at which jobs enter the K classes; they each equal α_1 for reentrant lines. The *traffic equations*

$$\lambda_\ell = \alpha_\ell + \sum_k \lambda_k P_{k,\ell}, \tag{1.6}$$

or, in vector form, $\lambda = \alpha + P^T\lambda$, are equivalent to (1.5), and are useful in certain situations. Employing m and λ, we define the *traffic intensity* ρ_j at station j to be

$$\rho_j = \sum_{k \in \mathcal{C}(j)} m_k \lambda_k. \tag{1.7}$$

(ρ_j is also known as the *nominal load*.) The traffic intensity is the rate at which work for the station j enters the network, and reduces to (1.3) for reentrant lines. We write ρ for the corresponding traffic intensity vector.

We continue the analogy with reentrant lines, and say that a station j is *subcritical* if $\rho_j < 1$, *critical* if $\rho_j = 1$, and *supercritical* if $\rho_j > 1$. When all stations of a network are either subcritical or are critical, we refer to the network as being subcritical or critical. We will sometimes abbreviate these conditions by writing $\rho < e$ and $\rho = e$, where $e = (1, \ldots, 1)^T$; we similarly write $\rho \leq e$,when $\rho_j \leq 1$ for each j. When at least one station is supercritical, we refer to the network as being supercritical. As in the reentrant line setting, a supercritical queueing network will not be stable, and the number of jobs in the network will increase linearly as $t \to \infty$. This is a bit tedious to show directly; it will follow quickly using fluid limits in Proposition 5.21. Of course, since a reentrant line with $\rho < e$ need not be stable, the same is the case for queueing networks with general routing. We will show in Chapters 4 and 5, though, that the condition $\rho < e$ is sufficient for the stability of queueing networks under various disciplines. Also, (1.4) suffices for stability irrespective of the discipline; a variant of this is shown in Example 1 at the end of Section 4.4.

In Chapters 4 and 5, we will also establish criteria for when a queueing network is *e-stable*. By this, we will mean that the underlying Markov process

is ergodic, i.e., it possesses a stationary distribution π to which the distribution at time t converges in total variation norm as $t \to \infty$, irrespective of the initial state. When the state space is countable, this is equivalent to the probabilities converging to $\pi(x)$ at each point x. As we will see, results on stability can be modified to results on e-stability with little additional work.

We have so far not used the term "unstable". Somewhat different definitions exist in the literature; in each case, they mean more than just "not stable". For us, a queueing network will be *unstable* if, for some initial state, the number of jobs in the network will, with positive probability, go to infinity as $t \to \infty$. (When the network has only a finite number of states with fewer than a given number of jobs, and all states communicate with one another, this is equivalent to saying that for each initial state, the number of jobs in the network goes to infinity almost surely as $t \to \infty$.) According to the paragraph before the last, a supercritical queueing network will be unstable; in fact, the number of jobs in the network grows linearly in time. This will typically be the case for unstable networks, such as those given in Chapter 3.

1.3 Queueing Network Equations and Fluid Models

One of the main themes in these lectures will be the application of fluid models to study the stability of queueing networks. Fluid models will be introduced and studied in detail in Chapter 4; they will then by applied to specific disciplines in Chapter 5. We provide here some of the basic motivation for fluid models.

In the last section, we gave various examples of queueing networks. Such systems are frequently complicated. They can be interpreted as Markov processes, but to derive specific results, one needs a means of expressing the properties of the specific queueing network. *Queueing network equations* provide an analytic formulation for this. After deriving these equations and taking appropriate limits, one obtains the corresponding fluid models.

Queueing network equations tie together random vectors, such as $A(t)$, $D(t)$, $T(t)$, and $Z(t)$, that describe the evolution of a queueing network. The individual coordinates of these K dimensional vectors correspond, respectively, to the cumulative number of arrivals $A_k(t)$ and departures $D_k(t)$ at a class k up to time t, the cumulative time $T_k(t)$ spent serving this class up to time t, and the number of jobs $Z_k(t)$ at this class at time t; we introduced $Z(t)$ in the last section. Examples of queueing network equations are

$$A(t) = E(t) + \sum_k \Phi^k(D_k(t)), \tag{1.8}$$

$$Z(t) = Z(0) + A(t) - D(t), \tag{1.9}$$

$$D_k(t) = S_k(T_k(t)), \quad k = 1, \ldots, K. \tag{1.10}$$

We are employing here the following terminology. The vector $E(t)$ is the cumulative number of jobs arriving by time t at each class from outside the network (i.e., *external arrivals*). When the interarrival times are exponentially distributed, $E(\cdot)$ will be a Poisson process. The vector $\Phi^k(d_k)$ is the number of the first d_k departures from class k that are routed to each of the K classes; $S_k(t_k)$ is the cumulative number of departures from class k after t_k units of service there. We will denote by Φ and S the matrix and vector corresponding to these quantities.

The quantities $E(\cdot)$, $S(\cdot)$, and $\Phi(\cdot)$ should be thought of as random, but known input into the system, from which the evolution of $A(\cdot), D(\cdot), T(\cdot)$, and $Z(\cdot)$ will be determined via (1.8)–(1.10) and other equations. The middle equation is easiest to read, and just says that the number of jobs at time t is equal to the original number, plus arrivals, and minus departures. The first equation says the total number of arrivals is equal to the number of external arrivals plus the number of arrivals from other classes; the last equation gives the number of departures at a class as a function of the time spent serving jobs there. We note that the first two equations hold irrespective of the discipline, whereas the last equation requires the discipline to be HL.

Other choices of variables, in addition to $A(\cdot), D(\cdot), T(\cdot)$ and $Z(\cdot)$, are frequently made. We will include other variables, such as the immediate workload $W(\cdot)$, in our detailed treatment in Section 4.3. Often, different formulations of the queueing network equations are equivalent, with the exact format being chosen for convenience. One can, for example, eliminate $A(t)$ and $D(t)$ in (1.8)–(1.10), and instead employ the single equation

$$Z(t) = Z(0) + E(t) + \sum_k \Phi^k(S_k(T_k(t))) - S(T(t)), \tag{1.11}$$

if one is just interested in the evolution of $Z(t)$ (which is most often the case). Note that, for multiclass networks, neither (1.8)–(1.10) nor (1.11) supplies enough information to solve for the unknown variables, since the discipline has not been specified, and so $T(\cdot)$ has not been uniquely determined. To specify the discipline, an additional equation (or equations) is required. For single class networks, this complication is not present. As an elementary example, note that for the $M/M/1$ queue, (1.11) reduces to the simple scalar equation

$$Z(t) = Z(0) + E(t) - S\left(\int_0^t \mathbf{1}\left\{Z(s) > 0\right\} ds\right), \tag{1.12}$$

since departing jobs are not rerouted into the queue and the network is nonidling.

Fluid model equations are the deterministic analog of queueing network equations, with one replacing the random quantities $E(\cdot), S(\cdot)$, and $\Phi(\cdot)$ by their corresponding means. The fluid model equations corresponding to (1.8)-(1.10) are then

$$A(t) = \alpha t + P^T D(t), \tag{1.13}$$
$$Z(t) = Z(0) + A(t) - D(t), \tag{1.14}$$
$$D_k(t) = \mu_k T_k(t), \quad k = 1, \ldots, K, \tag{1.15}$$

where α, P, and μ_k were defined in the previous section. By employing the matrix M, one can also write (1.15) as

$$D(t) = M^{-1}T(t). \tag{1.15$'$}$$

Similarly, the fluid model equation corresponding to (1.11) is

$$Z(t) = Z(0) + \alpha t + (P^T - I)M^{-1}T(t). \tag{1.16}$$

For the $M/M/1$ queue, this reduces to the scalar equation

$$Z(t) = Z(0) + \alpha t - \mu \int_0^t \mathbf{1}\{Z(s) > 0\}\, ds. \tag{1.17}$$

Fluid model equations can be thought of as belonging to a *fluid network* which is the deterministic equivalent of the given queueing network. Jobs are replaced by continuous fluid mass (or "job mass"), which follows the same routing as before. The constant rate at which such mass enters the network is given by the vector α. The rate at which mass is served for a class is μ_k and the service time per unit mass is $m_k = 1/\mu_k$. A set of fluid model equations, as in (1.13)–(1.15), is referred to collectively as a *fluid model*.

The solutions of fluid models are frequently much easier to analyze than are the corresponding queueing network equations. As we will see in Chapter 4, fluid model equations can be derived from the corresponding queueing network equations by taking limits of $Z(\cdot)$ and the other quantities after hydrodynamic scaling. (That is, "law of large numbers" scaling of the form $Z(st)/s$ as $s \to \infty$.) Fluid models will provide a valuable tool for demonstrating the stability of queueing networks, as we will see in Chapters 4 and 5. Fluid models are also an important tool for studying *heavy traffic limits*, which lie outside the scope of these lectures.

To analyze the stability of a queueing network, one introduces the following notion of stability for a fluid model. A fluid model is said to be *stable* if there exists an $N > 0$, so that for any solution of the fluid model equations, its $Z(\cdot)$ component satisfies

$$Z(t) = 0 \quad \text{for } t \geq N|Z(0)|. \tag{1.18}$$

(Here, $|\cdot|$ denotes the ℓ^1 norm.) We will show in Chapter 4 that, under certain conditions on the interarrival and service times, a queueing network will be stable if its fluid model is stable.

In the special case of the fluid model equation (1.17) corresponding to the $M/M/1$ queue, it is easy to explicitly solve for $Z(\cdot)$. For $\mu > \alpha$, the solution drifts linearly to 0 at rate $\mu - \alpha$, after which it remains there, and so

$$Z(t) = 0 \quad \text{for } t \geq Z(0)/(\mu - \alpha), \tag{1.19}$$

which is a special case of (1.18). This behavior of the solution of equation (1.17) corresponding to a subcritical $M/M/1$ queue is not surprising, since the solution $Z(\cdot)$ of (1.12) possesses the same negative drift $\alpha - \mu$ when $Z(t) > 0$. For more general networks, one typically constructs a Lyapunov function with respect to which the fluid model solution $Z(\cdot)$ of (1.16) exhibits a uniformly negative drift until hitting 0.

Despite the utility of fluid models, one needs to exercise some caution in their application. In particular, a fluid model need not have a unique solution for a given initial condition, since solutions might bifurcate. As we will see in Section 4.3, this can be the case even for certain standard disciplines. (Such behavior might occur at times when there are two or more empty multiclass stations.) On account of this, thinking of fluid model equations as belonging to a fluid network loses some of its appeal. In practice, it is often better to think directly in terms of the fluid model which is defined by the appropriate system of equations.

1.4 Outline of Lectures

We conclude the introduction with an outline of the subject matter we will be covering. The material can be broken into three parts, Chapter 2, Chapter 3, and Chapters 4 and 5, each with its own distinct character.

Chapter 2 discusses the "classical" queueing networks introduced in [BaCMP75] and the accompanying theory of quasi-reversible queueing networks in [Ke79]. The examples considered in [BaCMP75] include networks with the FIFO, PS, and LIFO disciplines that were introduced in Section 1.2, and an infinite server network, which we will introduce in Chapter 2. Exponentially distributed interarrival times, and in some cases, exponentially distributed service times, are assumed. For $\rho < e$, these networks are stable and explicit product-like formulas are given for their stationary distributions. This special structure is a generalization of that for the $M/M/1$ queue. Independent work of F. P. Kelly, leading to the book [Ke79], employed quasi-reversibility to show that these explicit formulas hold in a more general setting. These results have strongly influenced the development of queueing theory over the past several decades.

In the previous sections, we mentioned that even when $\rho < e$, a queueing network may be unstable. This behavior came as a surprise in the early 1990's. Various examples of instability have since been given in different settings, primarily in the context of SBP and FIFO disciplines. At this point, there is no comprehensive theory, and in fact, not much is known, in general, about how the Markov processes for such networks go to infinity as $t \to \infty$. Chapter 3 presents the best known examples of unstable subcritical queueing networks in more-or-less chronological order, with an attempt being made to provide

some feeling for the development of the subject. Examples include those from [LuK91],[RyS92], [Br94a], [Se94], and [Du97].

In Chapter 4, we give general sufficient conditions for the stability of a queueing network. The main condition is that the fluid model of the queueing network be stable; general conditions on the interarrival and service times of the queueing network are also needed. In the previous section, we gave a brief discussion of how the fluid model equations are obtained from the queueing network equations that describe the evolution of the queueing network. The material in Chapter 4 is largely based on [Da95] and the sources employed there. We go into considerable detail on the arguments leading to the main result, Theorem 4.16, because we feel that it is important to have this material accessible in one place.

The first part of Chapter 5 consists of applications of Theorem 4.16, where stability is demonstrated, under $\rho < e$, for a number of disciplines. In Sections 5.1, 5.2, and 5.3, the stability of single class networks, SBP reentrant lines with FBFS and LBFS priority schemes, and FIFO networks, with constant mean service times at a station, are demonstrated. In each case, the procedure is to demonstrate the stability of a fluid model; the stability of the queueing network then follows by applying the above theorem.

Sections 5.4 and 5.5 are different in nature from the previous three sections. Section 5.4 is concerned with the question of global stability. That is, when is a queueing network stable, irrespective of the particular discipline that is applied? Again applying Theorem 4.16, a queueing network will be globally stable if its fluid model is. For two-station fluid models with deterministic routing, a complete theory is given. Section 5.5 investigates the converse of Theorem 4.16, namely the necessity of fluid model stability for the stability of a given queueing network. It turns out that there is not an exact correspondence between the two types of stability, as examples show. Robust conditions for the necessity of fluid model stability are presently lacking. The material in Chapter 5 is taken from a member of sources, including [Br96a], [DaWe96], [Br99], [DaV00], and [DaHV04].

We conclude this section with some basic notation and conventions. We let $\mathbf{Z}_+$ and $\mathbf{R}_+$ denote the positive integers and real numbers; $\mathbf{Z}_{+,0}$ and $\mathbf{R}_{+,0}$ will denote the sets appended with $\{0\}$; and $\mathbf{Z}_+^d,\ \mathbf{R}_+^d,\ \mathbf{Z}_{+,0}^d$, and $\mathbf{R}_{+,0}^d$ will denote their d dimensional equivalents. For $x, y \in \mathbf{R}^d,\ x \leq y$ means $x_k \leq y_k$ for each coordinate k; $x < y$ means $x_k < y_k$ for each coordinate. Unless indicated otherwise, $|\cdot|$ denotes the ℓ^1 (or sum) norm, e.g., $|x| = \sum_{i=1}^d |x_i|$ for $x \in \mathbf{R}^d$. By $a \vee b$ and $a \wedge b$, we mean $\max\{a, b\}$ and $\min\{a, b\}$. For $x \in \mathbf{R}$, by $\lfloor x \rfloor$ and $\lceil x \rceil$, we mean the integer part of x and the smallest integer that is at least x. By $a(t) \sim b(t)$, we mean $a(t)/b(t) \to 1$ as $t \to \infty$, and by $a \approx b$, that a and b are approximately the same in some weaker sense.

When convenient, we will refer to a countable state Markov process as a Markov chain. (In the literature, Markov chain often instead refers to a discrete time Markov process.) We say that a Markov chain is positive recurrent if each state is positive recurrent and all states communicate. (The second

condition is sometimes not assumed in the literature.) As already mentioned, instead of the term "queueing network", we will often employ the shorter "network", when the context is clear. Throughout these lectures, continuous time stochastic processes will be assumed to be right continuous unless indicated otherwise.

2

The Classical Networks

In this chapter, we discuss two families of queueing networks whose Markov processes are positive recurrent when $\rho < e$, and whose stationary distributions have explicit product-like formulas. The first family includes networks with the FIFO discipline, and the second family includes networks with the PS and LIFO disciplines, as well as infinite server (IS) networks. We introduced the first three networks in Section 1.2; we will define infinite server networks shortly. These four networks are sometimes known as the "classical networks". Together with their generalizations, they have had a major influence on the development of queueing theory because of the explicit nature of their stationary distributions. For this reason, we present the basic results for the accompanying theory here, although only the FIFO discipline is HL. These results are primarily from [BaCMP75], and papers by F. P. Kelly (such as [Ke75] and [Ke76]) that led to the book [Ke79].

In Section 2.1, we state the main results in the context of the above four networks. We first characterize the stationary distributions for networks consisting of a single station, whose jobs exit from the network when service is completed, without being routed to another class. We will refer to such a station as a *node*. We then characterize the stationary distribution for networks with multiple stations and general routing. Since all states will communicate, the Markov processes for the networks will be positive recurrent, and hence the networks will be stable.

In the remainder of the chapter, we present the background for these results and the accompanying theory. In Section 2.2, we give certain basic properties of stationary and reversible Markov processes on countable state spaces that we will use later. Sections 2.3 and 2.4 apply this material to obtain generalizations of the node-level results in Section 2.1 to the two families of interest to us. The first family, homogeneous nodes, includes FIFO nodes under certain restrictions, and the second family, symmetric nodes, includes PS, LIFO, and IS nodes.

The concept of quasi-reversibility is introduced in Section 2.5. Using quasi-reversibility, the stationary distributions of certain queueing networks can be

M. Bramson, *Stability of Queueing Networks*. Lecture Notes in Mathematics 1950. 17

written as the product of the stationary distributions of nodes that correspond to the stations "in isolation". These queueing networks include the FIFO, PS, LIFO, and IS networks, and so generalize the network-level results in Section 2.1. Except for Theorem 2.9, all of the material in this chapter is for queueing networks with a countable state space.

The main source of the material in this chapter is [Ke79]. Section 2.2 is essentially an abridged version of the material in Chapter 1 of [Ke79]. Most of the material in Sections 2.3-2.5 is from Sections 3.1-3.3 of [Ke79], with [Wa88], [ChY01], [As03], and lecture notes by J.M. Harrison having also been consulted. The order of presentation here, starting with nodes in Sections 2.3 and 2.4 and ending with quasi-stationarity in Section 2.5, is different.

2.1 Main Results

In this section, we will give explicit formulas for the stationary distributions of FIFO, PS, LIFO, and infinite server networks. Theorems 2.1 and 2.2 state these results for individual nodes, and Theorem 2.3 does so for networks. In Sections 2.3-2.5, we will prove generalizations of these results.

The range of disciplines that we consider here is of limited scope. On the other hand, the routing that is allowed for the network-level results will be completely general. As in Chapter 1, routing will be given by a mean transition matrix $P = \{P_{k,\ell},\ k, \ell = 1, \ldots, K\}$ for which the network is open.[1] For all queueing networks considered in this section, the interarrival times are assumed to be exponentially distributed. When the service times are also exponentially distributed, the evolution of these queueing networks can be expressed in terms of a countable state Markov process. By enriching the state space, more general service times can also be considered in the countable state space setting. This will be useful for the PS, LIFO, and IS networks.

In Section 1.1, we introduced the $M/M/1$ queue with external arrival rate $\alpha = 1$. By employing the reversibility of its Markov process when $m < 1$, we saw that its stationary distribution is given by the geometric distribution in (1.1). Allowing α to be arbitrary with $\alpha m < 1$, this generalizes to

$$\pi(n) = (1 - \alpha m)(\alpha m)^n \quad \text{for } n = 0, 1, 2, \ldots. \tag{2.1}$$

For a surprisingly large group of queueing networks, generalizations of (2.1) hold, with the stationary distribution being given by products of terms similar to those on the right side of (2.1).

[1] In the literature (such as in [Ke79]), deterministic routing is frequently employed. For these lectures, we prefer to use random routing, which was used in [BaCMP75] and has been promulgated in its current form by J. M. Harrison. By employing sufficiently many routes, one can show the two approaches are equivalent. We find the approach with random routing to be notationally more flexible. The formulation is also more amenable to problems involving dynamic scheduling, which we do not cover here.

We first consider queueing networks consisting of just a single node. That is, jobs at the unique station leave the network immediately upon completion of service, without being routed to other classes. Classes are labelled $k = 1, \ldots, K$. The nodes of interest to us in this chapter fall into two basic families, depending on the discipline.

The first family, homogeneous nodes, will be defined in Section 2.3. FIFO nodes, which are the canonical example for this family, will be considered here. For homogeneous nodes, including FIFO nodes, we need to assume that the mean service times m_k at all classes k are equal. In order to avoid confusion with the vector m, we label such a service time by m^s (with "s" standing for "station"). We will refer to such a node as a *FIFO node of Kelly type*. In addition to assuming the interarrival times are exponentially distributed, we assume the same is true for the service times.

The state x of the node at any time will be specified by an n-tuple of the form

$$(x(1), \ldots, x(n)), \tag{2.2a}$$

where n is the number of jobs in the node and

$$x(i) \in \{1, \ldots, K\} \quad \text{for } i = 1, \ldots, n \tag{2.2b}$$

gives the class of the job in the i^{th} position in the node. We interpret $i = 1, \ldots, n$ as giving the order of arrival of the jobs currently in the node; because of the FIFO discipline, all service is directed to the job at $i = 1$. The state space S_0 will be the union of these states. The stochastic process $X(t),\ t \geq 0$, thus defined will be Markov with a countable state space. For consistency with other chapters, we interpret vectors as column vectors, although this is not needed in the present chapter (since matrix multiplication is not employed).

All states communicate with the empty state. So, if $X(\cdot)$ has a stationary distribution, it will be unique. Since there is only a single station, a stationary distribution will exist when the node is subcritical. Theorem 2.1 below gives an explicit formula for the distribution. As in Chapter 1, α_k denotes the external arrival rates at the different classes k; ρ denotes the traffic intensity and is in the present setting given by the scalar

$$\rho = m^s \sum_k \alpha_k. \tag{2.3}$$

As elsewhere in this chapter, when we say that a node (or queueing network) has a stationary distribution, we mean that its Markov process, on the chosen state space, has this distribution. (For us, "distribution" is synonymous with the somewhat longer "probability measure".)

Theorem 2.1. *Each subcritical FIFO node of Kelly type has a stationary distribution π, which is given by*

$$\pi(x) = (1 - \rho) \prod_{i=1}^{n} m^s \alpha_{x(i)}, \tag{2.4}$$

for $x = (x(1), \ldots, x(n)) \in S_0$.

The stationary distribution π in Theorem 2.1 can be described as follows. The probability of there being a total of n jobs in the node is $(1-\rho)\rho^n$. Given a total of n jobs, the probability of there being $n_1, \ldots, n_K$ jobs at the classes $1, \ldots, K$, with $n = n_1 + \ldots + n_K$ and no attention being paid to their order, is

$$\rho^{-n} \binom{n}{n_1, \ldots, n_K} \prod_{k=1}^{K} (m^s \alpha_k)^{n_k}. \tag{2.5}$$

Moreover, given that there are $n_1, \ldots, n_K$ jobs at the classes $1, \ldots, K$, any ordering of the different classes of jobs is equally likely. Note that since all states have positive probability of occurring, the process $X(\cdot)$ is positive recurrent. Consequently, the node is stable.

The other family of nodes that will be discussed in this chapter, symmetric nodes, will be defined in Section 2.4. Standard members of this family are PS, LIFO, and IS nodes. The PS and LIFO disciplines were specified in Chapter 1. In an *infinite server* (IS) node, each job is assumed to start receiving service as soon as it enters the node, which it receives at rate 1. One can therefore think of there being an infinite number of unoccupied servers available to provide service, one of which is selected whenever a job enters the node. All other disciplines studied in these lectures will have only a single server at a given station. We note that although the PS discipline is not HL, it is related to the HLPPS discipline given at the end of Section 5.3.

We consider the stationary distributions of PS, LIFO and IS nodes. As with FIFO nodes, we need to assume that the interarrival times of jobs are exponentially distributed. If we assume that the service times are also exponentially distributed, then the process $X(\cdot)$ defined on S_0 will be Markov. As before, we interpret the coordinates $i = 1, \ldots, n$ in (2.2) as giving the order of jobs currently in the node. For LIFO nodes, this is also the order of arrival of jobs there. For reasons relating to the definition of symmetric nodes in Section 2.4, we will instead assume, for PS and IS nodes, that arriving jobs are, with equal probability $1/n$, placed at one of the n positions of the node, where n is the number of jobs present after the arrival of the job. Since in both cases, jobs are served at the same rate irrespective of their position in the node, the processes $X(\cdot)$ defined in this manner are equivalent to the processes defined by jobs always arriving at the rear of the node.

The analog of Theorem 2.1 holds for PS, LIFO, and IS nodes when the service times are exponentially distributed. We no longer need to assume that the service times have the same means, so in the present setting, the traffic intensity ρ is given by

$$\rho = \sum_k m_k \alpha_k. \tag{2.6}$$

For subcritical PS and LIFO nodes, the stationary distribution π is given by

$$\pi(x) = (1-\rho)\prod_{i=1}^{n} m_{x(i)}\alpha_{x(i)}, \tag{2.7}$$

for $x = (x(1), \ldots, x(n)) \in S_0$. For any IS node, the stationary distribution π is given by

$$\pi(x) = \frac{e^{-\rho}}{n!}\prod_{i=1}^{n} m_{x(i)}\alpha_{x(i)}. \tag{2.8}$$

The stability of the infinite server node for all values of ρ is not surprising, since the total rate of service at the node is proportional to the number of jobs presently there.

For PS, LIFO, and IS nodes, an analogous result still holds when exponential service times are replaced by service times with more general distributions. One employs the "method of stages", which is defined in Section 2.4. One enriches the state space S_0 to allow for different stages of service for each job, with a job advancing to its next stage of service after service at the previous stage has been completed. After service at the last stage has been completed, the job leaves the node. Since the service times at each stage are assumed to be exponentially distributed, the corresponding process $X(\cdot)$ for the node, on this enriched state space S_e, will still be Markov. On the other hand, the total service time required by a given job will be the sum of the exponential service times at its different stages, which we take to be i.i.d. Such service times are said to have *Erlang distributions*.

Using the method of stages, one can extend the formulas (2.7) and (2.8), for the stationary distributions of PS, LIFO, and IS nodes, to nodes that have service distributions which are countable mixtures of Erlang distributions. This result is stated in Theorem 2.2. The state space here for the Markov process $X(\cdot)$ of the node is S_e, which is defined in Section 2.4. The analog of this extension for FIFO nodes is not valid.

Theorem 2.2. *Each subcritical PS node and LIFO node, whose service time distributions are mixtures of Erlang distributions, has a stationary distribution π. The probability of there being n jobs in the node with classes $x(1), \ldots, x(n)$ is given by (2.7). The same is true for any IS node with these service time distributions, but with (2.8) replacing (2.7).*

Mixtures of Erlang distributions are dense in the set of distribution functions, so it is suggestive that a result analogous to Theorem 2.2 should hold for service times with arbitrary distributions. This is in fact the case, although one needs to be more careful here, since one needs to replace S_e with an uncountable state space which specifies the residual service times of jobs at the node. More detail on this setting is given at the end of Section 2.4. Because of the technical difficulties for uncountable state spaces, (2.7) and (2.8) are typically stated for mixtures of Erlang distributions or other related distributions on countable state spaces. Moreover, quasi-reversibility, which we discuss shortly, employs a countable state space setting.

So far in this section, we have restricted our attention to nodes. As mentioned at the beginning of the section, the results in Theorems 2.1 and 2.2 extend to analogous results for queueing networks, which are given in Theorem 2.3, below. FIFO, PS, LIFO, and IS queueing networks are the analogs of the respective nodes, with jobs at individual stations being subjected to the same service rules as before, and, upon completion of service at a class k, a job returning to class ℓ with probability $P_{k,\ell}$, which is given by the mean transition matrix P. In addition to applying to these queueing networks, Theorem 2.3 also applies to networks that are mixtures of such stations, with one of the above four rules holding for any particular station. We also note that the formula in Theorem 2.3 holds for Jackson networks, as a special case of FIFO networks. The product formula for Jackson networks in [Ja63] predates those for the other networks.

In Theorem 2.3, we assume that the service times are exponentially distributed when the station is FIFO, and are mixtures of Erlang distributions in the other three cases. The distribution function π is defined on the state space

$$S = S^1 \times \ldots \times S^J,$$

where $S^j = S_0$ if j is FIFO and $S^j = S_e$ for the other cases, and π^j is defined on S^j. (The choice of K in each factor depends on S^j.)

Theorem 2.3. *Suppose that each station j of a queueing network is either FIFO of Kelly type, PS, LIFO, or IS. Suppose that in the first three cases, the station is subcritical. Then, the queueing network has a stationary distribution π that is given by*

$$\pi(x) = \prod_{j=1}^{J} \pi^j(x^j), \tag{2.9}$$

for $x = (x^1, \ldots, x^J)$. Here, each π^j is either of the form (2.4), (2.7), or (2.8), depending on whether the station j is FIFO of Kelly type, PS or LIFO, or IS, and α_k in the formulas is replaced by λ_k.

Theorem 2.3 will be a consequence of Theorems 2.1 and 2.2, and of the quasi-reversibility of the nodes there. Quasi-reversibility will be introduced in Section 2.5. Using quasi-reversibility, it will be shown, in Theorem 2.11, that the stationary distributions of certain queueing networks can be written as the product of the stationary distributions of nodes that correspond to the individual stations "in isolation". This will mean that when service of a job at a class is completed, the job will leave the network rather than returning to another class (either at the same or a different station). The external arrival rates α_k at classes are replaced by the total arrival rates λ_k of the network in order to compensate for the loss of jobs that would return to them. Quasi-reversibility can be applied to queueing networks whose stations are FIFO, PS, LIFO, or IS. By employing Theorem 2.11, one obtains Theorem 2.3 as a special case of Theorem 2.12.

2.2 Stationarity and Reversibility

In this section, we will summarize certain basic results for countable state, continuous time Markov processes. We define stationarity and reversibility, and provide alternative characterizations. Proposition 2.6, in particular, will be used in the remainder of the chapter.

The Markov processes $X(t),\ t \geq 0$, we consider here will be assumed to be defined on a countable state space S. The space S will be assumed to be irreducible, that is, all states communicate. None of the states will be instantaneous; we will assume there are only a finite number of transitions after a finite time, and hence no explosions. Sample paths will therefore be right continuous with left limits. The transition rate between states x and y will be denoted by $q(x, y)$; the rate at which a transition occurs at x is therefore $q(x) \stackrel{\text{def}}{=} \sum_{y \in S} q(x, y)$. The *embedded jump chain* has mean transition matrix $\{p(x, y),\ x, y \in S\}$, where $p(x, y) \stackrel{\text{def}}{=} q(x, y)/q(x)$ is the probability $X(\cdot)$ next visits y from the state x. The time $X(\cdot)$ remains at a state x before a transition occurs is exponentially distributed with mean $1/q(x)$.

A stochastic process $X(t),\ t \geq 0$, is said to be *stationary* if $(X(t_1), \ldots, X(t_n))$ has the same distribution as $(X(t_1 + u), \ldots, X(t_n + u))$, for each nonnegative $t_1, \ldots, t_n$ and u. Such a process can always be extended to $-\infty < t < \infty$ so that it is stationary as before, but with $t_1, \ldots, t_n$ and u now being allowed to assume any real values. When $X(\cdot)$ is a Markov process, it suffices to consider just $n = 1$ in order to verify stationarity.

A *stationary distribution* $\pi = \{\pi(x),\ x \in S\}$ for a Markov process $X(\cdot)$ satisfies the *balance equations*

$$\pi(x) \sum_{y \in S} q(x, y) = \sum_{y \in S} \pi(y) q(y, x) \quad \text{for } x \in S, \tag{2.10}$$

which say that the rates at which mass leaves and enters a state x are the same. We are assuming here that $\sum_{x \in S} \pi(x) = 1$. Since all states are assumed to communicate, π will be unique. If π exists, then it is the limit of the distributions of the Markov process starting from any initial state. If a measure π satisfying (2.10) with $\sum_{x \in S} \pi(x) < \infty$ exists, then it can be normalized so that $\sum_{x \in S} \pi(x) = 1$. If $\sum_{x \in S} \pi(x) = \infty$, then there is no stationary distribution, and for all $x, y \in S$,

$$P(X(t) = \ell \mid X(0) = x) \to 0 \quad \text{as } t \to \infty.$$

(See, e.g., [Re92] for such basic theory.)

A stochastic process $X(t),\ -\infty < t < \infty$, is said to be *reversible* if $(X(t_1), \ldots, X(t_n))$ has the same distribution as $(X(u - t_1), \ldots, X(u - t_n))$, for each $t_1, \ldots, t_n$ and u. This condition says that the process is stochastically indistinguishable, whether it is run forward or backwards in time. It is easy to see that if $X(\cdot)$ is reversible, then it must be stationary. When $X(\cdot)$

is Markov, it suffices to consider $n = 2$ in order to verify reversibility. The Markov property can be formulated as saying that the past and future states of the process $X(\cdot)$ are independent given the present. It follows that the *reversed process* $\hat{X}(t) \stackrel{\text{def}}{=} X(-t)$ is Markov exactly when $X(\cdot)$ is. If $X(\cdot)$ has stationary measure π, then π is also stationary for $\hat{X}(\cdot)$ and the transition rates of $\hat{X}(\cdot)$ are given by

$$\hat{q}(x, y) = \frac{\pi(y)}{\pi(x)} q(y, z) \quad \text{for } x, y \in S. \tag{2.11}$$

A stationary Markov process $X(\cdot)$ with distribution π is reversible exactly when it satisfies the *detailed balance equations*

$$\pi(x)q(x, y) = \pi(y)q(y, x) \quad \text{for } x, y \in S. \tag{2.12}$$

This condition says that the rate at which mass moves from x to y is the same rate at which it moves in the reverse direction. This condition need not, of course, be satisfied for arbitrary stationary distributions. When (2.12) holds, it often enables one to express the stationary distribution in closed form, as, for example, for the $M/M/1$ queue in Section 1.1. It will always hold for the stationary distribution of any birth and death process, and, more generally, for the stationary distribution of any Markov process on a tree. By summing over ℓ, one obtains the balance equations in (2.10) from (2.12).

An alternative characterization of reversibility is given by Proposition 2.4. We will not employ the proposition elsewhere, but state it because it provides useful intuition for the concept.

Proposition 2.4. *A stationary Markov process is reversible if and only if its transition rates satisfy*

$$\begin{aligned} q(x_1, x_2)q(x_2, x_3) \cdots q(x_{n-1}, x_n)q(x_n, x_1) \\ = q(x_1, x_n)q(x_n, x_{n-1}) \cdots q(x_3, x_2)q(x_2, x_1), \end{aligned} \tag{2.13}$$

for any $x_1, x_2, \ldots, x_n$.

The equality (2.13) says that the joint transition rates of the Markov process are the same along a path if it starts and ends at the same point, irrespective of its direction along the path.

Proof of Proposition 2.4. The "only if" direction follows immediately by plugging (2.12) into (2.13).

For the "if" direction, fix x_0, and define

$$\pi(x) = \prod_{i=1}^{n} [q(x_{i-1}, x_i)/q(x_i, x_{i-1})], \tag{2.14}$$

where $x_n = x$ and $x_0, x_1, \ldots, x_n$ is any path from x_0 to x, with $q(x_i, x_{i-1}) > 0$. One can check using (2.13) that the right side of (2.14) does not depend on the

particular path that is chosen, and so $\pi(x)$ is well defined. To see this, let $\mathscr{P}_1$ and $\mathscr{P}_2$ be any two paths from x_0 to x, and $\hat{\mathscr{P}}_1$ and $\hat{\mathscr{P}}_2$ be the corresponding paths in the reverse directions. Then, the two paths from x_0 to itself formed by linking $\mathscr{P}_1$ to $\hat{\mathscr{P}}_2$, respectively $\mathscr{P}_2$ to $\hat{\mathscr{P}}_1$, satisfy (2.13), from whence the uniqueness in (2.14) will follow.

Assume that for given y, $q(y,x) > 0$. Multiplication of both sides of (2.14) by $q(x,y)$ implies that

$$\begin{aligned}\pi(x)q(x,y) &= q(y,x)\left(\prod_{i=1}^{n} q(x_{i-1},x_i)/q(x_i,x_{i-1}\right)(q(x,y)/q(y,x)) \\ &= q(y,x)\pi(y).\end{aligned}$$

This gives (2.12), since the case $q(x,y) = q(y,x) = 0$ is trivial. Since the process is assumed to be stationary, $\sum_x \pi(x) < \infty$, and so π can be scaled so that $\sum_x \pi(x) = 1$. ■

The following result says that under certain modifications of the transition rates $q(x,y)$, a reversible Markov process will still be reversible. It will not be needed for later work, but has interesting applications.

Proposition 2.5. *Suppose that the transition rates of a reversible Markov process $X(\cdot)$, with state space S and stationary distribution π, are altered by changing $q(x,y)$ to $q'(x,y) = bq(x,y)$ when $x \in A$ and $y \notin A$, for some $A \subseteq S$. Then, the resulting Markov process $X'(\cdot)$ is reversible and has stationary distribution*

$$\pi'(x) = \begin{cases} c\pi(x) & \text{for } x \in A, \\ cb\pi(x) & \text{for } x \notin A, \end{cases} \tag{2.15}$$

where c is chosen so that $\sum_x \pi'(x) = 1$. In particular, when the state space is restricted to A by setting $b = 0$, then the stationary distribution of $X'(\cdot)$ is given by

$$\pi'(x) = \pi(x) \Big/ \sum_{y \in A} \pi(y) \quad \text{for } x \in A. \tag{2.16}$$

Proof. It is easy to check that q' and π' satisfy the detailed balance equations in (2.12). ■

The following illustration of Proposition 2.5 is given in [Ke79].

Example 1. *Two queues with a joint waiting room.* Suppose that two independent $M/M/1$ queues are given, with external arrival rates α_i and mean service times m_i, and $\alpha_i m_i < 1$. Let $X_i(t)$ be the number of customers (or jobs) in each queue at time t. The Markov processes are each reversible with stationary distributions as in (2.1). It is easy to check that the joint Markov process $X(t) = (X_1(t), X_2(t))$, $-\infty < t < \infty$, is reversible, with stationary distribution

$$\pi(n_1, n_2) = (1 - \alpha_1 m_2)(1 - \alpha_2 m_2)(\alpha_1 m_1)^{n_1}(\alpha_2 m_2)^{n_2} \quad \text{for } n_i \in \mathbf{Z}_{+,0}.$$

Suppose now that the queues are required to share a common waiting room of size N, so that a customer who arrives to find N customers already there leaves without being served. This corresponds to restricting $X(\cdot)$ to the set A of states with $n_1 + n_2 \leq N$. By Proposition 2.5, the corresponding process $X'(\cdot)$ is reversible, and has stationary measure

$$\pi'(n_1, n_2) = \pi(0, 0)(\alpha_1 m_1)^{n_1}(\alpha_2 m_2)^{n_2} \quad \text{for } (n_1, n_2) \in A. \qquad \blacksquare$$

It is often tedious to check the balance equations (2.10) in order to determine that a Markov process $X(\cdot)$ is stationary. Proposition 2.6 gives the following alternative formulation. We abbreviate by setting

$$q(x) = \sum_{y \in S} q(x, y), \quad \hat{q}(x) = \sum_{y \in S} \hat{q}(x, y), \tag{2.17}$$

where $q(x, y)$ are the transition rates for $X(\cdot)$ and $\hat{q}(x, y) \geq 0$ are for the moment arbitrary. When $X(\cdot)$ has stationary distribution π and $\hat{q}(x, y)$ is given by (2.11), it is easy to check that

$$q(x) = \hat{q}(x) \quad \text{for all } x. \tag{2.18}$$

The proposition gives a converse to this. As elsewhere in this section, we are assuming that S is irreducible.

Proposition 2.6. *Let $X(t)$, $-\infty < t < \infty$, be a Markov process with transition rates $\{q(x, y),\ x, y \in S\}$. Suppose that for given quantities $\{\hat{q}(x, y),\ x, y \in S\}$ and $\{\pi(x),\ x \in S\}$, with $\hat{q}(x, y) \geq 0$, $\pi(x) > 0$, and $\sum_x \pi(x) = 1$, that q, $\hat{q}$, and π satisfy (2.11) and (2.18). Then, π is the stationary distribution of $X(\cdot)$ and $\hat{q}$ gives the transition rates of the reversed process.*

Proof. It follows, by applying (2.11) and then (2.18), that

$$\sum_{x \in S} \pi(x) q(x, y) = \pi(y) \sum_{x \in S} \hat{q}(y, x) = \pi(y) \hat{q}(y) = \pi(y) q(y).$$

So, π is stationary for $X(\cdot)$. The transition rates of the reversed process are therefore given by (2.11). $\blacksquare$

Proposition 2.6 simplifies the computations needed for the demonstration of stationarity by replacing the balance equations, that involve a large sum and the stationary distribution π, by two simpler equations, (2.18), which involves just a large sum, and (2.11), which involves just π. On the other hand, the application of Proposition 2.6 typically involves guessing $\hat{q}$ and π. In situations where certain choices suggest themselves, the proposition can be quite useful. It will be used repeatedly in the remainder of the chapter.

2.3 Homogeneous Nodes of Kelly Type

FIFO nodes of Kelly type belong to a larger family of nodes whose stationary distributions have similar properties. We will refer to such a node as a *homogeneous node of Kelly type*. Such nodes are defined as follows.

Consider a node with K classes. The state $x \in S_0$ of the node at any time is specified by an n-tuple as in (2.2), when there are n jobs present at the node. The ordering of the jobs is assumed to remain fixed between arrivals and service completions of jobs. When the job in position i completes its service, the position of the job is filled with the jobs in positions $i+1, \ldots, n$ moving up to positions $i, \ldots, n-1$, while retaining their previous order. Similarly, when a job arrives at the node, it is assigned some position i, with jobs previously at positions $i, \ldots, n$ being moved back to positions $i+1, \ldots, n+1$. Each job requires a given random amount of service; when this is attained, the job leaves the node. As throughout this chapter, interarrival times are required to be exponentially distributed. As elsewhere in these lectures, all interarrival and service times are assumed to be independent.

We will say that such a node is a *homogeneous node* if it also satisfies the following properties:

(a) The amount of service required by each job is exponentially distributed with mean m_k, where k is the class of the job.

(b) The total rate of service supplied at the node is $\phi(n)$, where n is the number of jobs currently there.

(c) The proportion of service that is directed at the job in position i is $\delta(i,n)$. Note that this proportion does not depend on the class of the job.

(d) When a job arrives at the node, it moves into position i, $i = 1, \ldots, n$, with probability $\beta(i,n)$, where n is the number of jobs in the node including this job. Note that this probability does not depend on the class of the job.

When the mean m_k does not depend on the class k, we will say that such a node is a *homogeneous node of Kelly type*. We will analyze these nodes in this section. We use m^s when the mean service times of a node are constant, as we did in Section 2.1.

The rate at which service is directed to a job in position i is $\delta(i,n)\phi(n)$. So, the rate at which service at the job is completed is $\delta(i,n)\phi(n)/m^s$. We will assume that $\phi(n) > 0$, except when $n = 0$. The rate at which a job arrives at a class k and position i from outside the node is $\alpha_k \beta(i,n)$. We use here the mneumonics β and δ to suggest births and deaths at a node. Of course, $\sum_{i=1}^{n} \beta(i,n) = \sum_{i=1}^{n} \delta(i,n) = 1$.

We have emphasized in the above definition that the external arrival rates and service rates β and δ do not depend on the class of the job. This is crucial for Theorem 2.7, the main result in this section. This restriction will also be

needed in Section 2.4 for symmetric nodes, as will be our assumption that the interarrival times are exponentially distributed. On the other hand, the assumptions that the service times be exponential and that their means m_k be constant, which are needed in this section, are not needed for symmetric nodes. We note that by scaling time by $1/m^s$, one can set $m^s = 1$, although we prefer the more general setup for comparison with symmetric nodes and for application in Section 2.5.

In Section 2.5, we will be interested in homogeneous queueing networks of Kelly type. *Homogeneous queueing networks* are defined analogously to homogeneous nodes. Jobs enter the network independently at the different stations according to exponentially distributed random variables, and are assigned positions at these stations as in (d). Jobs at different stations are served independently, as in (a)-(c), with departing jobs from class k being routed to class ℓ with probability $P_{k,\ell}$ and leaving the network with probability $1 - \Sigma_\ell P_{k,\ell}$. Jobs arriving at a class ℓ from within the network are assigned positions as in (d), according to the same rule as was applied for external arrivals. The external arrival rates and the quantities in (a)-(d) are allowed to depend on the station. When the mean service times m_k are assumed to depend only on the station $j = s(k)$, we may write m_j^s; we refer to such networks as *homogeneous queueing networks of Kelly type*.

The most important examples of homogeneous nodes are FIFO nodes. Here, one sets $\phi(n) \equiv 1$,

$$\beta(i,n) = \begin{cases} 1 & \text{for } i = n, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\delta(i,n) = \begin{cases} 1 & \text{for } i = 1, \\ 0 & \text{otherwise,} \end{cases}$$

for $n \in \mathbf{Z}_+$. That is, arriving jobs are always placed at the end of the queue and only the job at the front of the queue is served. Another example is given in [Ke79], where arriving jobs are again placed at the end of the queue, but where L servers are available to serve the first L jobs, for given L. In this setting, $\phi(n) = L \wedge n$, β is defined as above, and

$$\delta(i,n) = \begin{cases} 1/n & \text{for } i \le n \le L, \\ 1/L & \text{for } i \le L < n, \\ 0 & \text{for } i > L. \end{cases}$$

The main result in this section is Theorem 2.7, which is a generalization of Theorem 2.1. Since the total service rate ϕ that is provided at the node can vary, the condition

$$B \stackrel{\text{def}}{=} \sum_{n=0}^{\infty} \left(\rho^n \Big/ \prod_{i=1}^{n} \phi(i) \right) < \infty \tag{2.19}$$

replaces the assumption in Theorem 2.1 that the node is subcritical. Here, $\rho = m^s \sum_{k=1}^{K} \alpha_k$ is the traffic intensity.

Theorem 2.7. *Suppose that a homogeneous node of Kelly type satisfies $B < \infty$ in (2.19). Then, it has a stationary distribution π that is given by*

$$\pi(x) = B^{-1} \prod_{i=1}^{n} (m^s \alpha_{x(i)} / \phi(i)), \tag{2.20}$$

for $x = (x(1), \ldots, x(n)) \in S_0$.

As was the case in Theorem 2.1, the structure of the stationary distribution π in Theorem 2.7 exhibits independence at multiple levels. The probability of there being a total of n jobs at the node is $\rho^n / B \prod_{i=1}^{n} \phi(i)$. Given a total of n jobs at the node, the probability of there being $n_1, \ldots, n_k$ jobs of classes $1, \ldots, K$, respectively, is given by the multinomial distribution in (2.5). Moreover, the ordering of the different classes of jobs is equally likely. As was the case in Theorem 2.1, all states communicate with the empty state, and the Markov process $X(\cdot)$ for the node is positive recurrent.

Demonstration of Theorem 2.7

The proof of Theorem 2.7 that we will give is based on Proposition 2.6. In order to employ the proposition, we need to choose quantities $\hat{q}$ and π so that (2.11) and (2.18) are satisfied for them and the transition rates q of the Markov process $X(\cdot)$ for the node. It will then follow from Proposition 2.6 that π is the stationary distribution for the node and $\hat{q}$ gives the transition rates for the reversed process $\hat{X}(\cdot)$. A similar argument will be used again for symmetric nodes in Section 2.4 and for networks consisting of quasi-reversible nodes in Section 2.5. We will summarize a more probabilistic argument for Theorem 2.7 at the end of the section.

In order to demonstrate Theorem 2.7, we write q and our choices for $\hat{q}$ and π explicitly in terms of α, β, δ, and ϕ. To be able to reuse this argument in Section 2.4 for symmetric nodes, we write m_k for the mean service times, which in the present case reduces to the constant m^s.

The nonzero transition rates $q(x, y)$ take on two forms, depending on whether the state is obtained from x by the arrival or exit of a job. In the former case, we write $y = a_{k,i}(x)$ if a class k job arrives at position i; in the latter case, it follows that $x = a_{k,i}(y)$, if i is the position of the exiting class k job. One then has

$$q(x, y) = \begin{cases} \alpha_k \beta(i, n_y) & \text{for } y = a_{k,i}(x), \\ m_k^{-1} \delta(i, n_x) \phi(n_x) & \text{for } x = a_{k,i}(y), \end{cases} \tag{2.21}$$

where n_x and n_y are the number of jobs at the node for states x and y.

Finding the transition rates $\hat{q}(x, y)$ of the reversed process involves some guessing, motivated by our idea of what $\hat{X}(\cdot)$ should look like. It is reasonable

to guess that $\hat{X}(\cdot)$ is also the Markov process for a homogeneous node. The external arrival rates α_k and $\hat{\alpha}_k$ will then be the same for both processes, since arrivals for $\hat{X}(\cdot)$ correspond to exits for $X(\cdot)$, and under the stationary distribution π, the two rates must be the same. The mean service times m_k and $\hat{m}_k$ will also be the same. It is reasonable to guess that $\hat{\phi}(n) = \phi(n)$; this would be the case if $X(\cdot)$ were reversible, as it is for the $M/M/1$ queue. Running $X(\cdot)$ backwards in time mentally, it is also tempting to set

$$\hat{\beta}(i,n) = \delta(i,n),\ \hat{\delta}(i,n) = \beta(i,n) \quad \text{for } n \in \mathbf{Z}_+,\ i \le n.$$

For instance, if the original node is FIFO, then jobs arrive at $i = n$ and exit at $i = 1$; if the node is run backwards in time, jobs arrive at $i = 1$ and exit at $i = n$. Substitution of these choices for α, β, δ, ϕ, and m in (2.21) yields

$$\hat{q}(x,y) = \begin{cases} \alpha_k \delta(i, n_y) & \text{for } y = a_{k,i}(x), \\ m_k^{-1} \beta(i, n_x) \phi(n_x) & \text{for } x = a_{k,i}(y). \end{cases} \tag{2.22}$$

We still need to choose our candidate for the stationary distribution π of both $X(\cdot)$ and $\hat{X}(\cdot)$. The equality (2.11), that is needed for Proposition 2.6, is equivalent to

$$\pi(x) q(x,y) = \pi(y) \hat{q}(y,x) \tag{2.23}$$

holding whenever $y = a_{k,i}(x)$ or $x = a_{k,i}(y)$. When $y = a_{k,i}(x)$ and $q(x,y) > 0$, substitution of (2.21) and (2.22) into (2.23) implies that

$$\pi(y)/\pi(x) = m_k \alpha_k / \phi(n_y). \tag{2.24}$$

The case $x = a_{k,i}(y)$ yields the same equality, but with the roles of x and y reversed. Reasoning backwards, it is not difficult to see that (2.23) also follows from (2.24).

Set $x = (x(1), \ldots, x(n))$, for $n \in \mathbf{Z}_{+,0}$. One can repeatedly apply (2.24) by removing jobs from x one at a time, starting from the last, until the empty state is reached. We therefore choose π so that

$$\pi(x) = B^{-1} \prod_{i=1}^{n} (m_{x(i)} \alpha_{x(i)} / \phi(i)), \tag{2.25}$$

where the normalizing constant $B = 1/\pi(\emptyset)$. Under (2.25), (2.24) must hold. We have therefore verified (2.11) for this choice of π. In particular, this holds for $m_k \equiv m^s$, as in Theorem 2.7.

In order to employ Proposition 2.6, we also need to verify (2.18). One can check that

$$\sum_y q(x,y) = \sum_k \alpha_k + \phi(n_x) \sum_i m_{x(i)}^{-1} \delta(i, n_x). \tag{2.26}$$

The first sum on the right side of (2.26) follows by summing the top line of (2.21) over all i and k. The relationship $x = a_{k,i}(y)$ implies that $x(i) = k$, and

so the last sum in (2.26) follows by summing the last line of (2.21) over all i. Using the same reasoning, one obtains the formula

$$\sum_y \hat{q}(x,y) = \sum_k \alpha_k + \phi(n_x) \sum_i m_{x(i)}^{-1} \beta(i, n_x) \tag{2.27}$$

from (2.22). As before, the first sum on the right side is obtained from arriving jobs and the last sum is obtained from exiting jobs.

We see from (2.26) and (2.27) that a sufficient condition for (2.18) is that

$$\sum_i m_{x(i)}^{-1} \delta(i, n_x) = \sum_i m_{x(i)}^{-1} \beta(i, n_x). \tag{2.28}$$

In Theorem 2.7, $m_k \equiv m^s$, which factors outside of the sum on both sides of (2.28). Since the resulting sums both equal 1, (2.28), and hence (2.18), holds in this setting. Note that this is the only point in the argument at which we need m_k to be constant.

We have shown that both (2.11) and (2.18) are satisfied for q and q' given by (2.21) and (2.22), and π given by (2.25), under the assumptions in Theorem 2.7. It therefore follows from Proposition 2.6 that π is the stationary distribution for the Markov process with transition rates q. This implies Theorem 2.7.

Some observations

One can generalize the above proof of Theorem 2.7 so that it applies to homogeneous queueing networks of Kelly type, rather than to just homogeneous nodes of Kelly type as in the theorem. Then, the stationary distribution π can be written as the product of stationary distributions π^j of nodes corresponding to the individual stations, when they operate "in isolation". This is done in Section 3.1 of [Ke79]. We prefer to postpone the treatment of homogeneous networks until Section 2.5, where they are considered within the context of quasi-reversibility.

One can give a more probabilistic proof of Theorem 2.7 that is based on the following intuitive argument. The rates $\beta(i,n)$, $\delta(i,n)$ and $\phi(n)$ governing the arrival and service rates of jobs, as well as the mean service time m^s, do not distinguish between classes of jobs. Jobs are therefore served as they would be for an $M/M/1$ queue modified to have the total rate of service $\phi(n)$, when there are n jobs, and having the arrival rate $\sum_k \alpha_k$. By randomly choosing the class of each job, with probability $\alpha_k / \sum_k \alpha_k$ for each k, at either time 0 or at some later time t, the distributions at time t of the two resulting processes will be the same. If the stationary distribution for the modified $M/M/1$ queue is chosen as its initial distribution, the resulting distribution π' for the K classes will therefore also be stationary.

One can show, by using reversibility, that the probability of there being n jobs for the stationary distribution of the modified $M/M/1$ queue is $\rho^n / B \prod_{i=1}^n \phi(i)$, where B is as in (2.19). Because of the random way in which

the classes of jobs are chosen above for π', the remaining properties in the alternative characterization of π in (2.20), that are given after the statement of Theorem 2.7, also hold. Therefore, $\pi' = \pi$, as desired.

We also note the following consequence of the proof of Theorem 2.7, that is a special case of phenomena that will be discussed in Section 2.5. (It also follows from the alternative argument that was sketched above.) The transition rates $\hat{q}$ of the reversed Markov process $\hat{X}(\cdot)$ in (2.22) are the rates for a homogeneous node of Kelly type. For this reversed node, arrivals are therefore given by K independent Poisson processes for the different classes, that are independent of the initial state. These arrivals correspond to exiting jobs for the original homogeneous node. It follows that the K different exit processes for the classes are also independent Poisson processes that are independent of any future state of the node.

2.4 Symmetric Nodes

PS, LIFO, and IS nodes all belong to the family of symmetric nodes. They are defined similarly to the homogeneous nodes of Kelly type in the previous section, with a few major differences. The basic framework is the same, with state space S_0 given by (2.2) and existing jobs being reordered as before upon the arrival and departure of jobs at the node. Moreover, the interarrival times are exponentially distributed.

In order for such a node to be a *symmetric node*, we require that it also satisfy the following properties:

(a) The amount of service required by each job is exponentially distributed with mean m_k. We will soon allow more general distributions, but this will require us to extend the state space.

(b) The total rate of service supplied at the node is $\phi(n)$, where n is the number of jobs currently there.

(c) The proportion of service that is directed at the job in position i is $\beta(i, n)$. Note that the proportion does not depend on the class of the job.

(d) When a job arrives at the node, it moves into position i, $i = 1, \ldots, n$, with probability $\beta(i, n)$, where n is the number of jobs in the node including this job. This function is the same as that given in (c).

In Section 2.5, we will also be interested in symmetric queueing networks. These networks are defined analogously, with properties (a)-(d) being assumed to hold at each station, and departing jobs from a class k being routed to a class ℓ with probability $P_{k,\ell}$. More detail is given in Section 2.3 for homogeneous networks, where the procedure is the same.

The properties (a)-(d) given here are more restrictive than the properties (a)-(d) in the previous section in that we now assume, in the notation of

Section 2.3, that $\delta = \beta$, in parts (c) and (d). These properties are more general in that the service time means m_k need no longer be equal at different classes. After comparing the stationary distributions of these nodes with those of Section 2.3, we proceed to generalize the exponential distributions of the service times in (a) in two steps, first to mixtures of Erlang distributions, and then to arbitrary distributions.

The PS, LIFO, and IS nodes are standard examples of symmetric nodes. For the PS discipline, one sets

$$\beta(i,n) = 1/n \quad \text{for } i \le n,$$

for $n \in \mathbf{Z}_+$, and for LIFO, one sets

$$\beta(i,n) = \begin{cases} 1 & \text{for } i = n, \\ 0 & \text{otherwise.} \end{cases}$$

In both cases, $\phi(n) \equiv 1$. For IS nodes, $\phi(n) = n$ and

$$\beta(i,n) = 1/n \quad \text{for } i \le n.$$

The analog of Theorem 2.7 holds for symmetric nodes when $B < \infty$, for B in (2.19), with the formula for the stationary distribution π,

$$\pi(x) = B^{-1} \prod_{i=1}^{n} (m_{x(i)} \alpha_{x(i)} / \phi(i)), \tag{2.29}$$

for $x = (x(1), \ldots, x(n))$, replacing (2.20). One can check that the same argument as before is valid. One applies Proposition 2.6, for which one needs to verify (2.11) and (2.18). As before, the formulas for q and $\hat{q}$ are given by (2.21) and (2.22), but with $\delta = \beta$. The formula for π is given by (2.25).

The argument we gave for (2.11) involved no restrictions on m_k, and therefore holds in the present context as well. The argument we gave for (2.18) consisted of equating (2.26) and (2.27), and therefore verifying (2.28). Previously, (2.28) held since m_k^{-1} was constant, and so could be factored out of both sums. It is now satisfied since $\delta = \beta$, and so the summands are identical. One can therefore apply Proposition 2.6, from which the analog of Theorem 2.7 for symmetric nodes follows, with (2.29) replacing (2.20).

The method of stages

As mentioned earlier, the assumption that the service times be exponentially distributed is not necessary for symmetric nodes. By applying the *method of stages*, one can generalize the service time distributions to mixtures of Erlang distributions. (These are gamma distributions that are convolutions of identically distributed exponential distributions.) We will show that the analog of the formula (2.29) for the stationary distribution continues to hold in this more general setting.

In order for the stochastic process corresponding to the node to remain Markov for more general service times, we need to enrich the state space S_0. For this, we replace each coordinate $x(i)$ in (2.2) by the triple $\mathbf{x}(i) = (x(i), s(i), v(i))$, where

$$x(i) \in \{1, \ldots, K\}, \quad s(i) \in \mathbf{Z}_+, \quad v(i) \in \{1, \ldots, s(i)\}. \tag{2.30}$$

Such a triple gives the *refined class* of a job, with $x(i)$ denoting its class as before. The third coordinate $v(i)$ gives the current *stage* of a job, with $s(i)$ denoting the total number of stages the job visits before leaving the node. The state space S_e will consist of such n-tuples $x = (\mathbf{x}(1), \ldots, \mathbf{x}(n))$, $n \in \mathbf{Z}_{+,0}$, under a mild restriction to ensure all states are accessible.

The basic dynamics of the node are the same as before, which satisfies the properties for symmetric nodes given at the beginning of the section, including properties (b)-(d); we are generalizing here the assumption in (a). Instead of entering the node at a class k with rate α_k, jobs enter at a refined class (k, s, s) with rate $\alpha_k p_k(s)$, where $\sum_s p_k(s) = 1$. Once at a refined class (k, s, v), such a job moves to $(k, s, v-1)$ after completing its service requirement, which is exponentially distributed with mean $m_k(s)$. After a job completes its service at the stage $v = 1$, it leaves the node. The current stage v of a job can therefore be thought of as the residual number of stages remaining before the job leaves the node. We note that the proportion of service that is directed at a job in position i is $\beta(i, n)$, which does not depend on its class or refined class.

The distribution of the service time that is required for the job between entering and leaving class k is a mixture of Erlang distributions, and has mean

$$m_k \stackrel{\text{def}}{=} \sum_s s p_k(s) m_k(s). \tag{2.31}$$

The state space S_e mentioned earlier is defined to consist of n-tuples whose components (k, s, v) satisfy $p_k(s) > 0$, in order to exclude inaccessible states. Under this restriction, all states will communicate. The state space is of course countable. When $p_k(1) = 1$ at all k, the service times are all exponentially distributed and the model reduces to the one considered at the beginning of the section. As with homogeneous and symmetric nodes with exponentially distributed service times, the networks corresponding to symmetric nodes with stages can be defined in the natural way.

We wish to show that the nodes just defined have stationary distributions π that generalize (2.29). This result is stated in Theorem 2.8.

Theorem 2.8. *Suppose that the service times of a symmetric node are mixtures of Erlang distributions, and that the node satisfies $B < \infty$ in (2.19). Then, the node has a stationary distribution π that is given by*

$$\pi(x) = B^{-1} \prod_{i=1}^{n} (p_{x(i)}(s(i)) m_{x(i)}(s(i)) \alpha_{x(i)} / \phi(i)), \tag{2.32}$$

where $x = (\mathbf{x}(1), \ldots, \mathbf{x}(n)) \in S_e$ and $\mathbf{x}(i) = (x(i), s(i), v(i))$, for $i = 1, \ldots, n$.

As was the case in Theorem 2.7 and in (2.29), the structure of the stationary distribution π is Theorem 2.8 exhibits independence at multiple levels. The probability of there being a total of n jobs at the node is $\rho^n / B \prod_{i=1}^{n} \phi(i)$. Given a total of n jobs at the node, the probability of there being $n_1, \ldots, n_k$ jobs at the classes $1, \ldots, K$ is given by the multinomial

$$\rho^{-n} \binom{n}{n_1, \ldots, n_K} \prod_{k=1}^{K} (m_k \alpha_k)^{n_k}. \tag{2.33}$$

The ordering of these classes is equally likely. Note that none of these quantities depends on the particular service time distributions, except for the means m_k.

The stationary distribution also has the following refined structure. Given the class of the job at each position i, the probability of the job at a given position, whose class is k, having refined class (k, x, v) is

$$p_k(s) m_k(s) / m_k,$$

and these events are independent at different i. Summing over all stages strictly greater than v and over all s, while keeping everything else fixed, this implies that the conditional probability of the job at position i being in a strictly earlier stage than v is

$$m_k^{-1} \sum_{s} (s - v) p_k(s) m_k(s). \tag{2.34}$$

Note that (2.34) depends on the actual service time distributions, and not just on their means.

Demonstration of Theorem 2.8

In order to demonstrate Theorem 2.8, we employ Proposition 2.6. To do so, we need to verify (2.11) and (2.18) for the transition rates q of the Markov process on S_e corresponding to the node, with an appropriate choice of the quantities $\hat{q}$ and π.

In order to specify q and $\hat{q}$, we modify the function $a_{k,i}(\cdot)$ we used for exponential service times on the state space S_0. Here, $a_{k,s,i}(x)$ will denote the state y obtained from state x by the arrival of a job at position i, with refined class (k, s, s). Since jobs exit from the node at refined classes of the form $(k, s, 1)$ (rather than at (k, s, s)), we need additional notation. With an eye on defining $\hat{q}$, we denote by $\hat{a}_{k,s,i}(x)$ the state y that is obtained from x by inserting a job with refined class $(k, s, 1)$ at i; the positions of jobs already at the node are shifted in the usual way. We also denote by $\mathfrak{s}_i(x)$ the state y obtained from a state x satisfying $2 \leq v(i) \leq s(i)$, when the stage at i advances to $v(i) - 1$. ($\mathfrak{s}_i(x)$ is not defined for other x.)

Using this notation, one can check that q is given by

$$q(x,y) = \begin{cases} \alpha_k p_k(s)\beta(i,n_y) & \text{for } y = a_{k,s,i}(x), \\ (m_k(s))^{-1}\beta(i,n_x)\phi(n_x) & \text{for } x = \hat{a}_{k,s,i}(y), \\ (m_k(s))^{-1}\beta(i,n_x)\phi(n_x) & \text{for } y = \mathfrak{s}_i(x), \end{cases} \tag{2.35}$$

with $q(x,y) = 0$ otherwise. Employing the same motivation as in Section 2.3, we choose $\hat{q}$ so that

$$\hat{q}(x,y) = \begin{cases} \alpha_k p_k(s)\beta(i,n_y) & \text{for } y = \hat{a}_{k,s,i}(x), \\ (m_k(s))^{-1}\beta(i,n_x)\phi(n_x) & \text{for } x = a_{k,s,i}(y), \\ (m_k(s))^{-1}\beta(i,n_y)\phi(n_y) & \text{for } x = \mathfrak{s}_i(y), \end{cases} \tag{2.36}$$

with $\hat{q}(x,y) = 0$ otherwise. The transition function $\hat{q}$ is the same as q, except that jobs arrive at the stage $v = 1$, exit at $v = s(i)$, with changes in stage occurring from $v-1$ to v, for $2 \le v \le s(i)$. We choose π as in (2.32).

The assumptions for Proposition 2.6 can be verified as they were in Section 2.3 for the space S_0, with only a small change in argument. The argument for (2.11) is the same when either $y = a_{k,s,i}(x)$ or $x = \hat{a}_{k,s,i}(y)$. For $y = a_{k,s,i}(x)$, the equality (2.24) is replaced by its analog

$$\pi(y)/\pi(x) = m_k(s)p_k(s)\alpha_k/\phi(n_y).$$

When $y \in \mathfrak{s}_i(x)$, one has

$$\pi(y)/\pi(x) = q(x,y)/\hat{q}(y,x) = 1, \tag{2.37}$$

in which case (2.11) is obvious. So, (2.11) holds in all cases. (Note that for the homogeneous nodes in Section 2.3 with distinct β and δ, the analog of (2.37) does not hold, and so the method of stages employed here for generalizing the exponential distributions will not work.)

The formula (2.18) holds for the same reasons as before, except that one now has

$$\sum_y q(x,y) = \sum_y \hat{q}(x,y) = \sum_k \alpha_k + \phi(n_x)\sum_i (m_{x(i)}(s(i)))^{-1}\beta(i,n_x),$$

with $m_{x(i)}(s(i))$ replacing $m_{x(i)}$. So, the assumptions for Proposition 2.6 hold. Application of the proposition therefore implies Theorem 2.8.

One can generalize the above argument so that it applies to symmetric queueing networks. Then, the stationary distribution π can be written as the product of stationary distributions π^j of nodes corresponding to the individual stations. As in the previous section, we choose to postpone the treatment of symmetric networks until Section 2.5, where they are considered within the context of quasi-reversibility.

Extensions to general distributions

We have employed the method of stages to generalize the formula (2.29), for the stationary distribution of symmetric nodes with exponentially distributed service times, to the formula (2.32), which holds for service times that are mixtures of Erlang distributions. The method of stages can also be employed to construct service times with other distributions. This approach is employed, for example, in Section 3.6 of [Wa88] and in Section 3.4 of [As03], where the more general *phase-type distributions* are constructed. [As03] also gives further background on the problem.

Let $\mathcal{H} = \bigcup_{N=1}^{\infty} \mathcal{H}_N$, where $\mathcal{H}_N$ denotes the family of mixtures of Erlang distributions, but with the restriction that $m_k(s) = 1/N$ for all k and s. It is not difficult to show that $\mathcal{H}$ is dense in the set of distribution functions, with respect to the weak topology; this result is given in Exercise 3.3.3 in [Ke79]. The basic idea is that the sum of Ns i.i.d. copies of an exponential distribution, with mean $1/N$, has mean s and variance s/N, and so, for large N, is concentrated around s. For large enough N, one can therefore approximate a given service time distribution function F_k as closely as desired, by setting

$$p_k^N(s) = F_k^N(s/N) - F_k^N((s-1)/N), \quad \text{for } s \in \mathbf{Z}_+, \tag{2.38}$$

equal to the probability that a job chooses a refined class with s stages, when it enters class k. Here, F_k^N is the distribution function satisfying $F_k^N(s') = F_k(s')$, for $Ns' \in \mathbf{Z}_+$, and which is constant off this lattice. For the same reason, the phase-type distributions that were mentioned in the previous paragraph are also dense.

Since the family $\mathcal{H}$ of mixtures of Erlang distributions is dense, it is tempting to infer that a stationary distribution will always exist for a symmetric node with any choice of service time distributions F_k satisfying (2.19), with m_k replacing m^s in the definition of ρ, and that this distribution has the same product structure as given below the statement of Theorem 2.8. Such a result holds, although the state space needs to be extended so that the last component v of the refined class (k, s, v) of a job can now take on any value in $(0, s]$; this corresponds to the residual service time of that job. The resulting state space S_∞ for the Markov process is uncountable; this causes technical problems which we discuss at the end of the section. Since the space is uncountable, it is most natural to formulate the result in a manner similar to that given below Theorem 2.8.

Theorem 2.9. *Suppose that a symmetric node satisfies $B < \infty$ in (2.19). Then, it has a stationary distribution π on S_∞ with*

$$\pi(x(i) = k(i),\ i = 1, \ldots, n) = B^{-1} \prod_{i=1}^{n} (m_{k(i)} \alpha_{k(i)} / \phi(i)). \tag{2.39}$$

Conditioned on any such set, the residual service times of the different jobs are independent, with the probability that a job of a class k has residual service time at most r being

$$F_k^*(r) = \frac{1}{m_k} \int_0^r (1 - F_k(s))ds. \tag{2.40}$$

One can motivate (2.39) by applying Theorem 2.8 to a sequence of nodes indexed by N, with p_k^N for each class k being given by (2.38). Since $F_k^N \Rightarrow F_k$ and $m_k^N \to m_k$ as $N \to \infty$, one should expect (2.39) to follow from (2.32). In order to motivate (2.40), one can reason as follows. Applying Theorem 2.8, one can check that, under the stationary distribution π^N and conditioned on the job at a given position being k, the probability that the stage there is strictly greater than v is

$$\sum_{s>v} (s - v) p_k^N(s) \Big/ \sum_s s p_k^N(s).$$

One can also check that $\sum_{s>v}(s-v)p_k^N(s) = \sum_{s>v} \bar{F}_k^N(s/N)$, where $\bar{F}_k^N(s) = 1 - F_k^N(s)$. So, the above quantity equals

$$\frac{1}{N} \sum_{s>v} \bar{F}_k^N(s/N) \Big/ \frac{1}{N} \sum_s \bar{F}_k^N(s/N).$$

By setting $v = Nr,\ r \geq 0$, and applying the Monotone Convergence Theorem to the numerator and denominator separately, one obtains the limit

$$\frac{1}{m_k} \int_r^\infty \bar{F}_k(s)ds. \tag{2.41}$$

On the other hand, the same reasoning as above (2.38) implies that, for large N, the stage v scaled by N typically approximates the residual service time. So, (2.41) will also give the limiting distribution of the residual service times as $N \to \infty$. Taking the complementary event, one obtains (2.40) from (2.41).

The same reasoning as above implies (2.40) is also the probability that the amount of service that has been received by a job is at least r. We point out that F_k^*, as in (2.40), is the distribution of the residual time for the stationary distribution of a renewal process, with lifetime distribution F_k.

The above reasoning, although suggestive, is not rigorous. In particular, implicit in the explanations for both (2.39) and (2.40) is the assumption that the stationary distribution π and the residual service time distributions $F_1^*, \ldots, F_K^*$ are continuous in $F_1, \ldots, F_K$. A rigorous justification for Theorem 2.9 is given in [Ba76]. There, the Markov processes $X^N(\cdot)$ corresponding to the above sequences of nodes are constructed on a common uncountable state space S, where the residual times of the jobs are included in the state. The Markov process that corresponds to the node in Theorem 2.9 is expressed as a weak limit of the processes $X^N(\cdot)$; this provides a rigorous justification for the convergence of π^N and $F_1^N, \ldots, F_K^N$ that is needed for the theorem. [Ba76] in fact demonstrates the analog of Theorem 2.9 in the more general context of symmetric queueing networks.

The above uncountable state space setting requires a more abstract framework than one typically wishes for a basic theory of symmetric networks. The countable state space setting is typically employed in the context of either mixtures of Erlang distributions, the more general phase-type distributions, or some other dense family of distributions. (See, for example, Section 3.4 of [As03], for more detail.) Quasi-reversibility, which we discuss in the next section, also employs a countable state space setting.

2.5 Quasi-Reversibility

In Sections 2.3 and 2.4, we showed that the stationary distributions of homogeneous nodes of Kelly type and symmetric nodes are of product form. Employing quasi-reversibility, it will follow that the stationary distributions of the corresponding queueing networks are also of product form, with the states at the individual stations being independent and the distributions there being given by Theorems 2.7 and 2.8.

Quasi-reversibility has two important consequences. When a queueing network can be decomposed in terms of nodes that are quasi-reversible, the stationary distribution of the network can be written as the product of the stationary distributions of these individual nodes. It will also follow from the "duality" present in quasi-reversibility that the exit processes of such networks are independent Poisson processes, a property that is inherited from the processes of external arrivals of the network. Quasi-reversibility does not depend on the routing in a network, but holds only under certain disciplines, like those mentioned in the first paragraph.

In this section, in order to avoid confusion, we will say that a departing job from a class that leaves the network *exits* from the network (as opposed to being routed to another class). For nodes, such as in the two previous sections, departures and exits are equivalent.

Before introducing quasi-reversibility, we first motivate the basic ideas with a finite sequence of $M/M/1$ queues that are placed in tandem:

$$\rightarrow 1 \rightarrow 2 \rightarrow \ldots \rightarrow k \rightarrow \ldots \rightarrow K \rightarrow . \tag{2.42}$$

Jobs are assumed to enter the one-class station, with $j = k = 1$, according to a rate-α Poisson process, and are served in the order of their arrival there. Upon leaving station 1, jobs enter station 2 and are served there, and so on, until leaving the network after having been served at station K. All jobs are assumed to have exponentially distributed service times which are independent, with means m_k so that $\alpha m_k < 1$.

An $M/M/1$ queue with external arrival rate α and mean service time m has stationary distribution given by (2.1), if $m\alpha < 1$. Under this distribution, the corresponding Markov process $X(t)$, $-\infty < t < \infty$, is reversible, and so is stochastically equivalent to its reversed process $\hat{X}(t) = X(-t)$. In particular,

the stationary process $X_1(\cdot)$ for the number of jobs at station 1 is stochastically equivalent to its reversed process $\hat{X}_1(\cdot)$. Interpreting $\hat{X}_1(\cdot)$ in terms of the arrival and departure of jobs, with the former corresponding to an increase and the latter a decrease of $\hat{X}_1(\cdot)$, jobs arrive according to a rate-α Poisson process. But, each arrival of a job for $\hat{X}_1(\cdot)$, at time t, corresponds to the departure of a job for $X_1(\cdot)$, at time $-t$. It follows that the departure process of jobs for $X_1(\cdot)$ is a Poisson rate-α process. Moreover, departures preceding any given time t_1 are independent of $X(t_1)$. What we are observing here, is that the specific nature of the Poisson *input* into station 1 results in an *output* of the same form.

Let $X_2(\cdot)$ denote the process for the number of jobs at station 2, and assume that $X_2(t_0)$ has the stationary distribution (2.1), with $m = m_2$, for a given t_0 and is independent of $X_1(t)$ for $t \geq t_0$. The arrival process of $X_2(\cdot)$ is also the departure process of $X_1(\cdot)$, which is a rate-α Poisson process, by the previous paragraph. It follows that $X_2(\cdot)$ is also the stationary process of an $M/M/1$ queue. Its arrivals, up to a given time t_1, with $t_1 \geq t_0$, are independent of $X_1(t_1)$ by the previous paragraph. Consequently, $X_1(t_1)$ and $X_2(t_1)$ are independent, each with the distribution in (2.1), with m_1 and m_2 replacing the mean m. Since t_1 was arbitrary, the joint process $(X_1(\cdot), X_2(\cdot))$ is Markov and stationary, for all $t \geq t_0$. Since t_0 was arbitrary, $(X_1(\cdot), X_2(\cdot))$ in fact defines a stationary Markov process over all t.

Continuing in this manner, one obtains a stationary Markov process $X(t) = (X_1(t), \ldots, X_K(t))$, $-\infty < t < \infty$, whose joint distribution at any time is a product of distributions of the form in (2.1). One therefore obtains the following result.

Theorem 2.10. *Assume that the interarrival time and the service times of the sequence of stations depicted in (2.42) are exponentially distributed, with $\alpha m_k < 1$ for each $k = 1, \ldots, K$. Then, the network has a stationary distribution π, which is given by*

$$\pi(n_1, \ldots, n_K) = \prod_{k=1}^{K} (1 - \alpha m_k)(\alpha m_k)^{n_k}, \tag{2.43}$$

for $n_k \in \mathbf{Z}_{+,0}$.

We note that although the components of the stationary distribution given by (2.43) are independent, this is not at all the case for the components $X_k(\cdot)$ of the corresponding stationary Markov process $X(\cdot)$. In particular, a departure at station k coincides with an arrival at station $k + 1$.

Results leading up to Theorem 2.10 and the above proof are given in [Ja54], [Bu56], and [Re57]. More detail on the background of the problem is given on page 212 of [Ke79].

The same formula as in (2.43) holds when the routing in (2.42) is replaced by general routing, if the total arrival rates λ_k are substituted for α. More precisely, suppose that a subcritical Jackson network (i.e., a single

class network with exponentially distributed interarrival and service times) has external arrival rates $\alpha = \{\alpha_k,\ k = 1, \ldots, K\}$ and mean routing matrix $P = \{P_{k,\ell},\ k, \ell = 1, \ldots, K\}$. Then, it has the stationary distribution π, with

$$\pi(n_1, \ldots, n_K) = \prod_{k=1}^{K} (1 - \lambda_k m_k)(\lambda_k m_k)^{n_k}, \tag{2.44}$$

for $n_k \in \mathbf{Z}_{+,0}$. This result is no longer as easy to see as is (2.43); it was shown in the important work [Ja63]. The result will follow as a special case of Theorems 2.3 and 2.12.

Basics of quasi-reversibility

The "input equals output" behavior of the network in (2.42) was central to our ability to write the stationary distribution of the network as a product of the stationary distributions at its individual stations. *Quasi-reversibility* generalizes this concept, and leads to similar results for more general families of networks. Quasi-reversibility was first identified in [Mu72] and has been extensively employed in work by F.P. Kelly. The property can be defined in different equivalent ways; we use the following analytic formulation.

We consider a node for which arrivals at its classes $k = 1, \ldots, K$ are given by independent Poisson processes, with intensities $\alpha = \{\alpha_k,\ k = 1, \ldots, K\}$, that do not depend on the state of the node at earlier times, and for which the exits occur only one at a time and do not coincide with an arrival. The evolution of the node is assumed to be given by a Markov process $X(\cdot)$ with stationary distribution π defined on a countable state space S. Assume that all states communicate. Also, let q denote the transition function of $X(\cdot)$ and $\hat{q}$ the transition function of the reversed process $\hat{X}(\cdot)$ satisfying (2.11).

Under the above assumptions, any change in the state of the node due to a transition from x to y must be due to an increase by 1 in the number of jobs at some class k, for which we write $y \in A_k(x)$; a decrease by 1 at some k, for which we write $y \in E_k(x)$; or a transition that involves neither an increase nor a decrease, for which we write $y \in I(x)$, and which we refer to as an *internal transition*. Note that $y \in I(x)$ and $x \in I(y)$ are equivalent. An example of an internal transition is the "advance in stage" $y = \mathfrak{s}_i(x)$ in Section 2.4, although in the current setting far more general changes of state are allowed, including the simultaneous swapping of positions by many jobs. General changes of state are also allowed with the arrival or exit of a job.

We will say the node is *quasi-reversible* if for each class k and state x,

$$\sum_{y \in A_k(x)} \hat{q}(x, y) = \beta_k \tag{2.45}$$

for some $\beta_k \geq 0$. The equality (2.45) says that the rate of arrivals at each class k for the reversed process $\hat{X}(\cdot)$ does not depend on the state x. It is equivalent to the apparently stronger

$$\sum_{y \in A_k(x)} \hat{q}(x,y) = \sum_{y \in A_k(x)} q(x,y) = \alpha_k \tag{2.46}$$

for each k and x, which states that $\beta_k = \alpha_k$. (Note that the last equality follows automatically from the definition of α_k.)

To see (2.46), we note that by (2.45), the arrival times of $\hat{X}(\cdot)$ form independent rate-β_k Poisson processes at the K classes. The same reasoning that was applied to the sequence of $M/M/1$ queues in (2.42) implies that the exit times of $X(\cdot)$ also form independent rate-β_k Poisson processes. By assumption, only one exit occurs at each such time. Moreover, under the stationary distribution π, the rates at which jobs enter and leave a class are the same. Since the former is α_k, this implies $\beta_k = \alpha_k$, as needed for (2.46).

In the preceding argument, we have shown that the exit processes for $X(\cdot)$ form independent rate-α_k Poisson processes. Comparison with $\hat{X}(\cdot)$ also shows that exits for $X(\cdot)$ preceding any given time t_1 are independent of $X(t_1)$. These are important properties of quasi-reversible nodes. We have already employed them in the proof of Theorem 2.10.

The term quasi-reversible can also be applied to a queueing network rather than just to a node, with equation (2.45) again being employed as the defining property. (One should interpret $A_k(x)$ in terms of external arrivals at k.) In this setting, the stronger (2.46) need not hold, since departures from a class, that are not exits, may occur because of a job moving to another class within the network, and the reasoning in the paragraph below (2.46) is not valid. Nevertheless, external arrivals for the reversed process $\hat{X}(\cdot)$ correspond to jobs exiting the network for $X(\cdot)$. The same reasoning that was employed for quasi-reversible nodes therefore implies that the exiting processes at the classes k are independent rate-β_k Poisson processes.

Although we will not use this here, we also note that the *partial balance equations*

$$\pi(x) \sum_{y \in A_k(x)} q(x,y) = \sum_{y \in A_k(x)} \pi(y) q(y,x), \tag{2.47}$$

for each k and x, are equivalent to (2.46), and hence to the quasi-reversibility of a node. This follows immediately from the definition of $\hat{q}$ in (2.11) and the assumption $\pi(x) \neq 0$ for all x. These equations are weaker than the detailed balance equations, which correspond to reversibility, but include information not in the balance equations. The partial balance equations are often used as an alternative to quasi-reversibility.

Construction of networks from quasi-reversible nodes and applications

The main result on quasi-reversibility is Theorem 2.11, which states that when a queueing network satisfies certain conditions involving quasi-reversible nodes, its stationary distribution can be written as the product of the stationary distributions of these nodes. These nodes typically correspond to the stations of the network in a natural way. Such a queueing network is itself

quasi-reversible. Examples of these queueing networks are the sequence of $M/M/1$ queues in (2.42), Jackson networks, the homogeneous networks of Kelly type that were defined in Section 2.3, and the symmetric networks that were defined in Section 2.4.

We will consider queueing networks in the following framework. The network will consist of J stations and K classes on a countable state space S of the form

$$S = S^1 \times \ldots \times S^J,$$

where S^j is the state space corresponding to the j^{th} station. We will typically write $x = (x^1, \ldots, x^J)$ for $x \in S$, where $x^j \in S^j$. For concreteness, we will assume that for each j, S^j is one of the two spaces S_0 and S_e that were employed in the last two sections, although the theory holds more generally. As usual, the queueing network is assumed to have transition matrix $P = \{P_{k,\ell},\ k, \ell = 1, \ldots, K\}$ and external arrival rates $\alpha = \{\alpha_k,\ k = 1, \ldots, K\}$. Recall that $\lambda = Q\alpha$ denotes the total arrival rate, and satisfies the traffic equations given in (1.6).

We will employ notation similar to what was used earlier in the section, with $A_k(x), E_k(x)$, $I_j(x)$, and $R_{k,\ell}(x)$ denoting the states y obtained from x by the different types of transitions. As before, $A_k(x)$, $E_k(x)$, and $I_j(x)$ will denote the states obtained by an arrival into the network at k, an exit from the network at k, and an internal state change at j. For $y \in A_k(x)$ or $y \in E_k(x)$, we will require that $y^j = x^j$ for $j \neq s(k)$, and that the number of jobs at k increase or decrease by 1, and elsewhere remain the same. For $y \in I_j(x)$, we will require that $y^j = x^j$ for $j \neq s(k)$, and that the number of jobs at each class remain the same. We let $R_{k,\ell}(x)$ denote the set of y obtained from x by a job returning to class ℓ after being served at class k. We require that $y^j = x^j$ for $j \neq s(k)$ and $j \neq s(\ell)$, and that the number of jobs at k decrease by 1, at ℓ increase by 1, and elsewhere remain the same.

The queueing networks we will consider will be assumed to satisfy properties (2.48)-(2.52), which are given in terms of prechosen quasi-reversible nodes. Before listing these properties, we provide some motivation, recalling the sequence of 1-class stations given in (2.42), with the stationary distribution in (2.43). When a given station j, with $j = k$, is viewed "in isolation", it evolves as an $M/M/1$ queue with mean service time m_k and external arrival rate α, and has as its stationary distribution the stationary distribution of the corresponding $M/M/1$ queue. The stationary distribution of the sequence of stations is given by the product of the stationary distributions of the individual queues. Because of the specific structure of the network, Poisson arrivals into a given station result in Poisson departures, which then serve as Poisson arrivals for the next station. This property allowed us to view the stations "in isolation".

We will show that queueing networks satisfying properties (2.48)-(2.52) will have stationary distributions that are the product of the stationary distributions of the quasi-reversible nodes given there. Each such node can be

interpreted as the corresponding station evolving "in isolation". Here, "in isolation" will also mean that routing between classes at the same station is not permitted. The network in the previous paragraph, consisting of a sequence of 1-class stations, will be a special case of this more general setup.

Because of the more abstract setting now being considered, we will not explicitly follow the evolution of individual jobs; instead, we will think of the quasi-reversible nodes as "black boxes", which have a given output for a given input, with a corresponding stationary distribution. Rather than employ the Poisson-in, Poisson-out property directly, we will use the definition of quasi-reversibility in (2.45). The external arrival rates α_k^j for the classes at a given node j will be given by the total arrival rate λ_k for the corresponding class in the network. This will be consistent with jobs always leaving the node after being served, without being routed to another class.

The nodes we employ are assumed to have state spaces $S^j,\ j = 1, \ldots, J$, which are the components of the state space S for the network. Therefore, for $x = (x^1, \ldots, x^J) \in S$ with $x^j \in S^j$, $j = 1, \ldots, J$, one can also interpret x^j as the state of the corresponding node. Jobs at a given node will have classes $k \in \mathcal{C}(j)$, which are in one-to-one correspondence with the classes of the correspondingly labelled station in the network. For $x^j \in S^j$ and $k \in \mathcal{C}(j)$, we employ notation introduced earlier in the section for quasi-reversible nodes, with $A_k^j(x^j),\ E_k^j(x^j)$, and $I^j(x^j)$ denoting those states y^j obtained from x^j by an arrival or exit at class k, or by an internal state change. We let $q^j,\ j = 1, \ldots, J$, denote the transition rates for the Markov processes $X^j(\cdot)$ of the nodes. The nodes are assumed to be quasi-reversible, with (2.45) being satisfied by $\hat{q}^j$, the transition rates of the reversed processes $\hat{X}^j(\cdot)$. As mentioned earlier, the external arrival rates of the nodes are given by $\alpha_k^j = \lambda_k$. As earlier in the section, we will assume that all states of a given node communicate.

We will assume that the transition rates $q(x, y)$ for the queueing network can be written in terms of the rates $q^j(x^j, y^j)$ for the nodes as follows. For $y \in A_k(x)$, we assume that

$$q(x, y) = (\alpha_k/\lambda_k) q^j(x^j, y^j). \tag{2.48}$$

Setting $p_k^j(x^j, y^j) = q^j(x^j, y^j)/\lambda_k$, this can be written as

$$q(x, y) = \alpha_k p_k^j(x^j, y^j), \tag{2.48$'$}$$

where $\sum_{y^j \in A_k^j(x^j)} p_k^j(x^j, y^j) = 1$ holds. (Here and later on, when j and k appear together, we implicitly assume that $k \in \mathcal{C}(j)$.) For $y \in E_k(x)$, we assume that

$$q(x, y) = q^j(x^j, y^j) P_{k,0}, \tag{2.49}$$

where $P_{k,0} \stackrel{\text{def}}{=} 1 - \sum_\ell P_{k,\ell}$. For $y \in R_{k,\ell}(x)$, we require the existence of an "intermediate" state z between x and y, with $z \in E_k(x)$ and $y \in A_\ell(z)$, and such that

$$q(x,y) = q^j(x^j, z^j)q^h(z^h, y^h)(P_{k,\ell}/\lambda_\ell), \tag{2.50}$$

where $h = s(\ell)$. One can also write this as

$$q(x,y) = q^j(x^j, z^j)P_{k,\ell}\, p^h_\ell(z^h, y^h). \tag{2.50'}$$

For $y \in I_j(x)$, we assume that

$$q(x,y) = q^j(x^j, y^j), \tag{2.51}$$

and finally, on the complement of the above sets, we assume that

$$q(x,y) = 0. \tag{2.52}$$

The equations (2.48)–(2.52) have the following interpretation in terms of the transition rates of the queueing network. The J different stations operate independently of one another, except for the movement of jobs between them. So, the transition rates in (2.48′), (2.49), and (2.51) depend on x^j and y^j instead of on the entire states x and y. In (2.50′), after a class k job is served, it moves to class ℓ with probability $P_{k,\ell}$, with the probability of the new state y depending on just z^h and y^h. When $h \neq j$, z is automatically given by $z^j = y^j$, $z^h = x^h$, and $z^{j'} = x^{j'} = y^{j'}$ for other values j'. The transition rates q^j in each display are those of node j, which does not permit returns. This node can be thought of as the one obtained from the corresponding station by replacing transitions to and from each class k by external arrivals and exits at the same rates.

We now employ the above terminology to state Theorem 2.11. When its hypotheses are satisfied, the theorem enables us to write the stationary distribution of a queueing network as the product of the stationary distributions of the corresponding nodes.

Theorem 2.11. *Suppose that the transition rates $q(x,y)$ of a queueing network satisfy (2.48)-(2.52), where the nodes with the transition rates $q^j(x^j, y^j)$ are quasi-reversible with stationary distributions π^j. Then, the queueing network has stationary distribution π given by*

$$\pi(x) = \prod_{j=1}^{J} \pi^j(x^j), \tag{2.53}$$

where $x = (x^1, \ldots, x^J)$. Moreover, the queueing network is itself quasi-reversible.

The proof of Theorem 2.11 will be given in the next subsection. We first note the following consequences of Theorem 2.11 and quasi-reversibility.

As an elementary illustration of Theorem 2.11, we return to the "sequence of $M/M/1$ queues" in (2.42). Equations (2.48)-(2.52) all hold in this setting, if q^j are the transition rates for the $M/M/1$ queues with $\alpha^j = \alpha$ and $m^j = m_j$.

All of these equations are easy to see and are nondegenerate in only a few cases. In (2.48), $q(x,y) \neq 0$ only for $k = j = 1$ and, in (2.49), $q(x,y) \neq 0$ only for $k = K = J$. In (2.50′), with $1 \leq k < K$, one has $P_{k,k+1} = p_{k+1}^{k+1}(z^{k+1}, y^{k+1}) = 1$ if z is chosen by removing a class k job from x; (2.51) is vacuous in this setting. Moreover, since the $M/M/1$ queues are reversible, they are quasi-reversible. These queues are assumed to be subcritical and have stationary distributions given by (2.1). Theorem 2.10 therefore follows as a special case of Theorem 2.11.

Equations (2.48)-(2.52) also hold for the more general Jackson networks (which are, in turn, special cases of FIFO networks of Kelly type). Again, q^j are the transition rates for $M/M/1$ queues, this time with $\alpha^j = \lambda_j$ and $m^j = m_j$. The equations are similar to those for the previous example, except that the mean transition matrix P is general, and so (2.48)-(2.50′) may be nonzero for arbitrary k. The formula for the stationary distribution in (2.44) is consequently an easy application of Theorem 2.11.

We now generalize Theorem 2.3 of Section 2.1. For homogeneous networks of Kelly type and symmetric networks, equations (2.48)-(2.52) all hold if the corresponding nodes are chosen in the natural way. Namely, each such node, for $j = 1, \ldots, J$, is obtained from the corresponding station by replacing transitions involving routing from one class to another by exits from the network, and by increasing the rate of external arrivals at each class k from α_k to λ_k to compensate for this. Then, (2.48) and (2.49) are immediate. In (2.50′), the state z is chosen by removing the served job at k from x. The equality (2.50′) then follows since a transition from x to y, with $y \in R_{k,\ell}(x)$, consists of a service completion at k, followed by the routing of the corresponding job to class l of a station h, with the job then being assigned a position i according to the rule p_ℓ^h. When $S^j = S_0$, the transition $q(x,y)$ in (2.51) does not occur; when, for a symmetric node, $S^j = S_e$, the transition corresponds to the advance of a stage. In either case, (2.51) is clear. Moreover, on account of (2.22) and (2.36) in Sections 2.3 and 2.4,

$$\sum_{y^j \in A_k^j(x^j)} \hat{q}^j(x^j, y^j) = \alpha_k \quad \text{for all } x^j \in S^j, \tag{2.54}$$

and so each such node is quasi-reversible. Note that this characterization continues to hold for networks that are of mixed type, with some stations being homogeneous of Kelly type and others being symmetric.

Recall that in Theorems 2.7 and 2.8, we saw that such nodes themselves have stationary distributions that are of product form, as given in (2.20) and (2.32). Theorem 2.7 was stated, for homogeneous nodes, in the context of service times that are exponentially distributed, and Theorem 2.8 was stated, for symmetric nodes, in the context of mixtures of Erlang distributions. Combining these results with Theorem 2.11, we therefore obtain Theorem 2.12. As in Theorems 2.7 and 2.8, when the node is homogeneous, $x^j \in S_0$ is assumed, whereas when the node is symmetric, $x^j \in S_e$. In either case, we employ the

condition

$$B_j \stackrel{\text{def}}{=} \sum_{n=0}^{\infty} \left(\rho_j^n \Big/ \prod_{i=1}^{n} \phi(i) \right) < \infty, \tag{2.55}$$

with $\rho_j = \sum_{k \in \mathcal{C}(j)} m_k \lambda_k$.

Theorem 2.12. *Suppose that each station j of a queueing network is either homogeneous of Kelly type or is symmetric, and satisfies (2.55). Then, the queueing network has a stationary distribution π that is given by*

$$\pi(x) = \prod_{j=1}^{J} \pi^j(x^j), \tag{2.56}$$

where each π^j is either of the form (2.20) or (2.32), depending on whether the station j is homogeneous of Kelly type or is symmetric, and α_k in these formulas is replaced by λ_k.

Theorem 2.11 shows that the stationary distribution of a queueing network that is composed of stations corresponding to quasi-reversible nodes has the product structure given in (2.53). Nevertheless, the processes that are associated with the stations are not independent. One can see this easily for the example at the beginning of the section consisting of a sequence of $M/M/1$ queues: a departure from one queue coincides with an arrival to the next. Similarly, the combined arrival processes at different classes (i.e., arrivals from other classes as well as external arrivals) are typically not independent, nor are the departure processes.

These processes are not to be confused with the processes of external arrivals or the processes of jobs exiting from the network, which are in either case independent. We note, though, that under the stationary distribution for a queueing network composed of quasi-reversible nodes, the conditional distribution at a station found by an arriving job is the same as the stationary distribution there. This is clearly the case for external arrivals, but is also true for arrivals in general. This is shown on page 70 of [Ke79] by considering the reversed Markov process for the network.

Demonstration of Theorem 2.11

In order to demonstrate Theorem 2.11, we will employ Proposition 2.6. We therefore need a candidate $\hat{q}$ for the transition rates of the reversed process $\hat{X}(\cdot)$ for the network under its stationary distribution. Letting $\hat{q}^j$ denote the reversed transition rates corresponding to q^j, we define $\hat{q}$ using the following analogs of (2.48)-(2.52). We set

$$\hat{q}(x, y) = (\hat{\alpha}_k / \lambda_k) \hat{q}^j(x^j, y^j) = P_{k,0} \hat{q}^j(x^j, y^j) \quad \text{for } y \in A_k(x), \tag{2.57}$$

$$\hat{q}(x, y) = \hat{q}^j(x^j, y^j) \hat{P}_{k,0} = \hat{q}^j(x^j, y^j)(\alpha_k / \lambda_k) \quad \text{for } y \in E_k(x). \tag{2.58}$$

For $y \in R_{k,\ell}(x)$, we set

$$\hat{q}(x,y) = \hat{q}^j(x^j,z^j)\hat{P}_{k,\ell}\frac{\hat{q}^h(z^h,y^h)}{\sum_{w^h\in A^h_\ell(z^h)}\hat{q}^h(z^h,w^h)} \tag{2.59}$$
$$= \hat{q}^j(x^j,z^j)\hat{q}^h(z^h,y^h)(P_{\ell,k}/\lambda_k).$$

We also set

$$\hat{q}(x,y) = \hat{q}^j(x^j,y^j) \quad \text{for } x \in I_j(y), \tag{2.60}$$

$$\hat{q}(x,y) = 0 \quad \text{for other values of } y. \tag{2.61}$$

Here, we are setting

$$\hat{\alpha}_k = \lambda_k P_{k,0}, \quad \hat{P}_{k,\ell} = \lambda_\ell P_{\ell,k}/\lambda_k, \quad \hat{P}_{k,0} = \alpha_k/\lambda_k. \tag{2.62}$$

The term $\lambda_k P_{k,0}$ is the rate at which jobs exit from the original network at class k, and so should be the rate they enter the reversed network at k. The second equality is obtained by reversing the direction of the mean transition matrix P; the third equality is obtained by setting $\hat{P}_{k,0} = 1 - \sum_\ell \hat{P}_{k,\ell}$, and applying the previous equality together with (1.6). We have implicitly set $\hat{\lambda}_k = \lambda_k$ in (2.57). The second equality in (2.59) needs to be justified; it follows from (2.62) together with

$$\sum_{w^h\in A^h_\ell(z^h)} \hat{q}^h(z^h,w^h) = \lambda_\ell. \tag{2.63}$$

Since each node is assumed to be quasi-reversible, (2.63) follows from (2.46) and $\alpha^h_\ell = \lambda_\ell$. In the proof of Theorem 2.11, we will also employ

$$\sum_k \hat{\alpha}_k = \sum_k \alpha_k, \tag{2.64}$$

which follows from the definition of $\hat{\alpha}_k$ and (1.6).

Proof of Theorem 2.11. The quasi-reversibility of the queueing network follows immediately from the first equality in (2.57) and from (2.63), since

$$\sum_{y\in A_k(x)} \hat{q}(x,y) = (\hat{\alpha}_k/\lambda_k) \sum_{y^j\in A^j_k(x^j)} \hat{q}^j(x^j,y^j) = (\hat{\alpha}_k/\lambda_k)\lambda_k = \hat{\alpha}_k,$$

which does not depend on x.

The remainder of the proof is devoted to showing that the distribution π in (2.53) is stationary. We wish to show that π satisfies

$$\pi(x)q(x,y) = \pi(y)\hat{q}(y,x) \quad \text{for all } x,y \in S \tag{2.65}$$

and

$$q(x) = \hat{q}(x) \quad \text{for all } x \in S, \tag{2.66}$$

where $\hat{q}$ is defined in (2.57)-(2.61). These are restatements of (2.11) and (2.18), and together with Proposition 2.6 imply that π is stationary. Since all states are assumed to communicate, this is the unique such distribution.

Demonstration of (2.65). In order to verify (2.65), one needs to check the different cases given by the formulas for q in (2.48)-(2.52). Each is straightforward, with the most involved case being $y \in R_{k,\ell}$. To check (2.65) for $y \in R_{k,\ell}(x)$, note that by (2.50) and the second part of (2.59), (2.65) reduces to

$$\pi^j(x^j)\pi^h(x^h)q^j(x^j,z^j)q^h(z^h,y^h) = \pi^j(y^j)\pi^h(y^h)\hat{q}^j(y^j,z^j)\hat{q}^h(z^h,x^h) \quad (2.67)$$

after cancelling the common terms $\pi^{j'}(x^{j'})$, with $j' \neq j$ and $j' \neq h$, and $P_{k,\ell}/\lambda_\ell$. This equality follows immediately from the definition of $\hat{q}^j$ and $\hat{q}^h$ in (2.11).

For the cases where $y \in A_k(x)$ and $y \in E_k(x)$, (2.65) reduces to analogs of (2.67), which are somewhat simpler since only one node rather than two is involved. The case $y \in I_j(x)$ follows from the definition of $\hat{q}^j$. For pairs x and y not covered in the preceding four cases, $q(x,y) = \hat{q}(y,x) = 0$ by (2.52) and (2.61). So, (2.65) holds in this last case as well.

Demonstration of (2.66). This part requires more work. We will show that

$$q(x) = \sum_k \alpha_k - \sum_k \lambda_k + \sum_j q^j(x^j) \quad (2.68)$$

and

$$\hat{q}(x) = \sum_k \hat{\alpha}_k - \sum_k \lambda_k + \sum_j \hat{q}^j(x^j). \quad (2.69)$$

The first sums in the two equalities are equal by (2.64), and the last sums are equal since (2.18) holds for each node. So, (2.68) and (2.69) together imply (2.66).

We first show (2.68). We rewrite $q(x)$ as

$$q(x) = \left(\sum_k \sum_{y \in A_k(x)} + \sum_k \sum_{y \in E_k(x)} + \sum_{k,\ell} \sum_{y \in R_{k,\ell}(x)} + \sum_j \sum_{y \in I_j(x)} \right) q(x,y), \quad (2.70)$$

and analyze the different parts. By (2.48′), the first double sum on the right equals

$$\sum_k \alpha_k \sum_{y^j \in A_k^j(x^j)} p_k^j(x^j,y^j) = \sum_k \alpha_k. \quad (2.71)$$

By (2.49), the second double sum equals

$$\sum_k \sum_{y^j \in E_k^j(x^j)} q(x^j,y^j) P_{k,0}. \quad (2.72)$$

By (2.50′), the third double sum equals

$$\sum_k \sum_{z^j \in E_k^j(x^j)} q^j(x^j, z^j) \sum_\ell P_{k,\ell} \sum_{y^h \in A_\ell^h(z^h)} p_\ell^h(z^h, y^h) \tag{2.73}$$
$$= \sum_k \sum_{z^j \in E_k^j(x^j)} q^j(x^j, z^j) \sum_\ell P_{k,\ell}.$$

By (2.51), the last double sum equals

$$\sum_j \sum_{y^j \in I^j(x^j)} q^j(x^j, y^j). \tag{2.74}$$

Summation of (2.72)-(2.74) gives

$$\left(\sum_k \sum_{y^j \in E_k^j(x^j)} + \sum_j \sum_{y^j \in I^j(x^j)} \right) q^j(x^j, y^j). \tag{2.75}$$

Also, note that

$$\sum_k \sum_{y^j \in A_k^j(x^j)} q^j(x^j, y^j) = \sum_k \alpha_k^j = \sum_k \lambda_k. \tag{2.76}$$

The sum of (2.75) and the left side of (2.76) is just $\sum_j q^j(x^j)$. On the other hand, the right side of (2.70) is equal to the sum of the left side of (2.71) and (2.75). So, $q(x)$ is equal to the sum of $\sum_k \alpha_k$ and (2.75), whereas $\sum_j q^j(x^j)$ is equal to the sum of $\sum_k \lambda_k$ and (2.75). Solving for this last term implies (2.68).

The argument for (2.69) is similar, and we employ the analog of the decomposition in (2.70), but for $\hat{q}(x)$ instead of $q(x)$. By the first equality in (2.57) and (2.63),

$$\sum_k \sum_{y \in A_k(x)} \hat{q}(x, y) = \sum_k (\hat{\alpha}_k / \lambda_k) \sum_{y^j \in A_k^j(x^j)} \hat{q}^j(x^j, y^j) = \sum_k \hat{\alpha}_k. \tag{2.77}$$

Also, using the first equalities in (2.58) and (2.59), and (2.60), the same reasoning as that leading to (2.75) implies that the sum of the terms corresponding to the last three double sums in (2.70) is

$$\left(\sum_k \sum_{y^j \in E_k^j(x^j)} + \sum_j \sum_{y^j \in I^j(x^j)} \right) \hat{q}^j(x^j, y^j). \tag{2.78}$$

On the other hand, it follows from (2.63) that

$$\sum_k \sum_{y^j \in A_k(x^j)} \hat{q}^j(x^j, y^j) = \sum_k \lambda_k. \tag{2.79}$$

The sum of (2.78) and the left side of (2.79) is just $\sum_j \hat{q}^j(x^j)$. Employing (2.77), (2.78), and (2.79) as we did (2.71), (2.75), and (2.76), the same reasoning as before implies (2.69). ■

Another proof for Theorem 2.11

Another proof for Theorem 2.11, that is more probabilistic, is given in [Wa82] and [Wa83] (see also [Wa88]). The basic idea of the proof is to modify the queueing network by imposing an ϵ delay, with $\epsilon > 0$, on all routing between classes. The corresponding stochastic process will be easier to analyze. It will not be Markov, but will have a distribution that is of product form and is invariant over time, and is the same for all values of ϵ. The limiting process as $\epsilon \downarrow 0$ will be the Markov process for the original queueing network, and its stationary distribution will be this distribution.

We now sketch the argument. Consider the J quasi-reversible nodes that are associated with the queueing network as in (2.48)-(2.52), but which have external arrival rates α_k instead of λ_k. We form a new network from these nodes by assuming that when jobs leave a node j from class k, they are routed back to class ℓ of node h with probability $P_{k,\ell}$, but with a fixed deterministic delay $\epsilon > 0$. During this delay, such jobs are assumed to not affect the transitions within the nodes, which now play the role of individual stations within the network.

One can construct the corresponding stochastic process inductively over time intervals of length ϵ, starting with $[0, \epsilon]$. One argues by first *assuming* that (a) the initial states at the J stations are independent of one another and are given by the stationary distributions of the isolated nodes with external arrival rates λ_ℓ and (b) over the time interval $(0, \epsilon]$, the jobs returning to classes ℓ constitute independent Poisson processes having rates $\lambda_k P_{k,\ell}$, which are independent of the initial states in (a). Jobs from outside the system arrive at class ℓ at rate α_ℓ, and so by (b) and the traffic equations (1.6), the combined arrivals at ℓ from these two sources of jobs are Poisson processes with rates λ_ℓ and are independent of one another. Because of the ϵ delay for returning jobs, jobs departing from nodes over $(0, \epsilon]$ will not return over this period, and so do not affect arrivals at ℓ.

On account of these arrival processes, the processes at the stations will be stationary over $(0, \epsilon]$ and independent of one another. Since the corresponding nodes are quasi-reversible, jobs depart from the classes k according to independent rate-λ_k Poisson processes over this period, which are independent of the states of the stations at time ϵ. Because of the deterministic ϵ delay required for returns, these jobs return to the classes ℓ as Poisson processes with rates $\lambda_k P_{k,\ell}$, over the period $(\epsilon, 2\epsilon]$. Consequently, the analogs of conditions (a) and (b) hold over the time interval $[\epsilon, 2\epsilon]$.

Iteration over the time intervals $(\epsilon, 2\epsilon]$, $(2\epsilon, 3\epsilon], \ldots$ produces a stochastic process on $[0, \infty)$ whose states at different stations at any fixed time are independent of one another, and whose distributions are the same as the

stationary distributions of the corresponding isolated nodes. This process can also be extended to all times $t \in (-\infty, \infty)$.

Letting $\epsilon \downarrow 0$, the sequence of these processes will converge to the Markov process corresponding to the original queueing network. Since each of these processes has the same joint distribution at any given time, this distribution will be stationary for the limiting Markov process. Since this distribution has the desired product form, this reasoning implies (2.53) of Theorem 2.11. By considering the exit processes of the sequence of processes, one can also show that the original queueing network is quasi-reversible.

3

Instability of Subcritical Queueing Networks

Until the early 1990's, the understanding of multiclass queueing networks was sketchy. In particular, relatively little thought had been given to the stability of such networks. The network, of course, needs to be subcritical. On the other hand, the "classical networks" considered in Chapter 2 are all stable when they are subcritical. Does this behavior hold in general, assuming all states of the corresponding Markov process communicate? Since one typically cannot explicitly compute the stationary distribution of a network, the direct approach in Chapter 2 needs to be replaced.

If stability holds universally for subcritical networks (or, in a broad enough setting), one should expect a reasonably simple proof of this; the argument would presumably be elementary because of its robustness. If, however, such a result holds in some settings but not in others, a general theory (if one exists) might be complicated. We now know through various examples that stability does not always hold. Whether a general theory is possible is still an open question. In this chapter, we present a number of examples exhibiting different situations in which the number of jobs in the network goes to infinity as $t \to \infty$. Chapter 4 will be devoted to positive results on the stability of subcritical networks.

This chapter is broken into three sections, according to the order of appearance and content of the examples. In Section 3.1, the first basic examples of unstable subcritical queueing networks are given. These consist of examples in [LuK91] and [RyS92] for static priority disciplines and [KuS90] for a clearing policy. In Section 3.2, examples of unstable subcritical FIFO networks are given, which consist of examples in [Br94a,b] and [Se94]. Section 3.3 discusses examples for other disciplines that illustrate features of interest. They consist of an unstable network of Kelly type from [DaWe96], and examples from [Du97] and [Ba98], where the region of stability of certain networks is examined in greater detail.

Somewhat different definitions of "unstable" exist in the literature. For us, a queueing network will be unstable if, for some initial state, the number of jobs in the network will, with positive probability, go to infinity as $t \to \infty$.

M. Bramson, *Stability of Queueing Networks.* Lecture Notes in Mathematics 1950. 53

This was mentioned in Section 1.2. When the network has only a finite number of states possessing fewer than a given number of jobs, and all states communicate with one another, this is equivalent to saying that, for each initial state, the number of jobs in the network goes to infinity almost surely as $t \to \infty$. Both conditions will be satisfied for the examples in this chapter with Poisson arrivals and exponential service times. Note that a network that is not stable is not necessarily unstable as defined above, since the corresponding Markov process can be null recurrent.

The intent of this chapter is to provide an elementary introduction to the subject and at the same time impart some feeling for the development of this subject in the 1990's. An omission with regard to the latter is the interaction with contemporary developments for heavy traffic limits. This includes the example in [DaWa93], which highlighted the general lack of understanding of the asymptotics for even simple multiclass queueing networks. This material requires additional terminology and new concepts. Since it lies outside the scope of these lectures, we omit it with some regret.

There has also been interest in examples exhibiting instability with regard to certain questions arising in computer science. In this setting, the models that are investigated and the relevant questions can take on a somewhat different flavor. For one such topic, *adversarial queueing*, stability under a "worst case" scenario is examined, where an all-knowledgeable adversary is allowed to modify the precise timing of input into the system. Service times are deterministic and are most often assumed to be the same everywhere. We omit this topic and instead refer the reader to [BoKRSW96] and [AnAFKLL96].

As mentioned in Section 1.2, it is sometimes convenient to employ different notation for the classes of a network, depending on its routing. Most of the examples in this chapter will be reentrant lines; we will usually label the classes sequentially based on the order in which they appear along the route. When the network has more than one deterministic route, this route will be indicated by the first coordinate. In the first and last examples in Section 3.2, we will find it convenient to use different notation because of the structure of the networks, and we include the station as one of the coordinates.

3.1 Basic Examples of Unstable Networks

An elementary example of an unstable subcritical SBP network is given by the *Lu-Kumar network*. The example is presented in a deterministic setting in [LuK91]. Its proof there is short, and requires just a couple paragraphs. We present here both this version and the model in a random setting, whose proof is not difficult but requires a bit more work. We also mention a related earlier model from [KuS90] for a clearing policy.

Independently, an unstable static priority network, the *Rybko-Stolyar network*, was analyzed in [RyS92]. Its behavior is almost identical to the Lu-Kumar network, with the routing differing in just one aspect. We discuss this

network briefly. Although insufficiently appreciated at the time, both networks have since had a substantial effect on thinking in queueing theory.

The Lu-Kumar network

This network is a reentrant line consisting of two stations, with two classes at each station. Jobs following the deterministic route first visit station 1 after entering the network, next visit station 2 twice, and then visit station 1 a second time, before exiting the network. The route is depicted in Figure 3.1; as indicated at the beginning of the chapter, we order the classes according to their appearance along the route. The system evolves according to a preemptive SBP discipline, with jobs at class 4 having priority over those at class 1, and jobs at class 2 having priority over those at class 3.

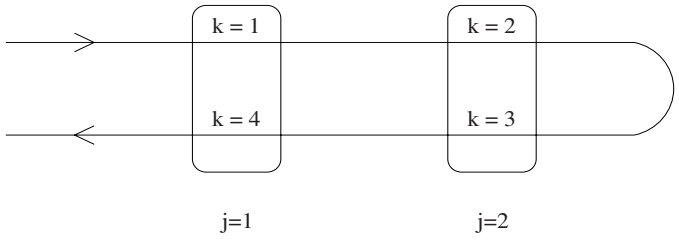

Fig. 3.1.

In [LuK91], a deterministic version of this model is given, with jobs entering periodically at the times $0, 1, 2, \ldots$. The deterministic service time at class k is given by m_k, with

$$m_2 = m_4 = 2/3 \tag{3.1a}$$

and

$$m_1 = m_3 = 0. \tag{3.1b}$$

This version of the model will be referred to as the *deterministic Lu-Kumar network*. As we will see, even though the service at classes 1 and 3 is instantaneous, their presence affects the evolution of the network. This model is somewhat artificial, but is easy to analyze. (A modification with more general m_k will be discussed shortly.)

Theorem 3.1. *The deterministic Lu-Kumar network is unstable.*

Note that because of the presence of instantaneous events at classes 1 and 3, we need to specify an ordering of service in case of "ties". For this, we assume jobs at all classes complete service at times $t-$ "just before" t; the assumption, in particular, allows jobs to be served at class 3 and move to class 4, before a new job enters the network at class 1 and is served there. The

changes in the long-term evolution of the system induced by other orderings are minor.

Proof of Theorem 3.1. We assume that there are initially, at $t = 0-$, $M \in \mathbf{Z}_+$ jobs at class 1 and no jobs elsewhere. Since $m_1 = 0$, the M jobs at class 1 immediately leave there and move to class 2, where they begin service. Because of the priority of class 2 over class 3, jobs at class 3 cannot be served until class 2 is empty, which means that class 4 must remain empty until then. Therefore, jobs at class 1 continue to be served until class 2 is empty. Since $m_2 = 2/3$, reasoning along these lines shows that at time $2M-$, all classes are empty except for class 3, which has $3M$ jobs (the M original jobs plus $2M$ new jobs).

Since $m_3 = 0$, these $3M$ jobs are immediately served at class 3 and move to class 4, where they begin service. Because of the priority of class 4 over class 1, jobs entering the network at class 1 cannot be served until all of these jobs depart from class 4. Since $m_4 = 2/3$, this occurs at time $4M-$. Over the elapsing time $2M$, $M' = 2M$ jobs have arrived at class 1; moreover, at time $4M-$, there are no jobs elsewhere in the network. The state at time $4M-$ therefore has the same form as it had initially, but with twice as many jobs; moreover, over $[0-, 4M-]$, there are never fewer than M jobs in the network. This cycle repeats itself indefinitely, resulting in always at least $2^n M$ jobs in the n^{th} cycle, which goes to infinity as $n \to \infty$. ■

We note that for $m_2 = m_4 = c$, any choice of $c > 1/2$ suffices for the above instability of the network; $c = 2/3$ was chosen above for convenience. The key feature of the network is that the main body of jobs is forced to remain "clumped together" as it moves from class 1 to class 4 during each cycle. This induces underutilization of (or "starvation" at) both stations 1 and 2, and so is responsible for the instability of the network. The above argument does not rule out the possibility of the number of jobs in the network remaining bounded over time when starting from different initial states, since such states need not communicate.

The above deterministic setting for the Lu-Kumar network and the restrictions on m_k, in (3.1), are not essential features of the model. Assume instead that jobs enter the network according to a rate-1 Poisson process and are served at all classes according to independent exponentially distributed random variables with means $m_k > 0$. We refer to this model as the *random Lu-Kumar network*.

Theorem 3.2. *The random Lu-Kumar network, with $m_2 + m_4 > 1$, is unstable.*

The proof of Theorem 3.2 is not difficult, but requires some preparation. Before proceeding, we first point out that if instead

$$m_1 + m_4 < 1, \quad m_2 + m_3 < 1 \tag{3.2}$$

(that is, the network is subcritical) and

$$m_2 + m_4 < 1, \tag{3.3}$$

then the network is stable. This is shown in [DaWe96]. The argument employs the machinery of fluid models, which are discussed in detail in Chapter 4. The exponential assumptions on the interarrival times and service times are important for neither direction, although in this framework, the Lu-Kumar network corresponds to a countable state Markov process. We recall the FBFS and LBFS reentrant lines, which were introduced in Chapter 1. It will be shown in Section 5.2 that when they are subcritical, such reentrant lines are always stable. The Lu-Kumar network, whose discipline is a mixture of these disciplines, is unstable.

We recall from Chapter 1 the following notation, which will reoccur more extensively later on in these lectures. Let $Z_k(t)$ denote the number of jobs at class k at time t, with $Z(t)$ being the corresponding vector. Since the interarrival and service times are all exponentially distributed, $Z(t)$ is a Markov process. We set $|Z(t)| = \sum_k Z_k(t)$. Also, let $W_j(t)$ denote the immediate workload at station j, that is, $W_j(t)$ is the amount of time required to serve all jobs currently at j, if one excludes other jobs from entering the station.

Before giving the proof of Theorem 3.2, we make the following two observations. Suppose that

$$Z_2(t) > 0 \text{ and } Z_4(t) > 0 \quad \text{for all } t \in (t_1, t_2], \tag{3.4}$$

for some $t_1 \leq t_2$. Then,

$$W_2(t_2) = W_2(t_1) - (t_2 - t_1). \tag{3.5}$$

The direction "$\leq$" follows from the priority of class 4 over class 1, which prevents any jobs from entering station 2 when $Z_4(t) > 0$, and hence over $(t_1, t_2]$. The other direction is automatic. Since $W_2(t_2) \geq 0$, (3.5) gives an upper bound on τ_0, the first time at which either $Z_2(t) = 0$ or $Z_4(t) = 0$, in terms of $W_2(0)$. In particular, $\tau_0 < \infty$ a.s.

Similarly, the priority of class 2 over class 3 prevents any jobs from entering class 4 as long as $Z_2(t) > 0$. Together with the sentence after (3.5), this implies that a.s.,

$$Z_2(t) = 0 \text{ or } Z_4(t) = 0 \quad \text{for all } t \geq \tau_0. \tag{3.6}$$

As an immediate consequence, we have:

Lemma 3.3. *For the random Lu-Kumar network, jobs in the classes 2 and 4 are a.s. not served simultaneously at any time $t \geq \tau_0$.*

This simple observation is the basis for the proof of Theorem 3.2. It says, in essence, that service at these classes is restricted as if they belonged to the same station, with the condition $m_2 + m_4 > 1$ implying that this "station" is supercritical. Consequently, the network is unstable. This observation was apparently first made in [BoZ92]. A more general version of it is used heavily in the work on *global stability* in [DaV00], which is discussed in Section 5.4.

Proof of Theorem 3.2. Set σ_i equal to the sum of the service times at classes 2 and 4 of the i^{th} job entering the network after time 0. Then, $\sigma_1, \sigma_2, \ldots$ are i.i.d. random variables with mean $m_2 + m_4$. The times at which these jobs are served are disjoint after time τ_0 because of Lemma 3.3. So, the time of departure from the network of the n^{th} of these jobs is at least $S_n - \tau_0$, where $S_n = \sum_{i=1}^{n} \sigma_i$. By the strong law of large numbers,

$$S_n/n \to m_2 + m_4 \quad \text{as } n \to \infty$$

holds a.s. It follows from this limit and the preceding observation that

$$\limsup_{t\to\infty} D(t)/t \le 1/(m_2 + m_4) \quad \text{as } t \to \infty \tag{3.7}$$

holds a.s., where $D(t)$ is the number of departures from the network over $(0, t]$.

Let $A(t)$ denote the number of arrivals in the network over $(0, t]$. Since $A(t)$ is given by a rate-1 Poisson process, it also follows from the strong law that

$$A(t)/t \to 1 \quad \text{as } n \to \infty \tag{3.8}$$

holds a.s. The difference $A(t) - D(t)$ gives a lower bound on $|Z(t)|$. So, by (3.7) and (3.8),

$$\liminf_{t\to\infty} |Z(t)|/t \ge 1 - 1/(m_2 + m_4) > 0 \quad \text{a.s.},$$

with the last inequality holding since $m_2 + m_4 > 1$. Consequently, $|Z(t)| \to \infty$ as $t \to \infty$. ■

The Kumar-Seidman network

A somewhat earlier example in [KuS90] exhibits behavior similar to the Lu-Kumar network. The model considered there consists of a reentrant line with two stations and four classes, with route again given by Figure 3.1. The model is again deterministic, with rate-1 arrivals and subcritical service times satisfying $m_2 + m_4 > 0$. The authors choose a continuous mass setting for their model, which is somewhat easier to work with. (This setting is also used for the model from [Se94] that is discussed in Section 3.2, where additional background is given.)

The discipline is a *clearing policy*. This requires each station to continue serving a class until there is no "job mass" left at that class, at which point the station begins service at one of the remaining nonempty classes, if there are any. (In the present example, each station has only two classes, and so there is only one such remaining class.) We refer to the network given in Figure 3.1 with this clearing policy as the *Kumar-Seidman network*. A clearing policy might be a practical choice for the discipline when there is a high start-up cost for switching the processing at a station from one task (i.e., class) to another.

The network was assumed to be subcritical, with $m_2 + m_4 > 1$. For a direct comparison with the deterministic Lu-Kumar network, we instead assume the

more restrictive (3.1). Using the same initial state as in Theorem 3.1, with job mass M initially at class 1, it is not difficult to check that this model is unstable in the same way as the Lu-Kumar network, with the job mass going to infinity as $t \to \infty$. Namely, mass starting at and entering class 1 is served at classes 1 and 2. When the mass at class 2 is exhausted, service takes place at class 3 and immediately afterwards at class 4, where all of the mass has moved. At this last step, service ceases at class 1, where mass now builds up while the mass at class 4 is served. By the time the last mass at class 4 has been served, the mass is $2M$ at class 1, which completes the cycle. Note that the clearing policy here plays the same role as the priority scheme in the Lu-Kumar network, with the main body of mass being forced to remain together as it moves from class 1 to class 4 during a cycle. This induces underutilization of both stations 1 and 2, and is responsible for the instability of the network.

The discipline of the Kumar-Seidman network depends on its earlier states rather than on the priorities of jobs currently in the network. So, as an example of instability, it is less convincing than the Lu-Kumar network. ([KuS90] also considered a "clear-a-fraction" discipline with the routing in Figure 3.2.) The paper pre-dates [LuK91]. For that reason, networks with the routing in Figure 3.1 (or Figure 3.2), under any discipline, are sometimes referred to as Kumar-Seidman networks.

The Rybko-Stolyar network

This network also consists of two stations, with two classes at each station. Jobs are assumed to follow one of two symmetric routes, visiting first one station and then the other, as indicated in Figure 3.2. The classes along the first route are labelled (1,1) and (1,2), in the order of their appearance, and the classes along the second route are labelled (2,1) and (2,2).

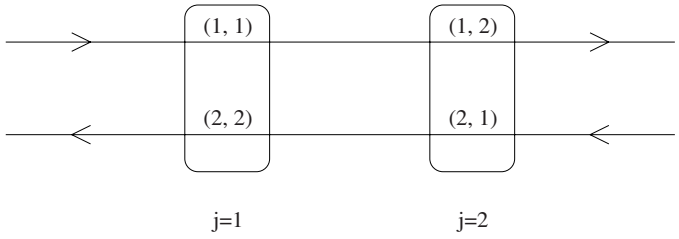

Fig. 3.2.

The system evolves according to a preemptive SBP discipline, with jobs at (2,2) having priority over jobs at (1,1) and jobs at (1,2) having priority over those at (2,1). That is, jobs at the last class along a route always have priority over jobs at the first class. Jobs are assumed to enter the network according to two independent rate-1 Poisson processes, and are served at the classes

(i,k) according to independent exponentially distributed random variables, with rates $m_{1,1} = m_{2,1}$ and $m_{1,2} = m_{2,2}$, and $m_{i,k} > 0$.

This model was examined in [RyS92], and is generally referred to as the *Rybko-Stolyar network*. Unlike the model in [LuK91], it is random. [RyS92] showed the following result.

Theorem 3.4. *The Rybko-Stolyar network, with* $m_{1,2} > \frac{1}{2}$*, is unstable.*

One can see, with some experimenting, that the Rybko-Stolyar network should evolve in the same basic manner as the random Lu-Kumar network, if one matches the classes (1,1), (1,2), (2,1), and (2,2) with the classes 1, 2, 3, and 4 of the latter network. Jobs entering the Rybko-Stolyar network at (1,1) have lower priority than jobs at (2,2) and jobs at (2,1) have lower priority than jobs at (1,2), and so, in either case, must wait until jobs at the latter classes are served. This is analogous to the relationship between classes 1 and 4, and 3 and 2 in the Lu-Kumar network. In fact, the Rybko-Stolyar network "becomes" the Lu-Kumar network if one connects the classes (1,2) and (2,1) as in Figure 3.3.

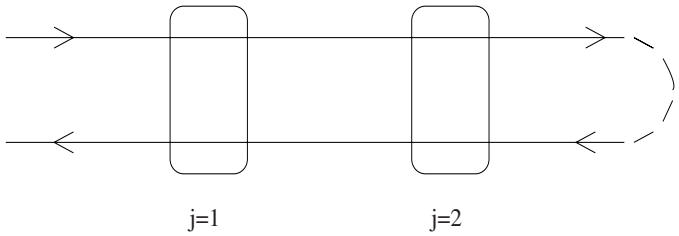

Fig. 3.3.

Both the Lu-Kumar and Rybko-Stolyar networks provide simple examples of unstable subcritical queueing networks, and can be used to motivate more complicated examples. The Lu-Kumar network has the advantage of being a reentrant line; the Rybko-Stolyar network is symmetric with respect to its two stations. A proof of Theorem 3.4 can be given along the same lines as that of Theorem 3.2, except that the step bounding τ_0 needs to be modified. (One can assume that station 2 is subcritical, from which it will follow that $\tau_0 < \infty$.) The proof in [RyS92] is different and does not rely directly on Lemma 3.3. We also note that the Rybko-Stolyar network is stable under the analogs of (3.2) and (3.3) for $m_{i,k}$. This was shown in [BoZ92].

3.2 Examples of Unstable FIFO Networks

The examples given in Section 3.1 are for SBP disciplines that have been specifically designed to impede the even flow of jobs, and therefore "starve"

their stations for work. These networks were initially regarded as artificial examples whose instability was not representative of "typical" disciplines, and so received insufficient attention. It turns out that even subcritical queueing networks with the FIFO discipline may be unstable. FIFO is a natural discipline and was a typical choice for the discipline of multiclass networks when multiclass networks started receiving attention. The demonstration of the existence of unstable subcritical FIFO queueing networks therefore had substantial influence on the stability theory of queueing networks.

Examples of subcritical FIFO queueing networks that are unstable were given in [Br94a], [Br94b], and [Se94]. We focus here primarily on the example in [Br94a]. It is the easiest to understand, and the mechanism that causes the uneven flow of jobs may be thought of as an extension of that in [LuK91].

An unstable FIFO example

The example in [Br94a] is a reentrant line consisting of two stations. Jobs following the deterministic route first visit station 1 after entering the network, next visit station 2 repeatedly for a total of K times (where K will be chosen large), and then visit station 1 a second time, before exiting the network. We employ the notation (j,k) here for a class, with $j = 1, 2$ denoting its station and k denoting the order this class is visited among classes of its station. In all, there are $K + 2$ classes in the network.

Jobs are assumed to enter the network according to a rate-1 Poisson process and have exponentially distributed service times with means $m_{j,k}$ corresponding to the k^{th} visit to the j^{th} station, with

$$m_{1,2} = m_{2,1} = c \tag{3.9a}$$

and

$$m_{j,k} = \delta \quad \text{for } (j,k) \neq (1,2) \text{ and } (j,k) \neq (2,1). \tag{3.9b}$$

The route and mean service times can be depicted as in (3.10), with the mean service times being given above the arrows pointing from the corresponding classes:

$$\rightarrow (1,1) \overset{\delta}{\rightarrow} (2,1) \overset{c}{\rightarrow} (2,2) \overset{\delta}{\rightarrow} \ldots \overset{\delta}{\rightarrow} (2,K) \overset{\delta}{\rightarrow} (1,2) \overset{c}{\rightarrow} \tag{3.10}$$

One requires c to be close to 1 and δ to be small. For the computations in [Br94a],

$$\frac{399}{400} \leq c < 1, \quad c^K \leq \frac{1}{50}, \quad \delta \leq \frac{1-c}{50K^2} \tag{3.11}$$

are used. For instance, one may choose

$$c = \frac{399}{400}, \quad K = 1,600, \quad \delta = 10^{-11}. \tag{3.12}$$

With additional effort, less extreme values can be chosen. Note that on account of (3.11),

$$\rho_1 = m_{1,1} + m_{1,2} = \delta + c < 1, \quad \rho_2 = \sum_{k=1}^{K} m_{2,k} = c + (K-1)\delta < 1, \quad (3.13)$$

and so these networks are subcritical.

One can demonstrate the following result.

Theorem 3.5. *FIFO queueing networks with the routing in (3.10) and mean service times in (3.9) and (3.11) are unstable.*

According to a simulation in [Da95], instability already occurs at $K = 4$ for appropriate mean service times, although this is likely difficult to show analytically. (The simulation is actually for a variant with slightly different service times than those in (3.9), so that $\rho_1 = \rho_2$ holds.) We also note that the network can be modified so that it remains unstable, without affecting the main structure of the corresponding proof. For instance, the long string of returns to station 2 in (3.10),

$$\rightarrow (2,1) \rightarrow (2,2) \rightarrow \ldots \rightarrow (2,K) \rightarrow$$

can be replaced by a route segment also involving a third station,

$$\rightarrow (2,1) \rightarrow (3,1) \rightarrow (2,2) \rightarrow (3,2) \rightarrow \ldots \rightarrow (2,K) \rightarrow (3,K) \rightarrow$$

where the service time at $(3,k)$, for $k = 1, \ldots, K$, is the same as at $(2,k)$, which is again chosen as in (3.9). So, consecutive returns to a station are not a crucial feature of the model. Similar modifications can be made to the SBP examples in Section 3.1 without affecting their instability.

In order to investigate the evolution of the queueing networks in Theorem 3.5, we employ the following notation. Here, $Z_{j,k}(t)$ will denote the number of jobs at the class (j,k) at time t, with $Z(t)$ denoting the corresponding vector and $|Z(t)|$ being the total number of jobs. (Recall that since $Z(t)$ does not reflect the order of jobs, more information is needed to specify the state of the corresponding Markov process.) By $(j,k)^+$, we will mean the set of classes occurring strictly after (j,k) along the route followed by jobs, and by $Z^+_{j,k}(t)$ the number of jobs in $(j,k)^+$.

Most of the work in demonstrating Theorem 3.5 is for the following induction step.

Proposition 3.6. *Consider a FIFO queueing network satisfying the routing in (3.10) and mean service times in (3.9) and (3.11), with*

$$Z_{1,1}(0) = M, \quad Z^+_{1,1}(0) \leq M/50. \quad (3.14)$$

Then for some $\epsilon > 0$, large enough M, and appropriate T (depending on M),

$$P(Z_{1,1}(T) \geq 100M,\ Z^+_{1,1}(T) \leq M) \geq 1 - e^{-\epsilon M} \quad (3.15)$$

and

$$P(|Z(t)| \geq M/4 \text{ for all } t \in [0,T]) \geq 1 - e^{-\epsilon M}. \quad (3.16)$$

We will later choose $T \approx 2cM/(1-c)$. Of course, the factor 50 in (3.14) is not special, although the ratio $Z^+_{1,1}(0)/Z_{1,1}(0)$ should be small.

Once Proposition 3.6 has been established, the proof of Theorem 3.5 is quick. To see this, suppose $Z(0)$ satisfies (3.14) for some large M. Repeated application of Proposition 3.6 implies that

$$P(|Z(t)| < M/4 \text{ for some } t \geq 0) \leq 2\sum_{i=0}^{\infty} e^{-100^i \epsilon M}, \tag{3.17}$$

which $\to 0$ as $M \to \infty$. All states of the Markov process $Z(t)$ corresponding to the network communicate with one another, and so, by (3.17), no state is recurrent. Hence, $|Z(t)| \to \infty$ a.s. as $t \to \infty$, for any $Z(0)$. This implies Theorem 3.5. Since $T \approx 2cM/(1-c)$, one can also employ Proposition 3.6 to show that the number of jobs increases linearly in t, i.e.,

$$0 = \liminf_{t\to\infty} |Z(t)|/t \leq \limsup_{t\to\infty} |Z(t)|/t < \infty.$$

Recall that the iterative procedure of viewing time intervals over which the number of jobs in a system grows geometrically, while the system returns to a "multiple" of its original state, was also employed in the proof of Theorem 3.1. Indeed, this is the most natural path to follow in attempting to analyze the asymptotic behavior of many unstable networks. One question one can ask for the FIFO queueing network in Theorem 3.5 is whether there are essentially different ways in which $Z(t)$ can approach infinity. More generally, one can ask about the nature of the Martin boundary of the Markov process associated with this or other unstable queueing networks. These questions remain essentially uninvestigated.

Outline of the proof of Proposition 3.6

We outline here the proof of Proposition 3.6. We begin by introducing a sequence of stopping times $S_1, S_2, \ldots, S_\ell, \ldots$ for the process $Z(t)$. Jobs at either station at a given time t are ordered according to the times at which they are next served, so we can talk about a "first" or "last" job in this sense. (Due to the multiple classes at each station, jobs entering the network earlier may nevertheless be ordered behind more recent arrivals.) Let S_1 denote the time at which the last of the original jobs (jobs at $t = 0$) at station 1 is served. Let $S_2, S_3, \ldots, S_\ell, \ldots$ denote the successive times at which the last jobs at station 2 are served, where the ordering is made at $t = S_{\ell-1}$. Set $S_L = S_{L+1} = S_{L+2} = \ldots$, where S_L is the time at which station 2 becomes empty. (On account of (3.13), $\rho_2 < 1$ and so $L < \infty$ a.s.) We can think of the intervals $(S_\ell, S_{\ell+1}]$, $\ell = 1, 2, \ldots$, as "cycles" at the end of which each job starting at $(2, k)$, $k < K$, is at $(2, k+1)$. Note that no job can be served twice at station 2 before every other job there is served once, due to the FIFO discipline. We also let T (which appears in Proposition 3.6) denote the time at which the last job at $(1, 2)$ at time S_{2K} leaves the network.

We break the outline of the proof into four main steps, corresponding to the evolution of $Z(t)$ over the intervals $(0, S_1]$, $(S_1, S_K]$, $(S_K, S_{2K}]$ and $(S_{2K}, T]$. We present here the intuition for each step, with the reader being referred to [Br94a] for a rigorous analysis.

Step 1. Behavior on $(0, S_1]$. We assume, as in (3.14), that $Z_{1,1}(0) = M$ with M large, and that there are few jobs elsewhere in the network. Since $m_{1,1} = \delta << 1$, one has $S_1 << M$ except on a set of small probability. Also, $m_{2,1} = c >> \delta$, and so at time S_1, nearly all of the original jobs in the network are still at $(2, 1)$. Moreover, comparatively few new jobs have entered the network up to time S_1, and so there are few jobs at $(1, 1)$. A schematic diagram for this and following steps is given in Table 3.1.

Table 3.1. This table gives the approximate number of jobs at each class of the network at the successive times $0, S_1, \ldots, S_{K+1}, T$. Classes marked with $*$ are classes having negligible numbers of jobs off sets of small probability. At $t = S_{K+1}$, the total number of jobs at $(1, 1)$ and $(2, 1)$ is approximately $c^K M$, which is itself negligible. Note that at $t = T$, the state is a "multiple" of that at time 0, with factor $c/(1-c)$. Since $S_K \approx cM/(1-c)$ and $T - S_K \approx cM/(1-c)$, one has $T \approx 2cM/(1-c)$.

$t \backslash (j,k)$	$(1,1)$	$(2,1)$	$(2,2)$	$(2,3)$	$\ldots$	$(2,K-1)$	$(2,K)$	$(1,2)$
0	M	$*$	$*$	$*$		$*$	$*$	$*$
S_1	$*$	M	$*$	$*$		$*$	$*$	$*$
S_2	$*$	cM	M	$*$		$*$	$*$	$*$
S_3	$*$	c^2M	cM	M		$*$	$*$	$*$
$\vdots$		$\vdots$	$\vdots$	$\vdots$				
S_K	$*$	$c^{K-1}M$	$c^{K-2}M$	$c^{K-3}M$	$\ldots$	cM	M	$*$
S_{K+1}	$*$	$*$	$c^{K-1}M$	$c^{K-2}M$	$\ldots$	c^2M	cM	M
S_{2K}	$*$	$*$	$*$	$*$		$*$	$*$	$M/(1-c)$
T	$cM/(1-c)$	$*$	$*$	$*$		$*$	$*$	$*$

Of course, since we are working with random events here, the above behavior is sometimes violated. However, such exceptional events occur with probabilities that are exponentially small in M, and one can show they can be ignored without affecting the basic nature of the evolution of $Z(t)$. Here and later on, we therefore neglect these exceptional probabilities. Needless to say, a rigorous proof requires accurate bookkeeping of such exceptional probabilities. We discuss this point after completing the description of $Z(t)$ over $[0, T]$.

Step 2. Behavior on $(S_1, S_K]$. We first consider the evolution of $Z(t)$ over $(S_1, S_2]$. Over this time interval, the (approximately) M jobs at $(2, 1)$ all move to $(2, 2)$. Since $m_{2,1} = c$, the time it takes to serve these jobs is (approximately) cM. The time required to serve other jobs is minimal, so $S_2 - S_1 \approx cM$. During this time, (approximately) cM new jobs enter the system, which quickly move

to (2,1). Thus, at $t = S_2$, there are (comparatively) few jobs in the system except at (2,2) and (2,1), where there are (approximately) M and cM jobs, respectively.

Continuing our reasoning along the same lines, we observe that over $(S_2, S_3]$, the jobs at (2,1) and (2,2) advance to (2,2) and (2,3), respectively. Since $m_{2,2} = \delta << 1$, the time required to serve the jobs at (2,2) is negligible; the time required for the jobs at (2,1) is c^2M, so $S_3 - S_2 \approx c^2M$. Over this time, c^2M new jobs enter the system, which quickly move to (2,1). So, at time S_3, there are few jobs in the system except at (2,3), (2,2), and (2,1), where there are M, cM and c^2M jobs, respectively. Proceeding inductively, we obtain that at time S_K, there are M jobs at $(2, K)$, cM jobs at $(2, K-1)$, and so on down to (2,1), where there are $c^{K-1}M$ jobs. At station 1, there are few jobs. The elapsed time $S_K - S_{K-1} \approx c^{K-1}M$. On account of (3.11), c^K is small, and so there are about

$$\sum_{\ell=0}^{K-1} c^\ell M \approx M/(1-c) \tag{3.18}$$

jobs in the system. Likewise, $S_K \approx cM/(1-c)$.

We point out that the large number of "quick" classes for station 2, in conjunction with the FIFO discipline, serves to trap most jobs within station 2, and prevent them from reaching class (1,2), until there are few remaining jobs at the "slow" class (2,1). The behavior of this network thus mimics that of the Lu-Kumar network, with the role of the single low priority "quick" class of station 2 being played by these many "quick" classes.

Step 3. Behavior on $(S_K, S_{2K}]$. Over the short period of time $(S_K, S_{K+1}]$, the evolution of the system changes. The M jobs from $(2, K)$ arrive at (1,2). Since $m_{2,1} = c$, these jobs require time cM to be served at station 1, during which time new arrivals at (1,1) will not be served. The cycles $(S_\ell, S_{\ell+1}]$, $\ell = K, \ldots, 2K-1$, are all of much shorter duration than cM, because of (3.11), as is their union $(S_K, S_{2K}]$. By the end of this period, the jobs already at station 2 at time S_K have already arrived at (1,2); because of (3.18), there are essentially $M/(1-c)$ such jobs. So, at time S_{2K}, there are essentially $M/(1-c)$ jobs at (1,2) and no jobs elsewhere. Of course, here and elsewhere, we are taking liberties in ignoring "negligible" quantities of jobs and probabilities.

Step 4. Behavior on $(S_{2K}, T]$. During $(S_{2K}, T]$, the $M/(1-c)$ jobs at (1,2) exit the system. The time required to serve these jobs is $cM/(1-c)$. So, $T - S_{2K} \approx cM/(1-c)$. During this time, $cM/(1-c)$ jobs enter the system. These new jobs are obliged to remain at (1,1) until time

$$T = S_{2K} + (T - S_{2K}) \approx 2cM/(1-c).$$

At this time, there are few jobs elsewhere in the system. So at time T, the state of the system is a "multiple", by the factor $c/(1-c)$, of the state at

time 0. This is the type of bound needed in (3.15) of Proposition 3.6. The bound $c \geq 399/400$, that we are assuming in (3.11), will be sufficient to derive (3.15) when the above argument is carried out rigorously. (Presumably, the system exhibits the same behavior when $c > 1/2$.)

We still need to demonstrate (3.16) of Proposition 3.6, which gives a lower bound on $|Z(t)|$ over $[0, T]$. The previous reasoning in fact shows that, except on a set of small probability, $|Z(t)|$ will not drop much below M on $[0, T]$. This is because, up to time S_K, most of the original M jobs at (1,1) remain in the system, with there being approximately $M/(1-c)$ jobs in the system at time S_K. Since $m_{1,2} = c$, before most of these jobs have left the system, an additional $cM/2(1-c) >> M$ jobs enter the system, which are trapped at (1,1) until time T. The bound in (3.16) follows from a rigorous version of this reasoning.

We stated during the outline of the proof of Proposition 3.6 that the probabilities of the exceptional events we neglected were exponentially small in M. Here, we provide a short summary of the approach used in [Br94a] to show this, referring the reader there for more detail.

Let $X_1, X_2, X_3, \ldots$ be i.i.d. mean-1 exponentially distributed random variables, with $Y_n = X_1 + \ldots + X_n$. Then, for each $\alpha > 0$, there exists $\beta > 0$, such that for $n \geq 1$,

$$P(|Y_n - n|/n > \alpha) \leq e^{-\beta n}. \tag{3.19}$$

This is a simple large deviations bound that can be demonstrated in the usual way, by applying Markov's Inequality to the moment generating function of Y_n. It extends immediately to i.i.d. exponentially distributed random variables with other means. The variables Y_n can also be inverted to obtain analogous exponential bounds on the number of exponentially distributed random variables occurring by a given time.

For a given class (j, k), one can write

$$Z_{j,k}(t) = Z_{j,k}(0) + A_{j,k}(t) - D_{j,k}(t), \tag{3.20}$$

where $A_{j,k}(t)$, respectively $D_{j,k}(t)$, are the total number of jobs arriving at, respectively departing from, (j, k) over $(0, t]$. By applying (3.19) repeatedly, one can derive upper and lower bounds on $A_{j,k}(t)$ and $D_{j,k}(t)$, and hence on $Z_{j,k}(t)$, over the times $S_1, S_2, \ldots, S_{2K}, T$ defined earlier. For instance, replacing 1/50 by η for readability, the first bounds employed in [Br94a] are exponentially small upper bounds in M for the probabilities that

$$S_1 > 2\eta M, \quad A_{1,1}(S_1) > 3\eta M, \quad D_{2,1}(S_1) > 3\eta M,$$

with the first of these bounds together with (3.19) being used for the last two bounds. The last two bounds are then applied to obtain exponentially small upper bounds on

$$Z_{2,1}(S_1) < (1-3\eta)M, \quad Z_{1,1}(S_1) + Z_{2,1}(S_1) > (1+4\eta)M,$$

$$Z^+_{2,1}(S_1) + D_{1,2}(S_1) > 4\eta M.$$

Similar bounds, such as on

$$Z^+_{2,k}(S_k) + D_{1,2}(S_k) > 4\eta M, \quad k = 2, \ldots, K,$$

are then obtained. These last bounds limit the rate at which jobs can move through the system. Together with further estimates, these bounds enable one to rigorously justify the reasoning employed in Steps 1-4.

Another unstable FIFO example

The following unstable FIFO reentrant line is given in [Se94]. It consists of four stations, each visited three times, with route given in Figure 3.4. As in Section 3.1, class k denotes the k^{th} class along the route.

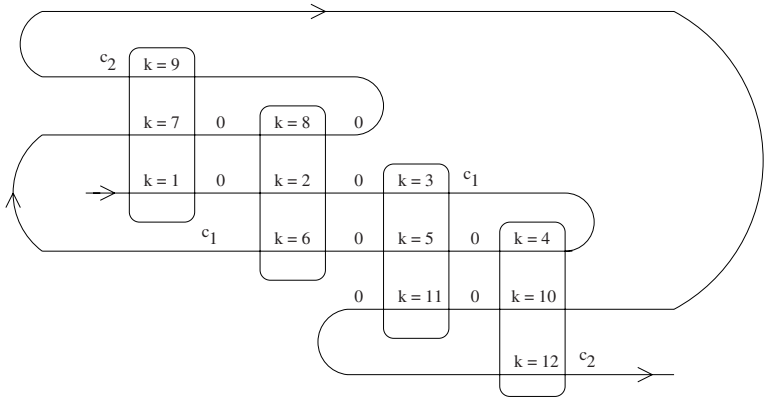

Fig. 3.4. The service time for a class is given along the route immediately after the class.

The model is a continuous, deterministic analog of a queueing network, whose state at each class is given by a nonnegative real number representing the quantity of "mass" there. As time evolves, this mass is served in a continuous and deterministic manner. This model is an example of a *fluid network*. Fluid networks were mentioned briefly in Section 1.3 and are similar to the fluid models that are considered in detail in Chapter 4; we omit a systematic discussion here. We denote by $Z_k(t)$ the amount of mass at class k at time t, by $Z(t)$ the corresponding vector, and by $|Z(t)|$ its magnitude. In analogy with our definition for queueing networks, we will say this fluid network is unstable if for some initial state, $|Z(t)| \to \infty$ as $t \to \infty$.

As indicated in the figure, each station has one "slow" class and two "quick" classes. The "slow" classes have service times c_1 and c_2, which are assumed to satisfy

$$2(c_1)^2 < c_2 < 1, \quad c_1 > 1/2; \tag{3.21}$$

the "quick" classes have 0 service times. The service rates are given by the reciprocals $1/c_i$ and ∞; service at the "quick" classes is therefore instantaneous. The presence of the "quick" classes nonetheless affects the flow of mass through the system since, according to the FIFO discipline, mass at such a class must wait until earlier arriving mass at the station's "slow" class is served. As with previous reentrant lines, we assume mass enters the system (in this case, deterministically) at rate 1. These features are similar to those in [KuS90] and [LuK91].

It is easy to see that this network is subcritical, since the sum of service times at each station is less than 1. The amount of mass in the network goes to ∞ as $t \to \infty$, however, for appropriate initial states. Such a state is given by

$$Z_2(0) = 2c_1 M, \quad Z_8(0) = M, \quad Z_{10}(0) = \frac{c_2 - 2(c_1)^2}{2c_1 c_2} M, \tag{3.22}$$

$$Z_{12}(0) = \frac{c_2 - 2(c_1)^2}{4(c_1)^2 c_2} M, \quad Z_k(0) = 0 \text{ elsewhere},$$

where the mass at class 12 is understood to have arrived before that at class 10. (Since the service at both classes 2 and 8 is instantaneous, the ordering there is not important.) From this, one obtains:

Theorem 3.7. *The FIFO fluid networks defined above are unstable.*

As with the networks considered so far in this chapter, the proof consists of an iterative argument, with the state of the system returning to a geometrically growing "multiple" of the original state after each iteration. As before, $|Z(t)|$ will grow linearly in t.

[Se94] also considers the variant of the above model with mass leaving the system after classes 3, 6, and 9 (as well as after class 12), and mass, at unit rate, entering the system at classes 4, 7, and 10 (as well as at class 1). Since the analog of Theorem 3.7 for this variant is somewhat easier to show, this is done in [Se94] and the proof of Theorem 3.7 is summarized. In both cases, replacement of "instantaneous" classes by "quick" classes, with service times $\delta > 0$, and the fluid network by the corresponding queueing network considerably complicates the bookkeeping necessary for keeping track of jobs. Although not published, this stochastic version is presumably doable.

An unstable FIFO network with quick service times

The examples of unstable subcritical networks given so far all have at least one class with mean service time greater than 1/2. In each of these examples, it follows that the traffic intensity ρ_j is greater than 1/2 at some station j. What happens when ρ_j is uniformly small at all stations? Must all such FIFO queueing networks be stable? The following example from [Br94b] shows this is not the case.

Jobs are assumed to follow one of two nearly identical routes, the "upper"and "lower" routes, at the end of which they exit from the system. As illustrated in (3.23), there are J stations along each route. (Because of space considerations, we label only the stations but not the classes along each route.)

$$\begin{gathered}\to 1 \to 2 \to \ldots \to 2 \to 3 \to \ldots \to 3 \to \ldots \to J \to \ldots \to J \to \\ \to 1 \to 1 \to 2 \to \ldots \to 2 \to 3 \to \ldots \to 3 \to \ldots \to J \to \ldots \to J \to 1 \to\end{gathered} \tag{3.23}$$

Jobs are assumed to enter each route at rate 1/2. Along each route, the stations $j = 2, \ldots, J$ are each visited seven times; at each such station, the first visit is "slow" and the remaining six visits are "quick". Only the visit to station 1 at the end of the lower route is "slow"; the three earlier visits to station 1 along both routes are "quick".

We now give a precise description of the network depicted in (3.23). We employ the notation (i, j, k) for a class, with $i = u, \ell$ denoting its route, $j = 1, \ldots, J$ denoting its station, and k denoting the order this class is visited, among classes of its station along this route. One has $k = 1, \ldots, 7$ for $j = 2, \ldots, J$; $k = 1$ for $i = u$ and $j = 1$; and $k = 1, 2, 3$ for $i = \ell$ and $j = 1$. The two types of jobs are assumed to enter the system according to independent rate-1/2 Poisson processes. Service times of jobs are independent and exponentially distributed, with means

$$\begin{aligned}&c \text{ at } (i, j, 1), \quad \text{for } i = u, \ell \text{ and } j = 2, \ldots, J, \\ &c \text{ at } (\ell, 1, 3), \\ &\delta \text{ at } (i, j, k), \quad \text{for } i = u, \ell,\ j = 2, \ldots, J \text{ and } k = 2, \ldots, 7, \\ &\delta \text{ at } (u, 1, 1), (\ell, 1, 1) \text{ and } (\ell, 1, 2).\end{aligned} \tag{3.24}$$

We assume that

$$0 < c \le \frac{1}{100}, \quad 0 < \delta \le c^8, \quad J = \lfloor 2c^{-1} \log(c^{-1}) \rfloor. \tag{3.25}$$

Each of the stations $2, \ldots, J$ therefore has one comparatively slow and six (very) quick classes for each type of job; station 1 has only the single quick class for the upper jobs, and two quick classes and one slow class for lower jobs. The choice of parameters is made for technical reasons. (The coefficient 2 in the definition of J has been chosen so that $(1 - c)^{-J} \sim c^{-2}$. The bound

$c \leq 1/100$ is somewhat arbitrary.) One can think of this family of networks as being constructed by piecing together J copies of the middle section of networks of the type in (3.10). (We only need $K = 7$ here, although $K \geq 7$ can instead be used.)

Under (3.25),

$$\rho_j \leq c + 6\delta \leq 2c \quad \text{for } j = 1, \ldots, J.$$

So, by choosing c small, the traffic intensity can be chosen as small as desired for each j. Nonetheless, the following is true.

Theorem 3.8. *FIFO queueing networks with the routing in (3.23) and mean service times in (3.24)-(3.25) are unstable.*

To demonstrate Theorem 3.8, one employs an appropriate analog of Proposition 3.6. One can then argue exactly the same way as immediately after Proposition 3.6 to finish the proof of Theorem 3.8. The spirit of the proof of this analog is similar to that of Proposition 3.6, although details are more involved. The purpose of the upper jobs is solely to restrict the flow of the lower jobs. Because of the multiple stations employed here, the same reasoning that shows, in the previous network, that the main body of jobs remains close together, cannot be applied directly. However, with the control resulting from the upper jobs, one can analyze the flow of lower jobs much as was done in Proposition 3.6. For more details, the reader is referred to [Br94b] or [Br95]. Presumably, the analog of Theorem 3.8 holds for an appropriate reentrant line, most likely with a route corresponding to that of the lower jobs in (3.23), although the reasoning given in [Br94b] no longer suffices.

Theorem 3.8 has the following interesting consequence. One can compare any FIFO queueing network satisfying (3.24)-(3.25) with the network that is obtained from it by replacing (3.24) with the assumption that the mean service time $m_{i,j,k} = c$ at every class. The lengths of the mean service times for the new network are, of course, everywhere at least as great as those of the original network. This new network is a subcritical FIFO network of Kelly type with $\rho_j \leq 7c$ for all j. Such a FIFO network has a stationary distribution that is given by (2.4) and (2.9). In particular, the stationary probability of there being n jobs at a given station is at most $(1 - 7c)(7c)^n,\ n \geq 1$, which means that the network is in fact "very stable" for small c. This comparison shows that decreasing the mean service times within a queueing network may result in making it unstable.

At present, there is a lack of general criteria for the stability of FIFO queueing networks. The difficulties are illustrated by the previous examples. They do not, however, rule out two possible criteria for stability. Set $m_j^{\min}$ ($m_j^{\max}$) equal to the minimum (maximum) over all $m_k,\ k \in \mathcal{C}(j)$, where j is fixed, and set $m_j^R = m_j^{\min}/m_j^{\max}$.

The first criterion is that, for given $m^R > 0$, there exists an $r > 0$, so that if a FIFO network satisfies $m_j^R \geq m^R$ and $\rho_j \leq r$, for all j, then it is stable.

The second criterion is that, for given $r < 1$, there exists an $m^R < 1$, so that if a FIFO network satisfies $m_j^R \geq m^R$ and $\rho_j \leq r$, for all j, then it is stable.

According to the first criterion, if the mean service times at a given station are not too different, then small enough traffic intensities suffice for stability independently of the specific structure of the network. Similarly, according to the second criterion, a subcritical network, with $\rho_j \leq 1 - \epsilon$ for all j and given $\epsilon > 0$, is stable as long as the ratios m_j^R are close enough to 1. Networks of Kelly type, with $m_j^R \equiv 1$, make up the limiting case in the latter scenerio. Note that the first criterion becomes elementary if the routing of the network is instead specified before r is chosen, since the total number of classes is then fixed. In that setting, r can be chosen small enough so that $\Sigma_j \rho_j < 1$, which implies the network is stable by the example at the end of Section 4.4. No progress has been made toward justifying or disproving these criteria.

3.3 Other Examples of Unstable Networks

In the previous two sections, we have given a number of examples of subcritical queueing networks that are unstable. Here, we give several other examples of unstable queueing networks. In some of these cases, more can be said about the region of stability as the traffic intensity ρ varies.

The first example we consider is from [DaWe96]. It consists of a subcritical SBP reentrant line of Kelly type that is unstable. (Recall that a network is of Kelly type if all classes at a given station have the same mean service time.) This contrasts, of course, with the stability of FIFO networks of Kelly type that was shown in Chapter 2.

The second example is from [Du97]. It exhibits a family of SBP networks for which the region of stability is nonmonotone in the parameter ρ, behavior that is shared with the last example in the previous section. The region of stability is, moreover, explicitly calculated and is not convex.

The last example is from [BaBo99]. The family of networks given there is FIFO, but with the added feature that the number of jobs on one of the given routes in the network is bounded at any given time. The region of stability here is also calculated explicitly and is neither monotone nor convex. The boundary is, moreover, self-similar around one of its boundary points.

An unstable network of Kelly type

In contrast to FIFO networks, it is not sufficient for a subcritical queueing network to be of Kelly type for it to be stable. It is not difficult to produce a reentrant line, with an appropriate SBP discipline, that exhibits this behavior. Such an example is provided in [DaWe96].

This example consists of two stations, with each station being visited three times, and is depicted in Fig 3.5. The priority scheme is (6,5,1) at station 1 and (3,2,4) at station 2, e.g., the last class visited at station 1 has the highest

priority and the first class visited there has the lowest priority. The discipline is preemptive. We assume, as usual, that jobs enter the network according to a rate-1 Poisson process. We also assume that the service times are exponentially distributed, with mean 0.3 at each class. The network is clearly subcritical, with $\rho = (.9, .9)$.

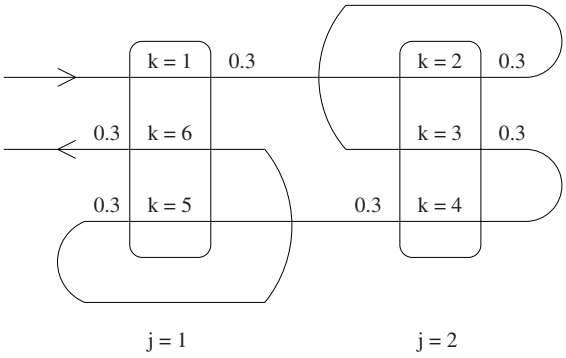

Fig. 3.5. This reentrant line has priority scheme (6,5,1) at station 1 and (3,2,4) at station 2. All service times have mean 0.3.

Theorem 3.9. *The static priority reentrant line of Kelly type in Figure 3.5 is unstable.*

We motivate Theorem 3.9 by comparing the evolution of the reentrant line there with the evolution of the Lu-Kumar network; a rigorous argument can be given by mimicking the proof of Theorem 3.2. First note that if one both "combines" the last two classes of station 1, $k = 5$ and $k = 6$, into a single class and "combines" the first two classes of station 2, $k = 2$ and $k = 3$, into a single class, one obtains the four-class reentrant line whose route and mean service times are given in Figure 3.6.

We assign to the classes of the new reentrant line the same priority scheme as in the Lu-Kumar network, with $k' = 4$ having priority over $k' = 1$, and= $k' = 2$ having priority over $k' = 3$. Adding the service times at the combined classes produces service times having gamma distributions with means 0.6, in both cases.

Some thought shows that this new network is a natural "projection" of that in Figure 3.5, in the sense that there is a pathwise correspondence between the evolution of jobs in the two networks, with the understanding that jobs at $k = 5$ and $k = 6$, respectively at $k = 2$ and $k = 3$, in the old network are combined into $k' = 4$, respectively $k' = 2$, in the new network. This correspondence relies on the assigned priority scheme for the old network: both classes 5 and 6, respectively classes 2 and 3, have higher priority than class 1, respectively class 4, and so the information lost in the projection is immaterial in assigning the priority of service to jobs in the combined class

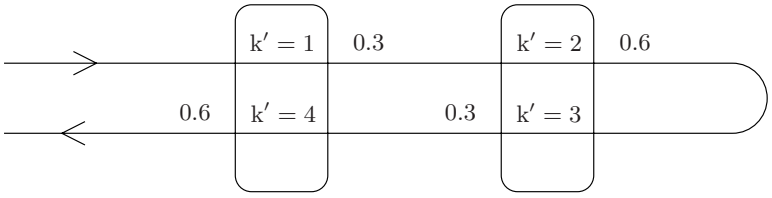

Fig. 3.6. "Combining" classes $k = 5$ and $k = 6$ into a single class and $k = 2$ and $k = 3$ into a single class produces this reentrant line, with the above mean service times.

with respect to the other remaining class. Moreover, since class 6 has higher priority than class 5 and class 3 has higher priority than class 2 in the old network, jobs currently in service in the new network at class 4 and at class 2 will not be preempted in the middle of their service by other jobs there. This allows us to maintain the pathwise correspondence between jobs in the two networks.

Because of the above relationship between the two networks, instability of one network implies instability of the other. The projected network is the same as the Lu-Kumar network in Theorem 3.2, except that the exponentially distributed service times at classes 2 and 4 are replaced by service times with gamma distributions there, each with mean 0.6. As mentioned after the statement of Theorem 3.2, the assumption that service times are exponential is not needed there, if one is willing to allow a more general state space. Since $0.6 > 0.5$, as required in the theorem, it will follow that the network in Figure 3.6 is unstable. One can also show Theorem 3.9 without referencing Theorem 3.2, but instead by mimicking its proof; the setup there has some similarity with the "method of stages" employed in Section 2.4.

Two examples with nonconvex regions of stability

The first example is an SBP queueing network from [Du97]. As in the Rybko-Stolyar network of Section 3.1, jobs travel along one of two routes that are oriented in opposite directions. In the present setting, there are three stations having two classes each, with the classes labelled according to the route and position along the route, as in Figure 3.7. Jobs of the second route are assumed to have priority over jobs of the first route at each of the first two stations, with this being reversed at the third station. The discipline is preemptive. Jobs enter the routes according to independent rate-α_k Poisson processes, $i = 1, 2$, and have independent exponentially distributed service times with means $m_{i,k}$.

Consider the functions $F_1(\rho)$, $F_2(\rho)$, and $F_3(\rho)$ defined by

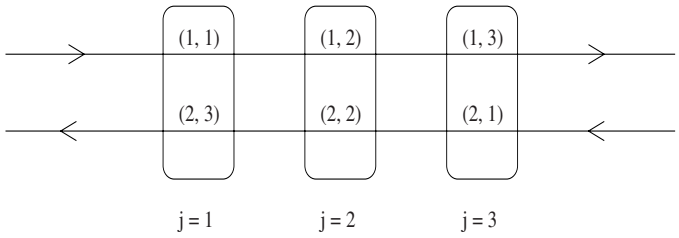

Fig. 3.7. The priority scheme favors the classes (2,3), (2,2), and (1,3), respectively, at the stations 1, 2, and 3.

$$
\begin{aligned}
F_1(\rho) &= (\rho_{1,1} + \rho_{2,3} - 1) \vee (\rho_{1,2} + \rho_{2,2} - 1) \vee (\rho_{1,3} + \rho_{2,1} - 1) \\
&\qquad \vee (\rho_{1,3} + \rho_{2,2} - 1), \\
F_2(\rho) &= (\rho_{1,3} + \rho_{2,3} - 1)(1 - \rho_{1,2} - \rho_{2,2}) - (\rho_{1,2} + \rho_{2,3} - 1)(1 - \rho_{1,3} - \rho_{2,1}), \\
F_3(\rho) &= (\rho_{1,3} + \rho_{2,3} - 1) \wedge [(\rho_{1,3} - \rho_{1,2}) \vee F_2(\rho)],
\end{aligned}
$$

where $\rho_{i,k} = \alpha_k m_{i,k}$. (Recall that $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$). Also, set $F(\rho) = F_1(\rho) \vee F_3(\rho)$. The regions of stability/instability for the network can be explicitly written in terms of these functions.

Theorem 3.10. *The SBP network in Figure 3.7 is stable if* $F(\rho) < 0$ *and unstable if* $F(\rho) > 0$.

The reader should not focus too much on the specifics of $F(\rho)$. For our purposes, it is enough to observe that the regions of stability/instability are explicit and are considerably more complicated for this relatively elementary network than for either the Lu-Kumar or Rybko-Stolyar network. Moreover, (a) the region of stability $F(\rho) < 0$ is not convex and (b) stability is not monotone in ρ, i.e., one can find $\rho < \rho'$ with $F(\rho) > 0$ and $F(\rho') < 0$. Both properties follow from Theorem 3.10 by setting $\rho_{1,1} = \rho_{2,2}$, $\rho_{1,2} = \rho_{2,3}$, and $\rho_{1,3} = \rho_{2,1}$, and restricting $F(\rho) < 0$ to this three dimensional subspace. Slices of this region at fixed values of $\rho_{1,1}$ can then be analyzed. As a quick check on the consistency of the instability condition $F(\rho) > 0$, note that $F_1(\rho) > 0$ if any of the three stations is supercritical.

In order to show both directions of Theorem 3.10, [Du97] employs ergodic/transience criteria developed in [MaM81] for reflected random walks on $\mathbf{Z}^d_{+,0}$. Motivation for the definition of $F(\rho)$ is provided by fluid equations that correspond to the original network. These equations are related to fluid models, which will be studied in the next chapter.

The second example we discuss is from [BaB99] and examines a family of networks consisting of a single route with *controlled jobs* and multiple transverse routes with *cross jobs*. ([BaB99] employs the terms *controlled customers* and *cross customers*.) Controlled jobs moving along their route visit each of the J stations in the network once; the cross jobs each visit only a single station before exiting the network. No more than L controlled jobs are allowed

in the network at a given time; when there are more than L such jobs, the excess jobs wait at an outside buffer until a controlled job leaves the network, at which point the first such job enters the network. There is no such restriction on cross jobs. (See Figure 3.8.) All jobs are served according to the FIFO discipline at each station. These networks can be thought of as a simple model with window flow control for packet-switched communication networks.

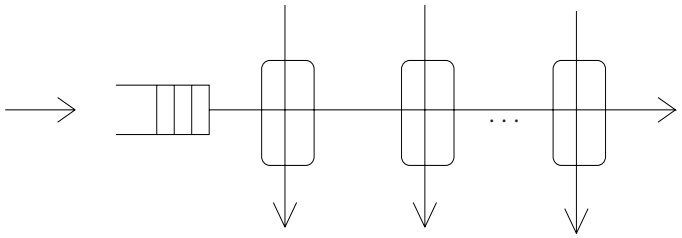

Fig. 3.8. The controlled jobs move horizontally and the cross jobs move vertically. The number of controlled jobs in the network at a given time is restricted, with excess controlled jobs required to wait at the outside buffer on the left. There are currently 3 jobs waiting at the outside buffer.

Detailed analysis of such a system is possible under certain restrictive assumptions. Set $J = 2$ and $L = 1$, and assume that the service times for all jobs (both controlled and cross) are deterministic and take value 1. Also, assume that the interarrival times of the cross jobs at station 1 are deterministic and take value τ, with $\tau \geq 1$, and that there are never any cross jobs at station 2. There are only minimal assumptions on the interarrival times for the controlled jobs, namely that they define a stationary and ergodic point process with some intensity λ.

On account of the outside control on jobs, this network differs from the other examples considered in this chapter. Also, because of the deterministic aspects of this model, it is more appropriate to weaken the definition of stable used elsewhere, and only require here that the model support a stationary process (that need not be ergodic). The definition of unstable remains the same as before.

This network is interesting because of how its region of stability depends on τ. For fixed τ, stability is monotone in the parameter λ, and does not otherwise depend on the arrival process for the controlled jobs; there is a $\bar{\lambda}$ so that for $\lambda < \bar{\lambda}$, the network is stable, while for $\lambda > \bar{\lambda}$, it is unstable. In Figure 3.9, $\bar{\lambda}$ is graphed as a function of $\lambda_1 = \tau^{-1}$, with the heavy line being both the graph of $\bar{\lambda}$ and the boundary between the stable and unstable regions.

As in the example in Figure 3.7, the region of stability of the network is not convex. Also, as in previous examples, it is not monotone. (This nonmonotonicity is in terms of the intensity of the cross traffic, though, instead of in terms of the mean service times, as in the previous examples.) Moreover,

the graph of $\bar{\lambda}$ is self-similar around $(0, \frac{1}{2})$, as is indicated in Figure 3.9. An open question is what part of this behavior remains when the deterministic interarrival and service times in the model are randomized.

Two other families of networks with nonmonotone regions of stability are given in[Br98a] and [DaHV99]. In both cases, decreasing the service time distributions can destabilize the network. In [Br98a], this is done by showing a network can be stabilized by inserting a single class station immediately before the visit to each class. (Related deterministic work was done in [Hu94].) In [DaHV99], this is done in the context of global stability, which is discussed in Section 5.4.

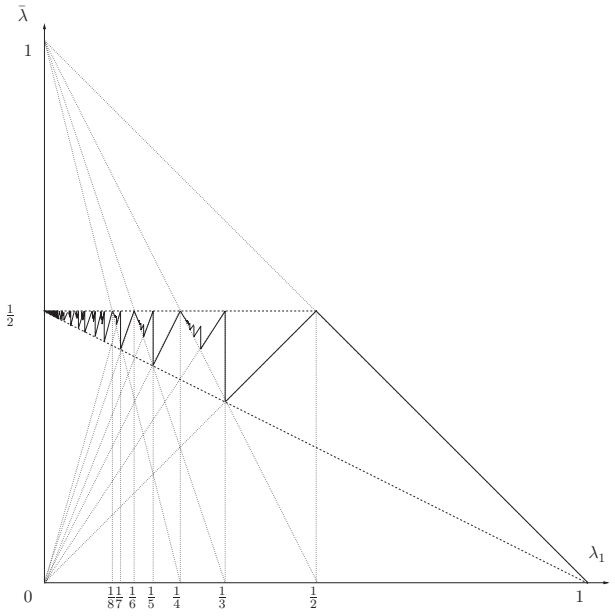

Fig. 3.9. This is the graph of $\bar{\lambda}$ as a function of $\lambda_1 = \tau^{-1}$, which is the boundary between the stable and unstable regions. (The heavy line is the graph of $\bar{\lambda}$; the dotted lines help one track the piecewise linear increments of $\bar{\lambda}$.)The graph of $\bar{\lambda}$ is self-similar around $(0, \frac{1}{2})$. (Example reprinted from [BaB99] with permission from Baltzer.)

4

Stability of Queueing Networks

In Chapter 2, we demonstrated the stability of several families of queueing networks by explicitly computing their stationary distributions. For queueing networks with other disciplines, or with nonexponential interarrival and service times, one cannot expect such explicit expressions. So, in order to investigate the stability of more general queueing networks, another approach is needed. Such an approach should be more qualitative and less computational in nature.

The approach we employ in this chapter to study the stability of queueing networks employs fluid limits and fluid models. Fluid models were discussed briefly in Section 1.3; we will examine both fluid limits and fluid models in detail here. Employing these tools, one can reduce the study of queueing networks to their simpler deterministic analogs. The basic theory is given here; applications will be given in Chapter 5.

In the next several paragraphs, we give some background on previous results on the stability of queueing networks. We recall that a queueing network is defined to be stable if its underlying Markov process is positive Harris recurrent. As mentioned earlier, when the state space is countable and all states communicate, this definition reduces to the usual definition of positive recurrence. Results in earlier works are typically stated within this more restrictive framework.

Interest in whether general families of queueing networks are stable has developed since the 1980's. Most early work was restricted to single class networks (see, e.g., [Bo86], [Si90], [BaF94], [ChTK94], and [MeD94]). Work on multiclass networks usually dealt with deterministic systems, with either discrete or continuous job mass (see, e.g., [PeK89], [KuS90], [LuK91], and [Ku93]); examples of both types of systems were given in Chapter 3. "Stability" for such deterministic systems was shown in various cases in the above literature, with stability, in this context, typically meaning that the quantity of job mass in the system converges to 0 or remains bounded over time. As in the stochastic setting, $\rho < e$ is the natural requirement for stability.

M. Bramson, *Stability of Queueing Networks*. Lecture Notes in Mathematics 1950.

Examples of unstable deterministic networks, with $\rho < e$, were given in [KuS90] and [LuK91]. An example of an unstable network in the stochastic setting was given in [RyS92]; these and other examples were discussed in Chapter 3. Such examples illustrate the importance of the discipline in determining whether a queueing network is stable.

It has long been believed that queueing networks should be stable under general assumptions. Inclusive conditions for stability are not known (and likely do not exist). Instead, stability is typically shown in the context of a specific discipline, under $\rho < e$ and perhaps other constraints. The approach we present in this chapter, using fluid limits and fluid models, reduces the study of such problems to a simpler, deterministic setting.

The foundation for this approach was given in [RyS92]. For a two station FIFO queueing network, with the same routing as in Figure 3.2 and with exponential interarrival and service times, the authors showed stability when $\rho < e$. Their argument involved showing the "stability" of solutions of a related deterministic system, and showing that rescaled solutions of the random system remain close to those of the deterministic system. As pointed out in [RyS92], this procedure is, in principle, quite general in nature. However, for all but the simplest systems, technical problems arise when comparing solutions of the random and deterministic systems.

Independently, [St95] and [Da95] developed criteria for the stability of queueing networks, in terms of the stability of limits of rescaled solutions of the network. (In the terminology of this chapter, [St95] used fluid limits and [Da95] used both fluid limits and fluid models.) [St95] assumed exponential distributions for the interarrival and service times; [Da95] considered more general distributions, but at the price of requiring the use of Markov processes with general state space. Applications illustrating this approach are given in [Da95].

The material in this chapter is based on the approach taken in [Da95]. The main theorem of the chapter, Theorem 4.16, corresponds to Theorem 4.2 in [Da95]. Our approach here is a modification of that in [Br98a]. Care has been taken to give a detailed presentation of the material, including that cited in [Da95]. As a consequence, the chapter is quite long; in the remainder of the introduction, we summarize its contents.

Summary of chapter

Chapter 4 consists of five sections. In Section 4.1, we present the foundations for the Markov processes we will need. We first give a detailed construction of the underlying Markov process for an HL queueing network. We next summarize relevant results from general Markov process theory. The third part of the section defines Harris recurrence and positive Harris recurrence, and gives an alternative formulation, in Theorem 4.1, that we will use.

We recall that a queueing network is e-stable, if its underlying Markov process is ergodic. In the last part of Section 4.1, we give general conditions

under which ergodicity holds. At the end of Section 4.4, we will use this to give criteria under which a queueing network is e-stable. In the continuous time, countable state space setting, positive Harris recurrence and ergodicity are equivalent.

Much of the material in Section 4.1 may be unfamiliar to readers. We point out that concepts such as positive Harris recurrent and petite are motivated by similar concepts in both the discrete time and countable state space settings. Various proofs either rely on, or are motivated by, similar results in these settings. Those readers who are interested in further background can refer to Section 4.5, which serves as an appendix to this section.

When the interarrival and service times are exponentially distributed, the underlying Markov process of the queueing network can, for many disciplines, be constructed on a countable state space. This considerably simplifies the preparation that is required to derive Theorem 4.16. In particular, one does not require the general machinery that is introduced in Section 4.1. For readers wishing such a "shortcut", we present a summary, at the end of the different sections, saying how the general approach presented here can be modified. (In Section 4.4, the summary is instead presented after Theorem 4.16.) Here in the introduction, we will also point out these shortcuts.

In Section 4.2, we present two results on bounded sets that we will need later on. The first, Proposition 4.6, is a variant of Foster's Criterion, which we refer to as the Multiplicative Foster's Criterion. The main condition is that, off of a bounded set, the Markov process $X(\cdot)$ have a uniformly negative drift on an appropriate time scale. One also requires that the bounded set be either petite or uniformly small, which are defined in Section 4.1. (In essence, petite means that all sets, weighted according to some measure, are "equally accessible" from any point in the petite set. Uniformly small is a somewhat stronger concept that is defined similarly.) One concludes that $X(\cdot)$ is positive Harris recurrent when the condition petite is assumed and ergodic when uniformly small is assumed. In the countable state space setting, the petite and uniformly small conditions can be dropped, since the empty state will be equally accessible from all points in the bounded set. Moreover, Theorem 4.1, from Section 4.1, is not needed for the proof of the Multiplicative Foster's Criterion in the countable state space setting.

The other result from Section 4.2, that we will need later, is Proposition 4.7. It says that when the interarrival times of a queueing network satisfy certain conditions, bounded sets will be uniformly small, and hence also petite; this enables us to employ the Multiplicative Foster's Criterion. The proposition is not needed in the countable state space setting.

Section 4.3 introduces fluid models and fluid limits. The first part of the section recalls the queueing network equations that were discussed briefly in Section 1.3; we provide more detail here. Fluid model equations are the deterministic analogs of the queueing network equations. They were also discussed briefly in Section 1.3; we provide further detail in the second part of Section 4.3. There, emphasis is placed on the basic fluid model equations, which are

the fluid model equations that do not depend on a specific discipline. Proposition 4.11, in the subsection, presents several elementary results on fluid models that will be used later. Two examples are given that illustrate nonuniqueness of fluid model solutions and instability when the system is subcritical.

Fluid limits are introduced in the last part of Section 4.3. They provide a rigorous connection between the queueing network and fluid model equations, with the latter being satisfied by limits of solutions of the former, under a "law of large numbers" scaling. Variants of fluid limits have been present in the literature at least since [Ne82] (see also [ChM91]).

The use of fluid limits, as in Proposition 4.12 of this section, makes the relationship between queueing network and fluid model equations precise. The connection between queueing network and fluid model equations will be important in Section 4.4, where we use the stability of fluid models (and, indirectly, of fluid limits) to show the stability of queueing networks. The approach taken in this section remains essentially the same when employed in the countable state space setting.

In Section 4.4, we demonstrate Theorem 4.16, which is the main result on the stability of queueing networks. As assumptions, one needs bounded sets to be petite and the fluid limits to be stable. Proposition 4.7, from Section 4.2, gives sufficient conditions for the former property to hold. The Multiplicative Foster's Criterion will be used, in conjunction with Proposition 4.7 and the bounds given in Section 4.4, to demonstrate Theorem 4.16. The work required for this simplifies considerably in the countable state space setting.

After the theorem, we provide various commentary, such as on modifications of its assumptions. One such modification, substituting uniformly small for petite, suffices for the stronger conclusion that the network is e-stable. The result, Theorem 4.17, follows by applying Theorem 4.3 from Section 4.1, and otherwise reasoning as in the proof of Theorem 4.16. As mentioned earlier, in the continuous time, countable state space setting, stability and e-stability are equivalent.

4.1 Some Markov Process Background

In this section, we first introduce the Markov processes that are associated with HL queueing networks. We next consider these Markov processes in an abstract setting, in which we define positive Harris recurrence. We then present a useful alternative characterization of positive Harris recurrence which will be applied to queueing networks in the following sections. Some of the background for this material is relegated to Section 4.5, which serves as an appendix to this section.

We note that in the countable state space setting, the material that is needed from this section is minimal. The construction of the Markov process simplifies, since sample paths are piecewise constant with finite jump rates.

Moreover, standard recurrence concepts from Markov chain theory apply, and so the discussion of Harris recurrence that is given here is not needed. More detail is given at the end of the section.

Definition of the Markov process

In this subsection, we give a careful construction of the Markov process that is associated with a given HL queueing network. Readers are encouraged to consider this material, but those who are not interested in the technical details should feel free to skip ahead to the next subsection. If the reader does so, he/she should keep in mind that we are defining continuous time Markov processes, on an appropriate space and with dynamics that correspond to the queueing networks of interest. One should note the "norm" $|x|$ of a state x, which is given in (4.3): it is the sum of components corresponding to the queue lengths, and residual interarrival and service times of the state. It will be employed later on in the chapter.

We begin by defining the state space $(S, \mathscr{S})$ of the Markov process. In order to motivate the different components of states of the space, we will repeatedly allude to the various queueing network quantities they will correspond to.

The state of the Markov process, at any time, will be given by a point $x = (y, r)$, where

$$y \in (\mathbf{Z} \times \mathbf{R})^{\infty} \times \mathbf{R}^{|\mathcal{A}|} \times \mathbf{R}^{K} \quad \text{and} \quad r \in \mathbf{R}^{K}, \tag{4.1}$$

and the coordinates are subject to appropriate positivity restrictions. Recall that $\mathcal{A}$ is the set of classes at which external arrivals are allowed. It is assumed that only a finite number of the pairs of coordinates of $(\mathbf{Z} \times \mathbf{R})^{\infty}$, indexed by i, are nonzero. For such a pair, the first coordinate k_i, $k_i = 1, \ldots, K$, is to be interpreted as the current class of a job in the network, with the second coordinate s_i, $s_i \geq 0$, measuring how long ago the job entered this class, where s_i is given in descending order. The job with the largest second coordinate, among those with first coordinate k, is thus the first or "oldest" job of class k. (For "ties" where two or more pairs have the same second coordinate, order the job with the smaller k coordinate first.)

We denote by $z = (z_1, \ldots, z_K)$ the number of jobs in each class, and set $|z| = \sum_{k=1}^{K} z_k$. The vector r is assumed to have coordinates $r_k \in [0, 1]$, $k = 1, \ldots, K$, with $r_k = 0$, for $z_k = 0$, and which sum to 1 over each nonempty station j. For each nonempty class k, the coordinate r_k is to be interpreted as the service rate of the oldest job of this class, with other jobs receiving no service. (This is the HL property.) The coordinates u_k, $u_k > 0$, of $\mathbf{R}^{|\mathcal{A}|}$ are to be interpreted as the *residual interarrival times* for classes k, with $k \in \mathcal{A}$. (That is, u_k is the remaining time before the next arrival at k of a job from outside the network.) The coordinates v_k of $\mathbf{R}^{K}$, for the y component, are the *residual service times* for the oldest job at each class k, $k = 1, \ldots, K$, with $v_k > 0$ except when $z_k = 0$, in which case we set $v_k = 0$. We denote by u and v the corresponding vectors, and set $|u| = \sum_{k \in \mathcal{A}} |u_k|$ and $|v| = \sum_{k=1}^{K} |v_k|$.

We denote by S the space given by (4.1) and the above restrictions on the coordinates. We wish to specify a metric on S. For this, it is convenient to denote by $\tilde{r}_i$ the service rate assigned to the i^{th} job, $i = 1, \ldots, |z|$. The metric is defined by adding up the contribution of each of the coordinates in (4.1), after taking differences for individual terms. Specifically, for $x, x' \in S$, with corresponding coordinates as denoted above, we set

$$d(x, x') = \sum_{i=1}^{\infty}((|k_i - k_i'| + |s_i - s_i'| + |\tilde{r}_i - \tilde{r}_i'|) \wedge 1) \qquad (4.2)$$
$$+ \sum_{k \in \mathcal{A}} |u_k - u_k'| + \sum_{k=1}^{K} |v_k - v_k'|.$$

For most purposes, we will not require the full metric, but just the associated "norm" $|\cdot|$, with

$$|x| = |z| + |u| + |v|. \qquad (4.3)$$

($|x|$ can be interpreted as the distance from x to the "origin", which is not in the space, however.) Note that $|x|$ is continuous as a function of x. We also equip S with the standard Borel σ-algebra inherited from the metric, which we denote by $\mathscr{S}$.

Since only a finite number of the coordinate pairs in $(\mathbf{Z} \times \mathbf{R})^{\infty}$ are nonzero, it is not difficult to see that the metric $d(\cdot, \cdot)$ is separable. It is also locally compact; with a bit of work, one can show this by showing that an open ball around a point x is homeomorphic to a finite product of intervals of the form $(0, 1)$ and $[0, 1)$. (Half-closed intervals are needed when either $s_i = 0$, $s_i = s_{i'}$ for some $i \neq i'$, or the coordinates r_k of r corresponding to a given station j are on the boundary of the simplex $\sum_{k \in \mathcal{C}(j)} r_k = 1$.) The metric is not complete, since one can choose Cauchy sequences x_n, with $u_n \to 0$ or $v_n \to 0$ as $n \to \infty$, that have no limit in S. It is, however, homeomorphic to the complete metric $d'(\cdot, \cdot)$ obtained by replacing the term $\sum_{k \in \mathcal{A}} |u_k - u_k'|$ in (4.2) by

$$\sum_{k \in \mathcal{A}} \left(|u_k - u_k'| + \left| \frac{1}{u_k} - \frac{1}{u_k'} \right| \wedge 1 \right),$$

and $\sum_{k=1}^{k} |v_k - v_k'|$ by the analogous term (where we set $\frac{1}{0} - \frac{1}{0} = 0$). We prefer to work with the simpler metric $d(\cdot, \cdot)$, since we will not require S to be complete.

We now formally define the Markov process *underlying* an HL queueing network as the stochastic process $X(t)$, $t \geq 0$, on S that undergoes the following evolution. We continue to use the same suggestive queueing vocabulary that was employed in motivating the construction of S.

We set $X(t) = (Y(t), R(t))$, with $Y(t)$ and $R(t)$ taking values y and r as in (4.1). (The coordinates s_i of y are allowed to exceed t, and so jobs may, in effect, arrive before time 0.) The evolution of $X(t)$, in between arrivals and departures of jobs at classes, is given by the service rates $R_k(t)$, which

are constant over such intervals. Upon an arrival or departure somewhere in the network at time t, the stochastic process is continued by assigning new service rates $R(t) = f(Y(t))$, where f is a measurable function. The choice of f corresponds to the discipline of the corresponding HL queueing network.

In order to describe the transition of $X(t)$ upon the arrival or departure of a job at a class, we introduce the sequences of positive i.i.d. random variables $\xi_k(i), k \in \mathcal{A}$, and $\gamma_k(i), k = 1, \ldots, K$, with $i = 1, 2, 3, \ldots$, which correspond to the interarrival and service times of the queueing network. We will also need the sequence of i.i.d. random vectors $\phi^k(i)$, $i = 1, 2, 3, \ldots$, with $\phi^k(i) = e_\ell$ for some $\ell = 1, \ldots, K$ or $\phi^k(i) = 0$, which give the routing of a job upon completion of its service at a class. (Here, $e_\ell \in \mathbf{R}^K$ is the unit vector in the positive ℓ direction.) We assume the sequences ξ, γ, and ϕ are mutually independent. These sequences, together with the function f and the initial state x, will determine the evolution of $X(\cdot)$ for all times along each sample path.

The process $X(t)$ can be constructed inductively as follows. At times t between arrivals and departures, $R(t) = r$ gives the rate of decrease of each component of the residual service time vector $V(t) = v$. When $V_k(t-) = 0$ occurs for some k, service at k for the oldest job of that class is assumed to be completed, with the job being routed to another class ℓ, if $\phi^k(i) = e_\ell$, and leaving the network, if $\phi^k(i) = 0$, where $\phi^k(i-1)$ is the previous routing vector applied at k. When this occurs, one sets $V_k(t) = \gamma_k(i)$ if class k is then nonempty, and sets $V_k(t) = 0$, if k is empty. Until a job leaves its class, its age $S_i(t) = s_i$ continues to increase at rate 1. (The label i of a job will typically change as it moves from one class to another.)

The components of the residual interarrival time vector $U(t) = u$ always decrease at rate 1 until hitting 0. At the time t for which this occurs at a given k, one includes the pair $(k, 0)$ in the state $Y(t)$ and sets $U(t) = \xi_k(i)$, where i is the index of the first unused interarrival time at k at time t. If the class k is empty at time $t-$, and hence $V_k(t-) = 0$, one sets $V_k(t) = \gamma_k(i')$, where i' is the index of the first unused service time at time t.

In various cases, the state space S can be simplified for queueing networks with specific HL disciplines, or specific interarrival and service time distributions. For instance, when the interarrival and service times are exponentially distributed, one can drop the u and v coordinates from the description of the state (unless r is chosen to depend on them).

Preemptive static buffer priority disciplines were introduced in Chapter 1. Such networks are HL, with $\tilde{r}_i = 1$ automatically holding for the oldest job of the highest ranked nonempty class at each station. One can therefore drop the age of jobs from the state space descriptor. As mentioned above, the residual interarrival and service times can also be dropped from the descriptor when the corresponding variables are exponentially distributed. The state space can then be reduced to points in $\mathbf{Z}^K$ with nonnegative coordinates, if one implicitly assumes the given priority scheme that orders classes at each station, since one is able to drop the coordinate r that governs service.

Networks with the FIFO discipline are also HL, since the oldest job at a station is always served first. If one chooses, one can drop the age of jobs from the state space descriptor, by instead ordering the ages at each station; one can also drop the coordinate r. As before, coordinates can also be eliminated when the interarrival and service times are exponentially distributed. The resulting state space will then be countable.

The state space S is large enough to contain the information needed for the queueing networks we will investigate. It can, however, be modified if one has other applications in mind. For instance, a state space descriptor for how long ago a job entered the network can be appended. This is needed for the FISFO networks mentioned briefly in Section 1.2 (see, e.g., [Br01]). Also, one can associate with each job in a class a residual service time, instead of with just the class itself (see, e.g. [Wi98b]). The definitions (4.2) and (4.3) then need to be modified accordingly.

Foundations and terminology

It is not difficult to see that the process $X(\cdot)$ just defined is time homogeneous and Markov, and that its sample paths are right continuous. Although $X(\cdot)$ is not a jump process, it evolves in a simple manner, having only isolated discontinuities and evolving deterministically in between. In particular, after a jump, the state of $X(\cdot)$ is explicitly known until its next jump, with its evolution being linear in its coordinates.

The process $X(\cdot)$ is an example of a *piecewise-deterministic Markov process* (PDP). Such processes are discussed in [Da84] and [Da93] in detail; we will rely on the latter in our discussion. In [Da93], PDPs are the more general family of processes whose evolution in between jumps, rather than being linear, is determined by a locally Lipschitz continuous vector field. Also, "killing" at a rate dependent on the position is allowed in [Da93]; after such killing, the process jumps according to a given random rule. In our setting, there is no such killing. Since, in between jumps, $X(\cdot)$ lives in a subset of S which is homeomorphic to an open ball in $\mathbf{R}^d$, for some d, one can append a "boundary" ∂S to S, which $X(\cdot)$ hits at a time $t-$ immediately before a jump. For the setting in [Da93], this approach is useful in assigning the jump rule for the process, and it also allows one to define the process on the space $D_S = D_S[0,\infty)$ of right continuous paths on S with left limits. However, since the process jumps instantly upon hitting ∂S, the introduction of ∂S has its own inconveniences. We avoid these complications and just stick with the space S, referring the reader to Sections 24 and 25 of [Da93] for the technical details in the more general setting.

So far, we have not defined the filtration for the process $X(\cdot)$. We let $\mathcal{F}_t^0$ denote the natural filtration

$$\mathcal{F}_t^0 = \sigma(X(s),\ 0 \le s \le t) \tag{4.4}$$

and let $\mathcal{F}_\infty^0$ be the σ-algebra generated by $\mathcal{F}_t^0$, for $t \ge 0$. Rather than employing $\mathcal{F}_t^0$ and $\mathcal{F}_\infty^0$ directly, we will use appropriate completions. As in [Da93],

for each initial probability measure μ on S, one can define a measure P_μ on the sample space corresponding to the process. Letting $\mathcal{F}_t^\mu$ be the completion of $\mathcal{F}_t^0$ obtained by including all P_μ-null sets of $\mathcal{F}_\infty^0$, one sets

$$\mathcal{F}_t = \bigcap_\mu \mathcal{F}_t^\mu, \tag{4.5}$$

and denotes by $\mathcal{F}_\infty$ the σ-algebra generated by $\mathcal{F}_t$, for $t \geq 0$. We henceforth employ $\{\mathcal{F}_t,\ t \geq 0\}$ as the filtration for the process $X(\cdot)$, and use P_μ to denote the corresponding probability measures. This setup is typical of general Markov process theory.

In addition to being complete in the sense of (4.5), the family $\{\mathcal{F}_t,\ t \geq 0\}$ is right continuous, that is,

$$\mathcal{F}_t = \mathcal{F}_{t+} \stackrel{\text{def}}{=} \bigcap_{\epsilon>0} \mathcal{F}_{t+\epsilon}. \tag{4.6}$$

Because $X(\cdot)$ is a PDP, this is not difficult to show. (See Theorem 25.3 in [Da93].) The properties (4.5) and (4.6) will be needed shortly for the Debut Theorem and the strong Markov property.

For $t \geq 0$, $x \in S$, and $A \in \mathscr{S}$, set

$$P^t(x, A) = P_x(X(t) \in A).$$

It is shown in [Da93] that $P^\cdot(\cdot, \cdot)$ is a probability transition kernel on $(S, \mathscr{S})$. That is, for fixed t and A, $P^t(\cdot, A)$ is $\mathscr{S}$-measurable; for fixed t and x, $P^t(x, \cdot)$ defines a probability measure on S; and $P^{s+t} = P^s \circ P^t$, for $s, t \geq 0$, defines a semigroup. However, $X(\cdot)$ need not be a Feller process. (Recall that $X(\cdot)$ is a Feller process if, in addition, $P^t : C(S) \to C(S)$, where $C(S)$ denotes the continuous bounded functions on S.) This may be the case even when $X(\cdot)$ corresponds to a queueing network with a FIFO or SBP discipline, since even a small change in the future arrival time of a job may cause it to be served after another job instead of before. This can induce a major change in the future evolution of $X(\cdot)$, which will imply that for $f \in C(S)$, $P^t f$ need not be continuous. Nevertheless, as shown in [Da93], $X(\cdot)$ is strong Markov.

In the next subsection, we will summarize the recurrence theory for Markov processes that we will use for queueing networks. Part of the appropriate literature assumes that these processes are *Borel right*. The intent there is to employ a well-studied general framework, but this comes at the cost of implicitly assuming some familiarity with a technical theory. Such knowledge will not be relevant for our applications, and so we relegate to Section 4.5 a brief discussion of the material. We summarize it here, noting that for a Borel right process, the σ-algebras satisfy (4.5) and (4.6), the process $X(\cdot)$ is defined on a "reasonable" state space, has a transition semigroup, is right continuous, and the process $(f \circ X)(\cdot)$ is right continuous when f is an α-excessive function. The last assumption can be replaced by the strong Markov property.

Harris recurrence of Markov processes

We discuss here basic criteria for positive recurrence of Markov processes on general state spaces. Since little is a priori assumed about either the state space or the process itself, one must provide some structure to be able to say anything of interest. There is a developed theory for the corresponding discrete time Markov processes that goes back to Doeblin (see [Do53] for an account), was developed in [Ha56], and includes contributions by [Or71], [AtN78], and [Nu78] among others. Standard references are [Nu84] and [MeT93d]. The basic approach in the discrete time setting is to formulate conditions that are general, but nonetheless enable one to mimic the machinery for discrete time Markov chains. With the aid of resolvents, problems in the continuous time setting can be reformulated in discrete time, which is the approach we summarize here.

Following [MeT93a-93c] and [KaM94], we will assume that the Markov processes $X(\cdot)$ are Borel right processes on a locally compact and separable metric state space $(S, \mathscr{S})$, where $\mathscr{S}$ is the Borel σ-algebra generated by the metric. The process $X(\cdot)$ will have a transition semigroup P^t acting on bounded measurable functions, its paths will be right continuous, and the process will be strong Markov. (These properties follow from the assumption that the processes are Borel right.) When working with discrete time Markov processes, we will assume the state space satisfies the same properties.

One wishes to formulate concepts that are the analogs of those for Markov chains, although there will be aspects not present in the countable state space theory. For $A \in \mathscr{S}$, let

$$\tau_A = \inf\{t \geq 0 : X(t) \in A\}, \quad \eta_A = \int_0^\infty \mathbf{1}\{X(t) \in A\}\, dt. \tag{4.7}$$

By the Debut Theorem (see e.g., [Sh88]), τ_A is a stopping time. A Markov process is said to be *φ-irreducible*, for a nontrivial σ-finite measure φ on $(S, \mathscr{S})$, if

$$\varphi(A) > 0 \quad \text{implies} \quad E_x[\eta_A] > 0 \quad \text{for all } x \in S; \tag{4.8}$$

φ is called an *irreducibility measure*.

If for some nontrivial σ-finite measure φ,

$$\varphi(A) > 0 \quad \text{implies} \quad P_x(\eta_A = \infty) = 1 \quad \text{for all } x \in S, \tag{4.9}$$

then $X(\cdot)$ is *Harris recurrent*. (This definition goes back to [AzKR67].) It is not difficult to show that it is equivalent to the condition that

$$\varphi(A) > 0 \quad \text{implies} \quad P_x(\tau_A < \infty) = 1 \quad \text{for all } x \in S \tag{4.10}$$

(see [KaM94] or [MeT93a]), although the choice of φ satisfying (4.10) need not satisfy (4.9). (Consider, for example, the process $X(t) = e^{it}$ on the unit circle, where φ is concentrated at a point.) Both formulations are useful in practice.

A σ-finite measure π on $(S, \mathscr{S})$ satisfying

$$\pi(A) = \pi P^t(A) \stackrel{\text{def}}{=} \int P^t(x, A)\, \pi(dx) \quad \text{for } A \in \mathscr{S},\ t \geq 0, \tag{4.11}$$

is *stationary* (or *invariant*). (Note that the definition does not involve φ.) It was shown in [Ge79] that if $X(\cdot)$ is Harris recurrent, then there is a unique stationary measure, up to a constant multiple. (We will discuss this result in Section 4.5.) If the stationary measure π is finite, it may be normalized to a probability measure. Harris recurrent processes with such π are *positive Harris recurrent*.

The reader should be aware that Harris recurrence and positive Harris recurrence have somewhat different implications than recurrence and positive recurrence, in the countable state space setting. For instance, if φ is concentrated at a point x, then a Markov chain can have x as an absorbing point and still be positive Harris recurrent. When all states communicate, the definitions are equivalent.

For discrete time Markov processes, φ-irreducibility, Harris recurrence, and positive Harris recurrence are defined by the analogs of (4.8), (4.9) and (4.11). (In this setting, (4.9) and (4.10) are clearly equivalent by the strong Markov property.) Since there is a wealth of theory available for such Markov processes, it is fruitful to be able to translate continuous time problems into the discrete time setting. This can be done by using the *resolvent* of the continuous time Markov process,

$$R(x, A) \stackrel{\text{def}}{=} \int_0^\infty e^{-t} P^t(x, A)\, dt \quad \text{for } x \in S,\ A \in \mathscr{S}. \tag{4.12}$$

The Markov process $\tilde{X}(n),\ n = 0, 1, 2, \ldots$, with one-step transition probability given by $R(\cdot, \cdot)$, is known as an *R-chain*. The R-chain $\tilde{X}(\cdot)$ can also be constructed directly from $X(\cdot)$ by setting

$$\tilde{X}(n) = X(\sigma_n), \quad n = 0, 1, 2, \ldots, \tag{4.13}$$

where the sequence $\sigma_0,\ \sigma_1, \sigma_2, \ldots$ of random variables, with $\sigma_0 = 0$, is independent of $X(\cdot)$ and has i.i.d. mean-1 exponentially distributed increments. (One enriches the sample space so as to include such a sequence.) It is easy to check that $\tilde{X}(\cdot)$ is φ-irreducible if and only if $X(\cdot)$ is. One can also show that the same is true for Harris recurrence and positive Harris recurrence, and that the same stationary measure is shared by both processes. The arguments are fairly quick although not immediate; they are given in Section 4.5.

Although the irreducibility measure for a given process $X(\cdot)$ is not unique, there exists a *maximal irreducibility measure* ψ, i.e., an irreducibility measure for the process such that $\varphi << \psi$ for any other irreducibility measure φ, and such that

$$\psi(\{x : P_x(\eta_A \neq 0) > 0\}) = 0 \tag{4.14}$$

for $\psi(A) = 0$ and $A \in \mathscr{S}$. ($\varphi << \psi$ means that φ is absolutely continuous with respect to ψ.) The measure ψ is equivalent to

$$\psi'(A) \stackrel{\text{def}}{=} \int_S R(x, A)\varphi'(dx) \quad \text{for } x \in S,\ A \in \mathscr{S}, \tag{4.15}$$

for any irreducibility measure φ'. (That is, $\psi << \psi'$ and $\psi' << \psi$.) Since the existence of a stationary measure π does not depend on the choice of φ, one is free to assume that φ is maximal when addressing such questions.

Maximal irreducibility measures are frequently used in discrete time, where (4.14) and (4.15) are replaced by

$$\psi(\{x : P_x(\tau_A < \infty) > 0\}) = 0 \tag{4.16}$$

and

$$\psi'(A) \stackrel{\text{def}}{=} \sum_{n=0}^{\infty} 2^{-(n+1)} \int_S P^n(x, A)\varphi'(dx) \quad \text{for } x \in S,\ A \in \mathscr{S}. \tag{4.17}$$

The existence of such a measure ψ and its equivalence to ψ' in the latter setting can be found on, e.g., page 88 of [MeT93d]. One can check that (4.14) and (4.15) hold for the process $X(\cdot)$ for a given irreducibility measure ψ if and only if (4.16) and (4.17) hold for its R-chain $\tilde{X}(\cdot)$, since the right side of (4.17) is equivalent to $\psi''(A) = \int_S P(x, A)\varphi'(dx)$ in this case.

The above definitions of Harris recurrent and positive Harris recurrent, while elegant, can be difficult to apply in practice. For applications, the following alternative formulation involving petite sets is very useful. A nonempty set $A \in \mathscr{S}$ is said to be *petite* if for some fixed probability measure a on $(0, \infty)$ and some nontrivial measure ν on $(S, \mathscr{S})$,

$$\nu(B) \leq \int_0^{\infty} P^t(x, B)a(dt) \tag{4.18}$$

for all $x \in A$ and all $B \in \mathscr{S}$; ν is then called a *petite measure*. A petite set A has the property that each set B is "equally accessible" from all points $x \in A$ with respect to the measure ν. Note that any nonempty measurable subset of a petite set is also petite. When (4.18) holds, with a being concentrated at a single point m_0, A is said to be *small*, and ν is called a *small measure*. Petite and small sets are defined analogously in the discrete time setting.

Let

$$\tau_A(\delta) = \inf\{t \geq \delta : X(t) \in A\}.$$

Theorem 4.1 below gives practical alternative characterizations of Harris recurrence and positive Harris recurrence in terms of petite sets. Versions of Theorem 4.1 are stated in [MeT93a-c], with that in [MeT93a] being used here. Discrete time analogs of the different parts of Theorem 4.1 are known. (See, e.g., [Or71], [Nu84], and [MeT93d].)

Theorem 4.1. *(a) A Markov process $X(\cdot)$ is Harris recurrent if and only if there exists a closed petite set A with*

$$P_x(\tau_A < \infty) = 1 \quad \textit{for all } x \in S. \tag{4.19}$$

(b) Suppose the Markov process $X(\cdot)$ is Harris recurrent. Then, $X(\cdot)$ is positive Harris recurrent if and only if there exists a closed petite set A such that for some $\delta > 0$ (or, equivalently, for any $\delta > 0$),

$$\sup_{x \in A} E_x[\tau_A(\delta)] < \infty. \tag{4.20}$$

We next make some general comments about Theorem 4.1. We then indicate how the theorem will be applied and discuss its proof. More detail on the proof will be supplied in Section 4.5.

We note that the irreducibility measure φ in (4.8) and the measure ν in (4.18) employed in the definitions of Harris recurrence and petite sets are different in general. In [MeT93a], petite sets rather than closed petite sets are employed for Harris recurrence, although closed petite sets are needed for positive Harris recurrence. We assume the sets are closed in both cases; this simplifies the proof of one of the steps. We note that the more useful direction (and the only one used in these lectures) is that Harris recurrence and positive Harris recurrence follow from (4.19) and (4.20), respectively. If one wishes, one can base the definition of Harris recurrence on (4.19), rather than on the irreducibility measure as in (4.9); this is done, for instance, in [As03] (and in [Dur96], in discrete time). This will simplify the work in showing the existence of a stationary measure.

Theorem 4.1 will prove very useful in conjunction with Section 4.2. There, we will show that for the underlying Markov process $X(\cdot)$ of a queueing network,

$$A = \{x : |x| \leq \kappa\} \quad \text{is a closed small set for each } \kappa > 0, \tag{4.21}$$

where $|x|$ is given by (4.3), if appropriate conditions hold for the distributions of the interarrival times for the queueing network. In Section 4.4, we will show that the conditions (4.19) and (4.20) in the theorem will follow from the stability of the associated fluid limits, which are introduced in Section 4.3.

We next briefly discuss the proof of Theorem 4.1. The demonstration that $X(\cdot)$ is Harris recurrent if (4.19) holds is elementary, if one sets $\varphi = \nu$. We summarize the argument here. Note that $X(\tau_A) \in A$, since A is closed. Starting from any $x \in S$ and applying the strong Markov property, one can therefore show that by a large enough fixed time T (depending on x), $X(\cdot)$ will, with at least a given positive probability that depends on A, hit a specified set B with $\nu(B) > 0$. Repetition of this reasoning, using appropriate random times T_n which depend on $X(T_{n-1})$, will imply that the probability B has not been hit after n iterations decreases exponentially quickly in n. This implies (4.10),

and hence that $X(\cdot)$ is Harris recurrent. ([MeT93a] gives a different argument that does not assume A is closed.)

The other direction in Part (a), and both directions in Part (b) of Theorem 4.1, require work. In Section 4.5, we will present a summary of the proofs. We will state there a discrete time analog of Theorem 4.1 and indicate how Theorem 4.1 can be shown using this, and the correspondence mentioned earlier between Harris recurrence, positive Harris recurrence, and the stationary measures for the Markov process and its R-chain. We will also provide a summary of the proof of the existence of a stationary measure for a discrete time recurrent Markov process, since it helps illustrate the nature of the discrete time theory.

Ergodicity

A continuous time Markov process $X(\cdot)$ is said to be *ergodic* if it possesses a stationary probability measure π for which

$$\lim_{t\to\infty} ||P^t(x,\cdot) - \pi(\cdot)|| = 0 \quad \text{for all } x \in S. \tag{4.22}$$

(Here, $||\cdot||$ denotes the total variation norm.) We recall from Section 1.2 that, when the underlying Markov process of a queueing network is ergodic, the queueing network is said to be *e-stable*. It is not difficult to see that positive Harris recurrence follows from ergodicity. Frequently, because of the following results, ergodicity of the Markov process also follows from positive Harris recurrence without much additional effort.

The first result is from Theorem 6.1 of [MeT93b].

Theorem 4.2. *Suppose that a Markov process $X(\cdot)$ is positive Harris recurrent. Then, $X(\cdot)$ is ergodic if and only if some skeleton chain is φ-irreducible for some measure φ.*

By a skeleton chain, we mean the Markov process defined by restricting $X(\cdot)$ to the times $n\Delta$, $n = 0, 1, 2, \ldots$, for some $\Delta > 0$. The necessity of irreducibility on some skeleton chain is clear. To show that this is also sufficient, one can employ the corresponding result for discrete time Markov processes and the fact that the norm of the difference in (4.22) is decreasing in t; we do not go into details here.

For our purposes, it will be more useful to have a sufficient condition for ergodicity in terms of small sets. When a set $A \in \mathscr{S}$ is small with respect to the same measure ν at each $m_0 \in [s_1, s_2]$, for some $0 < s_1 < s_2$, we will say that A is *uniformly small on* $[s_1, s_2]$, or, more briefly, *uniformly small*.

Theorem 4.3. *Suppose that a Markov process $X(\cdot)$ is positive Harris recurrent, and that some closed set A satisfies (4.19) and is uniformly small on $[s_1, s_2]$, for some $0 < s_1 < s_2$. Then, $X(\cdot)$ is ergodic.*

Theorem 4.3 follows from Theorem 4.2. One can show this by restarting $X(\cdot)$ at τ_A and applying the strong Markov property. Note that since A is closed, $X(\tau_A) \in A$. Theorem 4.3 is also proved in [As03] as Part (iii) of Proposition 3.8 on page 203, although in a somewhat different setting. Theorem 4.3 will be employed in Section 4.2.

Countable state space setting

As mentioned at the beginning of the section, the material that is needed from this section simplifies enormously in the countable state space setting. Since the interarrival and service times are exponentially distributed, one can drop the u and v coordinates from the description of the state of the process in the first subsection. The underlying jump process $X(\cdot)$ of the queueing network can be defined in the natural way. It will be constant in between arrivals and departures of jobs at the different classes, with arrivals occurring at rate α_k and departures at rate $R_k(t)\mu_k$ at class k. As before, the service rate vector $R(\cdot)$ remains constant until the next arrival or departure, at which time the new value $R(t) = f(X(t))$ is assigned, where the choice of f corresponds to the discipline. (The service rate vector is not part of the state space here.) In this setting, the norm in (4.3) is replaced by the simpler $|x| = |z|$, where z is the queue length vector. The state space is assigned the discrete topology.

The second subsection will no longer be needed, since $X(\cdot)$ is now a Markov chain, with bounded jump rates, and so standard Markov chain theory applies (see, e.g. [Re92]). For consistency with the general case, we choose its filtration to be $\{\mathcal{F}_t,\ t \geq 0\}$ as in (4.5), although (4.4) could also be used. In either case, the filtration will be right continuous and $X(\cdot)$ will be strong Markov.

The subsection on Harris recurrence can also be omitted. In the countable state space setting, Harris recurrence is equivalent to the existence of a recurrent state x that is accessible from all other states, i.e., $P_y(\tau_x < \infty) = 1$ for each $y \in S$ (where $\tau_x \stackrel{\text{def}}{=} \tau_{\{x\}}$). Positive Harris recurrence corresponds to $E_x[\tau_x(\delta)] < \infty$, for some $\delta > 0$. So, standard Markov chain theory can be applied. The concepts petite and small are no longer needed. They will be used later, in the general setting, only in the context of Proposition 4.6 for the sets $A = \{x : |x| \leq \kappa\},\ \kappa > 0$. These concepts are not needed in the countable state space setting since the state 0 will be uniformly accessible from A. We note that all points in a state space are small sets, and that Theorem 4.1 is elementary in the countable state space setting (in addition to no longer being needed).

In the countable state space, continuous time setting, it is routine to show that positive Harris recurrence implies ergodicity. So, Theorems 4.2 and 4.3 are no longer needed. (One can apply discrete time theory to any skeleton chain, which will be aperiodic, and then apply the monotonicity of $\|P^t(x,\cdot) - \pi(\cdot)\|$ in t.)

4.2 Results for Bounded Sets

Bounded sets will play an important role in Section 4.4 in showing the stability of queueing networks. In Section 4.2, we show two results involving bounded sets we will need there. Here and elsewhere in these lectures, a *norm* $|\cdot|$ denotes a nonnegative function on a state space S. A set $B \subseteq S$ is said to be *bounded* if $\sup_{x\in B} |x| < \infty$.

The first of these two results, Proposition 4.6, is a generalization of Foster's Criterion. As background, we first state Foster's Criterion, along with its proof. Both versions give useful criteria for positive Harris recurrence when the Markov process under consideration has a uniformly negative drift off of a bounded subset of the state space. In Proposition 4.6, we also give an analogous condition for ergodicity of the Markov process.

The second result, Proposition 4.7, gives criteria under which the bounded sets are uniformly small for the Markov process X underlying a queueing network. This condition is used in conjunction with Theorem 4.3 and Proposition 4.6 to show the Markov process is ergodic.

Foster's Criterion

Foster's Criterion is a simple, but very useful, criterion for demonstrating positive recurrence of Markov chains with a norm. We state it in the original discrete time context.

Proposition 4.4 (Foster's Criterion). *Let $X(n)$, $n = 0, 1, 2, \ldots$, be a Markov chain on which all states communicate. Suppose that*

$$E_x|X(1)| < \infty \quad \textit{for all } x \in A, \tag{4.23}$$

where $A = \{x : |x| \le \kappa\}$ for some $\kappa \ge 0$, and $|A| < \infty$. Also, suppose that for some $\epsilon > 0$,

$$E_x|X(1)| \le |x| - \epsilon \quad \textit{for all } x \notin A. \tag{4.24}$$

Then, $X(\cdot)$ is positive recurrent.

Foster's Criterion states, in essence, that if a Markov chain has a uniformly negative drift off of a bounded set, and if its behavior on the bounded set is not too bad, then the Markov chain will be positive recurrent. [Fo53] gave a computational proof when the state space is the nonnegative integers with the corresponding norm and $\kappa = 0$. Foster's Criterion is often employed in the above slightly more general setting of Proposition 4.4.

Proof of Proposition 4.4. Since all states communicate and $|A| < \infty$, in order to show that $X(\cdot)$ is positive recurrent, we claim it suffices to show $E_x[\tau_A] < \infty$ for $x \in A$. The claim is clear when A is a singleton. When A has more states, one can show it by noting that, by the strong Markov property, the expected number of returns to A grows linearly in n. The same is therefore true for at least one state in A, which must be positive recurrent.

To show that $E_x[\tau_A] < \infty$ for $x \in A$, we set

$$M(n) = |X(n)| + \epsilon n \quad \text{for all } n. \tag{4.25}$$

On account of (4.24), $M(n \wedge \tau_A)$ is a nonnegative supermartingale on the natural filtration of $X(\cdot)$. It follows by the Optional Sampling Theorem that

$$E_y[M(\tau_A)] \leq |y| \quad \text{for all } y \notin A.$$

It follows from this and (4.25) that

$$E_y[\tau_A] \leq \frac{1}{\epsilon} E_y[M(\tau_A)] \leq |y|/\epsilon \quad \text{for } y \notin A.$$

Together with (4.23), this implies that for $x \in A$,

$$E_x[\tau_A] \leq 1 + \sum_{y \notin A} p(x,y) E_y[\tau_A] \leq 1 + \frac{1}{\epsilon} \sum_{y \notin A} |y| p(x,y) < \infty,$$

where $p(\cdot,\cdot)$ denotes the one-step transitional probabilities. So, $E_x[\tau_A] < \infty$ for $x \in A$, and $X(\cdot)$ is positive recurrent. ■

In various cases, one might not know that (4.24) holds for a given Markov chain or Markov process, but rather that $|X(\cdot)|$ decreases linearly "on the average", over a much longer time interval. In our applications in Section 4.4, this will occur over time intervals of the form $[t, t + c|X(t)|]$, for large $|X(t)|$, where $c > 0$. To accommodate such a setting, we employ the following generalization of Foster's Criterion. Here, time is chosen to be continuous, although time could also be chosen to be discrete, if one omits the conclusion on ergodicity. If time is continuous and the state space is general, then $X(\cdot)$ is assumed to satisfy the regularity conditions for Markov processes given in Section 4.1, in the second paragraph of the subsection on Harris recurrence. Versions of Proposition 4.5 are given in [MaM81], [Fi89], [MeT94], and [FoK04].

Proposition 4.5 (Generalized Foster's Criterion). *Suppose that $X(\cdot)$ is a continuous time Markov process, such that for some $\epsilon > 0$, $\kappa > 0$, and measurable function $g : S \to \mathbf{R}$ with $g(x) \geq \delta > 0$,*

$$E_x|X(g(x))| \leq (|x| \vee \kappa) - \epsilon g(x) \quad \textit{for all } x. \tag{4.26}$$

Then,

$$E_x[\tau_A(\delta)] \leq \frac{1}{\epsilon}(|x| \vee \kappa) \quad \textit{for all } x, \tag{4.27}$$

where $A = \{x : |x| \leq \kappa\}$. In particular, if A is a closed petite set, then $X(\cdot)$ is positive Harris recurrent. If A is closed and is uniformly small on $[s_1, s_2]$, for some $0 < s_1 < s_2$, then $X(\cdot)$ is ergodic.

One can check that Proposition 4.4 is a special case of the discrete time version of Proposition 4.5, with $g \equiv 1$, since (4.23) and (4.24) follow from (4.26) after a new choice of κ, and since any finite set will be petite if all states communicate.

In these lectures, we will employ the following case of Proposition 4.5.

Proposition 4.6 (Multiplicative Foster's Criterion). *Suppose that $X(\cdot)$ is a continuous time Markov process, such that for some $c > 0$, $\epsilon > 0$, and $\kappa > 0$,*

$$E_x|X(c(|x| \vee \kappa))| \leq (1-\epsilon)(|x| \vee \kappa) \quad \textit{for all } x. \tag{4.28}$$

If $A = \{x : |x| \leq \kappa\}$ is a closed petite set, then $X(\cdot)$ is positive Harris recurrent. If A is closed and is uniformly small on $[s_1, s_2]$, for some $0 < s_1 < s_2$, then $X(\cdot)$ is ergodic.

Setting $g(x) = c(|x| \vee \kappa)$, Proposition 4.6 follows immediately from Proposition 4.5. The proof of Proposition 4.5 uses an elementary martingale argument together with Theorems 4.1 and 4.3 of the previous section. (This is the only place in these lectures where we will use Theorems 4.1 and 4.3.)

Proof of Proposition 4.5. Suppose that (4.27) holds. Then, clearly so does (4.19) of Theorem 4.1. Also by (4.27),

$$\sup_{x \in A} E_x[\tau_A(\delta)] \leq \kappa/\epsilon$$

must hold, and therefore so does (4.20). If A is assumed to be closed and petite, it therefore follows from both halves of Theorem 4.1 that $X(\cdot)$ is positive Harris recurrent. If A is assumed to be closed and uniformly small on some $[s_1, s_2], 0 < s_1 < s_2$, it follows from this and Theorem 4.3 that $X(\cdot)$ is ergodic. So, it suffices to show that (4.27) holds.

Set $\sigma_0 = 0$, and let $\sigma_1, \sigma_2, \ldots$ denote the stopping times defined inductively by

$$\sigma_n = \sigma_{n-1} + g(X(\sigma_{n-1})).$$

By (4.26) and the strong Markov property,

$$E_x[|X(\sigma_n)| \mid \mathcal{F}(\sigma_{n-1})] \leq (|X(\sigma_{n-1})| \vee \kappa) - \epsilon g(X(\sigma_{n-1}))$$

for all x, where $\mathcal{F}(T) \stackrel{\text{def}}{=} \mathcal{F}_T$ is the σ-algebra corresponding to the stopping time T. Set $M(0) = |x| \vee \kappa$ and

$$M(n) = |X(\sigma_n)| + \epsilon\sigma_n \quad \text{for } n \geq 1. \tag{4.29}$$

Also, set $\mathcal{G}_n = \mathcal{F}(\sigma_n)$ and note that $\sigma_n \in \mathcal{G}_{n-1}$. One can check that

$$E_x[M(n) \mid \mathcal{G}_{n-1}] \leq M(n-1) \quad \text{for } n \leq \rho,$$

where ρ is the first time $n > 0$ at which $M(n) \in A$. So, $M(n \wedge \rho)$ is a nonnegative supermartingale on $\mathcal{G}_n$.

It follows by the Optional Sampling Theorem that

$$E_x[M(\rho)] \leq |x| \vee \kappa.$$

Note that $\tau_A(\delta) \leq \sigma_\rho$. Therefore, by (4.29) and the above inequality,

$$\epsilon E_x[\tau_A(\delta)] \leq E_x[M(\rho)] \leq |x| \vee \kappa,$$

which implies (4.27), as desired. ■

Criteria for bounded sets to be petite or uniformly small

As mentioned earlier, we will employ Proposition 4.6 in Section 4.4 to establish criteria for when the Markov process $X(\cdot)$ underlying a queueing network is positive Harris recurrent or is ergodic. In order to cite Proposition 4.6, we need to be able to show bounded sets are petite or uniformly small. These conditions will not automatically hold, since states need not "communicate" in general. For instance, when the distributions of the interarrival times, at two classes k_1 and k_2 of a queueing network, are both integer valued, states x and x' for which the residual interarrival times satisfy

$$u_{k_2} - u_{k_1} \neq (u'_{k_2} - u'_{k_1}) \bmod 1$$

cannot both be visited along the same sample path.

In order to rule out such behavior, the following two conditions on the distributions of the interarrival times $\xi_k(1),\ k \in \mathcal{A}$, are often assumed. The first is that $\xi_k(1)$ is unbounded, that is, for each $k \in \mathcal{A}$,

$$P(\xi_k(1) \geq t) > 0 \quad \text{for all } t. \tag{4.30}$$

The second is that for some $\ell_k \in \mathbf{Z}_+$, the ℓ_k-fold convolution of $\xi_k(1)$ and Lebesgue measure are not mutually singular. That is, for $k \in \mathcal{A}$ and some nonnegative $q_k(\cdot)$ with $\int_0^\infty q_k(t)\,dt > 0$,

$$P(\xi_k(1) + \ldots + \xi_k(\ell_k) \in [c,d]) \geq \int_c^d q_k(t)dt \tag{4.31}$$

for all $c < d$. When arrivals in the network are permitted at only one class, e.g., as in reentrant lines, it is enough to assume just (4.30) to show bounded sets are petite.

It is annoying to need to assume either condition, especially the first, since they rule out reasonable distributions for which one should expect the underlying Markov process to be positive Harris recurrent. It appears difficult, however, to formulate simple criteria that are robust over networks with general routing structures and disciplines. For interarrival times not satisfying (4.30) and (4.31), one needs to show the existence of a petite or uniformly small set directly.

Proposition 4.7 is the main result in this subsection. It states that when (4.30) and (4.31) are satisfied for the interarrival times of a queueing network, then bounded sets will be uniformly small. It follows immediately from this that bounded sets are also petite. (No requirements are made on the service times.) A related result is given in Lemma 3.7 of [MeD94].

Proposition 4.7. *Assume that the interarrival times of an HL queueing network satisfy (4.30) and (4.31). Then, for each $\kappa > 0$, the set $A = \{x : |x| \leq \kappa\}$ is uniformly small on $[s_1, s_2]$ for some $0 < s_1 < s_2$.*

When arrivals in the network are permitted at only one class, one can instead use the following result to show bounded sets are petite.

Proposition 4.8. *Assume that the interarrival times of an HL queueing network, with $|\mathcal{A}| = 1$, satisfy (4.30). Then, for each $\kappa > 0$, the set $A = \{x : |x| \leq \kappa\}$ is petite.*

Since $|x|$ is continuous in x, the above sets A are closed. Once (4.28) has been verified, Proposition 4.8 can therefore be used in conjunction with Proposition 4.6 to show that the underlying Markov process $X(\cdot)$ of the queueing network is positive Harris recurrent, and hence that the queueing network is stable. Similarly, Proposition 4.7 can be used with the proposition to show $X(\cdot)$ is ergodic, and hence that the queueing network is e-stable.

We will prove Proposition 4.7, which requires some effort. The argument for Proposition 4.8 is simpler, an outline of which goes as follows. Choose $t_1 > 0$ and $\epsilon > 0$, so that for all $|x| \leq \kappa$, with κ fixed, the probability is at least ϵ that the queueing network will be empty over some interval $[t(x) - 1, t(x))$ but not remain empty over the entire interval $[t(x), t_1)$, where $t(x) \in [1, t_1]$. It is possible to do this because of (4.30). One can then show that the bounded set A will be petite, by choosing the probability measure a in (4.18) to be uniform over $(0, t_1)$, and choosing the petite measure ν to be uniform over the empty states of the queueing network with residual interarrival times in $(0, 1)$, so that its density is ϵ/t_1 with respect to this set. Details are similar to parts of the argument for the more complicated construction in the proof of Proposition 4.7.

Before beginning the proof of Proposition 4.7, we provide a summary of the argument and introduce some notation. Let $L_0 = \max_{k \in \mathcal{A}} \ell_k$, where ℓ_k is as in (4.31). Then, the distribution of $\Xi_k(L_0) = \sum_{i=1}^{L_0} \xi_k(i)$ has an absolutely continuous component for all k. If L is chosen large enough, then, for each k, the distribution of $\Xi_k(L)$ will uniformly cover some interval of length $\kappa+3$, say $[a_k - \kappa, a_k + 3]$, where κ is chosen as in the proposition. On account of (4.30), $\xi_k(L+1)$ can take arbitrarily large values, which we specify to be in $[b_k, b_k+1]$ for some b_k, with $b_k \geq N$ for some large N. It follows that, for $u_k \leq \kappa$, the distribution of $u_k + \Xi_k(L+1)$ uniformly covers $[a_k + b_k + 1, a_k + b_k + 3]$, with no external arrivals at k occurring over the long time period $[\bar{a}, N)$, where u_k is the initial residual interarrival time and $\bar{a} = \max_k\{a_k + 3\}$.

For large enough N, the network will be empty with positive probability by time $N/2$, and hence remain empty until time N, with the probability not depending on the initial state x, for $|x| \leq \kappa$. When a state y is empty, it is specified by its residual interarrival time vector u. The state is empty, with positive probability, at the times $s \in [N/2, N/2+1]$, and at these times, each coordinate u_k will have an absolutely continuous component covering

$$\mathcal{J}_k = [a_k + b_k + 1 - N/2, a_k + b_k + 2 - N/2].$$

These bounds do not depend on the initial state x, for $|x| \leq \kappa$. The set $A = \{x : |x| \leq \kappa\}$ will therefore be uniformly small on $[N/2, N/2+1]$, with the small measure ν in (4.18) being uniformly distributed over the empty states y with residual service times in the Cartesian product of $\mathcal{J}_k$, for $k \in \mathcal{A}$.

We proceed to demonstrate Proposition 4.7 along the lines outlined in the last two paragraphs. On account of (4.31) we may choose L large enough so that for some $\epsilon_1 > 0$ and $a_k > \kappa$,

$$P(\Xi_k(L) \in [t_1, t_2]) \geq \epsilon_1(t_2 - t_1) \quad \text{for } [t_1, t_2] \subseteq [a_k - \kappa, a_k + 3],$$

for all $k \in \mathcal{A}$. That is, $\Xi_k(L)$ has density at least ϵ_1 at all times in the interval $[a_k - \kappa, a_k + 3]$. For $|x| \leq \kappa$ (and hence $u_k \leq \kappa$), this implies

$$P(u_k + \Xi_k(L) \in [t_1, t_2]) \geq \epsilon_1(t_2 - t_1) \quad \text{for } [t_1, t_2] \subseteq [a_k, a_k + 3]. \tag{4.32}$$

Also, by (4.30), for any N, there exist times $b_k \geq N$ so that

$$P(\xi_k(L+1) \in [b_k, b_k + 1]) \geq \epsilon_2 \tag{4.33}$$

for some $\epsilon_2 > 0$; we will specify N later. We introduce the following terminology, setting

$$\begin{aligned}
&G_{1,k} = \{\omega : u_k + \Xi_k(L) \in [a_k, a_k + 3]\},\\
&G_{2,k}(t_{1,k}, t_{2,k}) = \{\omega : u_k + \Xi_k(L+1) \in [t_{1,k}, t_{2,k}]\},\\
&G_1 = \bigcap_{k \in \mathcal{A}} G_{1,k}, \quad G_2(\mathbf{t}_1, \mathbf{t}_2) = \bigcap_{k \in \mathcal{A}} G_{2,k}(t_{1,k}, t_{2,k}),\\
&G = G_1 \cap G_2(\mathbf{t}_1, \mathbf{t}_2),
\end{aligned}$$

where $\mathbf{t}_i = (t_{i,k},\ k \in \mathcal{A})$. Also, set $\mathcal{I}_k = [a_k + b_k + 1, a_k + b_k + 3]$.

We break most of the work in proving Proposition 4.7 into two lemmas. The first gives the following lower bound on $P(G)$, for $\mathbf{t}_i$ having coordinates $t_{i,k} \in \mathcal{I}_k$.

Lemma 4.9. *For given $\mathbf{t}_i$, $i = 1, 2$, with $t_{1,k} \leq t_{2,k}$ and $t_{i,k} \in \mathcal{I}_k$, for $k \in \mathcal{A}$,*

$$P(G) \geq (\epsilon_1 \epsilon_2)^{|\mathcal{A}|} \prod_{k \in \mathcal{A}} (t_{2,k} - t_{1,k}). \tag{4.34}$$

Proof. One has

$$\begin{aligned} P(G) &= P(G_1 \cap G_2(\mathbf{t}_1, \mathbf{t}_2)) = \prod_{k \in \mathcal{A}} P(G_{1,k} \cap G_{2,k}(t_{1,k}, t_{2,k})) \\ &\geq \prod_{k \in \mathcal{A}} \int_{b_k}^{b_k+1} P(u_k + \Xi_k(L) \in [t_{1,k} - s, t_{2,k} - s]) P(\xi_k(L+1) \in ds) \\ &\geq (\epsilon_1 \epsilon_2)^{|\mathcal{A}|} \prod_{k \in \mathcal{A}} (t_{2,k} - t_{1,k}). \end{aligned}$$

The second equality follows from the independence of the interarrival times $\xi_k(i)$ for different k. The first inequality is gotten by writing $\Xi_k(L+1)$ as a convolution of $\Xi_k(L)$ with $\xi_k(L+1)$, and noting that for $s \in [b_k, b_k + 1]$, $G_{1,k}$ occurs when $u_k + \Xi_k(L)$ is contained in $[t_{1,k} - s, t_{2,k} - s]$. The second inequality follows from the bounds in (4.32) and (4.33). ■

Let

$$\sigma = \inf\{t \geq \bar{a} : Z(t) = 0\},$$

where $\bar{a} = \max_k\{a_k + 3\}$ and $Z_k(t)$ is the number of jobs at class k at time t. The next result says that, given the event G, there is a uniform upper bound on σ that does not depend on the initial state x.

Lemma 4.10. *For given L and κ, and large enough N, there exists $\epsilon_3 > 0$ so that*

$$P_x(\sigma \leq N/2 \mid G) \geq \epsilon_3 \tag{4.35}$$

for all $|x| \leq \kappa$, and $\mathbf{t}_i$, $i = 1, 2$, with $t_{1,k} < t_{2,k}$ and $t_{i,k} \in \mathcal{I}_k$.

Proof. The reasoning behind (4.35) is not difficult. Since the notation that is involved can become cumbersome, we avoid it as much as possible, and argue in terms of basic queueing quantities.

We first note that for large enough M and small enough $\delta > 0$, for any class k, (a) There exist classes $k_1, k_2, \ldots, k_n$, with $k_1 = k$, $n \leq M$, and

$$\left(1 - \sum_{\ell} P_{k_n,\ell}\right) \prod_{i=1}^{n-1} P_{k_i,k_{i+1}} \geq \delta.$$

That is, a job starting at any k has positive probability δ of following a designated route and leaving the network in at most M steps. (b) $P(\gamma_k(1) \leq M) \geq 1/2$ for all k. That is, there is a uniform bound on the service time distributions. Let ζ be the total service time required by an arbitrary job that is either initially in the network or later enters it. In the former case, we know that its residual service time is at most κ. Therefore, by (a) and (b), and the independence of the corresponding events,

$$P(\zeta \leq M^2 + \kappa) \geq \delta 2^{-M}. \tag{4.36}$$

On the event G, (c) No jobs enter the network over $(\bar{a}, N)$. (d) At most $|\mathcal{A}|L$ customers enter the network over $(0, N)$, for a total of Λ jobs in the network up to time N, with

$$\Lambda \le |\mathcal{A}|L + |z| \le |\mathcal{A}|L + \kappa \stackrel{\text{def}}{=} L'.$$

(c) and (d) follow from the definitions of $G_1, b_k, \mathcal{I}_k$, and G_2. Let $\zeta_1, \zeta_2, \ldots, \zeta_\Lambda$ denote the total service times of these Λ jobs. By (4.36),

$$P_x\left(\sum_{\ell=1}^{\Lambda} \zeta_\ell \le L'(M^2 + \kappa)\right) \ge (\delta 2^{-M})^{L'}. \tag{4.37}$$

We now set

$$N = 4(L'(M^2 + \kappa) \vee \bar{a}) \quad \text{and} \quad \epsilon_3 = (\delta 2^{-M})^{L'}.$$

Then, $(N/4, N/2] \subseteq (\bar{a}, N/2]$, and so under the event in (4.37) and (c), $\sigma \le N/2$. Together with (4.37), this implies (4.35). ∎

Proposition 4.7 follows from Lemmas 4.9 and 4.10.

Proof of Proposition 4.7. The bounds in (4.34) and (4.35) imply that

$$P_x(\sigma \le N/2;\ G) \ge (\epsilon_1\epsilon_2)^{|\mathcal{A}|}\epsilon_3 \prod_{k\in\mathcal{A}} (t_{2,k} - t_{1,k}) \tag{4.38}$$

for $|x| \le \kappa$, and $\mathbf{t}_i$, $i = 1, 2$, chosen so that $t_{1,k} < t_{2,k}$ and $t_{i,k} \in \mathcal{I}_k$. For $s \in [N/2, N/2 + 1]$, it follows that

$$\begin{aligned} P_x(Z(s) = 0 \text{ and } U_k(s) \in [t_{1,k} - s, t_{2,k} - s],\ k \in \mathcal{A}) \\ \ge (\epsilon_1\epsilon_2)^{|\mathcal{A}|}\epsilon_3 \prod_{k\in\mathcal{A}} (t_{2,k} - t_{1,k}), \end{aligned} \tag{4.39}$$

since the event in (4.39) contains the event (4.38). To see this, note that if the network is empty at time $\sigma \le N/2$, it will, on G, remain empty until at least time $N \le \min_{k\in\mathcal{A}}\{a_k + b_k\}$. At the intermediate times s, the residual interarrival times will be the shifts, by s, of the times at which the events $u_k + \Xi_k(L + 1)$, given in $G_{2,k}(t_{1,k}, t_{2,k})$, occur.

When a state y is empty, it is specified by its residual interarrival time vector $u = \{u_k,\ k \in \mathcal{A}\}$. This is the case at time s for the event on the left side of (4.39). The inequality (4.39) states that, for $|x| \le \kappa$, the distribution of the residual time $U(s)$, for $s \in [N/2, N/2 + 1]$, has a component that is absolutely continuous, with density $\epsilon = (\epsilon_1\epsilon_2)^{|\mathcal{A}|}\epsilon_3$, with respect to $|\mathcal{A}|$-dimensional Lebesque measure λ that is restricted to the rectangle

$$\prod_{k\in\mathcal{A}} [a_k + b_k + 1 - N/2, a_k + b_k + 2 - N/2] \subseteq \prod_{k\in\mathcal{A}} [a_k + b_k + 1 - s, a_k + b_k + 3 - s].$$

It follows that, for all $s \in [N/2, N/2+1]$, the set A is small with respect to the measure $\nu = \epsilon\lambda$. ■

Countable state space setting

The work required in this section simplifies considerably in the countable state space setting. One still needs to demonstrate Proposition 4.5. The proof of (4.27) proceeds as before. In this setting, one can conclude directly from (4.27) that $X(\cdot)$ is ergodic (or is positive Harris recurrent), if $|A| < \infty$, by using

$$\inf_{x \in A} P_x(\tau_0 \leq 1) > 0;$$

the expected number of returns to the empty state 0 therefore grows linearly in t, which implies 0 is positive recurrent. (As before, $A = \{x : |x| \leq \kappa\}$.) So, neither petite nor uniformly small is needed as an assumption for the proposition. Since Proposition 4.6 is a direct consequence of Proposition 4.5, the same is also true there.

The concepts petite and uniformly small will be used later only in the context of Proposition 4.6. This, in particular, makes Propositions 4.7 and 4.8 unnecessary. The propositions are easy to show, however, since

$$\inf_{x \in A} P_x(X(s) = 0) \geq \epsilon(s)$$

for appropriate $\epsilon(s) > 0$, where $\epsilon(s)$ is uniformly bounded away from 0 on $[s_1, s_2]$, for $0 < s_1 < s_2$.

4.3 Fluid Models and Fluid Limits

In Section 1.3, we discussed fluid models and their connection with queueing network equations. The purpose there was to give a preview of these concepts. We return now to this material, this time giving a thorough presentation. The section consists of three subsections, covering queueing network equations, fluid models, and fluid limits, as well as a short comment on the countable state space setting.

Fluid models and fluid limits are studied in [Da95]. Modifications are given in [Ch95], [DaM95], [Br98a], and [Br98b] among other places. The approach taken here is closest to [Br98a], but with some further modification in the approach and in some of the definitions.

The fluid models of main interest to us will be subcritical. Fluid models are also an important tool in the study of heavy traffic limits, where the fluid models that are employed are critical; we will not discuss fluid models in the latter context here. (See [Br98b], [Wi98], and [BrD01] for background and further references.)

Queueing network equations

In the construction, in Section 4.1, of the Markov process $X(\cdot)$ underlying an HL queueing network, we introduced sequences of positive i.i.d. random variables $\xi_k(i), k \in \mathcal{A}$, and $\gamma_k(i)$, $k = 1, \ldots, K$, with $i = 1, 2, 3, \ldots$, which correspond to the interarrival and service times of the queueing network. We also introduced the sequence of i.i.d. random vectors $\phi^k(i)$, $i = 1, 2, 3, \ldots$, which give the routing of a job upon completion of its service at a class. The corresponding sequences ξ, γ, and ϕ were assumed to be mutually independent. Here, we will find it convenient to also denote by $\xi_k(0)$ and $\gamma_k(0)$ the initial residual interarrival and service times of the queueing network; they are included in the initial state $X(0) = x$.

We will employ the random quantities $E(\cdot)$, $\mathit{\Gamma}(\cdot)$, and $\mathit{\Phi}(\cdot)$, which are defined in terms of the partial sums of ξ, γ, and ϕ. The *external arrival process* $E(t) = \{E_k(t),\ k = 1, \ldots, K\}$, $t \geq 0$, counts the number of arrivals at each class from outside the network. That is, for $k \in \mathcal{A}$,

$$E_k(t) = \max\{n : \mathit{\Xi}_k(n) \leq t\},$$

where

$$\mathit{\Xi}_k(n) = \sum_{i=0}^{n-1} \xi_k(i).$$

The *cumulative service time process* $\mathit{\Gamma}(n) = \{\mathit{\Gamma}_k(n_k),\ k = 1, \ldots, K\}$, $n = (n_1, \ldots, n_K)$ with $n_k = 1, 2, \ldots$, is given by

$$\mathit{\Gamma}_k(n_k) = \sum_{i=0}^{n_k-1} \gamma_k(i).$$

The *routing process* $\mathit{\Phi}(n) = \{\mathit{\Phi}^k(n),\ k = 1, \ldots, K\},\ n = 1, 2, \ldots$, is given by

$$\mathit{\Phi}^k(n) = \sum_{i=1}^{n} \phi^k(i).$$

As mentioned in Section 4.1, the sequences ξ, γ, and ϕ, together with the initial state x and the discipline rule, determine the evolution of the process $X(\cdot)$ for all times along each sample path. The same is therefore true for $(E(\cdot), \mathit{\Gamma}(\cdot), \mathit{\Phi}(\cdot))$, which is referred to as the *primitive triple* of the queueing network.

As in Section 1.2, we will employ the means α_k m_k, and $P_{k,\ell}$ that are defined from ξ, γ, and ϕ. They are given by

$$\alpha_k = 1/E[\xi_k(1)] \ \text{ for } k \in \mathcal{A}, \quad m_k = E[\gamma_k(1)] \ \text{ for } k = 1, \ldots, K,$$

$$P_{k,\ell} = P(\phi^k(1) = e_\ell),$$

with $\alpha = \{\alpha_k,\ k = 1, \ldots, K\}$ being the *external arrival rate*, M being the diagonal matrix having the *mean service times* m_k at its diagonal entries, and

$P = \{P_{k,\ell},\ k,\ell = 1,\ldots,K\}$ being the *mean transition matrix* (or *mean routing matrix*). As before, $\mu_k = 1/m_k$ is the *service rate.* Throughout these lectures, we will implicitly assume that $E[\xi_k(1)] < \infty$ for $k \in \mathcal{A}$ and $E[\gamma_k(1)] < \infty$ for $k = 1,\ldots,K$, and so $\alpha_k, m_k, \mu_k \in (0,\infty)$. As in (1.2), $Q \stackrel{\text{def}}{=} (I - P^T)^{-1} = \sum_{n=0}^{\infty}(P^T)^n$, which is finite since the network is open. Also, the *total arrival rate* $\lambda = Q\alpha$ and the *traffic intensity* ρ, with $\rho_j = \sum_{k\in\mathcal{C}(j)} m_k\lambda_k$, are defined as in (1.5) and (1.7).

Queueing network equations tie together random vectors that describe the evolution of a given queueing network. Examples of such equations, with the vectors $A(t), D(t), T(t)$, and $Z(t)$, were given in Section 1.3. In the present more general setting, it will be more convenient to employ the 6-tuple

$$\mathfrak{X}(t) = (A(t), D(t), T(t), W(t), Y(t), Z(t)). \tag{4.40}$$

Here, the vector $W(t) = (W_1(t),\ldots,W_J(t))$ is the *immediate workload.* That is, $W_j(t)$ is the amount of time required to serve all jobs currently at station j, $j = 1,\ldots,J$, if all jobs arriving after time t are ignored. The vector $Y(t) = (Y_1(t),\ldots,Y_J(t))$ is the *cumulative idle time*, that is, the cumulative time that each of the stations $j = 1,\ldots,J$ is not working. Note that $A(t), D(t), T(t)$, and $Z(t)$ are class-level vectors, whereas $W(t)$ and $Y(t)$ are station-level vectors. From our perspective, $\mathfrak{X}(\cdot)$ contains all of the essential information on the evolution of the queueing network; it will be used as the starting point for our computations. With a slight abuse of notation, we will refer to $\mathfrak{X}(\cdot)$ as the *queueing network process.*

We note that $T(\cdot)$ and $Y(\cdot)$ are continuous and that $A(\cdot)$, $D(\cdot)$, $W(\cdot)$, and $Z(\cdot)$ are right continuous with left limits. All of the variables are nonnegative in each component, with $A(\cdot), D(\cdot), T(\cdot)$, and $Y(\cdot)$ being nondecreasing. By assumption, one has

$$A(0) = D(0) = T(0) = 0 \quad \text{and} \quad Y(0) = 0. \tag{4.41}$$

One can check that the components of $\mathfrak{X}(\cdot)$ satisfy the queueing network equations

$$A(t) = E(t) + \sum_k \Phi^k(D_k(t)), \tag{4.42}$$

$$Z(t) = Z(0) + A(t) - D(t), \tag{4.43}$$

$$W(t) = C\Gamma(A(t) + Z(0)) - CT(t), \tag{4.44}$$

$$CT(t) + Y(t) = et, \tag{4.45}$$

$$Y_j(t) \text{ can only increase when } W_j(t) = 0, \quad j = 1,\ldots,J, \tag{4.46}$$

for all $t \geq 0$. Here, C is the *constituency matrix*,

$$C_{j,k} = \begin{cases} 1 & \text{if } k \in \mathcal{C}(j), \\ 0 & \text{otherwise,} \end{cases}$$

and $e = (1, \ldots, 1)^T$.

The equations (4.42)-(4.46) are not difficult to verify. Equations (4.42) and (4.43) are the same as (1.8) and (1.9), and hold for the same reasons as before. The equality (4.44) states that the amount of current work at each station is equal to the sum of the cumulative amount of work having arrived at all of its classes less the sum of the cumulative service rendered at these classes. Equation (4.45) can be taken as the defining relation for the idletime $Y(t)$. In (4.46), we mean that $Y_j(t_2) > Y_j(t_1)$ implies $W_j(t) = 0$ for some $t \in [t_1, t_2]$, which reflects the nonidling property. Since $Y(\cdot)$ is continuous, it can also be written as

$$\int_0^\infty W_j(t) dY_j(t) = 0, \quad j = 1, \ldots, J.$$

The equations (4.42)-(4.46) hold for all disciplines. One can check that HL queueing networks also satisfy

$$\Gamma(D(t)) \leq T(t) < \Gamma(D(t) + e), \tag{4.47}$$

where the inequalities are componentwise and e denotes the K-vector of all 1's. (Whether e denotes a K-vector or J-vector will be clear from the context.) The equations (4.42)-(4.47) will be referred to as the *basic queueing network equations*.

Equations (4.42)-(4.47) do not specify the discipline of the queueing network. For multiclass queueing networks, there is consequently not enough information to solve for $\mathfrak{X}(\cdot)$. Later, when working with specific examples, an additional equation (or equations) will be introduced that correspond to the discipline. Such an equation will be referred to as an *auxiliary queueing network equation*. For example, for FIFO networks, this additional equation is

$$D_k(t + W_j(t)) = Z_k(0) + A_k(t) \quad \text{for } k = 1, \ldots, K. \tag{4.48}$$

For SBP networks with preemption, the equation is

$$t - T_k^+(t) \text{ can only increase when } Z_k^+(t) = 0 \quad \text{for } k = 1, \ldots, K. \tag{4.49}$$

Here, $Z_k^+(t)$ denotes the sum of the queue lengths at the station $j = s(k)$ of classes having priority at least as great as k, and $T_k^+(t)$ denotes the corresponding sum of cumulative service times. The order of the priorities is given by the specific SBP discipline.

As mentioned in Section 1.3, there is some flexibility in the choice of the components of $\mathfrak{X}(\cdot)$ and the corresponding queueing network equations. For specific disciplines, one typically eliminates one or more of these components. For instance, for HL queueing networks, it is generally not necessary to employ both $D(\cdot)$ and $T(\cdot)$. Since different variables will be more natural in different settings, we employ the flexible formulation given above.

The discerning reader might note that in Section 1.3, we employed equation (1.10) rather than (4.47) to relate $D(t)$ and $T(t)$. This has the advantage of

leading to the formula in (1.11), but does not incorporate the HL property. Both formats lead to the same fluid model equation (4.55) given below, if the HL property is implicitly assumed in conjunction with (1.10).

Fluid models

Fluid model equations were discussed in Section 1.3. They are the deterministic analog of queueing network equations, with the random quantities $E(\cdot), \Gamma(\cdot)$, and $\Phi(\cdot)$ being replaced by their respective means α, M, and P. The fluid model equations corresponding to (4.42)-(4.46) are

$$A(t) = \alpha t + P^T D(t), \tag{4.50}$$

$$Z(t) = Z(0) + A(t) - D(t) \tag{4.51}$$

$$W(t) = CM(A(t) + Z(0)) - CT(t), \tag{4.52}$$

$$CT(t) + Y(t) = et, \tag{4.53}$$

$$Y_j(t) \text{ can only increase when } W_j(t) = 0,\ j = 1, \ldots, J, \tag{4.54}$$

for all $t \geq 0$. In the HL setting, one includes

$$T(t) = MD(t), \tag{4.55}$$

which corresponds to (4.47). For a given choice of α, M, and P, the fluid model equations (4.50)-(4.55) will be referred to as the *basic fluid model equations*. Similarly, the fluid model consisting of the equations (4.50)-(4.55) will be referred to as the *basic fluid model*.

Equations (4.50)-(4.55) do not specify the discipline of the corresponding queueing network. So, as was the case for the queueing network equations, an additional fluid model equation (or equations) still needs to be added. Such an equation will be a deterministic expression involving $A(\cdot)$, $D(\cdot)$, $T(\cdot)$, $W(\cdot)$, $Y(\cdot)$, and $Z(\cdot)$, and will be referred to as an *auxiliary fluid model equation*. For networks with FIFO and SBP disciplines, the auxiliary fluid model equations are given by (4.48) and (4.49); two examples involving particular SBP networks will be given shortly. In these lectures, a *fluid model* will be a set of fluid model equations that includes the basic fluid model equations (4.50)-(4.55). Solutions of such equations are *fluid model solutions*. In Section 1.3, we were a bit vague on what is meant by a fluid model corresponding to a queueing network. We will make this precise in the next subsection where fluid limits are introduced.

The same notation was used in equations (4.50)-(4.55) as in (4.42)-(4.47) for the unknown variables $A(\cdot)$, $D(\cdot)$, $T(\cdot)$, $W(\cdot)$, $Y(\cdot)$, and $Z(\cdot)$. When convenient, we will employ the same vocabulary for the fluid model analogs of queueing network quantities, such as the immediate workload $W(\cdot)$ and the queue length $Z(\cdot)$. We will employ $\mathfrak{X}(t)$, given in (4.40), for solutions of fluid models of (4.50)-(4.55) and the auxiliary equations that may be added. The use of the same notation for the queueing network and fluid model variables is in general helpful; the one that is meant will be clear from the context.

Equations such as (4.50)-(4.55) are also referred to as *fluid model equations without delay*, since one is, in effect, setting $u = v = 0$ here, where u and v are the residual interarrival and service times of the initial state x that were introduced in Section 4.1. For general residual times, the corresponding equations are referred to as *fluid model equations with delay*. In that setting, one needs to modify (4.50), (4.52) and (4.55). The resulting equations are a bit awkward to work with. We will not require these more general equations here, and unless indicated to the contrary, the fluid model equations considered here will always be assumed to be without delay.

We will assume that all of the components of $\mathfrak{X}(\cdot)$ are nonnegative, with $A(\cdot)$, $D(\cdot)$, $T(\cdot)$, and $Y(\cdot)$ being nondecreasing. Using (4.50)-(4.53), one can check that

$$A(0) = D(0) = T(0) = 0 \quad \text{and} \quad Y(0) = 0, \tag{4.56}$$

which is the analog of (4.41). Employing (4.51), (4.52), and (4.55), one can also show the useful relationship between the queue length and immediate workload vectors,

$$W(t) = CMZ(t) \quad \text{for all } t. \tag{4.57}$$

Using the basic fluid model equations, it is not difficult to check that knowledge of any of $D(t)$, $T(t)$, or $Z(t)$, at a given t, and knowledge of $Z(0)$ are enough to determine all of the components of $\mathfrak{X}(t)$. Simple examples show this is not true for either $A(t), W(t)$, or $Y(t)$.

Starting first with $T(\cdot)$ and $Y(\cdot)$ in (4.53), it is easy to show that $A(\cdot), D(\cdot), T(\cdot), W(\cdot), Y(\cdot)$, and $Z(\cdot)$ are all Lipschitz continuous. That is, for some $N > 0$ (depending on the triple (α, M, P)),

$$|f(t_2) - f(t_1)| \leq N|t_2 - t_1| \quad \text{for all } t_1, t_2 \geq 0, \tag{4.58}$$

if $f(\cdot)$ is any of the above functions. (Recall that, when dealing with vectors, we always employ the sum norm, although this is a matter of convenience.) Consequently, these functions are absolutely continuous, and so $f'(t)$ exists a.e., with

$$f(b) - f(a) = \int_a^b f'(t)dt \quad \text{for all } a, b. \tag{4.59}$$

Times at which the derivative exists for all of the components of $\mathfrak{X}(\cdot)$ will be referred to as *regular points*.

The representation in (4.59) will be quite useful later on. Assume, for instance, that the dot product $(Z'(t), w) \leq -\epsilon$, for some fixed $\epsilon > 0$ and fixed vector w with nonnegative coordinates, whenever $Z(t) \neq 0$ and t is a regular point. Then, it is not difficult to see, using (4.59), that

$$Z(t) = 0 \quad \text{for } t \geq (Z(0), w)/\epsilon. \tag{4.60}$$

In analogy with queueing networks, one can envision the components of $\mathfrak{X}(\cdot)$ for fluid models in terms of continuous "job mass" flowing through the

system. Also in analogy with queueing networks, for the prescribed triple (α, M, P), stations are defined as *subcritical* or *critical*, if $\rho_j < 1$ or $\rho_j = 1$, where ρ_j is given by (1.7), for Q and λ defined as in (1.2) and (1.5). The fluid model is labelled correspondingly if all stations are of the same type.

We recall from (1.18) that a fluid model is *stable* if there exists an $N > 0$, so that for any solution of its fluid model equations, the $Z(\cdot)$ component satisfies

$$Z(t) = 0 \quad \text{for } t \geq N|Z(0)|. \tag{4.61}$$

The main result in Section 4.4, Theorem 4.16, gives general criteria for the stability of the corresponding queueing network. An important condition is that its fluid model be stable. (Other conditions involve the interarrival time distributions of the queueing network.) Ascertaining whether a fluid model is stable is itself not an elementary problem in general, and will be discussed in Chapter 5.

The following results are easy to derive using our present machinery, and will be useful for showing analogous results for queueing networks. Parts (a), (b), (c) and (d) of Proposition 4.11 will be used, respectively, for Example 1 of Section 4.4, Corollaries 1 and 2 of Proposition 4.12 of this section, and Proposition 5.21 on Section 5.5. As usual, the inequalities in Part (b) are to be interpreted componentwise.

Proposition 4.11. *(a) Any fluid model with $\sum_j \rho_j < 1$ is stable. (b) For any solution of a fluid model,*

$$\liminf_{t\to\infty} Y(t)/t \geq e - \rho.$$

The rate of convergence is uniform over bounded $|Z(0)|$ for these solutions. When $Z(0) = 0$, $Y(t) \geq (e - \rho)t$ for all t. (c) Suppose that for some solution of a fluid model, $Z_k(t) < Z_k(0)$ for some t and all k. Then, $\rho < e$. (d) Suppose that for a fluid model, $\rho_j > 1$ for some j. Then, for some $\epsilon > 0$, $|Z(t)| \geq \epsilon t$ for all t and all fluid model solutions.

Proof. By (4.50) and (4.51),

$$Z(t) - Z(0) = \alpha t - (I - P^T)D(t).$$

Multiplying both sides by CMQ gives

$$CMQ(Z(t) - Z(0)) = \rho t - CMD(t) = \rho t - CT(t). \tag{4.62}$$

The first equality employs $\rho = CM\lambda = CMQ\alpha$, which follows from (1.5) and (1.7), and the second equality follows from (4.55). By (4.53), (4.62) can be rewritten as

$$CMQ(Z(t) - Z(0)) = (\rho - e)t + Y(t). \tag{4.63}$$

Parts (c) and (d) follow quickly from the resulting inequality

$$CMQ(Z(t) - Z(0)) \geq (\rho - e)t. \tag{4.64}$$

Under the assumption in (c), the left side of (4.64) is negative in each coordinate for that t, and so $\rho < e$. (Note that $Q \geq I$.) Let $\rho_j > 1$ for a given j, as in (d). The j coordinate of the left side of (4.64) is bounded below by $(\rho_j - 1)t$, and the matrix CMQ is constant. So, (d) holds for an appropriate choice of $\epsilon > 0$.

The display in Part (b) follows from (4.63), after dividing both sides by t and taking limits. Since $CMQZ(0)/t \to 0$ as $t \to \infty$, the rate of convergence for the lim inf in (b) is also uniform over bounded $|Z(0)|$. The case where $Z(0) = 0$ is an immediate consequence of (4.63).

For Part (a), multiplication of both sides of (4.62) by e^T and taking derivatives gives

$$e^T CMQZ'(t) = e^T(\rho - CT'(t))$$

at all regular points. By (4.53), for each choice of j, this is

$$\leq \sum_{j'} \rho_{j'} - \sum_{k \in \mathcal{C}(j)} T_k'(t) = \sum_{j'} \rho_{j'} - 1 + Y_j'(t).$$

By (4.54) and (4.57), $Y_j'(t) = 0$ for at least one choice of j when $Z(t) \neq 0$. Setting $\epsilon = 1 - \sum_{j'} \rho_{j'} > 0$, it follows that

$$e^T CMQZ'(t) \leq -\epsilon$$

at such points. By the bound in (4.60),

$$Z(t) = 0 \quad \text{for } t \geq e^T CMQZ(0)/\epsilon,$$

and so the fluid model is stable. ■

Even though fluid model equations are simplifications of the corresponding queueing network equations, we will need to exercise some caution in their application. In particular, a fluid model need not have a unique solution for given initial data. This is not surprising for the basic fluid model equations, since the solutions may depend on the discipline, which is not included in these equations. However, it is not difficult to show this is also sometimes the case when either equation (4.48) (corresponding to the FIFO discipline) or equation (4.49) (corresponding to an SBP discipline) is added to the basic queueing network equations. So, nonuniqueness can persist even when the discipline has been specified.

We conclude this subsection with two examples of fluid models which exhibit this nonuniqueness. Both examples employ fluid models that correspond to certain parameter values for the Rybko-Stolyar network that was introduced in Section 3.1. The routes, which are given in Figure 3.2, each possess two stations, which are each visited once along the route. The discipline is preemptive SBP, with priority at each station given to the class of jobs that

are about to leave the network. Classes are labelled by (i,k), where $i=1,2$ denotes the route and $k=1,2$ denotes the sequential ordering of the class along the route. The fluid model is assumed to consist of the basic fluid model equations (4.50)-(4.55) together with (4.49). The routing matrix P here is given by the above routing. The external arrival rates are given by

$$\alpha_{1,1}=\alpha_{2,1}=1, \quad \alpha_{1,2}=\alpha_{2,2}=0;$$

and the service times are assumed to satisfy

$$m_1 \stackrel{\text{def}}{=} m_{1,1}=m_{2,1}>0, \quad m_2 \stackrel{\text{def}}{=} m_{1,2}=m_{2,2}>0,$$

which will be further specified in the examples. The SBP equation (4.49) is assumed to incorporate the above priority scheme.

Example 1. *A fluid model with nonunique solutions.* In this example, we assume that $m_1<m_2$ and assign the initial data

$$Z_{1,1}(0)=Z_{2,1}(0)=1, \quad Z_{1,2}(0)=Z_{2,2}(0)=0.$$

On account of the discipline, when either $Z_{1,2}(t_0)\neq 0$ or $Z_{2,2}(t_0)\neq 0$, this is enough to uniquely determine the evolution of $\mathfrak{X}(t)$ for small times after t_0. This is not the case for the given initial data since, as one can check, there exist distinct solutions over $t\in[0,1/(1-\mu_1)]$, with

$$\begin{gathered} Z_{i,1}(t)=1+(1-\mu_1)t, \quad Z_{i,2}(t)=(\mu_1-\mu_2)t, \\ Z_{i',1}(t)=1+t, \quad Z_{i',2}(t)=0, \end{gathered} \tag{4.65}$$

for either $i=1$ or $i=2$, where i' denotes the other route. (As usual, $\mu_k=1/m_k$.)

One can interpret the above evolution of the fluid model as follows. The initial job mass in route i "gets an infinitesimal lead" over that in route i', with job mass starting to flow into the class $k=2$ along the route before this starts to occur along route i'. Once this flow begins, since $\mu_2<\mu_1$, there will be mass at class $(i,2)$. The mass at $(i,2)$ has priority of service over that at $(i',1)$. This prevents service at $(i',1)$, and so the class $(i',2)$ remains empty. Mass from $(i,1)$ continues to flow to $(i,2)$ until at least time $1/(\mu_1-1)$, after which class $(i,1)$ is empty, and so the rate at which mass enters $(i,2)$ slows to $\alpha_{i,1}=1$. Over these times, there will be mass at $(i,2)$ and no service at $(i',1)$.

In addition to the solutions of the fluid model with $Z(t)$ given by (4.65), there is the symmetric solution over $t\in[0,(\mu_1+\mu_2)/(\mu_1\mu_2-\mu_1-\mu_2)]$, with

$$Z_{i,1}(t)=1+(1-\mu_1\mu_2/(\mu_1+\mu_2))t, \quad Z_{i,2}(t)=0, \tag{4.66}$$

for $i=1,2$. Here, the fraction of effort allocated to the classes with $k=1$ is $\mu_2/(\mu_1+\mu_2)$, with the fraction allocated to $k=2$ being $\mu_1/(\mu_1+\mu_2)$. This

allocation of effort keeps the high priority classes with $k = 2$ empty, and so allows continual service at both classes with $k = 1$.

The solution given by (4.66) is unstable in the sense that, at any time t_0, there are other solutions emanating from it. This occurs when the classes along one of the routes start to "monopolize" the service at their stations. After time t_0, these solutions evolve as in (4.65), except for the lag in time and the different starting mass at the classes with $k = 1$. With a bit of effort, one can check that all of the solutions of the fluid model in a neighborhood of time 0 and with the assigned initial data are given by the above solutions. That is, the solution in (4.66) is followed until some assigned time, after which the solutions evolve as in (4.65).

We point out that for $m_1 + m_2 < 1/2$, one has $\rho_1 + \rho_2 < 1$, and so the assumptions of Part (a) of Proposition 4.11 are satisfied. Therefore, under this additional condition, the above fluid model is stable. As mentioned above the proposition, this is sufficient for the corresponding queueing networks to be stable (under appropriate conditions on the interarrival time distributions). So, the nonuniqueness of solutions for the fluid model is not connected with the stability of either the fluid model or the corresponding queueing network.

We also point out that, although we have not bothered to construct the above fluid model solutions for all time, these solutions can always be extended past the times that are given. One way of doing this is to employ fluid limits, which are introduced in the next section. ■

Example 2. *A fluid model that has a nonzero solution with* $Z(0) = 0$. In this example, we assume that $m_1 < 1/3$ and $m_2 = 2/3$. (We fix m_2 in order to simplify the coefficients in our computations.) By Theorem 3.4, the corresponding queueing network is unstable, if the interarrival and service times are exponentially distributed. The fluid model exhibits similar behavior, which we show for $Z(0) = 0$. We do this by constructing a self-similar solution, in the spirit of the proof of Theorem 3.1. We give the values of $T(\cdot)$ and $Z(\cdot)$; using (4.50)-(4.55), the other components of $\mathfrak{X}(\cdot)$ can be calculated from these. ($T(\cdot)$ and $Z(\cdot)$ can also be calculated from each other.)

We proceed in two steps. We first construct a solution $\tilde{\mathfrak{X}}(\cdot)$ of the fluid model on $[0, 2]$, with $|\tilde{Z}(0)| = 1$ and $|\tilde{Z}(2)| = 2$, where all the mass at $t = 0$ is at class $(1, 1)$ and that at $t = 2$ is at class $(2, 1)$. Because of the symmetry of the network and the doubling of mass by $t = 2$, the same reasoning allows us to extend the solution up until $t = 6$, when the total mass is 4 and all the mass is again at class $(1, 1)$. We then piece together scaled versions of this solution so that the resulting solution is defined over $[0, \infty)$, and grows linearly starting at $Z(0) = 0$. We note that since $\rho_1, \rho_2 < 1$, $Z(t) \equiv 0$ gives another solution of the fluid model. So, this construction gives another example of the nonuniqueness of fluid model solutions under an SBP discipline.

Set $b_1 = 1/(\mu_1 - 1) = m_1/(1 - m_1)$. We construct $\tilde{T}(t)$ and $\tilde{Z}(t)$, with $t \in [0, 2]$, piecewise over the time intervals $[0, b_1]$ and $[b_1, 2]$. We choose $\tilde{T}'(t)$ so that it is constant over each interval, with

$$\begin{aligned} \tilde{T}'_{1,1}(t) &= \begin{cases} 1 & \text{for } t \in (0, b_1), \\ m_1 & \text{for } t \in (b_1, 2), \end{cases} \\ \tilde{T}'_{1,2}(t) &= 1 \quad \text{for } t \in (0, 2), \\ \tilde{T}'_{2,1}(t) &= \tilde{T}'_{2,2}(t) = 0 \quad \text{for } t \in (0, 2). \end{aligned} \tag{4.67}$$

Integration then gives $\tilde{T}(t)$. The function $\tilde{Z}(t)$ will be linear over these intervals, with values at the endpoints given by

$$\begin{aligned} \tilde{Z}_{1,1}(0) = 1, \quad \tilde{Z}_{1,1}(b_1) = \tilde{Z}_{1,1}(2) = 0, \\ \tilde{Z}_{1,2}(0) = 0, \quad \tilde{Z}_{1,2}(b_1) = b_1(\mu_1 - 3/2), \quad \tilde{Z}_{1,2}(2) = 0, \\ \tilde{Z}_{2,1}(0) = 0, \quad \tilde{Z}_{2,1}(2) = 2, \quad \tilde{Z}_{2,2}(0) = \tilde{Z}_{2,2}(2) = 0. \end{aligned} \tag{4.68}$$

Note that the value of μ_1, for $\mu_1 > 3/2$, does not affect the value of $\tilde{Z}(2)$.

The choice of $\tilde{T}(\cdot)$ and $\tilde{Z}(\cdot)$ as in (4.67)-(4.68) corresponds to the assignment of all effort at both stations to the classes along the first route. This is consistent with the discipline, because class $(1, 2)$ has priority over $(2, 1)$, and there is never any mass at $(2, 2)$ to impede service at $(1, 1)$. The time interval is divided into two parts, with the latter beginning when class $(1, 1)$ first becomes empty. As claimed earlier, all of the 2 units of mass at $t = 2$ is at class $(2, 1)$. A repetition of this reasoning, this time over the interval $[2, 6]$, shows that at $t = 6$, all of the mass is again at class $(1, 1)$, with

$$\tilde{Z}_{1,1}(6) = 4.$$

The second step consists of piecing together scaled versions of the above construction. For $t \in [\frac{1}{2}4^{i+1}, \frac{1}{2}4^{i+2}]$, set

$$T(t) = 4^i(\tilde{T}(4^{-i}t - 2) + U), \quad Z(t) = 4^i\tilde{Z}(4^{-i}t - 2), \tag{4.69}$$

where U is the constant vector with

$$U_{1,1} = m_1, \quad U_{2,1} = 2m_1, \quad U_{1,2} = U_{2,2} = 4/3,$$

and set $T(0) = Z(0) = 0$. It is straightforward to check that over the interval $[\frac{1}{2}4^{i+1}, \frac{1}{2}4^{i+2}]$, with $i \in \mathbf{Z}$, $T(\cdot)$ and $Z(\cdot)$ give a solution of the fluid model, since the effect of the scaling terms 4^i and 4^{-i} cancel each other out and since the translation terms do not affect the evolution of solutions. ($i = 0$ corresponds to the solution $\tilde{\mathfrak{X}}(\cdot)$, after a time shift.) By (4.68), $\tilde{Z}(6) = 4\tilde{Z}(0)$; therefore, $Z(\cdot)$ is consistently defined at the endpoints $\frac{1}{2}4^i$. One can also check that the same is true for $T(\cdot)$. So, $T(\cdot)$ and $Z(\cdot)$ define a solution over $(0, \infty)$. Since

$$\lim_{t \downarrow 0} T(t) = \lim_{t \downarrow 0} Z(t) = 0,$$

$T(\cdot)$ and $Z(\cdot)$ are continuous at 0, which implies that $T(\cdot)$ and $Z(\cdot)$ define a solution over $[0, \infty)$, as desired.

It follows from $|Z(2)| = |\tilde{Z}(0)| = 1$ and (4.69), evaluated $t = \frac{1}{2}4^i$, $i \in \mathbf{Z}_+$, that

$$\limsup_{t\to\infty} |Z(t)|/t \geq 1/2.$$

One can, in fact, check with a little more work that $\lim_{t\to\infty} |Z(t)|/t = 1/2$. ·

■

Fluid limits

The basic fluid model equations (4.50)-(4.55) are obtained from the corresponding queueing network equations (4.42)-(4.47) by substituting the means α, M, and P for E, Γ, and Φ. Auxiliary equations that specify the discipline, such as (4.48) and (4.49), are obtained similarly from the corresponding queueing network equations. In addition to being obtained by this formal substitution, fluid model equations can be derived from a "law of large numbers" scaling of the form $\mathfrak{X}(st)/s$ as $s \to \infty$. This interpretation will be important in Section 4.4 when we use the stability of fluid models to show the stability of the corresponding queueing networks. Fluid limits make this limiting procedure precise.

We first define the set G on which we will take fluid limits. Let G be the set on which the strong law of large numbers holds for the external arrivals, service times, and routing of a queueing network. That is, for $\omega \in G$,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \xi_k(i) = 1/\alpha_k, \quad \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \gamma_k(i) = m_k, \tag{4.70}$$
$$\lim_{n\to\infty} \frac{1}{n}(\Phi^k(n))_\ell = P_{k,\ell},$$

where $k \in \mathcal{A}$ in the first term and $k = 1, \ldots, K$ elsewhere, and $\ell = 1, \ldots, K$ in the last term. One has $P(G) = 1$. (Other sets G', with $P(G') = 1$ and satisfying (4.70), can also be used.)

We define fluid limits as follows. Let (a_n, x_n) be a sequence of pairs, with

$$\lim_{n\to\infty} a_n = \infty, \quad \limsup_{n\to\infty} |z_n|/a_n < \infty, \tag{4.71}$$
$$\lim_{n\to\infty} |u_n|/a_n = \lim_{n\to\infty} |v_n|/a_n = 0.$$

Here, $a_n \in \mathbf{R}_+$ and $x_n \in S$, with z_n, u_n, and v_n denoting the queue length, residual interarrival time, and residual service time vectors, with $|x_n| = |z_n| + |u_n| + |v_n|$ as in (4.3). (Throughout the remainder of this section and the next, we will employ subscripts for terms in sequences as well as for coordinates of vectors; which one is meant will be clear from the context.) A *fluid limit* of the queueing network, with queueing network process $\mathfrak{X}(\cdot)$, is any limit

$$\bar{\mathfrak{X}}(t) = \lim_{n\to\infty} \frac{1}{a_n}\mathfrak{X}^{x_n}(a_n t), \tag{4.72}$$

for any choice of $\omega \in G$ and any sequence (a_n, x_n) satisfying (4.71). Convergence is required to be uniform on compact sets (u.o.c.), and is to be interpreted componentwise with respect to each of the six components of $\bar{\mathfrak{X}}(\cdot)$. Here and later on, the superscript gives the initial state of a process, e.g., $\mathfrak{X}^x(\cdot)$ indicates that $\mathfrak{X}^x(0) = x$.

We will say that the family of these fluid limits is *associated* with the queueing network. Such a family will be *stable* if there exists an $N > 0$, so that for any of its fluid limits, the $\bar{Z}(\cdot)$ component satisfies

$$\bar{Z}(t) = 0 \quad \text{for } t \geq N|\bar{Z}(0)|. \tag{4.73}$$

This definition of stability is the analog of that for fluid models in (4.61). (The scaled sequences $\{\mathfrak{X}^{x_n}(a_n t)/a_n,\ n \in \mathbf{Z}_+\}$ that are obtained from a queueing network process, and their limits are frequently referred to as a fluid limit model. We do not use that terminology here.)

We next employ fluid limits to specify the relationship between queueing networks and fluid models that we will use in the succeeding sections. Let $\mathcal{M}$ be a fluid model, that is, a set of fluid model equations including the basic fluid model equations (4.50)-(4.55). Then, $\mathcal{M}$ is *associated* with a queueing network if, for $\omega \in G$, each sequence (a_n, x_n) satisfying (4.71) possesses a subsequence (a_{i_n}, x_{i_n}) (depending on ω), on which the components of the scaled sequences $\mathfrak{X}^{x_{i_n}}(a_{i_n} t)/a_{i_n}$ obtained from the queueing network process converge u.o.c., and this limit satisfies the fluid model equations of $\mathcal{M}$. Such a limit will automatically satisfy (4.56) and the positivity and monotonicity properties given immediately before (4.56).

We note that these limits are, by definition, fluid limits of the queueing network. We are thus requiring here that each sequence possess a subsequence with a fluid limit that satisfies the fluid model equations. It also follows from this definition that every fluid limit must satisfy the fluid model equations. (The definitions given here for fluid limits and associated fluid models differ somewhat from those in the literature.)

Consider a fluid model whose triple (α, M, P) corresponds to the triple $(E(\cdot), \Gamma(\cdot), \Phi(\cdot))$ of a queueing network. (That is α, M, and P are the means corresponding to $E(\cdot), \Gamma(\cdot)$, and $\Phi(\cdot)$.) In Proposition 4.12, we will show that, for HL queueing networks, a converging subsequence always exists that satisfies the basic fluid model equations. So, this basic fluid model is always associated with the queueing network. Consequently, in order to show that a fluid model (with corresponding triple (α, M, P)) is associated with a given queueing network, it suffices to show that its auxiliary fluid model equations are also satisfied by all fluid limits.

The basic fluid model for a given queueing network is uniquely specified. Typically, there will be a particular "canonical" fluid model that is associated with the queueing network, which includes the basic fluid model equations, together with an appropriate equation (or equations) that describe the discipline. The choice of these auxiliary equations is usually fairly natural. We will, for example, employ (4.48) as the auxiliary fluid model equation for

the canonical fluid model for FIFO queueing networks and (4.49) for SBP queueing networks. Note that there exist other associated fluid models; for instance, the basic fluid model for a queueing network is always associated with it. We will also use associated in the opposite direction and say that a queueing network is associated with a fluid model, although there will be many such queueing networks for a given fluid model, since different choices of the distributions of $\xi(1)$ and $\gamma(1)$, with the same means, are possible.

We now state Proposition 4.12, which was cited above.

Proposition 4.12. *For each HL queueing network and $\omega \in G$, every sequence of pairs (a_n, x_n) satisfying (4.71) possesses a subsequence (a_{i_n}, x_{i_n}) on which the limit in (4.72) exists and is u.o.c. This limit satisfies the basic fluid model equations (4.50)-(4.55) with the corresponding triple. Hence, each HL queueing network is associated with its basic fluid model.*

Proof. We need to show the existence of a limit $\bar{\mathfrak{X}}(\cdot)$ that satisfies (4.50)-(4.55) and is u.o.c. These properties will follow from the queueing network equations (4.42)-(4.47) and (4.70). First note that for a given $\omega \in G$, one can restrict the sequence x_n to a subsequence x_{i_n}, so that for some $\bar{T}(\cdot)$,

$$\frac{1}{a_{i_n}} T^{x_{i_n}}(a_{i_n} t) \to \bar{T}(t) \quad \text{as } n \to \infty, \tag{4.74}$$

on a dense set of t, say, the nonnegative rationals. On account of (4.45), the terms on the left side of (4.74) are all Lipschitz continuous with coefficient 1, and so the sequence is uniformly equicontinuous. The limit in (4.74) therefore holds u.o.c. for all t, if $\bar{T}(\cdot)$ is replaced by its continuous extension to all of $\mathbf{R}_{+,0}$. Applying (4.45) again, it follows that

$$\bar{Y}(t) = \lim_{n \to \infty} \frac{1}{a_{i_n}} Y^{x_{i_n}}(a_{i_n} t)$$

also exists, with convergence being u.o.c. Moreover, (4.53) is satisfied and both $\bar{T}(\cdot)$ and $\bar{Y}(\cdot)$ are Lipschitz continuous.

To show (4.55), note that on account of (4.47),

$$\begin{aligned} \frac{1}{a_{i_n}} \gamma_k(0) + \frac{1}{a_{i_n}} \sum_{i=1}^{D_k^{x_{i_n}}(a_{i_n} t)-1} \gamma_k(i) &\le \frac{1}{a_{i_n}} T_k^{x_{i_n}}(a_{i_n} t) \\ &< \frac{1}{a_{i_n}} \gamma_k(0) + \frac{1}{a_{i_n}} \sum_{i=1}^{D_k^{x_{i_n}}(a_{i_n} t)} \gamma_k(i), \end{aligned} \tag{4.75}$$

for given k, t, and n. Since $\omega \in G$, it follows from (4.70) that, for given $\epsilon > 0$ and large n, the quantity to the right of the strict inequality is at most

$$\frac{1}{a_{i_n}} (m_k + \epsilon) D_k^{x_{i_n}}(a_{i_n} t) + \epsilon.$$

A similar lower bound, with ϵ replaced by $-\epsilon$, holds for the quantity to the left of the other inequality. Letting $n \to \infty$, it follows that $m_k D_k^{x_{i_n}}(a_{i_n}t)/a_{i_n}$ converges to the limit of $T_k^{x_{i_n}}(a_{i_n}t)/a_{i_n}$, which is $\bar{T}_k(t)$. Since $\bar{T}(\cdot)$ is continuous and $D^{x_{i_n}}(\cdot)$ is nondecreasing, this convergence is u.o.c. So,

$$\bar{D}(t) = \lim_{n\to\infty} \frac{1}{a_{i_n}} D^{x_{i_n}}(a_{i_n}t)$$

also exists, with convergence being u.o.c. and (4.55) holding. Since $\bar{T}(\cdot)$ is Lipschitz continuous, so is $\bar{D}(\cdot)$.

We next show (4.50). Here, one uses the limits involving $\sum_{i=1}^{n} \xi(i)$ and $\Phi(n)$ in (4.70), and $|u_n|$ in (4.71), as $n \to \infty$. Together with the above limit $\bar{D}(\cdot)$, they imply that

$$\frac{1}{a_{i_n}} E^{x_{i_n}}(a_{i_n}t) \to \alpha t, \quad \frac{1}{a_{i_n}} \sum_k \Phi(D_k^{x_{i_n}}(a_{i_n}t)) \to P^T \bar{D}(t) \tag{4.76}$$

for all t, as $n \to \infty$. Consequently, the limit $\bar{A}(t)$ exists and (4.50) holds. Since $E^{x_{i_n}}(\cdot)$ and $\Phi^k(\cdot)$ are monotone and convergence to $\bar{D}(\cdot)$ is u.o.c., convergence to $\bar{A}(\cdot)$ is also u.o.c. Since $\bar{D}(\cdot)$ is Lipschitz continuous, so is $\bar{A}(\cdot)$.

Equation (4.51) and convergence to $\bar{Z}(\cdot)$ follow quickly from (4.43) and the previous limits, on a further subsequence that is chosen so that $\bar{Z}(0)$ exists. The derivation of (4.52) from (4.44) and convergence to $\bar{W}(\cdot)$ are similar to the previous steps.

We still need to show (4.54). Suppose that for some j and all $t \in [t_1, t_2]$, $\bar{W}_j(t) > 0$. Since $\bar{W}_j(\cdot)$ is continuous, it is bounded away from 0 on the interval. Convergence to $\bar{W}(\cdot)$ is u.o.c., and so for sufficiently large n, $W_j^{x_{i_n}}(a_{i_n}t) > 0$ on $[t_1, t_2]$ as well. Because of (4.46), one must have $Y_j^{x_{i_n}}(t_1) = Y_j^{x_{i_n}}(t_2)$. Since the same equality also holds in the limit as $n \to \infty$, one obtains (4.54), as desired. ■

Proposition 4.12 will be applied to Theorem 4.16 in Section 4.4. We mention here the following quick applications of the proposition and Part (b) of Proposition 4.11. The first application says that for a queueing network, the limiting proportion of time is positive that a given subcritical station j is empty. We note that this, by itself, does not imply the queueing network is stable. As the examples in Chapter 3 show, different stations can be empty at nonoverlapping times, with $|Z(t)| \to \infty$ as $t \to \infty$ nevertheless holding. The second application is a modification of the one just described, and will be employed in Proposition 4.15 of Section 4.4.

Corollary 1. *For each HL queueing network and any initial state x,*

$$\liminf_{t\to\infty} Y^x(t)/t \geq e - \rho \quad \text{on } G. \tag{4.77}$$

Moreover, for each $\epsilon > 0$, there exists a c_0, so that

$$\liminf_{|x|\to\infty} Y^x(c|x|)/c|x| \geq (1-\epsilon)e - \rho \quad \textit{on } G, \tag{4.78}$$

for all $c \geq c_0$.

Proof. We demonstrate (4.77). Suppose that for some $\omega \in G$, the above limit is false. That is, along some sequence of times a_n, with $a_n \to \infty$ as $n \to \infty$,

$$\lim_{n\to\infty} Y_j^x(a_n)/a_n < 1 - \rho_j \quad \text{for some } j. \tag{4.79}$$

By Proposition 4.12,

$$\bar{\mathfrak{X}}(t) = \lim_{n\to\infty} \mathfrak{X}^x(a_{i_n}t)/a_{i_n}$$

exists along some subsequence a_{i_n}, and satisfies the basic fluid model equations. Since $\bar{Z}(0) = 0$, it follows from Part (b) of Proposition 4.11 that $\bar{Y}(t) \geq (e-\rho)t$ for all t. Therefore,

$$\liminf_{n\to\infty} Y^x(a_{i_n}t)/a_{i_n} \geq (e-\rho)t, \tag{4.80}$$

which contradicts (4.79).

The argument for (4.78) is almost the same. In (4.79), one replaces a_n by $t|x_n|$ and $1-\rho_j$ by $1-\epsilon-\rho_j$, for given $\epsilon > 0$. The fluid limit will now satisfy $|\bar{Z}(0)| \leq 1$. One obtains from Proposition 4.11 that $\bar{Y}(t) \geq ((1-\epsilon)e-\rho)t$ for given ϵ and $c \geq c_0$, for large enough c_0. Substitution of the limiting sequence gives the analog of (4.80), which contradicts the analog of (4.79). ■

We also note that Proposition 4.12 provides a means for showing the existence of fluid model solutions with given initial data $\mathfrak{X}(0) = x$. Suppose that (a_n, x_n) satisfies (4.71) and that $z_n/a_n \to z$ as $n \to \infty$, where z denotes the queue length vector for x. Then, for $\omega \in G$, any fluid limit $\bar{\mathfrak{X}}(\cdot)$ will satisfy the basic fluid model equations (4.50)-(4.55), with $\bar{Z}(0) = z$. The same conclusion will hold for more general fluid models that are associated with some queueing network.

Similar reasoning, together with Part (c) of Proposition 4.11, leads to the following elementary result, which will be used in the proof of Proposition 4.15.

Corollary 2. *Suppose that the family of fluid limits associated with an HL queueing network is stable. Then, the queueing network is subcritical.*

Proof. Consider a sequence of pairs (a_n, x_n) satisfying (4.71), where for each coordinate $(z_n)_k$ of z_n, $\liminf_{n\to\infty}(z_n)_k/a_n > 0$. For any fluid limit $\bar{\mathfrak{X}}(\cdot)$ of this sequence, $\bar{Z}(0) > 0$. Also, since the family of fluid limits is stable, $\bar{Z}(N|\bar{Z}(0)|) = 0$ for N chosen as in (4.73). By Proposition 4.12, such a fluid limit will exist and will satisfy the basic fluid model equations. It therefore follows from Part (c) of Proposition 4.11 that the fluid model is subcritical, and hence so is the queueing network. ■

Countable state space setting

There are only minor simplifications in this section when the state space S of a queueing network is countable. There are no changes in the approach to queueing network equations and fluid models. The comment on fluid models with delay is now vacuous, since the residual times u and v are no longer part of the state space descriptor. A few statements in the subsection on fluid limits simplify slightly. In (4.71), one omits the conditions on u_n and v_n. Also, in the proof of Proposition 4.12, the limits obtained from (4.75) and (4.76) no longer depend on assumptions on the residual times.

4.4 Demonstration of Stability

In this chapter, we have stated a number of results for Markov processes and queueing networks. Here, we employ these results to demonstrate the stability and e-stability of queueing networks in Theorems 4.16 and 4.17. Proposition 4.6 of Section 4.2 and Proposition 4.12 of Section 4.3 will be the main ingredients for this. In addition, we will need two further results, Propositions 4.14 and 4.15, which we provide in this section. Our approach here is a modification of that in [Br98a], which is based on that in [Da95].

Bounds on $|Z(t)|$, $|U(t)|$, and $|V(t)|$

Proposition 4.6 of Section 4.2 is the Multiplicative Foster's Criterion; it requires bounds on the expectation of $E_x|X(c(|x| \vee \kappa))|$, which are given in (4.28). In order to obtain these bounds, we will need bounds on $|Z(t)|$, $|U(t)|$, and $|V(t)|$, and their expectations. These bounds will be provided in Propositions 4.14 and 4.15. To obtain them, we will use the following lemma.

Lemma 4.13. *Let $\beta(1), \beta(2), \beta(3), \ldots$ be i.i.d. positive random variables with finite mean. Set $B_1(t) = \frac{1}{t} \max\{n : \sum_{i=1}^{n-1} \beta(i) \le t\}$. Then,*

$$\sup_{t \ge 1} E[B_1(t); B_1(t) \ge M] \to 0 \quad \text{as } M \to \infty. \tag{4.81}$$

Set $B_2(t) = \frac{1}{t} \max\{\beta(n) : \sum_{i=1}^{n-1} \beta(i) \le t\}$, and let G_β denote the set on which $\frac{1}{n} \sum_{i=1}^{n} \beta(i) \to m$ as $n \to \infty$. Then,

$$B_2(t) \to 0 \quad \text{on } G_\beta \text{ as } t \to \infty, \tag{4.82}$$

and

$$E[B_2(t)] \to 0 \quad \text{as } t \to \infty. \tag{4.83}$$

Proof. We first show (4.81). For $i = 1, 2, 3, \ldots$, set

$$\tilde{\beta}(i) = \begin{cases} \delta & \text{for } \beta(i) \ge \delta, \\ 0 & \text{for } \beta(i) < \delta, \end{cases}$$

where $\delta \in (0,1]$ is chosen small enough so that $P(\beta(1) > \delta) > 0$. Define $\tilde{B}_1(t)$ analogously to $B_1(t)$, but with respect to $\tilde{\beta}(i)$ instead of $\beta(i)$. Since $\tilde{B}_1(t) \geq B_1(t)$, to show (4.81), it suffices to show the analogous limit for $\tilde{B}_1(t)$.

For $\ell = 0, 1, 2, \ldots$, let $\zeta(\ell)$ denote the number of indices n, $n = 0, 1, 2, \ldots$, for which $\sum_{i=1}^{n} \tilde{\beta}(i) = \delta\ell$. Then, $\zeta(0), \zeta(1), \zeta(2), \ldots$ are i.i.d. random variables with finite means and

$$\tilde{B}_1(t) \leq \frac{1}{t} \sum_{\ell=0}^{\lceil t/\delta \rceil} \zeta(\ell). \tag{4.84}$$

Using Markov's Inequality, one can check that

$$\sup_{t \geq 1} P\left(\frac{1}{t} \sum_{\ell=0}^{\lceil t/\delta \rceil} \zeta(\ell) \geq M \right) \leq \frac{3}{\delta M} E[\zeta(0)] \to 0 \quad \text{as } M \to \infty. \tag{4.85}$$

By (4.84) and the exchangeability of $\zeta(\ell)$,

$$\sup_{t \geq 1} E[\tilde{B}_1(t); \tilde{B}_1(t) \geq M] \leq \sup_{t \geq 1} \frac{3}{\delta} E\left[\zeta(0); \frac{1}{t} \sum_{\ell=0}^{\lceil t/\delta \rceil} \zeta(\ell) \geq M \right].$$

Because of (4.85), this $\to 0$ as $M \to \infty$, which is the desired limit.

To show (4.82), note that on G_β,

$$\frac{1}{n} \sum_{i=1}^{\lfloor cn \rfloor} \beta(i) \to cm \quad \text{as } n \to \infty,$$

where convergence is uniform on $c \in [0,1]$. Therefore, for given $\epsilon > 0$ and large enough n,

$$\max_{i \in [c_1 n, c_2 n)} \beta(i) \leq \sum_{i=\lfloor c_1 n \rfloor}^{\lfloor c_1 n \rfloor} \beta(i) \leq (c_2 - c_1 + \epsilon) mn$$

for all $0 \leq c_1 \leq c_2 \leq 1$. Setting $c_2 = c_1 + \epsilon$ and $n = \lfloor 2t/m \rfloor$, one has, for large enough t, that

$$B_2(t) \leq \frac{1}{t} \sup_{c_1 \in [0,1]} \max_{i \in [c_1 n, (c_1+\epsilon) n)} \beta(i) \leq 2\epsilon mn/t \leq 4\epsilon.$$

Since $\epsilon > 0$ is arbitrary, (4.82) follows.

We will show that

$$\sup_{t \geq 1} E[B_2(t); B_2(t) \geq M] \to 0 \quad \text{as } M \to \infty. \tag{4.86}$$

Together with (4.82), this will imply (4.83) since $P(G_\beta) = 1$. For this, we again choose $\delta \in (0,1]$ small enough so $P(\beta(1) > \delta) > 0$, and let $\beta'(1)$,

$\beta'(2)$, $\beta'(3), \ldots$ denote the subsequence obtained by restricting the original sequence to terms with $\beta(i) \geq \delta$. Clearly, $B_2(t) \leq \frac{1}{t} \max_{1 \leq i \leq \lceil t/\delta \rceil} \beta'(i)$. Also, by Markov's Inequality,

$$\sup_{t \geq 1} P\left(\frac{1}{t} \max_{1 \leq i \leq \lceil t/\delta \rceil} \beta'(i) \geq M\right) \leq \sup_{t \geq 1} \frac{1}{Mt} E\left[\max_{1 \leq i \leq \lceil t/\delta \rceil} \beta'(i)\right]$$
$$\leq \frac{2}{\delta M} E[\beta'(1)],$$

which $\to 0$ as $M \to \infty$, since $\beta'(1)$ has finite mean. It follows that the left side of (4.86) is at most

$$\frac{2}{\delta} \sup_{t \geq 1} E\left[\beta'(1);\ \frac{1}{t} \max_{1 \leq i \leq \lceil t/\delta \rceil} \beta'(i) \geq M\right],$$

which $\to 0$ as $M \to \infty$. This shows (4.86) and hence (4.83). ■

By applying the first part of Lemma 4.13, it is easy to obtain uniform bounds on the growth of the queue length $Z^x(t)$ of an HL queueing network. (The result is not dependent on the HL property and so holds for more general networks, once defined.) Recall that the set G is given in (4.70).

Proposition 4.14. *For each HL queueing network, $c > 0$, and $\omega \in G$,*

$$\limsup_{|x| \to \infty} |Z^x(c|x|)|/|x| \leq c|\alpha| + 1. \tag{4.87}$$

Moreover,

$$\sup_{t \geq 1} \sup_{|x| \leq t} E[|Z^x(t)|/t;\ |Z^x(t)| \geq Mt] \to 0 \quad \text{as } M \to \infty; \tag{4.88}$$

in particular,

$$\sup_{t \geq 1} \sup_{|x| \leq t} E[|Z^x(t)|]/t < \infty. \tag{4.89}$$

Proof. At any time t,

$$|Z^x(t)| \leq |E^x(t)| + |z|, \tag{4.90}$$

that is, the sum of the queue lengths is bounded by the number of external arrivals plus the sum of the original queue lengths. Also, on G,

$$\limsup_{|x| \to \infty} E_k^x(c|x|)/|x| \leq c\alpha_k,$$

for any k, since the number of external arrivals will only be reduced, at $k \in \mathcal{A}$, as the initial residual interarrival time u_k of x increases. Summation of the components in this inequality and application of (4.90) implies (4.87), since $|z| \leq |x|$.

To obtain the uniform integrability bound in (4.88), set $\beta(i) = \xi_k(i)$ in Lemma 4.13, where $\xi_k(1), \xi_k(2), \ldots$ is the sequence of interarrival times at $k \in \mathcal{A}$ defined earlier in the chapter. Since

$$E_k^x(t)/t \le B_1(t)$$

for all t and x, it follows from (4.81) that

$$\sup_{t\ge 1} \sup_x E[E_k^x(t)/t;\ E_k^x(t) \ge Mt] \to 0 \quad \text{as } M \to \infty.$$

This limit holds trivially for $k \notin \mathcal{A}$. Since the sum of uniformly integrable random variables is uniformly integrable,

$$\sup_{t\ge 1} \sup_x E[|E^x(t)|/t;\ |E^x(t)| \ge Mt] \to 0 \quad \text{as } M \to \infty.$$

Together with (4.90), this implies (4.88). The limit in (4.89) is an immediate consequence of (4.88). ■

By applying the last two limits in Lemma 4.13, one can obtain the following bounds on the growth of the residual times $U^x(t)$ and $V^x(t)$, as $|x| \to \infty$.

Proposition 4.15. *For each subcritical HL queueing network, there exists a $c_0 \ge 1$, so that for $c \ge c_0$,*

$$\frac{1}{|x|}|U^x(c|x|)| \to 0,\ \frac{1}{|x|}|V^x(c|x|)| \to 0 \quad \text{on } G \text{ as } |x| \to \infty. \tag{4.91}$$

Moreover,

$$\sup_{t\ge 1} \sup_{|x|\le t} \frac{1}{t} E|U^x(t)| < \infty, \quad \sup_{t\ge 1} \sup_{|x|\le t} \frac{1}{t} E|V^x(t)| < \infty \tag{4.92}$$

and

$$\frac{1}{|x|} E|U^x(c|x|)| \to 0,\ \frac{1}{x} E|V^x(c|x|)| \to 0 \quad \text{as } |x| \to \infty. \tag{4.93}$$

In particular, (4.91)-(4.93) all hold when the family of fluid limits that is associated with the queueing network is stable.

Proof. The last assertion is an immediate consequence of Corollary 2 to Proposition 4.12, which asserts that the queueing network is subcritical when the fluid limits are stable.

For the first half of (4.91), we set $\beta(i) = \xi_k(i)$ in Lemma 4.13, for a given $k \in \mathcal{A}$. Clearly, for all $t > 0$, x, and ω,

$$\frac{1}{t} U_k^x(t) \le B_2(t) \vee \frac{1}{t} u_k. \tag{4.94}$$

For $c \ge 1$ and $t = c|x|$, one has

$$\frac{1}{|x|}U_k^x(c|x|) \le cB_2(c|x|), \tag{4.95}$$

since $u_k \le |x| \le c|x|$, and so the initial interarrival residual times have expired by then. The first half of (4.91) follows from (4.82) and (4.95), since $G \subseteq G_\beta$.

Taking expectations in (4.94) implies

$$\frac{1}{t}E[U_k^x(t)] \le E[B_2(t)] + \frac{1}{t}u_k.$$

Summing over $k \in \mathcal{A}$ implies the first half of (4.92), since $\sum_{k\in\mathcal{A}} u_k = |u| \le x \le t$, for $|x| \le t$, and $\sup_{t\ge 1} E[B_2(t)] < \infty$ because of (4.83). The first half of (4.93) follows from (4.83) and (4.95).

The argument for the second half of (4.91) is similar to that just given for the first half, but with $\beta(i) = \gamma_k(i)$ in Lemma 4.13, for any k. For all x and ω,

$$\frac{1}{t}V_k^x(t) \le B_2(t) \vee \frac{1}{t}v_k. \tag{4.96}$$

Setting $t = c|x|$, for given c, this implies that

$$\frac{1}{|x|}V_k^x(c|x|) \le cB_2(c|x|) \quad \text{on } \sigma_k^x \le c|x|, \tag{4.97}$$

where σ_k^x is the time at which the initial service residual time expires.

The time at which this occurs is not obvious, since the amount of service that class k receives will depend on the discipline. Note, however, that for large enough c and $|x|$ (where the latter depends on ω), and for $\omega \in G$, (4.78) of the first corollary to Proposition 4.12 implies that the station $j = s(k)$ is idle, and hence empty, at some time before $c|x|$. Here, $\epsilon > 0$ is chosen small enough so that $\rho_j < 1 - \epsilon$, which is possible since the network is subcritical. So,

$$\sigma_k^x \le c|x| \quad \text{for large enough } |x|,$$

for $\omega \in G$. Consequently, (4.97) will eventually hold on G, without the restriction that $\sigma_k^x \le c|x|$. The second half of (4.91) follows from this and (4.82), since $G \subseteq G_\beta$.

The argument for the second half of (4.92) is the same as that for the first half of (4.92), where one uses (4.96) instead of (4.94). We still need to demonstrate the second half of (4.93). Taking expectations in (4.97), after restricting to the set $\sigma_k^x \le c|x|$, and applying (4.83) gives

$$\frac{1}{|x|}E[V_k^x(c|x|)\mathbf{1}\{\sigma_k^x \le c|x|\}] \to 0 \quad \text{as } |x| \to \infty.$$

On the other hand, $V_k^x(c|x|) \le v_k \le |x|$ on $\sigma_k^x > c|x|$. So, by (4.91) and bounded convergence, one also has

$$\frac{1}{|x|}E[V_k^x(c|x|)\mathbf{1}\{\sigma_k^x > c|x|\}] \to 0 \quad \text{as } |x| \to \infty.$$

Together, these limits imply that $E[V_k^x(c|x|)]/|x| \to 0$ as $|x| \to \infty$, which implies (4.93). ■

The main theorem

We now have the needed background to show the main result on stability for queueing networks, Theorem 4.16. A similar result for e-stability will be shown at the end of the section.

Theorem 4.16. *For a given HL queueing network, suppose that* $A = \{x : |x| \leq \kappa\}$ *is petite for each* $\kappa > 0$. *If the family of fluid limits that is associated with the queueing network is stable, then the queueing network is stable. In particular, the queueing network is stable whenever an associated fluid model is stable.*

Before proving the theorem, we first provide some motivation for its assumptions. After the proof, we discuss the theorem further and provide some extensions.

Petite was defined in the last part of Section 4.1, and says, in essence, that each set, after being weighted according to some measure ν, is "equally accessible" from all states in a petite set A. We require the above sets $A = \{x : |x| \leq \kappa\}$ to be petite in order to be able to apply Proposition 4.6 in the proof of the theorem. On account of Proposition 4.7, the assumptions (4.30) and (4.31) on the distributions of the interarrival times suffice for these sets to be petite. By Proposition 4.8, (4.30) suffices when $|\mathcal{A}| = 1$. For the sake of concreteness, results on the stability of queueing networks in the literature have often made these assumptions, rather than assuming the sets A are petite. It is not clear what broader conditions might replace them.

In the statement of the theorem, one also needs to assume that either the associated fluid limits or associated fluid model is stable. Demonstration of either of these conditions requires most of the effort, in practice, when applying the theorem. The assumption on the stability of fluid limits is more general than the corresponding assumption on fluid models.

The latter assumption is the more tractable version, and is the one typically used in applications. It reduces a random problem, the stability of a queueing network, to a (presumably simpler) deterministic problem, the stability of a fluid model. This problem consists of analyzing the solutions of the deterministic equations that constitute the fluid model. In the first three sections of Chapter 5, we will present several examples that illustrate the power of this approach. In such applications, one also needs to show that a specific fluid model is associated with a given queueing network. This is typically routine, with the argument for justifying the auxiliary fluid model equations proceeding along the same lines as the proof of Proposition 4.12.

Unfortunately, there is no satisfactory converse to Theorem 4.16, where the stability of an appropriately chosen fluid model follows from the stability of the queueing network. This will be discussed in Section 5.5.

Proof of Theorem 4.16. We first note that the family of fluid limits that is associated with a queueing network is a subset of the solutions of any associated fluid model. This is an immediate consequence of the definition of an associated fluid model. Stability of an associated fluid model therefore implies stability of the associated fluid limits. So, the second claim in the statement of the theorem follows from the first.

Recall that for any x in the state space, $|x| = |z| + |u| + |v|$. Setting $t = c(|x| \vee 1)$, with $c \geq 1$, in (4.89) of Proposition 4.14 and (4.92) of Proposition 4.15, one has

$$E_x|X(c(|x| \vee 1))| \leq bc(|x| \vee 1) \tag{4.98}$$

for some b and any x. We will demonstrate here that for large enough c,

$$\frac{1}{|x_n|} E_{x_n}|X(c|x_n|)| \to 0 \quad \text{as } n \to \infty, \tag{4.99}$$

for any sequence x_n, with $x_n \to \infty$ as $n \to \infty$. This is equivalent to

$$\frac{1}{|x|} E_x|X(c|x|)| \to 0 \quad \text{as } |x| \to \infty. \tag{4.100}$$

Together, (4.98) and (4.100) imply (4.28) for an appropriate choice of κ. This is the basic assumption for the Multiplicative Foster's Criterion (Proposition 4.6).

Since $|x|$ is continuous in x, the set $A = \{x : |x| \leq \kappa\}$ is closed. By assumption, A is petite. So, it will follow from (4.28) that the underlying Markov process $X(\cdot)$ is positive Harris recurrent, and hence the queueing network is stable. The remainder of the proof is devoted to the demonstration of (4.99).

We break the demonstration of (4.99) into two main steps. Step 1 consists of translating the condition (4.73), for fluid limit stability, into a related limit (4.104) on the expected value of $|Z(\cdot)|$ for the queueing network. We restrict our attention here to sequences of pairs (a_n, x_n) as in (4.71), since this condition is assumed for fluid limits. In Step 2, we start the process at general x_n. After restarting it at times $c_1|x_n|$ where (4.71) is satisfied, we employ (4.104) from Step 1 to get (4.110). Using Proposition 4.15, we finish the proof by showing the corresponding limit, but with $|X(\cdot)|$ replacing $|Z(\cdot)|$, which gives us (4.99).

Step 1. Assume that the sequence of pairs (a_n, x_n) satisfies (4.71), and choose $\hat{z}$ so that $\hat{z} \geq \limsup_{n\to\infty} |z_n|/a_n$. We first show that

$$\frac{1}{a_n}|Z^{x_n}(N\hat{z}a_n)| \to 0 \quad \text{on } G \text{ as } n \to \infty, \tag{4.101}$$

for appropriate $N > 1$ not depending on the sequence.

By Proposition 4.12, any subsequence $\mathfrak{X}^{x_{i_n}}(\cdot)$ of $\mathfrak{X}^{x_n}(\cdot)$ has a fluid limit $\bar{\mathfrak{X}}(\cdot)$ along some further subsequence (a_{ℓ_n}, x_{ℓ_n}), as in (4.72). Since the family

of fluid limits that is associated with the queueing network is assumed to be stable and $\hat{z} \geq |\bar{Z}(0)|$, this implies

$$\bar{Z}(t) = 0 \quad \text{for } t \geq N\hat{z} \tag{4.102}$$

and appropriate N, where we can assume $N > 1$. Restating the limit in terms of the original process, after setting $t = N\hat{z}$, one obtains

$$\frac{1}{a_{\ell_n}} |Z^{x_{\ell_n}}(N\hat{z}a_{\ell_n})| \to 0 \quad \text{as } n \to \infty. \tag{4.103}$$

Since a subsequence (a_{ℓ_n}, x_{ℓ_n}) satisfying (4.103) exists for every subsequence (a_{i_n}, x_{i_n}) of (a_n, x_n), the limit holds along the entire sequence as in (4.101).

We claim that

$$\frac{1}{a_n} E|Z^{x_n}(N\hat{z}a_n)| \to 0 \quad \text{as } n \to \infty. \tag{4.104}$$

On account of (4.101), it suffices to show that the sequence there is uniformly integrable for large enough n. This follows from (4.88) of Proposition 4.14, since $\limsup_{n\to\infty} |x_n|/a_n < N\hat{z}$.

Step 2. We now allow x_n to be arbitrary, assuming only that $|x_n| \to \infty$ as $n \to \infty$. We wish to restart the processes $X^{x_n}(\cdot)$ at times $c_1|x_n|$, where $c_1 \geq 1$ is nonrandom, and at which the following three limits hold for $\omega \in G$:

$$\limsup_{n\to\infty} \frac{1}{|x_n|} |Z^{x_n}(c_1|x_n|)| \leq c_1|\alpha| + 1, \tag{4.105}$$

$$\frac{1}{|x_n|} |U^{x_n}(c_1|x_n|)| \to 0 \quad \text{as } n \to \infty, \tag{4.106}$$

$$\frac{1}{|x_n|} |V^{x_n}(c_1|x_n|)| \to 0 \quad \text{as } n \to \infty. \tag{4.107}$$

The first limit follows from (4.87) of Proposition 4.14, with arbitrary c_1, and the last two limits follow from (4.91) of Proposition 4.15, with $c_1 \geq c_0$, where c_0 is given in Proposition 4.15.

We set

$$a'_n = |x_n|, \quad x'_n = X^{x_n}(c_1|x_n|), \quad z'_n = Z^{x_n}(c_1|x_n|),$$

$$u'_n = U^{x_n}(c_1|x_n|), \quad v'_n = V^{x_n}(c_1|x_n|).$$

On account of (4.105)-(4.107), the random sequence (a'_n, x'_n) satisfies (4.71) on the set G. For $\hat{z}' \stackrel{\text{def}}{=} c_1|\alpha| + 1$, one has $\hat{z}' \geq \limsup_{n\to\infty} |z'_n|/a'_n$ on all such sequences. It follows from (4.104) that

$$\frac{1}{|x_n|} E_{x'_n} |Z(N\hat{z}'|x_n|)| \to 0 \quad \text{on } G \text{ as } n \to \infty. \tag{4.108}$$

(In (4.108), we write the initial state x'_n outside the expectation since x'_n is random; for nonrandom initial states as in (4.104), the two formulations are of course equivalent.)

Set $c = c_1 + N\hat{z}' = c_1 + N(c_1|\alpha| + 1)$. Since $X(\cdot)$ is Markov and $P(G) = 1$, it follows from (4.108) that

$$\frac{1}{|x_n|} E_{x_n}[Z(c|x_n|) \mid \sigma(X(c_1|x_n|))] \to 0 \quad \text{as } n \to \infty \tag{4.109}$$

holds a.s. It is not difficult to check that, because of (4.109),

$$\frac{1}{|x_n|} |Z^{x_n}(c|x_n|)| \to 0 \quad \text{as } n \to \infty$$

holds in probability. On account of (4.88) of Proposition 4.14, the sequence is uniformly integrable. So, in fact,

$$\frac{1}{|x_n|} E|Z^{x_n}(c|x_n|)| \to 0 \quad \text{as } n \to \infty. \tag{4.110}$$

On the other hand, by (4.93) of Proposition 4.15,

$$\frac{1}{|x_n|} E|U^{x_n}(c|x_n|)| \to 0, \ \frac{1}{|x_n|} E|V^{x_n}(c|x_n|)| \to 0 \quad \text{as } n \to \infty. \tag{4.111}$$

Together with (4.110), these two limits imply that

$$\frac{1}{|x_n|} E|X^{x_n}(c|x_n|)| \to 0 \quad \text{as } n \to \infty,$$

which is (4.99). This completes the proof of Theorem 4.16. ■

The reader should note that, in the proof of Theorem 4.16, the only places where the stability of the fluid limits is used are (a) in (4.102), and (b) where Proposition 4.15 is applied. Stability is used in (b) only to conclude the queueing network is subcritical. The latter use of stability can of course be avoided by directly assuming the queueing network is subcritical. The condition in (a) is the major "given" in the theorem, with verification of the stability of the fluid limits, as needed in (4.102), being left to specific applications.

The definition of fluid limit, that is given in Section 4.3, differs somewhat from that given in [Da95]. There, one sets $a_n = |x_n|$, and, unlike (4.71), makes no assumptions on u_n and v_n. This permits the limiting residual times $\bar{u}$ and $\bar{v}$ to be different than 0. The argument for deriving (4.99) is then essentially that given in Step 1 of the above proof together with the limits in (4.111), except that one now requires the stability of the corresponding fluid limits with delay (which will satisfy the corresponding fluid model equations with delay). The approach used here and in [Br98a] allows one to avoid dealing with fluid limits and fluid models with delay altogether. (An intermediate approach for fluid models was given in [Ch95].)

As noted at the end of the preceding sections, analyzing the Markov process $X(\cdot)$ becomes easier when the state space is countable. We point out that the same is true here, although in this context, the absence of coordinates corresponding to the exponential residual times, rather than countability, is what is important. In this simpler setting, the last two displays in Lemma 4.13 are not needed, nor is Proposition 4.15. The proof of Theorem 4.16 simplifies, with Step 2 no longer being needed. In Step 1, one can instead set $a_n = |x_n| = |z_n|$. Then, (4.104) immediately implies (4.99), with $c = N$.

The stability assumptions in Theorem 4.16 are sufficient for most applications. On occasion, somewhat greater flexibility in the definition of fluid limits is helpful. One can, for instance, replace the assumption of fluid limit stability with

$$\frac{1}{a_n}|Z^{x_n}(N\hat{z}a_n)| \to 0 \quad \text{in probability as } n \to \infty, \tag{4.112}$$

for all pairs of sequences (a_n, x_n) satisfying (4.71). This generalizes pointwise convergence in (4.101), at the beginning of Step 1 in the proof of the theorem, to convergence in probability. The conclusion (4.104), in Step 1, follows as before from the uniform integrability condition on $|Z(t)|$ in (4.88). In [Br99], the concept *asymptotic stability* is used; this is a weaker variant of fluid limit stability that implies (4.112). As in the previous paragraph, matters simplify in the exponential setting when employing (4.112).

Throughout this chapter, we have assumed that queueing networks are head-of-the-line. If one wishes, one can replace this assumption by the assumption that only a bounded number of jobs at any class and time have already received some service there. The approach is then essentially the same as before, with only the obvious modifications being made at certain points. One needs to replace the residual service time vector, in the construction of the state space and the Markov process in the first part of Section 4.1, by a vector with correspondingly more components. Under the corresponding new definition of the norm $|x|$, the same results leading up to Theorem 4.16 hold as before. The basic fluid model equations (4.50)-(4.55) are still valid. (The queueing network inequality (4.47) leading to (4.55) needs to be modified, though.) Other results where the reasoning proceeds as before include Proposition 4.12, where the fluid limits satisfy the basic fluid model equations, and Proposition 4.15, where the bounds on $|V^x(\cdot)|$ continue to hold.

This approach fails when the number of jobs that have received partial service is allowed to become unbounded. In particular, the bounds on $|V^x(\cdot)|$ in Proposition 4.15 can fail, and $|V^x(c|x|)|$ can be of the same order of magnitude as $|x|$. Changing the sum norm on $|v|$ to the max norm does not help, since the relationship between fluid limits and the basic fluid model equations in Proposition 4.12 need not continue to hold.

e-stability

By strengthening the assumptions in Theorem 4.16 so that the sets A there are uniformly small, one can show, in Theorem 4.17, that the queueing

network is e-stable. Theorem 4.17 follows by using the same argument as in Theorem 4.16, but instead applying the last part of Proposition 4.6 to it. A similar result is given in [Du96].

Theorem 4.17. *Assume that an HL queueing network satisfies the same conditions as in Theorem 4.16, except that the condition that $A = \{x : |x| \leq \kappa\}$ be petite is replaced by the condition that, for each $\kappa > 0$, A be uniformly small on an interval $[s_1, s_2]$, with $0 < s_1 < s_2$. Then, the queueing network is e-stable.*

Proof. Whenever a set is small, it is also petite. So, all of the assumptions in the statement of Theorem 4.16 are satisfied. Consequently, all steps in the proof remain valid, which include the bound and limit given in (4.98) and (4.100). As before, they together imply the basic assumption (4.28) for the Multiplicative Foster's Criterion (Proposition 4.6), for some choice of κ.

Since $|x|$ is continuous in x, the set A is closed. By assumption, it is uniformly small. Employing the last part of the Multiplicative Foster's Criterion, it follows that $X(\cdot)$ is ergodic. Hence the queueing network is e-stable. ■

Under the assumptions (4.30) and (4.31) on the interarrival times, it follows from Proposition 4.7 that the sets $A = \{x : |x| \leq \kappa\}, \kappa > 0$, are uniformly small. It therefore follows from Theorem 4.17 that, under (4.30) and (4.31), a queueing network will be e-stable if either its associated fluid limits or associated fluid model is stable.

In Chapter 5, we will employ Theorems 4.16 and 4.17 to demonstrate the stability/e-stability for queueing networks with different disciplines. In each case, the main part of the argument consists of showing the stability of an associated fluid model. The amount of work involved depends on the discipline. Here, as an elementary illustration of the procedure, we show the following.

Example 1. *Assume that (4.30)-(4.31) holds for an HL queueing network with $\sum_j \rho_j < 1$. Then, the queueing network is e-stable.*

Proof. By Proposition 4.7 and Theorem 4.17, it suffices to show that an associated fluid model is stable. We do this for the basic fluid model, which, by Proposition 4.12, is always associated with the network. By Part (a) of Proposition 4.11, this fluid model is stable. Consequently, the queueing network is e-stable. ■

The assumption $\sum_j \rho_j < 1$ is sufficiently strong so that the above example can be proven directly, with a bit of work. For this, one can apply the Multiplicative Foster's Criterion to the total workload of the queueing network. (The total workload consists of all future work required of all jobs currently in the network.) As one should expect, the assumption that the discipline be HL is not necessary here, although one then needs to reformulate the definition

of the underlying Markov process given in Section 4.1 and its accompanying state space. For less elementary examples, a direct proof of stability or e-stability will be quite tedious, if feasible.

4.5 Appendix

The purpose of this section is to serve as an appendix for Section 4.1 and to go into more detail on background material that was omitted there. We cover here four topics. We first discuss the connection between Borel right processes and piecewise-deterministic Markov processes, which was only mentioned briefly in Section 4.1. We then go into more detail on three topics connected with recurrence that were mentioned in the subsection on Harris recurrence. In Proposition 4.18, we show the equivalence of Harris recurrence and of positive Harris recurrence for a Markov process X and its R-chain $\tilde{X}$, and that X and $\tilde{X}$ have the same stationary measures. We next summarize how one can demonstrate Theorem 4.1 for a process X by employing its R-chain. We then summarize the argument showing the existence of a stationary measure for a discrete time Markov process with a petite set A that satisfies the discrete time analog of (4.19).

Borel right processes and PDPs

We first state the definition of a Borel right process, and then show that the class of PDPs that are associated with HL queueing networks are Borel right processes. For the definition of Borel right, we follow Section 27 of [Da93]. Other references are [Ge79], [Kn84], and [Sh88]; [Da93] relies on the approach given in [Sh88].

We assume that the σ-algebras associated with the continuous time Markov process X satisfy the completeness and right continuity conditions given in (4.5) and (4.6). The conditions on page 77 of [Da93] state that $X(\cdot)$ is Borel right if, in addition:

(a) The state space S is a Lusin space.
(b) X has a semigroup P^t that maps $B(S)$ into $B(S)$.
(c) The sample paths $t \mapsto X(t)$ are a.s. right continuous.
(d) If f is an α-excessive function for P^t, for some $\alpha \geq 0$, then the sample path $t \mapsto f(X(t))$ is a.s. right continuous.

We proceed to explain the terms that are employed above, and show that these four conditions are satisfied for the Markov processes X underlying the HL queueing networks we have introduced. The first three conditions are quite general, and do not pose any serious constraints. Condition (d) is needed for the regularity of the process. In our setting, none is difficult to show.

A topological space is called a Lusin space if it is homeomorphic to a Borel subset of a compact metric space. As pointed out in [Sh88], a locally

compact Hausdorff space with countable basis is a Lusin space, since its one point compactification is compact and metrizable. The space S introduced in Section 4.1 satisfies these properties, and so is a Lusin space. In (b), $B(S)$ denotes the bounded measurable functions on S. The property holds since P^t is a probability transition kernel. As mentioned in Section 4.1, this is showed in [Da93]. The argument consists of writing $P^t f$, $f \in B(S)$, as the limit of the n-fold iterates of $G^t f$, where G^t is the truncated operator obtained by stopping X at the time of its first jump, and employing the measurability of G^t. In (c) and (d), "a.s." means almost surely with respect to all initial probability measures μ. Since the sample paths of X are right continuous by construction, property (c) is automatic.

A function f is *α-super-mean-valued*, for some $\alpha \geq 0$, if $f(x) \geq 0$ and $e^{-\alpha t} P^t f(x) \leq f(x)$ for all $t \geq 0$ and $x \in S$. The function f is *α-excessive* if it is α-super-mean-valued and $e^{-\alpha t} P^t f(x) \uparrow f(x)$ as $t \downarrow 0$. In our setting, it is easy to see that (d) holds for the larger family of functions where $P^t f(x) \to f(x)$ as $t \downarrow 0$. To see this, let $t < s$ be close enough so that no jumps occur in $(t, s]$ along a given sample path. Then, since $X(\cdot)$ is piecewise deterministic,

$$f(X(s)) = f(P^{s-t}(X(t))) = f(P^{s-t}(x)),$$

for $x = X(t)$. As $s \downarrow t$, the last quantity converges to $f(x) = f(X(t))$, which shows $f(X(\cdot))$ is right continuous, and hence shows (d).

By (i) of Theorem 9.4 of [Ge79], under the conditions (a)-(c), the assumption that X is strong Markov is equivalent to condition (d). So, (d) may be replaced by

(d′) The process X is strong Markov.

As mentioned in Section 4.1, PDPs are strong Markov, and therefore so are our processes X that are associated with HL queueing networks. The demonstration of (d) is of course quicker than (d′) for such processes, but most readers will presumably find the latter condition more familiar.

Recurrence for a Markov process and its R-chain

Let $X(t)$, $t \geq 0$, be a Markov process and $\tilde{X}(n)$, $n = 0, 1, 2, \ldots$, be its R-chain. The process X inherits certain properties from $\tilde{X}$, which we used in Section 4.1. We demonstrate them now in Proposition 4.18. For this, we employ the following standard result from discrete time Markov process theory, which can be found, for instance, in Chapter 10 of [MeT93d]: if a Markov process $Y(n)$, $n = 0, 1, 2, \ldots$, is Harris recurrent, then it has a stationary measure, and this measure is unique up to a constant multiple. (We will outline part of the argument in the last subsection.) Much of the proof for the proposition is from [MeT93a].

Proposition 4.18. *The process X is Harris recurrent, respectively positive Harris recurrent, if and only if $\tilde{X}$ is. In either case, X and $\tilde{X}$ have the same stationary measures. Consequently, X has a stationary measure, which is unique up to a constant multiple.*

Proof. The existence and uniqueness of a stationary measure for X follows immediately from the preceding statement in Proposition 4.18, and the comment before the proposition about discrete time Markov processes.

It is immediate from the characterization of recurrence in (4.10) that when $\tilde{X}$ is recurrent with respect to the measure φ, then so is X. For the other direction, it suffices to show that for $\varphi(A) > 0$,

$$P_x(\tilde{\tau}_A < \infty) = 1 \quad \text{for all } x,$$

where $\tilde{\tau}_A$ is the hitting time of A for $\tilde{X}$ constructed as in (4.13). This follows immediately from (4.9) and the formula

$$P_x(\tilde{\tau}_A < \infty) = 1 - E_x[e^{-\eta_A}], \tag{4.113}$$

which is given in Theorem 2.3 of [MeT93a]. Intuitively, (4.113) is not difficult to see since, conditioning on the process X, the number of indices n, at which $\tilde{X}(n) = X(\sigma_n) \in A$, will be Poisson with mean η_A, and so

$$P_x(\tilde{\tau}_A = \infty \mid \sigma(X)) = e^{-\eta_A} \quad \text{for all } x.$$

Suppose now that a measure π is stationary for P^t. Then, $\pi P^t = \pi$, and integration against e^{-t} on both sides gives $\pi R = \pi$. So, π is also stationary for R.

It remains to show that if π is stationary for R, then it is also stationary for P^t. Assuming π is stationary for R, we first show that $\pi^t \stackrel{\text{def}}{=} \pi P^t,\ t > 0$, is also stationary for R, that is,

$$\pi^t = \pi^t R. \tag{4.114}$$

Arguing as in Theorem 3.1 of [MeT93a], for $A \in \mathscr{S}$, one has

$$\begin{aligned} \pi P^t R(A) &= \int_0^\infty e^{-s} \pi P^t P^s(A) ds = \int_0^\infty e^{-s} \pi P^s P^t(A) ds \\ &= \pi R P^t(A) = \pi P^t(A), \end{aligned}$$

where the first and third equalities come from reversing the order of integration, and the fourth holds since π is stationary for R. So, (4.114) also holds.

Since the stationary measure for R is unique up to a constant multiple, and both π and π^t are stationary for R, one has

$$\pi^t = c(t)\pi \quad \text{for } t \geq 0,$$

for some measurable function $c(t) \geq 0$. Since mass is conserved, one must have $\pi^t = \pi$ when $\tilde{X}$ is positive Harris recurrent (and hence $\pi(S) < \infty$). So, in this case, π is stationary for P^t. The same holds in general, because $P^{s+t} = P^s \circ P^t$, and so $c(s+t) = c(s)c(t)$, which implies $c(t) = e^{Ct}$ for some C. Therefore, for $A \in \mathscr{S}$,

$$\pi(A) = \pi R(A) = \int_0^\infty e^{-t}\pi^t(A)dt = \pi(A)\int_0^\infty e^{(C-1)t}dt,$$

which implies $C = 0$, and so, in fact, $\pi^t = \pi$, as desired. ■

The following proposition is a variant of Proposition 4.18, and will be used in the last subsection. It employs the resolvent

$$K(x, A) = \sum_{n=1}^{\infty} 2^{-n} P^n(x, A) \quad \text{for } x \in S,\ A \in \mathscr{S}, \tag{4.115}$$

for a discrete time Markov process Y. (Typically, summation is chosen to start at $n = 0$, with the coefficients modified accordingly.) The corresponding discrete time Markov process, with transition function K, will be referred to as a K-chain. On account of Proposition 4.19, it will be easier to work with the K-chain than with Y directly.

Proposition 4.19. *Suppose that a discrete time Markov process Y has a small set A that satisfies*

$$P_x(\tau_A < \infty) = 1 \quad \textit{for all } x \in S. \tag{4.116}$$

Then, A also satisfies (4.116) for the K-chain, with respect to which A will be small with $m_0 = 1$. Moreover, Y and its K-chain have the same stationary measures.

Proof. The argument that uses (4.113) can also be used to show (4.116) holds for the K-chain. (The number of indices in A, after conditioning on Y, will now be binomial instead of Poisson.) It is easy to see that A will also be small for the K-chain: one can choose $m_0 = 1$ and small measure $\nu = 2^{-m_0'}\nu'$, where ν' is the small measure for Y and m_0' is the time at which it is employed.

Suppose now that a measure π is stationary for P. Then, $\pi P = \pi$, and summation against 2^{-n} on both sides gives $\pi K = \pi$, so π is also stationary for K. For the other direction, assume that $\pi K = \pi$. One has the string of equalities

$$\pi P = \pi K P = 2\pi R - \pi P = 2\pi - \pi P.$$

Comparison of the first and last terms then implies $\pi P = \pi$, as desired. The first and third equalities follow from the stationarity of π with respect to K, and the second equality follows quickly from the definition of K. ■

Proof of Theorem 4.1

We summarize here the proof of Theorem 4.1. Our approach will be to first state a discrete time analog of Theorem 4.1, Theorem 4.20, and say a little about how it is shown. We will then summarize how Theorem 4.1 follows from a slight modification of Theorem 4.20 and Proposition 4.18. At various points, this subsection relies on ideas from [MeT93a].

Theorem 4.20 is similar to Theorem 4.1. Differences are that the set A is no longer required to be closed and, in (4.118), $\tau_A(\delta)$ is replaced by τ_A. (These simpler quantities suffice since $Y(\tau_A) \in A$ is automatic and $\tau_A(\delta) = \tau_A$ for $\delta \leq 1$.) In one direction of Part (a), the set A is chosen to be small (rather than just petite), since this is needed in the next subsection. The state space $(S, \mathscr{S})$ is not assumed to have a topological structure, although the σ-algebra $\mathscr{S}$ is assumed to be countably generated.

Theorem 4.20. *(a) If a discrete time Markov process Y is Harris recurrent, then there exists a small set A with*

$$P_x(\tau_A < \infty) = 1 \quad \textit{for all } x \in S. \tag{4.117}$$

Conversely, if (4.117) holds for a petite set A, then Y is Harris recurrent. (b) Suppose the discrete time Markov process Y is Harris recurrent. Then, Y is positive Harris recurrent if and only if there exists a petite set A for which

$$\sup_{x \in A} E_x[\tau_A] < \infty. \tag{4.118}$$

We discuss briefly the reasoning for the different parts of Theorem 4.20. As in the case of Theorem 4.1, the argument for the converse direction of Part (a) is elementary. It consists of repeatedly restarting Y at an increasing sequence of random times, between which Y has at least some probability $\epsilon > 0$ of hitting a specified set B, for which $\varphi(B) > 0$, where φ is the irreducibility measure. Here, in the discrete time setting, less theoretical justification is required, since the strong Markov property and the measurability of hitting times of sets are elementary.

The argument for the other direction of Theorem 4.20(a), namely the construction of a small set satisfying (4.117), is longer, and requires cleverness. It is sometimes referred to as Orey's C-Set Theorem. For a proof, one can consult pages 18-19 in [Nu84] or Theorem 5.2.1, on page 107 of [MeT93d]. The construction involves setting $\nu = \psi \mathbf{1}\{D\}$, where ψ is a maximal irreducibility measure for Y with total mass 1, and D is an appropriately defined "high density" set. We note here that in the setting of Section 4.1, where $\mathscr{S}$ is generated by a locally compact, separable metric, one can choose the set A so that it is also closed. (By Proposition 5.2.4(iii) of [MeT93d], there is a small set A' with $\psi(A') > 0$, for the maximal irreducibility measure ψ, and therefore a closed subset A of A' for which $\psi(A) > 0$ as well.)

In Part (b) of Theorem 4.20, the argument that positive Harris recurrence follows from the existence of a petite set A satisfying (4.118) is quite quick, if one assumes the following formula in (4.119). It is of interest in its own right, and says that the stationary measure π of a Harris recurrent discrete time Markov process Y satisfies

$$\pi(B) = \int_A E_x \left[\sum_{n=0}^{\tau_A - 1} \mathbf{1}\{Y(n) \in B\} \right] \pi(dx) \quad \text{for } B \in \mathscr{S}, \tag{4.119}$$

for any $A \in \mathscr{S}$ with $\psi(A) > 0$. Often, the indices in the sum are taken from $n = 1$ to τ_A (as in Theorem 10.0.1 in [MeT93d]). Setting $B = S$, we will employ the special case,

$$\pi(S) = \int_A E_x[\tau_A]\pi(dx). \tag{4.120}$$

The formula (4.119) may be motivated by the following informal argument. For $x \in S$ and $A, B \in \mathscr{S}$, with $\psi(A) > 0$, set

$$\pi_A(B) = \pi(A \cap B) \quad \text{and} \quad {}^A P(x, B) = P(x, B \cap A^c). \tag{4.121}$$

Then, (4.119) may be rewritten as $\pi = \pi_A \sum_{n=0}^{\infty} {}^A P^n$. Assume that $(I - {}^A P)^{-1} = \sum_{n=0}^{\infty} {}^A P^n$ exists. Then, this is equivalent to

$$\pi(I - {}^A P) = \pi_A,$$

that is,

$$\pi\, {}^A P = \pi_{A^c},$$

which follows immediately from the definition in (4.121) and the stationarity of π.

In order to employ (4.120), we first note that $\pi(A) < \infty$ must hold if the set A is petite with respect to any measure ν. Otherwise, it is not difficult to check that, from the definition of petite, $\pi(B) = \infty$ must hold for any B with $\nu(B) > 0$. Since π is assumed to be σ-finite, this is not possible. Applying (4.118) to (4.120), one therefore obtains

$$\pi(S) \leq \pi(A) \sup_{x \in A} E_x[\tau_A] < \infty.$$

So, Y is positive Harris recurrent, as desired.

The argument for the other direction of Theorem 4.20(b), which involves the construction of a petite set satisfying (4.118) is, not surprisingly, longer. The result is stated in Theorem 11.0.1 of [MeT93d]. It involves the use of *regular sets*, with Theorem 11.1.4 providing the key decomposition of S. We note here again that, in the topological setting of Section 4.1, one can choose A so that it is also closed. (In Theorem 11.0.1, one can choose a regular set A' with $\psi(A') > 0$, and therefore a closed regular subset A of A' for which $\psi(A) > 0$; both A' and A will be petite.)

This completes our discussion of Theorem 4.20. We now summarize how Theorem 4.1 follows from a slight modification of Theorem 4.20 and Proposition 4.18. As in Theorem 4.20, there are two parts to show, each with two directions. Recall that X denotes a continuous time Markov process as in Section 4.1 and $\tilde{X}$ denotes the corresponding R-chain.

The argument that the Harris recurrence of X follows from the existence of a petite set satisfying (4.19) is elementary, and was summarized in Section 4.1. To show the other direction of Theorem 4.1(a), we first note that, by

Proposition 4.18, if X is Harris recurrent, then so is $\tilde{X}$. By Theorem 4.20(a), there exists a small set A for $\tilde{X}$, with corresponding time m_0, for which (4.117) is satisfied. As mentioned in the outline of the proof of Theorem 4.20, A can be chosen so that it is also closed. Choosing a in (4.18) to be the m_0-fold convolution of the exponential distribution, it follows that A is petite for X. Using (4.13), it is easy to see that since (4.117) holds for $\tilde{X}$, its analog (4.19) holds for X. So, A has the properties stated in Theorem 4.1(a).

In order to show Part (b) of Theorem 4.1, we again employ Theorem 4.20 and Proposition 4.18. We also need the following comparisons between the bounds on the hitting times of petite sets A, in (4.20) and (4.118), for the processes X and $\tilde{X}$. These comparisons are taken from Section 4 of [MeT93a].

Proposition 4.21. *(a) Suppose that for the continuous time Markov process X, (4.20) is satisfied for some closed petite set A. Then, there is a petite set A' for X, such that (4.118) is satisfied for the R-chain $\tilde{X}$. (b) Suppose that for the R-chain $\tilde{X}$ of X, (4.118) is satisfied for some petite set A of $\tilde{X}$. Then, there is a closed petite set A' for the process X which satisfies (4.20).*

We note that, in [MeT93a], it is shown that the set A in Part (b) is regular for X, with $\psi(A) > 0$ for the maximal irreducibility measure ψ. The existence of a closed petite subset A' satisfying (4.20) then follows. (This application of regularity is analogous to that mentioned at the end of the discussion of the proof of Theorem 4.20.)

Using Proposition 4.21, we now show Theorem 4.20(b). Suppose that (4.20) is satisfied for some closed petite set A. Then, the petite set A' in Part (a) of the proposition must have $\pi(A') < \infty$, for the reasons mentioned in the demonstration of Theorem 4.20(b). Here, π is the stationary measure for both X and $\tilde{X}$. Since A' satisfies (4.118), one may apply the formula (4.120), as in the demonstration of Theorem 4.20(b), to conclude that $\pi(S) < \infty$. Hence, X is positive Harris recurrent, as desired.

For the other direction of Theorem 4.1(b), we assume that X is positive Harris recurrent. By Proposition 4.18, $\tilde{X}$ is also positive Harris recurrent, and so by Theorem 4.20(b), there is a petite set A of $\tilde{X}$ that satisfies (4.118). Applying Proposition 4.21(b), it follows that there exists a closed petite set A', for the process X, which satisfies (4.20). This completes our discussion of Theorem 4.1.

Existence and uniqueness of stationary measures

One of the fundamental properties of discrete time Harris recurrent Markov processes is the existence of a stationary measure that is unique up to a constant multiple. This is a standard result in discrete time Markov process theory and can be found in a number of places, such as [Nu84] and [MeT93d]. A short account is given in [Dur96], which we follow here in summarizing the ideas behind the proof. Because of Proposition 4.18, this result immediately extends to continuous time Markov processes.

On account of Part (a) of Theorem 4.20, in order to show the existence of a stationary measure for a discrete time Harris recurrent Markov process, it suffices to show the following result. We will discuss uniqueness of the stationary measure afterwards.

Theorem 4.22. *Suppose that Y is a discrete time Markov process with a small set A satisfying (4.117). Then, there exists a stationary measure π.*

The reasoning for Theorem 4.22 can be broken into two main steps. The first step consists of showing that the conclusion of the theorem holds when S contains a *recurrent atom* α for the process Y. By this, we mean that $P_y(\tau_\alpha < \infty) = 1$ for all y, where $\tau_\alpha \stackrel{\text{def}}{=} \tau_{\{\alpha\}}$. (The use here of the term recurrent atom is not standard.) The second step shows that there is an appropriate "lumping together" of points (sometimes referred to as "splitting" in the literature) that allows one to construct a process $\bar{Y}$ with such an atom and which has the same recurrence behavior as Y.

Before beginning, we note that, because of Proposition 4.19, one can assume without loss of generality that $m_0 = 1$ for the small set A. Also, by scaling the corresponding small measure ν by $\epsilon = 1/\nu(S)$, one can replace the inequality (4.18) by

$$\nu(B) \le \epsilon P(y, B) \quad \text{for } y \in A,\ B \in \mathscr{S}, \tag{4.122}$$

where ν is now a probability measure.

For the first step, we assume that S has a recurrent atom α. Then, $P_\alpha(\tau_\alpha < \infty) = 1$, and it is not difficult to show that the measure π defined by

$$\pi(B) = E_\alpha \left[\sum_{n=0}^{\tau_\alpha - 1} \mathbf{1}\{Y(n) \in B\} \right] \tag{4.123}$$

is σ-finite. (See Exercise 6.8 on page 331 of [Dur96].)

We claim that π is stationary for the process Y. The proof is essentially the same as that for Markov chains, which is given for Theorem 4.3, on page 303 of [Dur96]. It can be motivated by using the "cycle trick", noting that the number of visits to any $y \in S$ over the time set $\{0, \ldots, \tau_\alpha - 1\}$ is always the same as over $\{1, \ldots, \tau_\alpha\}$, and taking expectations over both sides. It follows from this that $\pi P = \pi$. Formula (4.123) is a special case of (4.119), with $A = \{\alpha\}$. (Note that the representation of π in (4.119) is not explicit, whereas the right side of (4.123) does not involve π.)

For the second step of the proof, one wishes to compare the given Markov process Y with one having a recurrent atom. The small set A will be used to construct the atom, and (4.117) will be used to show the atom is recurrent. We begin by appending a point α to the state space S, thus creating a new space $\bar{S}$ and a corresponding Borel σ-algebra $\bar{\mathscr{S}}$,

$$\bar{S} = S \cup \{\alpha\} \quad \text{and} \quad \bar{\mathscr{S}} = \{B, B \cup \{\alpha\} : B \in \mathscr{S}\}.$$

One can define a probability transition kernel $\bar{P}$ on $(\bar{S}, \bar{\mathscr{S}})$ by

$$\begin{aligned}
\bar{P}(y, \{\alpha\}) &= \begin{cases} \epsilon & \text{for } y \in A, \\ 0 & \text{for } y \in S - A, \end{cases} \\
\bar{P}(y, B) &= \begin{cases} P(y, B) - \epsilon\nu(B) & \text{for } y \in A,\ B \in \mathscr{S}, \\ P(y, B) & \text{for } y \in S - A,\ B \in \mathscr{S}, \end{cases} \\
\bar{P}(\alpha, B) &= \int \bar{P}(y, B)\nu(dy) \quad \text{for } B \in \bar{\mathscr{S}},
\end{aligned} \tag{4.124}$$

where ν and $\epsilon > 0$ are chosen as in (4.122). We denote by $\bar{Y}$ the corresponding Markov process.

We also find it convenient to introduce the probability transition kernel V on $(\bar{S}, \bar{\mathscr{S}})$, with

$$V(y, \{y\}) = 1 \ \text{ for } y \in S, \quad V(\alpha, B) = \nu(B) \ \text{ for } B \in \mathscr{S}. \tag{4.125}$$

Note that $V(y, \alpha) = 0$ for all $y \in \bar{S}$. One can check that

$$V\bar{P} = \bar{P} \quad \text{and} \quad (\bar{P}V)\,|_S = P. \tag{4.126}$$

Here, $|_S$ denotes the restriction to S in both the domain and range. As usual, we are employing the convention that the transition kernel on the left is the first to be applied to a given initial measure. We will demonstrate a version of (4.126) below (4.128).

The construction of $\bar{P}$ can be motivated as follows. One wishes to "lump together" points in S in some way so as to form an atom α, while not changing, in essence, the transition rule P. If one is to lump together points according to some weight ν, transitions to and from α, by the new transition rule $\bar{P}$, must both be done in a manner consistent with this weight. Since A is assumed to be small, with $m_0 = 1$ and with (4.122) holding, one can choose $\bar{P}(y, \{\alpha\})$ as on the top two lines of (4.124); the bottom line of (4.124) respects the weight ν. The third and fourth lines of (4.124) define $\bar{P}$ according to the old transition rule P on the mass that has not been directed to α.

To develop some feel for (4.124), one can consider the case where S is discrete. (In this case, there is of course no need to lump together points to create an atom.) Denoting the transition densities by $p(y, y')$ and setting $q(y) = \nu(\{y\})$, (4.124) becomes

$$\begin{aligned}
\bar{p}(y, \alpha) &= \begin{cases} \epsilon & \text{for } y \in A, \\ 0 & \text{for } y \in S - A, \end{cases} \\
\bar{p}(y, z) &= \begin{cases} p(y, z) - \epsilon q(z) & \text{for } y \in A,\ z \in S \\ p(y, z) & \text{for } y \in S - A,\ z \in S, \end{cases} \\
\bar{p}(\alpha, z) &= \sum_i \bar{p}(y_i, z) q(y_i) \quad \text{for } z \in \bar{S}.
\end{aligned} \tag{4.127}$$

Here $\{y_i\}$ denotes an enumeration of the points in S. Similarly, one can define v by

$$v(y,y) = 1,\ v(\alpha,y) = q(y) \quad \text{for } y \in S.$$

The equations in (4.126) become

$$(v\bar{p})(y,z) = \bar{p}(y,z) \quad \text{and} \quad (\bar{p}v)(y,z) = p(y,z), \tag{4.128}$$

with $y,z \in \bar{S}$ in the first equation and $y,z \in S$ in the second. The first equation follows from the invariance of v on $y \in S$ and

$$(v\bar{p})(\alpha,z) = \sum_i \bar{p}(y_i,z)q(y_i) = \bar{p}(\alpha,z).$$

For the second equation, one has

$$\begin{aligned}(\bar{p}v)(y,z) &= \sum_i \bar{p}(y,y_i)v(y_i,z) + \bar{p}(y,\alpha)v(\alpha,z),\\ &= \bar{p}(y,z) + \bar{p}(y,\alpha)q(z)\end{aligned}$$

for $y,z \in S$. For both $y \in A$ and $y \in S - A$, it follows from the definition of $\bar{p}$ that the last line equals $p(y,z)$, as desired. The equations in (4.126) can be derived similarly.

We still need to show that the atom α is recurrent. For this, it suffices to show

$$\sum_n \bar{P}^n(x,\{\alpha\}) = \infty \quad \text{for all } x \in \bar{S}. \tag{4.129}$$

We employ the string of inequalities

$$\begin{aligned}\sum_n \bar{P}^n(x,\{\alpha\}) &= \sum_n V(\bar{P}V)^{n-1}\bar{P}(x,\{\alpha\}) \geq \epsilon \sum_n V(\bar{P}V)^{n-1}(x,A)\\ &\geq \epsilon \min_{x\in S} \sum_n (\bar{P}V)^{n-1}(x,A) \geq \epsilon \min_{x\in S} \sum_n P^{n-1}(x,A) = \infty.\end{aligned}$$

The first equality follows from the first half of (4.126), the first inequality follows from the first line of (4.124), and the second inequality follows from (4.125). The last inequality is a consequence of the second half of (4.126), since restricting the domain decreases the corresponding integrals. (Because of (4.125), one actually has equality.) The equality at the end follows from Borel-Cantelli and (4.117), since A will be hit infinitely often with probability 1. So, (4.129) holds.

We now tie together the two previous steps involving the creation and manipulation of the atom α. If one assumes that the Markov process Y given in Theorem 4.22 has a petite set A satisfying (4.117), it follows that the process $\bar{Y}$ defined by (4.124) has a recurrent atom at α. The measure $\bar{\pi}$ given in (4.123) for $\bar{Y}$ is then stationary for $\bar{Y}$. We claim that the measure $\pi \stackrel{\text{def}}{=} (\bar{\pi}V)|_S$,

given by the restriction of $\bar{\pi}V$ to S, is stationary for Y. Theorem 4.22 follows immediately from this.

To show the claim, it suffices to verify the string of equalities

$$\pi P = (\bar{\pi}V)\,|_S\, P = (\bar{\pi}\bar{P}V)\,|_S = (\bar{\pi}V)\,|_S = \pi. \tag{4.130}$$

Here, the third equality follows from the stationarity of $\bar{\pi}$ and the first and fourth follow from the definition of π. The second equality follows from

$$(\bar{\pi}V)\,|_S\, P = (\bar{\pi}V)\,|_S\, (\bar{P}V)\,|_S = (\bar{\pi}V\bar{P}V)\,|_S = (\bar{\pi}\bar{P}V)\,|_S, \tag{4.131}$$

where one applies first the second half of (4.126) and last the first half of (4.126). The middle equality in (4.131) holds since $\bar{\pi}V(\alpha) = 0$, and so no mass is lost by the restriction $(\bar{\pi}V)\,|_S$. This demonstrates (4.131), and hence (4.130) and the stationarity of π, as desired.

One can also show that the stationary measure of a discrete time Harris recurrent Markov process is unique, up to a constant multiple. As before, on account of Part (a) of Theorem 4.20, it suffices to show the following version.

Theorem 4.23. *Suppose that Y is a discrete time Markov process with a small set satisfying (4.117). Then, up to a constant multiple, the stationary measure of Y is unique.*

We note that some sort of condition, corresponding to (4.117), is of course necessary. Without it, as in the countable state space setting, states may not communicate and there may be many stationary measures.

We summarize here the argument for Theorem 4.23. As in the demonstration of the existence of a stationary measure, we can assume without loss of generality that $m_0 = 1$ for the small set A, because of Proposition 4.19. The following argument, for $m_0 = 1$, can be found in Theorem 6.7, on page 331 of [Dur96].

As before, one employs the process $\bar{Y}$ corresponding to (4.124), which has a recurrent atom. Let ρ be a σ-finite measure on $(S, \mathscr{S})$, and denote by $\rho^{\bar{S}}$ its extension to $(\bar{S}, \bar{\mathscr{S}})$ given by

$$\rho^{\bar{S}}(B) = \rho(B \cap S) \quad \text{for } B \subseteq \bar{S}.$$

If ρ is stationary for Y, one can show that $\bar{\rho} = \rho^{\bar{S}}\bar{P}$ is stationary for the process $\bar{Y}$. This is done in Lemma 6.6, on page 331 of [Dur96].

In the countable state space setting when all states communicate, it is not difficult to show that $\bar{\rho} = \bar{\rho}(\alpha)\bar{\pi}$, where $\bar{\pi}$ is the stationary measure for $\bar{Y}$ defined in (4.123). This is done in Theorem 4.4, on page 305 of [Dur96], and the argument in the general state space setting is essentially the same. Set $\pi = (\bar{\pi}V)\,|_S$. One then has

$$\begin{aligned} \rho = \rho P &= \rho((\bar{P}V)\,|_S) = (\rho^{\bar{S}}\bar{P}V)\,|_S \\ &= (\bar{\rho}V)\,|_S = (\bar{\rho}(\alpha)\bar{\pi}V)\,|_S = \bar{\rho}(\alpha)\pi. \end{aligned} \tag{4.132}$$

Here, the first equality follows from the stationarity of ρ, the fourth and sixth equalities from the definitions of $\bar{\rho}$ and π, and the second equality from the second half of (4.126). The third equality holds since $\bar{\rho}(\{\alpha\}) = 0$. By (4.132), $\rho = \bar{\rho}(\alpha)\pi$, which shows that ρ is a constant multiple of π, as desired.

5

Applications and Some Further Theory

Theorem 4.16, from Section 4.4, gives conditions under which an HL queueing network will be stable. Theorem 4.17, from the same section, gives similar conditions under which a network will be e-stable. In the first three sections of Chapter 5, we will demonstrate stability/e-stability for three families of queueing networks by using these criteria. In Section 5.1, we do this for a family of a networks that includes single class networks, and in Section 5.2, we do this for two families of SBP reentrant lines, FBFS and LBFS. In Section 5.3, we apply the same criteria to FIFO networks of Kelly type.

In Sections 5.4 and 5.5, we address other topics connected with stability. Global stability is introduced in Section 5.4. The condition says that the network is stable irrespective of the HL discipline that is employed. It will follow directly from Theorem 4.16 that a sufficient condition for global stability is the stability of the associated basic fluid model. For two stations, there is a developed theory for these fluid models, which we summarize. Important underlying concepts include virtual stations and the push start phenomenon. We also discuss briefly rate stability and global rate stability for queueing networks, and the corresponding concept of weak stability for fluid models.

In Section 5.5, we investigate the converse direction to Theorem 4.16. Namely, does queueing network stability, under reasonable side conditions, imply fluid model stability? In particular, are queueing network and fluid model stability in some sense equivalent? Most of this section is spent on two examples that show this is not always the case. Virtual stations and push starts are again useful tools in this context.

Our approach in Sections 5.1-5.3 will be based on the following considerations involving Theorems 4.16 and 4.17. The conditions in the theorems are of two types, where (a) one needs the bounded sets $A = \{x : |x| \leq \kappa\}$, $\kappa > 0$, to be either petite or uniformly small and (b) one needs the fluid limits or fluid model, which is associated with the queueing network, to be stable. As noted in Section 4.4, the latter of the two stability conditions is the more tractible one, and is the one that is typically applied in practice. In the next three sections, we will demonstrate the stability of fluid models that are associated

M. Bramson, *Stability of Queueing Networks*. Lecture Notes in Mathematics 1950.

with the queueing networks there; the desired results then follow by applying Theorems 4.16 and 4.17. The first two cases are fairly quick, and the third takes a bit longer to show.

In order to avoid repetition later on, we will, in each case, explicitly assume that

$$A = \{x : |x| \leq \kappa\} \text{ is petite for each } \kappa > 0. \tag{5.1}$$

This condition suffices, in Theorem 4.16, for stability of the queueing network, once the fluid model has been shown to be stable. In order to conclude that the queueing network is e-stable, by using Theorem 4.17, one needs to replace (5.1) here with the assumption that, for each $\kappa > 0$,

$$A = \{x : |x| \leq \kappa\} \text{ is uniformly small on } [s_1, s_2], \text{ for some } 0 < s_1 < s_2. \tag{5.2}$$

Whether the above sets A are either petite or uniformly small may depend on the discipline of the queueing network. We recall that the conditions (4.30) and (4.31) on the interarrival times are sufficient for both (5.1) and (5.2) under all HL disciplines, and that (4.30) alone is sufficient for (5.1) when $|\mathcal{A}| = 1$. In this chapter, as was the case in Chapter 4, it will be implicitly assumed that the interarrival and service times have finite means.

We recall that in Section 4.3, we defined regular points of a fluid model solution $\mathfrak{X}(\cdot)$ to be those times t at which the derivatives of all components of $\mathfrak{X}(\cdot)$ exist. Since the components of $\mathfrak{X}(\cdot)$ are all Lipschitz continuous, the derivatives at regular points determine $\mathfrak{X}(\cdot)$. In our computations in Sections 5.1-5.3, we will often restrict our attention to regular points, without always explicitly saying so.

The material in this chapter relies on a number of papers. The material in the first three sections is mostly from [Br98a], [DaWe96], and [Br96a]. The material in Section 5.4 is mostly from [DaV00], and that in Section 5.5 is mostly from [Da96], [Br99], [DaHV04], and [GaH05].

5.1 Single Class Networks

In Section 2.5, we gave the explicit formula (2.44) for the stationary distributions of subcritical single class HL queueing networks with exponentially distributed interarrival and service times, which we referred to as Jackson networks. These networks are stable (in fact, e-stable). Does stability still hold when the interarrival and service times are not exponential? This was considered a difficult question, and was answered in the affirmative in a number of papers under various assumptions on the interarrival and service time distributions ([Bo86], [Si90], [BaF94], [ChTK94], [MeD94]). Fluid models, together with Theorem 4.16, provide a quick proof of stability for these networks.

To show stability of the associated fluid models, we employ the proof given in [Br98a]; another argument was given in [Da95]. The proof relies on the inequality

$$D_k'(t) \geq \lambda_k + \epsilon \tag{5.3}$$

for some $\epsilon > 0$ and all k, at all regular points of the fluid model solutions $\mathfrak{X}(\cdot)$ where $Z_k(t) > 0$. It follows from (4.53) and (4.55) that (5.3) is satisfied for all subcritical single class fluid models.

Since the restriction to one class per station is otherwise not used in the proof of stability of the fluid model, we will instead assume that the HL queueing networks we consider here satisfy the related condition

$$R_k(t) \geq m_k(\lambda_k + \epsilon), \tag{5.4}$$

for some $\epsilon > 0$ and all k, whenever the k^{th} class is not empty. Recall that, as in the first part of Section 4.1, $R_k(t)$ is the proportion of service at station j allocated to the first job in class k. For subcritical single class networks, (5.4) holds for small enough ϵ, since $m_k\lambda_k < 1$. We will show shortly that (5.3), for fluid limits, follows from (5.4). Needless to say, (5.4) is a severe restriction, and is not satisfied by the usual multiclass disciplines. (In [Br98a], the condition is used to show that any HL queueing network can be "stabilized" by inserting sufficiently quick single class stations between visits to classes.)

The following theorem is the main result in this section.

Theorem 5.1. *Any HL queueing network satisfying (5.1) and (5.4) is stable. Consequently, any subcritical single class HL queueing network satisfying (5.1) is stable.*

We will employ Theorem 4.16 to show Theorem 5.1. As the associated fluid model in Theorem 4.16, we choose the fluid model consisting of the basic fluid model equations (4.50)-(4.55), together with (5.3). We already know, from Proposition 4.12, that (on the set G) every fluid limit of the queueing network satisfies the basic fluid model equations. So, in order to show the fluid model is associated with the queueing network in Theorem 5.1, we only need to verify that (5.3) is satisfied by all fluid limits.

The reasoning for (5.3) is analogous to that in Proposition 4.12 for the other fluid model equations. By (5.4), for given x,

$$T_k^x(t_2) - T_k^x(t_1) \geq m_k(\lambda_k + \epsilon)(t_2 - t_1)$$

for all k and $t_1 \leq t_2$, if $Z_k(t) > 0$ on $[t_1, t_2]$. For the same reason,

$$\frac{1}{a_n} T_k^{x_n}(a_n t_2) - \frac{1}{a_n} T_k^{x_n}(a_n t_1) \geq m_k(\lambda_k + \epsilon)(t_2 - t_1) \tag{5.5}$$

for any sequence of pairs (a_n, x_n) satisfying (4.71), if $Z_k^{x_n}(t) > 0$ on $[a_n t_1, a_n t_2]$. On the other hand, for any fluid limit $\bar{\mathfrak{X}}(\cdot)$, its component $\bar{Z}(\cdot)$ is continuous. So, if it is assumed that $\bar{Z}_k(t) > 0$ on an interval $[t_1, t_2]$, then it is bounded away from 0 on this interval. Convergence to $\bar{Z}(\cdot)$ is u.o.c., and so for sufficiently large n, $Z_k^{x_n}(t) > 0$ on $[a_n t_1, a_n t_2]$. Together with (5.5), this implies

$$\bar{T}_k(t_2) - \bar{T}_k(t_1) \geq m_k(\lambda_k + \epsilon)(t_2 - t_1).$$

By (4.55), the inequality is equivalent to

$$\bar{D}_k(t_2) - \bar{D}_k(t_1) \geq (\lambda_k + \epsilon)(t_2 - t_1),$$

which is, of course, equivalent to (5.3).

Since the system of fluid model equations (4.50)-(4.55) together with (5.3) is associated with the queueing network in Theorem 5.1, it suffices to demonstrate the following result in order to show Theorem 5.1.

Theorem 5.2. *Any fluid model satisfying (5.3) is stable.*

We break the proof of Theorem 5.2 into two lemmas. The first lemma gives a lower bound on the rate of departures from any class, empty or not.

Lemma 5.3. *Assume that a fluid model satisfies (5.3). Then,*

$$D(t_2) - D(t_1) \geq \lambda(t_2 - t_1) \tag{5.6}$$

for $0 \leq t_1 \leq t_2$.

In order to show Lemma 5.3, we require some notation. At each time $t \geq 0$, let $\mathcal{K}_0(t)$ denote those classes where $Z_k(t) = 0$, let $\mathcal{K}_+(t)$ denote its complement, and let $|\mathcal{K}_0(t)|$ and $|\mathcal{K}_+(t)|$ be the number of elements in each set. The subscripts 0 and + will denote the restrictions to $\mathcal{K}_0(t)$ and $\mathcal{K}_+(t)$, respectively, for vectors such as α, λ, $D'(t)$, and $Z'(t)$. Similarly, P_0 and P_+ will denote the $|\mathcal{K}_0(t)| \times |\mathcal{K}_0(t)|$ and $|\mathcal{K}_+(t)| \times |\mathcal{K}_0(t)|$ matrices obtained by these restrictions.

Proof of Lemma 5.3. By (5.3),

$$D'(t)_+ \geq \lambda_+. \tag{5.7}$$

In order to show (5.6), it therefore suffices to show

$$D'(t)_0 \geq \lambda_0. \tag{5.8}$$

First note that $I_0 - P_0^T$ is invertible, with

$$Q_0 \stackrel{\text{def}}{=} (I_0 - P_0^T)^{-1} = I_0 + P_0^T + (P_0^T)^2 + \dots$$

having nonnegative entries. This is the analog of (1.2). Also, the analog of (1.6) holds,

$$\lambda_0 = \alpha_0 + P_0^T \lambda_0 + P_+^T \lambda_+,$$

which implies

$$\lambda_0 = Q_0(\alpha_0 + P_+^T \lambda_+). \tag{5.9}$$

On the other hand, combining (4.50) and (4.51), and taking derivatives, one gets

$$Z'(t) = \alpha + (P^T - I)D'(t).$$

Restriction of the coordinates to $\mathcal{K}_0(t)$ implies

$$Z'(t)_0 = \alpha_0 + (P_0^T - I_0)D'(t)_0 + P_+^T D'(t)_+.$$

Multiplying by Q_0 and applying (5.9), one therefore gets

$$Q_0 Z'(t)_0 = \lambda_0 - D'(t)_0 + Q_0 P_+^T (D'(t)_+ - \lambda_+).$$

Since $Z(t)_0 = 0$, if t is a regular point, one must have $Z'(t)_0 = 0$. Hence, $Q_0 Z'(t)_0 = 0$ there. Applying this to the left side of the above equality, and (5.3) to the last term on the right side of the equality, implies (5.8), as desired. ∎

To show Theorem 5.2, we will employ the Lyapunov function

$$f(t) = e^T Q Z(t). \tag{5.10}$$

The function f counts the "average" number of present and future visits within the network, under the transition matrix P, for jobs already in the network at time t. By combining (4.50) and (4.51), one has

$$Z(t) = Z(0) + \alpha t - (I - P^T)D(t),$$

which, after multiplying both sides by Q, gives

$$QZ(t) = QZ(0) + \lambda t - D(t).$$

Substitution into (5.10) shows that

$$f(t) = f(0) + te^T\lambda - e^T D(t). \tag{5.11}$$

By applying Lemma 5.3 to $f(t)$, we will obtain the following result, Lemma 5.4. Since $f(t)$ and $|Z(t)|$ are at most bounded multiples of one another, Theorem 5.2 follows immediately from Lemma 5.4.

Lemma 5.4. *Assume that (5.3) is satisfied. Then,*

$$f(t) \le [f(0) - \epsilon t]^+ \quad \textit{for all } t \ge 0. \tag{5.12}$$

Proof. Lemma 5.3 implies that

$$e^T(D(t_2) - D(t_1)) \ge (t_2 - t_1)e^T\lambda$$

for $0 \le t_1 \le t_2$. So, by (5.11), $f(t)$ is nonincreasing. Consequently, if $Z(t_0) = 0$ for a given t_0, then $f(t) = 0$ for all $t \ge t_0$.

Consider now $Z(t)$ for

$$t < t_0 \stackrel{\text{def}}{=} \inf\{t : Z(t) = 0\}.$$

By (5.3), $D_k'(t) \ge \lambda_k + \epsilon$ for some k (which depends on t). Together with Lemma 5.3, this implies that

$$e^T D(t) \ge t(e^T\lambda + \epsilon) \quad \text{for } t < t_0.$$

Plugging this into (5.11) implies (5.12) for $t < t_0$ as well. (One can instead, if one wishes, give a somewhat different proof by using (4.60).) ∎

5.2 FBFS and LBFS Reentrant Lines

The static buffer priority (SBP), first-buffer-first-served (FBFS), and last-buffer-first-served (LBFS) disciplines were introduced in Chapter 1. We review their definitions here. For SBP disciplines, classes at each station are assigned a strict ranking, with jobs of higher ranked (or priority) classes always being served before jobs of lower ranked classes, irrespective of when they arrived at the station. In this section, we assume the disciplines are preemptive, so that arriving higher ranked jobs interrupt lower ranked jobs currently in service; when the service of such higher ranked jobs has been completed, service of the lower ranked jobs continues where it left off. The first job of each class always receives all of the service allocated to the class, so the discipline is HL.

We focus here on two SBP disciplines for networks that are reentrant lines, FBFS and LBFS. For the FBFS discipline, jobs at earlier classes along the route have priority over later classes. For the LBFS discipline, the priority is reversed, with jobs at later classes along the route having priority over earlier classes. In both cases, we assign classes the values $k = 1, \ldots, K$ according to the order of their appearance along the route. With this ordering, classes with smaller k have priority under FBFS and classes with larger k have priority under LBFS.

In order to specify the evolution of an SBP network, one requires another equation in addition to the basic queueing network equations (4.42)-(4.47). This is given by

$$t - T_k^+(t) \text{ can only increase when } Z_k^+(t) = 0 \quad \text{for } k = 1, \ldots, K, \tag{5.13}$$

for all $t \geq 0$. Here, $Z_k^+(t)$ denotes the sum of the queue lengths at the station $j = s(k)$ of classes having priority at least as great as k, and $T_k^+(t)$ denotes the corresponding sum of cumulative service times. It is easy to verify (5.13). (Recall that the discipline is assumed to be preemptive.) Note that in this setting, (4.46) is redundant, since it is equivalent to (5.13) when k is the lowest ranked class at its station.

Arguing as in the last part of the proof of Proposition 4.12, it is not difficult to show that (5.13) is satisfied for all fluid limits of SBP queueing networks. We already know from Proposition 4.12 that the basic fluid model equations (4.50)-(4.55) are satisfied by all fluid limits of the queueing network. So, the fluid model given by the equations (4.50)-(4.55) and (5.13) is associated with the SBP queueing network with corresponding parameters. We refer to these equations as the *SBP fluid model equations*, and to the corresponding fluid model as the *SBP fluid model*. When the priority scheme among the different classes at each station corresponds to the FBFS or LBFS disciplines, we refer to these equations as the *FBFS fluid model equations* or the *LBFS fluid model equations*, respectively. The respective fluid models are defined analogously. We note that (5.13) is equivalent to

$$(T_k^+)'(t) = 1 \text{ when } Z_k^+(t) > 0 \quad \text{for } k = 1, \ldots, K, \tag{5.14}$$

at all regular points t.

We saw in Chapter 3 that there exist subcritical SBP reentrant lines that are not stable. This is not the case for FBFS and LBFS reentrant lines. Our goal in this section is to show the following two results. Both are done in [DaWe96], whose reasoning we follow. [LuK91] showed analogous results for discrete deterministic systems.

Theorem 5.5. *Any subcritical FBFS reentrant line satisfying (5.1) is stable.*

Theorem 5.6. *Any subcritical LBFS reentrant line satisfying (5.1) is stable.*

As in Section 5.1, we will use Theorem 4.16 to demonstrate stability for the disciplines of interest. Since we have already shown that the FBFS and LBFS fluid models are associated with the reentrant lines in Theorems 5.5 and 5.6, it suffices to demonstrate the following two results.

Theorem 5.7. *Any subcritical FBFS fluid model is stable.*

Theorem 5.8. *Any subcritical LBFS fluid model is stable.*

The remainder of the section will be devoted to demonstrating Theorems 5.7 and 5.8. We observe that, in both cases, it suffices to show stability, as in (4.61), but for fluid model solutions that also satisfy $|Z(0)| = 1$. This additional assumption will simplify the bookkeeping somewhat. As is typically the case for fluid models, $\tilde{\mathfrak{X}}(t) \stackrel{\text{def}}{=} \mathfrak{X}(ct)/c,\ c > 0$, also satisfies the same fluid model equations as $\mathfrak{X}(t)$. It follows from this that the conditions (4.61), with and without $|Z(0)| = 1$, are equivalent.

For the proofs of both Theorem 5.7 and 5.8, we find it convenient to set $d_k(t) = D_k'(t)$ for the departure rate from a class k in the fluid model. As mentioned earlier, we only need to consider the behavior of solutions $\mathfrak{X}(\cdot)$ at regular points t.

Proof of Theorem 5.7. We will use induction to prove that, for each $k = 1, \ldots, K$, there exists a $t_k \geq 0$ so that, for any fluid model solution with $|Z(0)| = 1$,

$$Z_\ell(t) = 0 \text{ on } [t_k, \infty), \quad \text{for } \ell = 1, \ldots, k.$$

For the induction step, we assume that $Z_\ell(t) = 0$ on $[t_{k-1}, \infty)$ for $\ell = 1, \ldots, k-1$. It follows that for $t \geq t_{k-1}$,

$$d_{k-1}(t) = d_{k-2}(t) = \ldots = d_1(t) = \alpha_1. \tag{5.15}$$

Set $\mathcal{H}_k = \{\ell \leq k : s(\ell) = s(k)\}$. If $Z_k(t) > 0$, then by (4.55) and (5.14),

$$\sum_{\ell \in \mathcal{H}_k} m_\ell d_\ell(t) = (T_k^+)'(t) = 1.$$

So, for $Z_k(t) > 0$ and $t \geq t_{k-1}$,

$$d_k(t) = \mu_k \left(1 - \alpha_1 \sum_{\ell \in \mathcal{H}_k \setminus \{k\}} m_\ell\right). \tag{5.16}$$

It follows from (5.15) and (5.16) that

$$Z_k'(t) = d_{k-1}(t) - d_k(t) = \alpha_1 - \mu_k \left(1 - \alpha_1 \sum_{\ell \in \mathcal{H}_k \setminus \{k\}} m_\ell\right).$$

Since $\alpha_1 \sum_{\ell \in \mathcal{H}_k} m_\ell < 1$ by assumption, it is not difficult to see that the right side of this equation is strictly negative. Setting

$$t_k' = t_{k-1} + \frac{Z_k(t_{k-1})}{\mu_k \left(1 - \alpha_1 \sum_{\ell \in \mathcal{H}_k \setminus \{k\}} m_k\right) - \alpha_1},$$

it follows that $Z_k(t) = 0$ for $t \geq t_k'$.

Since $|Z(0)| = 1$ is assumed, one has $Z_{k-1}(t) \leq |Z(t)| \leq 1 + \alpha_1 t$ for all t. So, we can choose $t_k \geq t_k'$ independently of the particular fluid model solution $\mathfrak{X}(\cdot)$, so that $Z_k(t) = 0$ for $t \geq t_k$. Since, by induction, this holds for all k, it follows that for any fluid model solution with $|Z(0)| = 1$, one has $Z(t) = 0$ for $t \geq N$ and some fixed N. So, the fluid model is stable. ■

We now demonstrate Theorem 5.8.

Proof of Theorem 5.8. We will show that for some N, one must have $Z_k(t) = 0$ for $t \geq N$ and all k, for any fluid model solution with $|Z(0)| = 1$. To show this, assume that at a given k and t, $Z_k(t) > 0$, with $Z_\ell(t) = 0$ for $\ell = k+1, \ldots, K$. It follows that if t is a regular point,

$$d_k(t) = d_{k+1}(t) = \ldots = d_K(t). \tag{5.17}$$

Set $\mathcal{H}_k = \{\ell \geq k : s(\ell) = s(k)\}$. Since $Z_k(t) > 0$, it follows from (5.17), (4.55), and (5.14), that

$$d_K(t) \sum_{\ell \in \mathcal{H}_k} m_\ell = \sum_{\ell \in \mathcal{H}_k} m_\ell d_\ell(t) = (T_k^+)'(t) = 1.$$

So, $d_K(t) = 1/\sum_{\ell \in \mathcal{H}_k} m_\ell$. Because the system is subcritical,

$$\alpha_1 \sum_{\ell \in \mathcal{H}_k} m_\ell \leq \alpha_1 \max_j \sum_{\ell \in \mathcal{C}(j)} m_\ell = \max_j \rho_j \stackrel{\text{def}}{=} \rho_{\max} < 1.$$

In particular, $d_K(t) \geq \alpha_1/\rho_{\max} > \alpha_1$.

Set $f(t) = |Z(t)|$. At regular points of $f(t)$,

$$f'(t) = \alpha_1 - d_K(t).$$

From the previous paragraph, we know this is at most $\alpha_1(1 - 1/\rho_{\max}) < 0$ when $Z(t) \neq 0$. For $|Z(0)| = 1$, it follows that $Z(t) = 0$ for $t \geq N$, where

$$N = 1/\alpha_1(1/\rho_{\max} - 1).$$

So, the fluid model is stable. ■

5.3 FIFO Networks of Kelly Type

We saw, in Section 2.5, that the explicit formula (2.44) for the stationary distribution of subcritical single class HL networks generalizes to the related formula for subcritical FIFO networks of Kelly type in (2.4) and (2.9). In both cases, the interarrival and service times of the network were assumed to be exponentially distributed. We showed, in Section 5.1, that subcritical single class queueing networks are stable under more general distributions by using the machinery of fluid limits. Here, we do the same for FIFO networks of Kelly type. At the end of the section, we briefly discuss similar behavior of HLPPS queueing networks, which are an HL variant of PS networks.

In order to specify the evolution of such queueing networks, one requires an equation corresponding to the FIFO discipline, in addition to the basic queueing network equations (4.42)-(4.47). The equation we employ is

$$D_k(t + W_j(t)) = Z_k(0) + A_k(t), \quad k = 1, \ldots, K, \tag{5.18}$$

for all $t \geq 0$. To verify (5.18), note that $t + W_j(t)$ is the time at which the service of the last of the jobs currently at j will be completed, since jobs arriving after time t will have lower priority under the FIFO discipline.

Together, (4.42)-(4.47) and (5.18) form the *FIFO queueing network equations*; the corresponding 6-tuple $\mathfrak{X}(\cdot)$ is the *FIFO queueing network process*. One can check that the triple $(E(\cdot), \Gamma(\cdot), \Phi(\cdot))$, together with

$$\{D_k(t) \text{ for } t \leq W_j(0),\ k = 1, \ldots, K\}, \tag{5.19}$$

determines $\mathfrak{X}(\cdot)$, for all $t \geq 0$, if $\mathfrak{X}(\cdot)$ evolves according to the FIFO queueing network equations. The information in (5.19) serves the role of the *initial data* for solutions of these equations. This additional information is needed, since $Z(0)$ is by itself not enough to determine the order in which the original jobs are served.

Arguing as in the proof of Proposition 4.12, it is not difficult to show that (5.18) is satisfied for all fluid limits of FIFO queueing networks. (Recall that the component $\bar{D}(\cdot)$ of a fluid limit is continuous and nondecreasing.) We already know from Proposition 4.12 that the basic fluid model equations (4.50)-(4.55) are satisfied by all fluid limits of the queueing network. So, the fluid model given by the equations (4.50)-(4.55) and (5.18) is associated with the FIFO queueing network.

We refer to these equations as the *FIFO fluid model equations*, and to the corresponding fluid model as the *FIFO fluid model*. Solutions of the fluid model equations are denoted by the 6-tuple $\mathfrak{X}(\cdot)$. The *initial data* is again given by (5.19). Using the basic fluid model equations, one can show that

$$\sum_{k \in \mathcal{C}(j)} m_k D_k(t) = t \quad \text{for } t \leq W_j(0) \tag{5.20}$$

must hold; (5.20) serves as a consistency condition on the initial data. In keeping with the definition for queueing networks, we say that a FIFO fluid model is of *Kelly type*, if $m_k = m_\ell$ whenever $k, \ell \in \mathcal{C}(j)$ for some j. We will then write m_j^s for m_k.

We saw, in Chapter 3, that there exist subcritical FIFO queueing networks that are not stable. On the other hand, subcritical FIFO queueing networks of Kelly type that have exponentially distributed interarrival and service times are stable, as was shown in Chapter 2. The following result from [Br96a] shows that stability continues to hold for more general interarrival and service times. We follow the presentation that is given there.

Theorem 5.9. *Any subcritical FIFO queueing network of Kelly type satisfying (5.1) is stable.*

As in the previous two sections, we will use Theorem 4.16 to demonstrate this stability. Since we already know that FIFO fluid models are associated with the FIFO queueing networks, it suffices to demonstrate the following result.

Theorem 5.10. *Any subcritical FIFO fluid model of Kelly type is stable.*

Demonstration of Theorem 5.10

In order to demonstrate Theorem 5.10, we introduce a form of entropy. Let

$$h(x) = x \log x, \quad x \geq 0 \tag{5.21}$$

and

$$h_k(x) = \lambda_k h(x/\lambda_k) = x \log(x/\lambda_k), \quad x \geq 0, \tag{5.22}$$

for $k = 1, \ldots, K$. Note that $h(x)$ and $h_k(x)$ are convex, with

$$h(0) = h(1) = 0, \quad h_k(0) = h_k(\lambda_k) = 0, \quad h'(1) = h_k'(\lambda_k) = 1. \tag{5.23}$$

The basic tool for analyzing the asymptotic behavior of $Z(t)$ will be the *entropy function* $\mathcal{H}(t)$,

$$\mathcal{H}(t) = \sum_k \int_t^{t+W_j(t)} h_k(D_k'(r))\, dr, \quad t \geq 0. \tag{5.24}$$

Since $D(\cdot)$ and $W(\cdot)$ are Lipschitz continuous, it is not difficult to check that $\mathcal{H}(\cdot)$ is also Lipschitz continuous. We will analyze $\mathcal{H}'(t)$ at regular points t.

One can think of $\mathcal{H}(t)$ as measuring the "distance" at time t from the "equilibrium" $D_k'(r) \equiv \lambda_k$, averaged over the above values of r. One can check that, for $\rho < e$, such equilibria only occur when $Z(r) = 0$. (For $\rho = e$, there are other solutions.) We will not need this fact here, although it motivates our approach.

Since $h_k(\cdot)$ is convex and $\rho < e$, the following lemma is not difficult to show with the help of Jensen's Inequality. We will use here the expression

$$m_j^s \sum_{k\in\mathcal{C}(j)} D_k'(t) = 1 \quad \text{on } \{t : W_j(t) \neq 0\}, \tag{5.25}$$

which follows from (4.53)-(4.55).

Lemma 5.11. *Suppose $\mathfrak{X}(\cdot)$ is any FIFO fluid model solution of Kelly type, with $\rho_j \leq 1$ for all j. Then, $\mathcal{H}(t) \geq 0$ for all t.*

Proof. Rewriting (5.24) gives

$$\mathcal{H}(t) = \sum_{j\in F_t} \lambda_j^\Sigma \int_t^{t+W_j(t)} \sum_{k\in\mathcal{C}(j)} (\lambda_k/\lambda_j^\Sigma) h(D_k'(r)/\lambda_k)\, dr, \tag{5.26}$$

where

$$F_t = \{j : W_j(t) \neq 0\} \quad \text{and} \quad \lambda_j^\Sigma = \sum_{k\in\mathcal{C}(j)} \lambda_k.$$

By Jensen's Inequality, (5.26) is at least

$$\sum_{j\in F_t} \lambda_j^\Sigma \int_t^{t+W_j(t)} h\left(\sum_{k\in\mathcal{C}(j)} D_k'(r)/\lambda_j^\Sigma\right) dr.$$

On account of (5.25) and $\rho_j \leq 1$, which holds for all j, the integrand

$$h\left(\sum_{k\in\mathcal{C}(j)} D_k'(r)/\lambda_j^\Sigma\right) = h(1/\rho_j) \geq 0 \tag{5.27}$$

at regular points $r \in [t, t + W_j(t)]$. It follows from (5.26)-(5.27) that $\mathcal{H}(t) \geq 0$ for all t. ∎

If we knew that $\mathcal{H}'(t) \leq -c_1,\ c_1 > 0$, whenever $\mathcal{H}(t) > 0$, it would follow immediately that $\mathcal{H}(t) = 0$ for $t \geq \mathcal{H}(0)/c_1$. We will instead show the weaker inequality that, for appropriate $c_2, c_3 > 0$,

$$\mathcal{H}(t) - \mathcal{H}(t + c_2 W^M(t)) \geq c_3 W^M(t) \quad \text{for all } t, \tag{5.28}$$

where $W^M(t) = \max_j W_j(t)$. From this, we will show that

$$\mathcal{H}(t) = 0 \quad \text{for } t \geq c_4 \mathcal{H}(0), \tag{5.29}$$

and some c_4 not depending on $\mathfrak{X}(\cdot)$. It will then follow quickly that

$$Z(t) = 0 \quad \text{for } t \geq c_5 |Z(0)| \tag{5.30}$$

and appropriate c_5, which implies Theorem 5.10.

Proposition 5.12 is the most important step in deriving (5.28). The representation for $\mathcal{H}'(t)$ given on the right side of (5.31) will enable us to compute upper bounds on $\mathcal{H}'(t)$ without too much difficulty. As usual, statements on $\mathcal{H}'(t)$ refer to regular points t.

Proposition 5.12. *For any FIFO fluid model of Kelly type,*

$$\mathcal{H}'(t) = \sum_k [h_k(A'_k(t)) - h_k(D'_k(t))] - \sum_j \frac{1}{m_j^s} h(1 + W'_j(t)). \tag{5.31}$$

Proof. Differentiation of (5.24) gives

$$\mathcal{H}'(t) = \sum_k [(1 + W'_j(t)) h_k(D'_k(t + W_j(t))) - h_k(D'_k(t))]. \tag{5.32}$$

By (5.18), this

$$= \sum_k [(1 + W'_j(t)) h_k(A'_k(t)/(1 + W'_j(t))) - h_k(D'_k(t))].$$

Using the definition of $h_k(\cdot)$, one can check this

$$= \sum_k [h_k(A'_k(t)) - h_k(D'_k(t))] - \sum_j \left(\sum_{k \in \mathcal{C}(j)} A'_k(t) \right) \log(1 + W'_j(t)). \tag{5.33}$$

If $Y'_j(t) \neq 0$ and $W'_j(t)$ exists, it follows from (4.54) that $W'_j(t) = 0$. For each j, the summand on the right side of (5.33) is therefore equal to

$$\left(\sum_{k \in \mathcal{C}(j)} A'_k(t) + \frac{1}{m_j^s} Y'_j(t) \right) \log(1 + W'_j(t)). \tag{5.34}$$

On the other hand, combining (4.52), (4.53), and (4.55), one gets

$$t + W_j(t) = W_j(0) + m_j^s \sum_{k \in \mathcal{C}(j)} A_k(t) + Y_j(t).$$

Plugging the derivative of the above quantity into the first factor in (5.34), shows that (5.34) is equal to $h(1 + W'_j(t))/m_j^s$. Consequently, (5.33) equals

$$\sum_k [h_k(A'_k(t)) - h_k(D'_k(t))] - \sum_j \frac{1}{m_j^s} h(1 + W'_j(t)).$$

Together with (5.32), this implies (5.31). ∎

The following proposition provides bounds for the two terms on the right side of (5.31), and hence for $\mathcal{H}'(t)$.

Proposition 5.13. *For any FIFO fluid model of Kelly type, both*

$$\sum_k [h_k(A'_k(t)) - h_k(D'_k(t))] \leq \sum_k Z'_k(t) \tag{5.35}$$

and, for some $c_6 > 0$,

$$\sum_j \frac{1}{m_j^s} h(1 + W_j'(t)) \geq \sum_k Z_k'(t) + c_6 \sum_j (W_j'(t))^2. \tag{5.36}$$

Consequently,

$$\mathcal{H}'(t) \leq -c_6 \sum_j (W_j'(t))^2. \tag{5.37}$$

The proof of (5.36) is just a few lines. The proof of (5.35) is a bit longer. It relies on the convexity of $h_k(\cdot)$ and on Jensen's Inequality, as well as on the general equality

$$Z(t) = Z(0) + (I - P^T)(\lambda t - D(t)), \tag{5.38}$$

which follows from (4.51), (4.50), and (1.6). The inequality (5.35) reflects the randomness present in the mean transition matrix P. For instance, for reentrant lines that are closed (i.e., there are no arrivals to or departures from the system), this randomness is absent and both sides of (5.35) are equal to 0. (The sum on the left side telescopes, with all terms cancelling.)

Proof of Proposition 5.13. The inequality (5.37) immediately follows from Proposition 5.12 and (5.35)-(5.36). For (5.36), note that since $h(1) = 0$, $h'(1) = 1$, $h''(x) = 1/x$, and $W_j'(t)$ is bounded,

$$h(1 + W_j'(t)) \geq W_j'(t) + c_7 (W_j'(t))^2$$

for some $c_7 > 0$. Also, by (4.57),

$$W_j'(t) = m_j^s \sum_{k \in \mathcal{C}(j)} Z_k'(t).$$

Combining these two expressions and summing over j shows (5.36).

To show (5.35), note that by (4.50),

$$h_k(A_k'(t)) = h_k \left(\alpha_k + \sum_\ell P_{\ell,k} D_\ell'(t) \right) \tag{5.39}$$

for each t, which can be rewritten as

$$\lambda_k h \left(\lambda_k^{-1} \left[\alpha_k + \sum_\ell (\lambda_\ell P_{\ell,k})(\lambda_\ell^{-1} D_\ell'(t)) \right] \right).$$

By (1.6), $\lambda_k^{-1}(\alpha_k + \sum_\ell \lambda_\ell P_{\ell,k}) = 1$. So, by Jensen's Inequality and $h(1) = 0$, this is

$$\leq \alpha_k h(1) + \sum_\ell \lambda_\ell P_{\ell,k} h(\lambda_\ell^{-1} D_\ell'(t)) = \sum_\ell P_{\ell,k} h_\ell(D_\ell'(t)). \tag{5.40}$$

It follows from (5.39)-(5.40) that, for each k,

$$h_k(A'_k(t)) \le \sum_\ell P_{\ell,k} h_\ell(D'_\ell(t)).$$

Summation over k shows that

$$\sum_k [h_k(A'_k(t)) - h_k(D'_k(t))] \le -\sum_k \left(1 - \sum_\ell P_{k,\ell}\right) h_k(D'_k(t)). \tag{5.41}$$

Since $h_k(\cdot)$ is convex with $h_k(\lambda_k) = 0$ and $h'_k(\lambda_k) = 1$, the right side of (5.41) is

$$\le -\sum_k \left(1 - \sum_\ell P_{k,\ell}\right)(D'_k(t) - \lambda_k) = \sum_k Z'_k(t), \tag{5.42}$$

with the equality following from (5.38). The inequality (5.35) is an immediate consequence of (5.41) and (5.42). ■

Let $\tau_j(t)$ denote the additional time, starting at t, until a station j is next empty. We will need the following general bound for showing (5.28).

Lemma 5.14. *For each subcritical fluid model,*

$$\tau_j(t) \le c_2 W^M(t) \quad \textit{for all } t \tag{5.43}$$

and some constant c_2.

Proof. On account of (4.57), (5.43) is equivalent to

$$\tau_j(t) \le c_8 |Z(t)|, \tag{5.44}$$

for appropriate c_8. By the general equality (4.63),

$$CMQ(Z(t') - Z(t)) = (\rho - e)(t' - t) + Y(t') - Y(t)$$

for $t' \ge t$. Therefore, for each j,

$$m_j^s \sum_{k \in \mathcal{C}(j)} Z_k(t') \le (CMQZ(t))_j + (\rho_j - 1)(t' - t) + Y_j(t') - Y_j(t).$$

This implies (5.44), with $c_8 = (CMQZ(t))_j/(1 - \rho_j)$. ■

The bound in (5.28) follows from (5.37), Jensen's Inequality, and (5.43).

Proof of (5.28). By (5.37) and Jensen's Inequality,

$$\mathcal{H}(t) - \mathcal{H}(t') \ge c_6 \int_t^{t'} (W'_j(r))^2 dr \ge \frac{c_6}{t' - t}(W_j(t') - W_j(t))^2 \tag{5.45}$$

for $0 \leq t < t'$ and any j. Suppose that $W_j(t) \neq 0$ for given t and j. Setting $t' = t + \tau_j(t)$, it follows from (5.43) and (5.45) that

$$\mathcal{H}(t) - \mathcal{H}(t + \tau_j(t)) \geq c_3 (W_j(t))^2 / W^M(t),$$

where $c_3 = c_6/c_2$. Choosing j so that $W_j(t) = W^M(t)$, it follows from the monotonicity of $\mathcal{H}(\cdot)$ that

$$\mathcal{H}(t) - \mathcal{H}(t + c_2 W^M(t)) \geq c_3 W^M(t),$$

which is (5.28). ■

We now finish the demonstration of Theorem 5.10.

Proof of Theorem 5.10. We will iterate along the times $t_{i+1} = t_i + c_2 W^M(t_i)$, $i = 0, 1, 2, \ldots$, where $t_0 = 0$. This gives

$$\mathcal{H}(0) - \mathcal{H}(t_i) \geq c_3 t_i / c_2 \quad \text{for all } i,$$

by (5.28). Since $\mathcal{H}(t_i) \geq 0$, it follows that

$$t_\infty \stackrel{\text{def}}{=} \lim_{i \to \infty} t_i \leq c_2 \mathcal{H}(0)/c_3.$$

Consequently, by the continuity of $W(t)$, $W(t_\infty) = 0$. Moreover, because of (5.24), $\mathcal{H}(t_\infty) = 0$, which implies that $\mathcal{H}(t) = 0$ for $t \geq t_\infty$. This is equivalent to (5.29). It follows from this and (5.28) that

$$W(t) = 0 \quad \text{for } t \geq t_\infty.$$

Since $D(t)$ is Lipschitz continuous, it is not difficult to see that

$$\mathcal{H}(0) \leq c_9 W^M(0)$$

for appropriate c_9. Together, the last three displays imply

$$W(t) = 0 \quad \text{for } t \geq c_{10} W^M(0),$$

with $c_{10} = c_2 c_9 / c_3$. It follows from this and (4.57) that

$$Z(t) = 0 \quad \text{for } t \geq c_5 |Z(0)|,$$

for appropriate c_5, which is (5.30). Theorem 5.10 follows. ■

HLPPS queueing networks

The *head-of-the-line proportional processor sharing* (HLPPS) discipline is a variant of processor sharing, which was discussed in Chapter 2. Under this discipline, all nonempty classes present at a station are served simultaneously, with the fraction of time spent serving a class being proportional to the number of jobs of the class currently there, and all of the service going into the

first job of each class. The discipline is clearly HL. When the service times are exponentially distributed, the queueing network process $\mathfrak{X}(\cdot)$ of an HLPPS network coincides with that of the corresponding processor sharing network, where all jobs of a class receive equal service instead of the first receiving all of the service. The HLPPS discipline can be thought of as a simpler variant of processor sharing that exhibits the HL property. It can be preferable to processor sharing in certain situations when a penalty is attached to sharing service among too many jobs.

The following result is shown in [Br96b].

Theorem 5.15. *Any subcritical HLPPS queueing network satisfying (5.1) is stable.*

As in the previous examples, Theorem 4.16 can be employed to demonstrate stability of these networks. In order to use the accompanying fluid model machinery, we observe that the HLPPS property can be expressed as

$$T(t) = \int_0^t Z^P(s)\, ds \tag{5.46}$$

for all $t \geq 0$, where

$$Z_k^P(s) = \begin{cases} Z_k(s)/Z_j^\Sigma(s) & \text{for } Z_j^\Sigma(s) > 0, \\ 0 & \text{for } Z_j^\Sigma(s) = 0, \end{cases}$$

and

$$Z_j^\Sigma(s) = \sum_{k \in \mathcal{C}(j)} Z_k(s).$$

The *HLPPS queueing network equations* are then (4.42)-(4.47) together with (5.46). The *HLPPS fluid model equations* are (4.50)-(4.55) together with

$$T_k'(t) = Z_k^P(t) \text{ when } Z_j^\Sigma(t) > 0, \quad k = 1, \ldots, K. \tag{5.47}$$

Arguing as in the proof of Proposition 4.12, it is not difficult to show that (5.47) is satisfied for all fluid limits of HLPPS queueing networks. So, the fluid model given by (4.50)-(4.55) and (5.47) is associated with the HLPPS queueing network. In order to show Theorem 5.15, it therefore suffices to show its fluid model analog.

Theorem 5.16. *Any subcritical HLPPS fluid model is stable.*

The demonstration of Theorem 5.16 exhibits similarities to that of Theorem 5.10 for FIFO fluid models of Kelly type. The argument employs an entropy function

$$\mathcal{H}(t) = \sum_k Z_k(t) \log(D_k'(t)/\lambda_k), \quad t \geq 0.$$

This entropy function is equivalent to

$$\sum_k m_k Z_j^{\Sigma}(t) h_k(D_k'(t)), \quad t \geq 0,$$

which is similar to that in (5.24).

As before, the goal is to show (5.29), from which (5.30) will follow. In the present setting, one can show that

$$\mathcal{H}'(t) \leq -c_{11} < 0$$

until $\mathcal{H}(t) = 0$, in place of (5.28). The individual steps of the argument differ from those for Theorem 5.10. We note that Theorem 5.16 (and hence Theorem 5.15) holds when $m_k = m_\ell$ for $s(k) = s(\ell)$ is not assumed, unlike its FIFO analog. This might be expected, because the stationary distribution for processor sharing networks given by (2.7) and (2.9) also does not require this condition.

5.4 Global Stability

In the last chapter and in the first three sections of this chapter, we have addressed the question on when an HL queueing network is stable. Our main technique has been the employment of fluid models: when an associated fluid model of a queueing network is stable, so is the queueing network. In this section, we introduce a different notion of stability, global stability. As before, fluid models will provide the main technique for demonstrating global stability.

We say that a queueing network is *globally stable* if it remains stable when the discipline is replaced by any HL discipline. That is, under any HL discipline (e.g., an SBP or FIFO discipline), the resulting network, with its given routing, and interarrival and service distributions, is positive Harris recurrent.

Global stability is clearly a more restrictive requirement, in general, than is stability. For example, the subcritical two-station reentrant lines, with route given in Figure 3.1, may or may not be stable, depending on the discipline. The FBFS discipline is stable, by Theorem 5.5 (provided (5.1) holds). On the other hand, the Lu-Kumar network, which has the priority scheme (4,1) and (2,3) at the two stations, is unstable for the range of mean service times given in Theorem 3.2.

We recall that the basic fluid model equations (4.50)-(4.55) do not reflect the discipline of a queueing network. By Proposition 4.12, each HL queueing network is associated with its basic fluid model. The following result is therefore a direct consequence of Theorem 4.16.

Proposition 5.17. *Assume that a given HL queueing network satisfies (5.1) and that its basic fluid model is stable. Then, the queueing network is globally stable.*

As we have seen earlier in this chapter, it is easier to study the stability of fluid models rather than directly investigating the stability of the corresponding queueing networks. We will take this approach here, and study the stability

of the basic fluid model. A complete, explicit theory is given in [DaV00] when the network has two stations and routing is deterministic. Most of this section is devoted to providing a summary of these results. The remainder of the section briefly presents related results on weak stability and global weak stability.

We note that our terminology is somewhat different than that in [DaV00], and in related work [Ha97] and [DaHV99]. There, the authors employ the term "fluid networks" rather than "fluid models", and employ the term "global stability" for when the fluid networks are stable. The latter definition corresponds to stability for the basic fluid model here, although somewhat different fluid model equations are used in those papers. Another difference arises from the usage of the term "queueing network" in our lectures, which assumes that a discipline has already been assigned; in the above references, this is not assumed.

Virtual stations and push starts

An appealing theory for the stability of basic fluid models with two stations is developed in [DaV00] for networks with deterministic routing, with necessary and sufficient conditions for stability being given. Key ingredients include the concepts of *virtual stations* and *push starts*, in terms of which stability can be phrased. The theory does not extend to networks with three or more stations.

We will later give a systematic definition of virtual stations and push starts, but the concepts are better illustrated by an example. We will employ the example from [DaHV04], which is a reentrant line with route given in Figure 5.1.

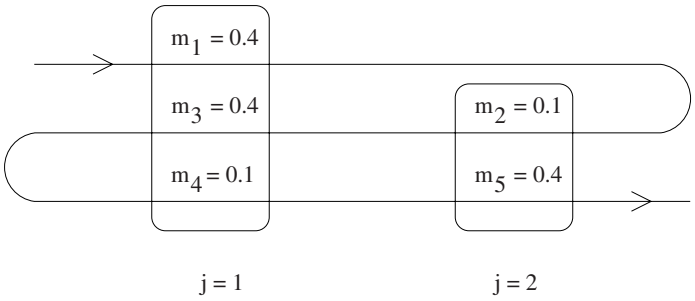

Fig. 5.1. The five classes along the route are labelled in the order of their appearance; the mean service time of each is given. The priority scheme is (1,3,4) at station 1 and (5,2) at station 2.

The queueing network has two stations, with three and two classes each, as illustrated in the figure. The mean service times at the classes are

$$m_1 = m_3 = m_5 = 0.4 \quad \text{and} \quad m_2 = m_4 = 0.1, \tag{5.48}$$

and jobs are assumed to enter the network at rate 1. The network is therefore subcritical. The actual distributions of the interarrival and service times will not be important to us. A preemptive SBP discipline is assigned, with priority scheme (1,3,4) at station 1 and (5,2) at station 2. That is, the discipline is FBFS at station 1 and LBFS at station 2. We will exhibit a virtual station for this queueing network.

We also consider the fluid model obtained by adding the SBP equation (5.13) to the basic fluid model equations, where the classes have the same priorities as above. The SBP fluid model thus obtained is associated with the above queueing network; this was shown in Section 5.2. Examination of the basic fluid model suffices presently, although we will need the SBP fluid model later when examining push starts. In order to exhibit a virtual station for either fluid model, we will need to employ the fluid limits from the above queueing network.

The interaction between classes 3 and 5 is important for understanding the evolution of this queueing network. If one assumes that $Z_3(0) = 0$ or $Z_5(0) = 0$, it then follows that

$$Z_3(t) = 0 \text{ or } Z_5(t) = 0 \quad \text{for each } t \geq 0. \tag{5.49}$$

That is, if at least one of these two classes is initially empty, then this condition persists for all time. The reason is that since class 3 has higher priority than class 4, no job from the latter class can enter class 5 as long as class 3 is not empty. Moreover, if class 5 is not empty, then no job can enter class 3, since class 5 has higher priority than class 2. This behavior relies on the assumption that the discipline is preemptive. (The corresponding behavior for classes 2 and 4 of the Lu-Kumar network was used in the proof of Theorem 3.2.)

It follows from (5.49) that, under such initial data, classes 3 and 5 can never be served simultaneously. Consequently,

$$T_3(t_2) - T_3(t_1) + T_5(t_2) - T_5(t_1) \leq t_2 - t_1 \tag{5.50}$$

whenever $t_1 \leq t_2$, or equivalently,

$$T_3'(t) + T_5'(t) \leq 1 \tag{5.51}$$

wherever the derivative is defined. One can think of classes 3 and 5 as forming a "virtual station", with their service being constrained, as in (5.51), as if they actually belonged to a single station.

Let $\bar{\mathfrak{X}}(\cdot)$ be a fluid limit obtained from a sequence (a_n, x_n) as in (4.71), with $Z_3^{x_n}(0) = a_n$ and $Z_5^{x_n}(0) = 0$. Then, (5.50) holds for each term of the sequence, and so the same is also true for $\bar{T}(\cdot)$. On the other hand, $\bar{\mathfrak{X}}(\cdot)$ is a solution of the SBP fluid model equations. So, classes 3 and 5 form a "virtual station" for the SBP fluid model as well. Note that not all SBP fluid model solutions need satisfy (5.50) or (5.51), since fluid mass can possibly be served

and pass through a class k, with $Z_k(t) = 0$ nonetheless holding over an entire time interval. This contrasts with the behavior for the associated queueing network.

For the SBP fluid model (and hence for the basic fluid model) to be stable,

$$m_3 + m_5 < 1 \tag{5.52}$$

needs to hold. The argument is similar to that for Part (c) of Proposition 4.11, and relies on (5.50). (One can employ the analogs of (4.62) and (4.63), but with C, ρ, and Y augmented to include the virtual station.) For either fluid model with the parameters given by (5.48), $m_3 + m_5 = 0.8 < 1$, which does not preclude stability. However, if m_5 is replaced by $m_5' = 0.7$, then $m_3 + m_5' = 1.1 > 1$, and so the resulting fluid model will not be stable even though the network is still subcritical. One can then show, in fact, that some solutions satisfy $\liminf_{t\to\infty} Z(t)/t \geq 1/11$. As we will see, both fluid models with the original service rates are already not stable. For this, we will need to use push starts.

The push start phenomenon relies on the presence of a virtual station, and provides a stability condition which is an amplification of that provided by the virtual station. It is caused by a higher priority class that shares a station with one of the classes of the virtual station, and always receives at least a fixed proportion of the service at that station. In contrast to virtual stations, it is a fluid model phenomenon, rather than a queueing network phenomenon.

In the setting of the example in Figure 5.1, the stability condition (5.52) can be replaced by

$$\rho_{\text{push}} = \frac{m_3}{1 - m_1} + m_5 < 1, \tag{5.53}$$

by using push starts. We first note that if one deletes class 1 from the reentrant line in Figure 5.1, then the four-class reentrant line that remains is equivalent to the Lu-Kumar network in Section 3.1. If one retains the labelling of the original five-class reentrant line, then one can check that the bound (5.52) is needed for the associated SBP fluid model to be stable, for the same reasons as for the five-class fluid model.

Returning to the SBP fluid model for the five-class network in Figure 5.1, we observe that class 1 has the highest priority at its station. Since fluid enters the reentrant line at rate 1 and $m_1 < 1$, class 1 empties in finite time and remains empty thereafter. In keeping class 1 empty, station 1 spends proportion $m_1 = 0.4$ of its effort in processing fluid at class 1. The remaining proportion $1 - m_1$ of its effort can be spent on fluid in classes 3 and 5. This occurs for all solutions of the SBP fluid model.

Since class 1 remains empty, its sole effect on the remainder of the system is to reduce the amount of effort available at station 1 for the other two classes. Removing this class and expanding the service times at classes 3 and 4 by the factor $1/(1 - m_1)$ to compensate for this reduced effort, one can show with a little work that the resulting SBP fluid model is identical to the four-station

SBP fluid model mentioned above, but with the service times at classes 3 and 4 expanded by $1/(1-m_1)$. The push start stability condition (5.53) therefore replaces the condition (5.52), as desired. With the choice of service times given in (5.48), $\rho_{\text{push}} = \frac{2}{3} + \frac{2}{5} > 1$, and so (5.53) is violated. Hence, the SBP fluid model in Figure 5.1 is not stable. This also implies that the corresponding basic fluid model is not stable.

Systematic presentation of virtual stations and push starts

We will provide an abridged version of the construction given in [DaV00], referring sometimes to the virtual station and push start example given earlier for motivation. In order to simplify matters somewhat, we will restrict consideration to networks with just a single deterministic route, i.e., to reentrant lines.

We employ the following terminology. An *excursion* is a maximal set of consecutive classes along the route that belong to a single station. A *last class* of an excursion is the last class visited there, and a *first class* denotes all of the remaining classes; if the excursion contains only one class, then it has no first class. A set S of excursions is *strictly separating* if it contains no consecutive excursions and does not contain the first excursion. For each such set S, the *virtual station* $V(S)$ consists of the classes in the excursions of S, together with the first classes of excursions for which the immediately preceding excursion is not in S. Also, let $k_1, k_2, \ldots, k_L$ denote the last classes visited for each of the L excursions, and let $F^{\leq}(\ell)$ denote all of the classes along the route visited up to and including k_ℓ; then $F^{<}(\ell) \stackrel{\text{def}}{=} F^{\leq}(\ell) - \{k_\ell\}$ is called a *push start set*. One sets $V_j(S)$, $F_j^{\leq}(\ell)$, and $F_j^{<}(\ell)$, $j = 1, 2$, equal to the corresponding classes restricted to the stations 1 and 2, respectively.

For the example in Figure 5.1, there are four excursions consisting of the sets of classes, $\{1\}$, $\{2\}$, $\{3,4\}$, and $\{5\}$. The virtual stations are $\{2\}$, $\{2,5\}$, $\{3,4\}$, and $\{3,5\}$. The only one that is not a subset of a station, and is therefore of interest, is $V(\{5\}) = \{3,5\}$. For future reference, we record that

$$V(\{5\}) = \{3,5\}, \quad F^{\leq}(2) = \{1,2\}, \quad F^{<}(2) = \{1\}. \tag{5.54}$$

We now state the main result in [DaV00], restricted to reentrant lines. Here, we use the abbreviation $m(A) = \sum_{k \in A} m_k$ for any set A of classes. The rate at which fluid enters the first class is denoted by α_1.

Theorem 5.18. *A two-station basic fluid model is stable if and only if*

$$\rho_j < 1 \quad \textit{for } j = 1, 2, \tag{5.55}$$

and for each strictly separating set S and $\ell = 1, \ldots, L$,

$$\sum_{j=1}^{2} \frac{\alpha_1 m(V_j(S) \backslash F_j^{\leq}(\ell))}{1 - \alpha_1 m(F_j^{<}(\ell))} < 1. \tag{5.56}$$

In [DaV00], it is also shown that if the left side of (5.56) is strictly greater than 1 for some strictly separating set S and some ℓ, then there exists a fluid model solution whose fluid mass $|Z(t)| \to \infty$ as $t \to \infty$. (The authors actually consider the limit for the "work in progress", i.e., the total workload, which is equivalent to this limit.)

We note that, in order to be of interest in (5.56), virtual stations need to include classes from both stations. Otherwise, (5.56) follows from the subcriticality of each station.

One can interpret the example in Figure 5.1 in terms of Theorem 5.18. Substitution of (5.54) into (5.56) reduces the latter to the inequality in (5.53). Since this is violated for the choice of m in the example, Theorem 5.18 implies that the basic fluid model is not stable. There is, moreover, a fluid model solution with $|Z(t)| \to \infty$ as $t \to \infty$. We saw earlier that such a solution is given by the SBP discipline with priorities (1,3,4) and (5,2) at the two stations. For arbitrary two-station fluid models, the calculation of the left side of (5.56) for all strictly separating sets S and all ℓ will in general be tedious.

There is, for arbitrary two-station reentrant lines, a strong connection between stability of the associated basic fluid model and stability of the corresponding family of fluid models with SBP disciplines. Namely, when the basic fluid model is not stable, there must also exist an SBP fluid model that is not stable. Another way of phrasing this is the following. Let $\mathcal{D} \subset \mathbf{R}_+^K$ denote the set of service time vectors, with coordinates m_k, on which the basic fluid model is stable for a given route and choice of α_1. (In [DaHV99] and [DaV00], $\mathcal{D}$ is referred to as the global stability region.) Similarly, let $\mathcal{D}_S \subset \mathbf{R}_+^K$ denote the set on which all fluid models with SBP disciplines are stable. Clearly, $\mathcal{D} \subseteq \mathcal{D}_S$. The above assertion is that, in fact,

$$\mathcal{D} = \mathcal{D}_S. \tag{5.57}$$

This is shown in [DaV00], while demonstrating Theorem 5.18.

We also point out that, as a consequence of Theorem 5.18, the region $\mathcal{D}$ is monotone. That is, reducing service times maintains stability. This follows immediately from the conditions (5.55) and (5.56).

The proofs of the two directions of Theorem 5.18 are of different levels of difficulty. The necessity of the conditions (5.55) and (5.56) is reasonably straightforward. If (5.56) is violated for some value of ℓ, one can make the same basic type of argument we sketched for the SBP fluid model in Figure 5.1, once an appropriate SBP discipline has been chosen. One wants a discipline that (a) assigns the highest priority to classes in $F^<(\ell)$, with classes in $F^<(\ell)$ being ordered according to the FBFS discipline, (b) assigns a high priority to the remaining classes in $V(S)$, and (c) assigns a low priority to the other classes, including the class in $F^\leq(1)\backslash F^<(1)$. (The particular priority is not important in each of (b) and (c).) Using Theorem 5.5, one can then show that $F^<(\ell)$ will be empty after large times, like the first class in Figure 5.1. By applying the same type of argument we sketched for that network, one can also show that once all of the classes in either $V_1(S)$ or $V_2(S)$ are empty, this

condition persists for all time. This argument heavily uses the structure of $V(S)$, which was defined with appropriate "gaps" between its classes so that service at $V_1(S)$ will prevent service at $V_2(S)$ and vice versa, because of the presence of low priority classes in between. More detail is given in [DaV00] and in [Ha97], which looks at a related problem. (The reference [DaV96] cited in [DaV00] never appeared due to an unrelated flaw in extending push starts to the queueing network setting. This flaw also affects the example on page 756 in [Da96].)

The demonstration of the sufficiency of (5.55) and (5.56) is more involved and requires most of the work in [DaV00]. The paper employs linear programming techniques to construct a piecewise linear Lyapunov function, from which the stability of the fluid model will follow. (Piecewise linear Lyapunov functions have also been employed in [BoZ92], [DoM94], and [DaWe96].) The duality between minimum flows and maximum cuts is used to obtain the explicit formulation in (5.55) and (5.56). We will not go into details here.

It is natural to ask whether Theorem 5.18 extends to fluid models with more than two stations. [DaHV99] shows this is not the case, in general, by analyzing the fluid models whose routing is given by Figure 5.2. Not only is there no analog of Theorem 5.18, but $\mathcal{D} \neq \mathcal{D}_S$ and $\mathcal{D}$ is not monotone. There is presently no developed theory for stability for more than two stations.

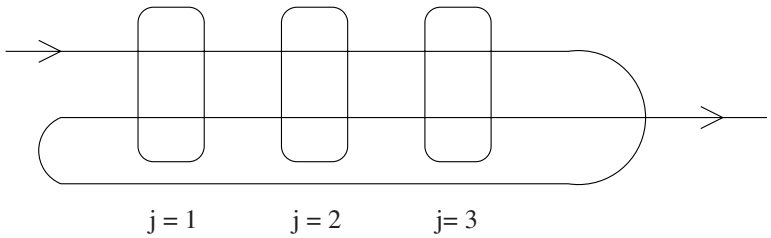

Fig. 5.2. The basic fluid model for this three-station reentrant line has irregular behavior with regard to stability, when m and the discipline are varied. This behavior is not present for two-station reentrant lines.

Rate and global stability

We conclude this section with two other types of stability. We will say that an HL queueing network is *rate stable* (or *pathwise stable*) if for any given initial state x,

$$\lim_{t\to\infty} Z^x(t)/t = 0 \quad \text{a.s.} \tag{5.58}$$

(Alternative definitions are often given. For instance, $\lim_{t\to\infty} D^x(t)/t = \lambda$ a.s. is equivalent to (5.58).) An HL queueing network will be *globally rate stable* if (5.58) continues to hold irrespective of the discipline. Rate stability for queueing networks differs from stability in that only the first order behavior

of $Z^x(t)$ enters into (5.58); an unstable queueing network might conceivably be rate stable, with $\lim_{t\to\infty} |Z^x(t)| = \infty$ a.s., but with $|Z^x(t)| = o(t)$.

As is the case for stability, fluid models and fluid limits may be employed to demonstrate rate stability. A fluid model is *weakly stable*, if for each solution of the fluid model equations with $Z(0) = 0$, one has $Z(t) = 0$ for all $t \geq 0$. Clearly, stability of a fluid model implies weak stability. Also, as was done in the context of global stability, the basic fluid model may be employed to demonstrate global rate stability. (In the literature, the term *globally weakly stable* is used when the basic fluid model is weakly stable.)

For queueing networks, rate stability and global rate stability are less satisfying properties than are stability and global stability, but they are easier to show. One has the following analog of Theorem 4.16.

Theorem 5.19. *Assume that an associated fluid model of a given HL queueing network is weakly stable. Then, the queueing network is rate stable.*

Proof. Suppose on the contrary that the queueing network is not rate stable. Then,

$$\limsup_{t\to\infty} |Z^x(t)|/t > 0$$

for some $\omega \in G$, where G is given in (4.70). Let $a_n \to \infty$, as $n \to \infty$, be a sequence on which $\liminf_{t\to\infty} |Z^x(a_n)|/a_n > 0$. The sequence (a_n, x) satisfies (4.71). So, there is a subsequence (a_{i_n}, x) along which $\mathfrak{X}^x(a_{i_n}t)/a_{i_n}$ has a limit $\bar{\mathfrak{X}}(\cdot)$ that satisfies the associated fluid model equations.

Since the initial state is constant, $\bar{Z}(0) = 0$. Because the fluid model is assumed to be weakly stable, it follows that $\bar{Z}(t) \equiv 0$. In particular, $\bar{Z}(1) = 0$. Therefore,

$$\lim_{n\to\infty} |Z^x(a_{i_n})|/a_{i_n} = 0,$$

which is a contradiction. So, the queueing network is, in fact, rate stable. ■

By Proposition 4.12, the basic fluid model is associated with its queueing network. The following corollary is therefore an immediate consequence of Theorem 5.19. Both Theorem 5.19 and the corollary are related to Theorem 4.1 in [Ch95].

Corollary 1. *Assume that the basic fluid model of a given HL queueing network is weakly stable. Then, the queueing network is globally rate stable.*

The following weak stability analog of Theorem 5.18 is given in [DaV00]. As before, we restrict the result to reentrant lines from queueing networks with deterministic routing. We employ the same notation as before involving strictly separating sets of excursions, virtual stations, and push start sets. Note that the conditions (5.59) and (5.60) for weak stability are the same as (5.55) and (5.56) in Theorem 5.18, except that strict inequalities in (5.55) and (5.56) are replaced by inequalities.

Theorem 5.20. *A two-station basic fluid model is weakly stable if and only if*

$$\rho_j \leq 1 \quad \text{for } j = 1, 2, \tag{5.59}$$

and for each strictly separating set S and $\ell = 1, \ldots, L$,

$$\sum_{j=1}^{2} \frac{\alpha_1 m(V_j(S) \backslash F_j^{\leq}(\ell))}{1 - \alpha_1 m(F_j^{<}(\ell))} \leq 1. \tag{5.60}$$

The proof of Theorem 5.20 is not spelled out there, but, according to [DaV00], is analogous to that of Theorem 5.18.

5.5 Relationship Between QN and FM Stability

The material in the first four sections of this chapter has relied heavily on the stability of fluid models that are associated with a given queueing network. In the first three sections, stability of such fluid models enabled us to demonstrate stability for a number of disciplines when the queueing network is subcritical. In the last section, this approach was applied to global stability.

We have so far avoided the question in the opposite direction, of whether stability of a queueing network implies the stability of its associated fluid model. On account of Theorem 4.16, this would imply that the two concepts of stability are equivalent, modulo certain side conditions. If the above implication is not correct, how "close" are the two concepts? Besides being aesthetically pleasing, a two-directional relationship would allow reduction of questions involving the stability of queueing networks to the less complex setting of fluid models. Results, such as Theorem 5.18 of the previous section, would also take on added significance. Of course, for disciplines such as those in the first three sections of the chapter, this equivalence is already clear, if both the queueing networks and fluid models are stable whenever they are subcritical.

The question should not be taken in its most naive form. For instance, the basic fluid model for a queueing network (which has no equations specifying the discipline) need not be stable even if the queueing network is. So, the fluid model needs to include an appropriate equation (or equations) corresponding to the discipline; we have already seen that there are often "canonical" equations that suggest themselves in this context. Also, one should exclude certain "exotic" disciplines. For instance, if the priority rule favoring different classes is allowed to change with the total queue length $|Z|$, such a discipline might consist of priority rules that are decreasingly stable as $|Z| \to \infty$. The fluid model would then correspond to the limiting rule, which is not stable, whereas the queueing network itself could be stable.

As we will see, even for certain "standard" disciplines, the above two forms of stability are not equivalent: there exist stable queueing networks whose fluid models are not stable. Moreover, there are no general results in this direction. Nonetheless, one must work to produce such examples, which seem to be, in some sense, "borderline". So, at this point, one can claim that the reduction to fluid models "works well in practice". The same should hold for the basic fluid model in the context of global stability. For global weak stability, there is, in fact, a partial result, which we mention at the end of the section.

This section is divided into four parts. We first present an elementary condition for the instability of a queueing network from [Da96]. We then present examples of stable queueing networks with unstable fluid models from [Br99] and [DaHV04], which together constitute most of the section. We conclude with the global weak stability result mentioned above, which is from [GaH05].

An elementary condition for instability

Proposition 5.21 gives an elementary condition for the instability of a queueing network in terms of its fluid limits. The result relies on Proposition 4.11 and is a variant of a result from [Da96]. Related results, with more involved conditions, are given in [Me95] and [PuR00]. Similar reasoning also shows that an HL queueing network with a supercritical station is unstable; the result is included in Proposition 5.21.

Proposition 5.21. *(a) Assume that for every fluid limit $\bar{\mathfrak{X}}(\cdot)$ of a given HL queueing network, with $\bar{Z}(0) = 0$, that $\bar{Z}(\delta) \neq 0$ for some fixed $\delta > 0$. Then, for every initial state x,*

$$\liminf_{t\to\infty} |Z^x(t)|/t > 0 \quad \text{on } G. \tag{5.61}$$

(b) Assume that for a given HL queueing network, $\rho_j > 1$ at some j. Then, for some $\epsilon > 0$ and every initial state x,

$$\liminf_{t\to\infty} |Z^x(t)|/t \geq \epsilon \quad \text{on } G. \tag{5.62}$$

Proof. The argument for (5.61) is almost the same as that used in the proof of Theorem 5.19. Suppose on the contrary that, for some $\omega \in G$,

$$\liminf_{t\to\infty} |Z^x(\delta t)|/t = 0,$$

where $\delta > 0$ is chosen as in the statement of the proposition. Let $a_n \to \infty$, as $n \to \infty$, be a sequence on which this limit holds. Since the sequence (a_n, x) satisfies (4.71), by Proposition 4.12, there is a subsequence (a_{i_n}, x) along which $\mathfrak{X}^x(a_{i_n}t)/a_{i_n}$ has a limit $\bar{\mathfrak{X}}(\cdot)$. Since the initial state is constant, $\bar{Z}(0) = 0$, and so, by assumption, $\bar{Z}(\delta) \neq 0$. Therefore,

$$\lim_{t\to\infty} |Z^x(\delta a_{i_n})|/a_{i_n} > 0,$$

which is a contradiction. Consequently, (5.61) holds.

Suppose $\rho_j > 1$ for some j. The basic fluid model is associated with the queueing network. By Part (d) of Proposition 4.11, $|Z(1)| \geq \epsilon$ for some $\epsilon > 0$ and all solutions of the basic fluid model. Reasoning analogous to that for (5.61) then implies (5.62). ■

The assumption in Part (a) of Proposition 5.21, that $\bar{Z}(\delta) \neq 0$ for *all* fluid limits with $\bar{Z}(0) = 0$, is unfortunately too strong for most applications, as is the assumption, in Part (b), that $\rho_j > 1$ for some j. One can, in fact, wonder whether the proposition has any applications to subcritical networks. Note, for instance that, under $Z(0) = 0$, the conditions $Z(t) \equiv 0$ and $D(t) = \lambda t$ are equivalent for any fluid model. Since none of the stations, in this case, is overloaded if the network is subcritical, these equations provide a solution for fluid models such as the basic fluid model and the other fluid models that have appeared in this chapter. So, the assumptions in Part (a) will not be satisfied if one considers all fluid model solutions (rather than just fluid limits).

The situation is different if one considers only the fluid limits of the queueing network. For instance, a subcritical queueing network might contain supercritical virtual stations, as in the previous section. In this setting, $\bar{Z}(\delta) = 0$ will no longer be possible for any fluid limit. Hence, Part (a) of Proposition 5.21 will be applicable. Other situations where $\bar{Z}(\delta) \neq 0$ for all fluid limits will also occur.

An example of a stable queueing network with unstable fluid model

In this subsection and the next, we will present two examples of stable queueing networks with unstable fluid models. In this subsection, the example consists of a network with routing that is a modification of that given in Figure 5.3.

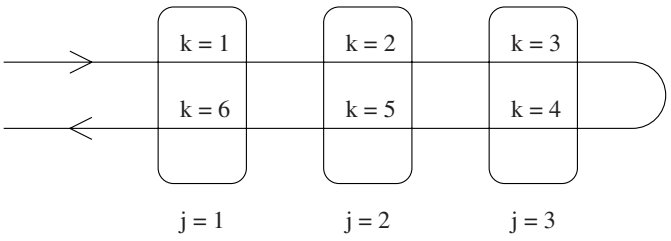

Fig. 5.3. The six classes along the route are labelled in the order of their appearance; the mean service times are given in (5.63). The priority scheme is (6,1), (5,2), and (3,4) at the three stations.

The network portrayed in Figure 5.3 is a reentrant line with three stations, each possessing two classes. The discipline is a preemptive SBP, with priority scheme (6,1), (5,2), and (3,4) at the three stations. Interarrival and service

times are assumed to be exponentially distributed, with the interarrival times having mean 1 and the service times having means

$$m_2 = m_3 = m_6 = 3/4, \quad m_1 = m_5 = \gamma, \quad m_4 = \gamma/L^2, \tag{5.63}$$

where $\gamma \in (0, 1/8)$ and $L \in \mathbf{Z}_+$. (One can, for example, set $\gamma = 1/16$.) The queueing network is subcritical with

$$\rho_1 = \rho_2 = \frac{3}{4} + \gamma < \frac{7}{8} \quad \text{and} \quad \rho_3 = \frac{3}{4} + \frac{\gamma}{L^2} < \frac{7}{8}. \tag{5.64}$$

The process $Z(t)$ corresponding to the queueing network is Markov because of the SBP discipline and the exponential interarrival and service times.

The modified queueing network we will employ is defined by "splitting" the station 3 into L separate two-class stations, $(3,1), \ldots, (3,L)$, one of which is randomly chosen along the route. That is, the higher priority class of each of these stations is entered with probability $1/L$, by a job leaving class 2. After service at this class is completed, the job passes to the lower priority class of this station, and then to class 5 of the original network; the queueing network otherwise evolves as before (see Figure 5.4). The new service times are exponentially distributed with means $\frac{3}{4}L$ and γ/L. The modified network is subcritical, with traffic intensity again given by (5.64). The process $Z(t)$ corresponding to the modified queueing network is again Markov. The original network can be obtained by "collapsing" the stations $(3,1), \ldots, (3,L)$ into a single station 3.

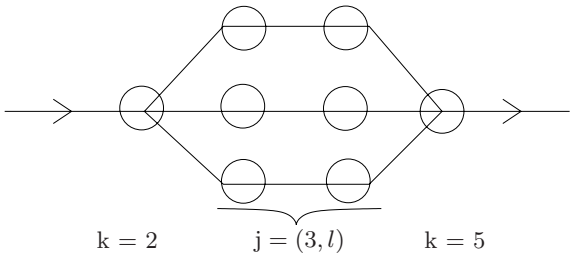

Fig. 5.4. In the modified network, jobs leaving class 2 are randomly routed to one of the L stations, $(3,l)$. After service at the two classes at the station, jobs are routed to class 5. In the figure, L = 3.

Both the original and modified queueing networks have preemptive SBP disciplines. As in Section 5.2, the fluid models consisting of the basic fluid equations, together with the SBP equation (5.13), are associated with these queueing networks. In [Br99], the stability of the modified queueing network and its fluid model are characterized as follows.

Theorem 5.22. *(a) For sufficiently large L, the modified queueing network defined above is stable. (b) For any L, the associated fluid model is not stable. In particular, there is a solution of the fluid model, with*

$$\liminf_{t\to\infty} |Z(t)|/t = 1/3. \tag{5.65}$$

Sketch of proof. The argument for Part (b) consists of explicitly constructing a solution of the fluid model equations that satisfies (5.65). The construction is similar to that for the simpler two-station, four-class network in Example 2 of the second part of Section 4.3. Rather than doing this here, we will instead motivate the evolution of the solution. The interested reader can refer to page 825 of [Br99] for a precise treatment.

The above solution is constructed so as to take the same values at each of the stations $(3,\ell), l = 1,\ldots,L$. The sum of the contributions of these stations therefore undergoes the same evolution as the corresponding solution for the original network, which we henceforth consider. For the original network, $m_5 \le m_6$. Since class 5 is the higher priority class at station 2, this implies that the fluid mass there is always passed to class 6 at least as fast as it can be served at class 6. Also, since $m_2 = m_3$, fluid mass for this solution is also passed from class 2 to class 3 as fast as it can be served at class 3, provided no mass is concurrently served at the higher priority class 5. Such will be the case for this particular solution, as can easily be verified by checking its explicit construction.

Using these two observations, one can see that this solution will become a solution of the two-station network obtained by deleting station 2, if one combines the mass at class 2 with that at class 3, and the mass at class 5 with that at class 6. The resulting network is just the Lu-Kumar network in Figure 3.1, with priorities $(6,1)$ and $(3,4)$ at the remaining stations 1 and 3, and service times

$$m_3 = m_6 = 3/4, \quad m_1 = \gamma, \quad m_4 = \gamma/L^2. \tag{5.66}$$

The classes 3 and 6 are the high priority classes at their stations. For the same reasons as given between (5.49) and (5.52) for the fluid model considered there, the classes 3 and 6 form a virtual station, with

$$\rho_{\text{virtual}} = 3/2 > 1. \tag{5.67}$$

So, the two-station fluid model is not stable. For our particular solution, fluid mass will only be processed at this virtual station at 2/3 the rate at which it arrives there, and so 1/3 will be a lower bound for the limit in (5.65). From the actual construction of the solution, the limit in fact equals 1/3, as claimed in Part (b) of the theorem.

The argument for Part (a) of Theorem 5.22 is more involved. The result might also be true for the original queueing network, but one needs to employ the modified queueing network, with large L, to obtain the bounds that are used here.

The basic reason for the different behavior in Parts (a) and (b) of the theorem is the different behavior at the stations $(3, \ell)$. For the unstable fluid model solution that was just discussed, as in the examples of the unstable queueing networks in Chapter 3, the flow of mass through the network is "cyclic", with, in particular, an increasingly large periodic buildup of mass at the different stations. This includes the stations $(3, \ell)$, $\ell = 1, \ldots, L$. (Or equivalently, the station 3 for the original network.)

At the stations $(3, \ell)$, $\ell = 1, \ldots, L$, the behavior of the queueing network is different. One can show that, for large L, the effective service rate at class 2 is slower than the combined service rate at the classes $(3, \ell)$, $\ell = 1, \ldots, L$, when the combined number of jobs at the classes $(3, \ell)$ is large. The basic idea is that for large L, the probability is close to 1 that, relatively frequently, one of the classes, say $(3, \ell_0)$, becomes empty. Once this occurs, service begins on the jobs at the quick low priority class $(4, \ell_0)$ that follows $(3, \ell_0)$. The served jobs from there continue to the high priority class 5, which interrupts service at class 2. This interruption prevents jobs from entering *any* of the $(3, \ell)$ classes until class 5 empties, which only occurs after it stops receiving jobs from the different $(4, \ell)$ classes. As additional $(3, \ell)$ classes empty, this allows service to begin at the corresponding $(4, \ell)$ classes. Without this interference from class 5, the service rate at class 2 is the same as the combined service rate of all of the $(3, \ell)$ classes, which is 4/3. This interference, however, creates idle periods for class 2, which causes it to have a slower effective service rate than the combined service rate of the $(3, \ell)$ classes. This service rate is also slower than the combined service rate of the $(3, \ell)$ stations, since service at the $(4, \ell)$ classes is quick.

The above behavior creates a strong bias for the total number of jobs $Z_3^{\Sigma}(t)$ at all of the $(3, \ell)$ stations to drift toward 0. Using this, one can show that $Z_3^{\Sigma}(t)$ typically remains close to 0 on the relevant time scale after it first hits 0. It is therefore reasonable to guess that the presence of the $(3, \ell)$ stations should have only a negligible effect on the evolution of the queueing network on this time scale, and that the qualitative behavior of the network should not change if these stations are omitted. This is, in fact, correct. By using fluid limits, one can make this statement precise. (The fluid limits used in [Br99] are slightly more general than those introduced in Section 4.3. They are mentioned briefly in the discussion after Theorem 4.16.)

After a fixed time (depending on the initial state), all such fluid limits will have no mass at station 3. They will satisfy the SBP fluid model consisting of the remaining stations 1 and 2, whose classes have priorities (6,1) and (5,2), and service times

$$m_2 = m_6 = 3/4 \quad \text{and} \quad m_1 = m_5 = \gamma. \tag{5.68}$$

This SBP discipline is LBFS. Since $\rho_1 = \rho_2 < 7/8 < 1$, one knows from Theorem 5.6 that this two-station fluid model is stable. It follows that the fluid limits for the entire modified queueing network are stable. From this,

it in turn follows that the modified queueing network is stable. So, Part (a) holds. ■

The queueing network in Theorem 5.22 is an example of a stable SBP queueing network for which the associated fluid model consisting of the basic fluid model equations and (5.13) is not stable. The discipline here is an SBP discipline and not an "exotic" discipline one would wish to exclude from consideration. This example therefore casts doubt on a robust equivalence between queueing network and fluid model stability.

Despite this example, one can still ask whether there is some general correspondence between queueing network stability and some notion similar to fluid model stability. This question is, of course, vague, and there are different possible approaches. One approach is to focus on the stability of fluid limits instead of fluid models. There is no known correspondence here either. Moreover, the set of fluid limits for a queueing network will be difficult to describe in general.

Another approach is more philosophical. How does one know that the fluid model employed in Theorem 5.22 is the "right one"? Perhaps fluid limits of the queueing network automatically satisfy further "hidden" fluid model equations, and under these additional equations, all fluid model solutions will be stable. Maybe the same is true for queueing networks in general. It is unclear how to disprove such a thesis. However, even if the thesis is correct, one will still need a way of finding such hidden equations, in order for fluid models to provide a general practical framework for reformulating queueing network stability.

The literature on related work includes two papers, [FoK99] and [StR99], that give examples of stable networks with unstable fluid limits for polling models. The model in [FoK99] consists of two stations and two servers, which switch back and forth between the stations when the work is exhausted; there is also a switchover period. The paper also introduces a less restrictive criterion of stability under which the fluid limits are stable.

An example where stability depends on the distributions of the queueing network

At the end of the last subsection, we wondered how one could discount the possibility that an unstable fluid model, which is associated with a stable queueing network, merely lacks equations that are implicit in the evolution of the queueing network. One convincing response to this would be to find two queueing networks, one stable and the other not, that differ only in their interarrival and service time distributions, but for which everything else, including the interarrival and service time means, is the same. The queueing networks would then have the same fluid model, which would show that stability of a queueing network cannot always be determined at the fluid model level.

This approach is taken in [DaHV04]. The queueing network that is analyzed is the two-station SBP reentrant line in Figure 5.1 of the previous sec-

tion. It has priority scheme $(1, 3, 4)$ and $(5, 2)$ at the two stations, interarrival rate 1, and mean service times given by (5.48). One version of the queueing network is assumed to have deterministic interarrival and service times. Hence, there is no randomness in the evolution of the network. The other version of the queueing network is assumed to have exponentially distributed interarrival and service times. Both versions are assumed to be nonpreemptive.

In order to describe the evolution of the deterministic network, abbreviated notation is employed in [DaHV04] to designate certain specific states. One uses the 6-tuples, $(z_1, \ldots, z_5; a)$, where z_k is the number of jobs in class k and a is the remaining interarrival time until the next job enters the network. Only certain states occurring at the instant of a service completion are designated this way, and they are viewed at the time $t-$ just prior to the completion of service. For example, by

$$(0, 0, 0, 1, 0; a), \tag{5.69}$$

one means the "state" at a time $t-$ if service of the class 4 customer is completed at time t. Strictly speaking, these are not states for the state space we introduced in Section 4.1, but they suffice for describing the evolution of the network from the specific "states" mentioned above. One can, in particular, verify that the deterministic network starting from (5.69), with $a \in (0, 0.1]$, returns to this state exactly one unit of time later. Thus, such a trajectory forms an *orbit*; for a given $a \in (0, 0.1]$, this orbit is called an *a orbit*.

Employing this terminology, it is shown in [DaHV04] that the queueing network with deterministic distributions has the following behavior.

Theorem 5.23. *For any initial state of the nonpreemptive deterministic queueing network specified above, there exists a finite time at which the network enters an a orbit, with* $a \in (0, 0.1]$.

On the other hand, in [DaHV04], it is also shown that the following result holds for the queueing network with exponentially distributed interarrival and service times.

Theorem 5.24. *For any initial state of the nonpreemptive exponential queueing network specified above,*

$$|Z(t)| \to \infty \quad \text{as } t \to \infty, \tag{5.70}$$

with probability 1.

The exponential queueing network in Theorem 5.24 is unstable. (This is also true for the preemptive version of the discipline.) On the other hand, the trajectories of the deterministic queueing network in Theorem 5.23 all eventually enter a fixed bounded set. Since not all states communicate with one another, its underlying (deterministic) Markov process is not positive Harris recurrent, and so the network is not stable in the sense we are using in these lectures. Nevertheless, the two examples show distinctly different behavior with regard to their "stability", despite having interarrival and service time

distributions with the same means, and therefore not being distinguishable at the fluid model level.

In [DaHV04], it is also shown that the above deterministic queueing network, but with preemptive rather than nonpreemptive discipline, is unstable, with $|Z(t)| \to \infty$ linearly as $t \to \infty$, for appropriate initial states. So, whether or not a queueing network is preemptive can also influence its stability.

The natural question arises as to whether the behavior of the nonpreemptive deterministic network is an anomaly. Simulations in [DaHV04] seem to indicate that when the deterministic interarrival and service times are replaced by those with uniform distributions having the same means as before, the (nonpreemptive) network is stable when the width of the uniform distributions is 0.001, but is not stable when the width is 0.1.

It is tricky to attempt to formulate rules as to which distributions should be stable and which should not, for more general networks. On the basis of the above examples, it is tempting to say that both a more deterministic distribution and a nonpreemptive discipline should be "good" with regard to stability, whereas a more random distribution and a preemptive discipline should be "bad". But, the three-station example in Theorem 5.22 of the previous subsection consists of a stable preemptive queueing network with exponential distributions and a fluid model that is not stable, which make this less clear. In [DaHV04], it is suggested that exponential distributions should be "bad" for a large family of two-station networks, based on the belief that the proof of Theorem 5.24 should generalize.

Not surprisingly, the argument for Theorem 5.23 is primarily computational, on account of the system's deterministic evolution. The reader is referred to [DaHV04] for details. The argument for Theorem 5.24 is more involved. In the remainder of this subsection, we will discuss some of the ideas behind the proof of the theorem that relate to virtual stations and push starts.

Virtual stations and push starts were introduced in the previous section in the context of the associated fluid model for the preemptive version of the queueing network in Theorem 5.24. Since the push start condition (5.53) is violated for the choice of service times in (5.48), one already knows from the previous section that the fluid model is not stable. One would like to be able to apply similar reasoning to show that the queueing network in Theorem 5.24 is unstable.

We recall that for the same queueing network, but where the discipline is preemptive, the classes 3 and 5 form a virtual station. As explained in the previous section, the virtual station owes its presence to the constraint in (5.49), which does not allow simultaneous service at the two classes if either is initially empty.

Unfortunately, for the nonpreemptive version of interest to us here, (5.49) need not hold. For instance, a job at class 4 can complete service there and continue to class 5 after another job has already arrived at class 3. However, at most one job at class 5 can coexist with jobs at class 3; the same is true for jobs at class 5 coexisting with those at class 3. For the nonpreemptive

network, one can therefore replace (5.49) with

$$(Z_3(t) - 1)^+ = 0 \text{ or } (Z_5(t) - 1)^+ = 0 \quad \text{for each } t \geq 0, \tag{5.71}$$

assuming (5.49) holds at $t = 0$. In contrast to (5.49), (5.71) allows simultaneous service at classes 3 and 5. One might hope that there is not "too much" such joint service, so that the classes together still behave like a virtual station, up to a small error.

There are also difficulties in applying the push start condition in (5.53) to the queueing network. The inequality applies to the fluid model, where it relies on the proportion of effort devoted to class 1 being constant over time. For the queueing network (either preemptive or nonpreemptive), there may or may not be a job there receiving service at a given time. So, service at class 1 may or may not interfere with service at class 3. The reasoning leading up to (5.53) will therefore not hold in the strict sense. However, if there is sufficient independence between the times of external arrivals to class 1 and times when class 3 is not empty, one might expect the same behavior to hold up to a small error. Since the arrivals to the network are Poisson, one might expect that to be the case in the present setting. It is certainly not the case for the deterministic network in Theorem 5.23.

Much of the work in [DaHV04] is devoted to carrying out the ideas that are summarized in the last two paragraphs, in order to demonstrate Theorem 5.24. The queueing network itself must be analyzed, which involves deriving estimates based on its random evolution. However, the virtual station and push start behavior of the associated fluid model provide guidance for these steps.

Global rate stability

The examples in the previous two subsections cast doubt on the equivalence, under general conditions, of queueing network and fluid model stability. This does not preclude analogous positive results in the context of global stability, though. In particular, the presence of a basic fluid model that is not stable might imply the existence of an associated queueing network that is not stable, under *some* discipline. This is an open question even for two-station reentrant lines. If one knows this equivalence in the two-station reentrant line setting, one can then apply Theorem 5.18 to obtain necessary and sufficient conditions on the global stability of such reentrant lines.

At the end of Section 5.4, we introduced the concepts of rate stability, global rate stability, and weak stability. As in the two previous subsections, one can attempt to show the equivalence of rate stability for queueing networks and weak stability for fluid models. One does not meet with greater success here than with stability. The deterministic queueing network in Theorem 5.23, for example, is rate stable, but its fluid model (with auxiliary equation (5.13)) is not.

The question of global rate stability is more approachable. We recall, from the corollary to Theorem 5.19, that an HL queueing network is globally rate

stable if its basic fluid model is weakly stable. The following partial converse is shown in [GaH05] for a family of queueing networks. The family consists of two-station networks with deterministic routing. The interarrival and service distributions there are assumed to satisfy a large deviation condition.

Theorem 5.25. *Assume that a two-station queueing network with the above properties is globally rate stable. Then, its basic fluid model is weakly stable.*

The proof of Theorem 5.25 is by contradiction. The basic idea is to show the existence of a fluid model solution that diverges linearly to infinity, and to construct a discipline for which the sample paths of the queueing network almost surely "track" this solution. The large deviation condition is employed in this construction.

The queueing networks in the theorem differ, in two ways, from the HL queueing networks we defined in Section 4.1 and have been employing since. Changes in service allocation are allowed between the arrival and departure times of jobs at stations, in order to facilitate tracking. More seriously, the discipline that was constructed need not be time homogeneous. It should be possible to modify the construction so as to eliminate both of these differences.

Since the discipline of a queueing network is not reflected in the basic fluid model equations, the corollary to Theorem 5.19 continues to hold in the setting of the queueing networks in Theorem 5.25. Together with the theorem, this implies the equivalence of global rate stability for these two-station queueing networks and weak stability for their basic fluid models. As an immediate consequence of this and Theorem 5.20, one obtains the following necessary and sufficient conditions for the global rate stability of two-station queueing networks. Since Theorem 5.20 was stated for reentrant lines, we make the same restriction here. We also make the same assumptions on the networks as were made in Theorem 5.25.

Theorem 5.26. *A two-station reentrant line with the above properties is globally rate stable if and only if both (5.59) and (5.60) hold.*

References

[AnAFKLL96] Andrews, M., Awerbuch, B., Fernandez, A., Kleinberg, J., Leighton, T., and Liu, Z. (1996). Universal stability results for greedy contention-resolution protocols. In *Symposium on Computer Science*, 1996.

[As03] Asmussen, S. (2003). *Applied Probability and Queues*, second edition. Springer, New York.

[AzKR67] Azema, J., Kaplan-Duflo, M. and Revuz, D. (1967). Mesure invariante sur les classes récurrentes des processus de Markov. *Z. Wahrscheinlichkeitstheorie verw. Geb.* **8**, 157-181.

[BaB99] Baccelli, F. and Bonald, T. (1999). Window flow control in FIFO networks with cross traffic. *Queueing Systems* **32**, 195-231.

[BaF94] Baccelli, F. and Foss, S.G. (1994). Ergodicity of Jackson-type queueing networks I. *Queueing Systems* **17**, 5-72.

[Ba76] Barbour, A.D. (1976). Networks of queues and the method of stages. *Adv. Appl. Prob.* **8**, 584-591.

[BaCMP75] Baskett, F., Chandy, K.M., Muntz, R.R., and Palacios, F.G. (1975). Open, closed and mixed networks of queues with different classes of customers. *J. Assoc. Comput. Mach.* **22**, 248-260.

[BoKRSW96] Borodin, A., Kleinberg, J., Raghavan, P., Sudan, M., and Williamson, D.P. (1996). Adversarial queueing theory. In *Symposium on Computer Science*, 1996.

[BoZ92] Botvich, D.D. and Zamyatin, A.A. (1992). Ergodicity of conservative communication networks. *Rapport de Recherche* **1772**, INRIA.

[Br94a] Bramson, M. (1994). Instability of FIFO queueing networks. *Ann. Appl. Probab.* **4**, 414-431. (Correction **4** (1994), 952.)

[Br94b] Bramson, M. (1994). Instability of FIFO queueing networks with quick service times. *Ann. Appl. Probab.* **4**, 693-718.

[Br95] Bramson, M. (1995) Two badly behaved queueing networks. *Stochastic Networks*, 105-116. IMA Math. Appl. **71**. Springer, New York.

[Br96a] Bramson, M. (1996). Convergence to equilibria for fluid models of FIFO queueing networks. *Queueing Systems* **22**, 5-45.

[Br96b] Bramson, M. (1996). Convergence to equilibria for fluid models of head-of-the-line proportional processor sharing queueing networks. *Queueing Systems* **23**, 1-26.

[Br98a] Bramson, M. (1998). Stability of two families of queueing networks and a discussion of fluid limits. *Queueing Systems* **28**, 7-31.

[Br98b] Bramson, M. (1998). State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems* **30**, 89-148.

[Br99] Bramson, M. (1999). A stable queueing network with unstable fluid model. *Ann. Appl. Probab.* **9**, 818-853.

[Br01] Bramson, M. (2001). Earliest-due-date, first-served queueing networks. *Queueing Systems* **39**, 79-102.

[BrD01] Bramson, M. and Dai, J.G. (2001). Heavy traffic limits for some queueing networks. *Ann. Appl. Probab.* **11**, 49-90.

[Bu56] Burke, P.J. (1956). The output of a queueing system. *Operat. Res.* **4**, 699-704.

[ChTK94] Chang, C.S., Thomas, J.A., and Kiang, S.H. (1994). On the stability of open networks: a unified approach by stochastic dominance. *Queueing Systems* **15**, 239-260.

[Ch95] Chen, H. (1995). Fluid approximations and stability of multiclass queueing networks: work-conserving disciplines. *Ann. Appl. Probab.* **5**, 637-655.

[ChM91] Chen, H. and Mandelbaum, A. (1991). Discrete flow networks: Bottleneck analysis and fluid approximations. *Math. Oper. Res.* **16**, 408-446.

[ChY01] Chen, H. and Yao, D.D. (2001). *Fundamentals of Queueing Networks*. Springer, New York.

[Da95] Dai, J.G. (1995). On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *Ann. Appl. Probab.* **5**, 49-77.

[Da96] Dai, J.G. (1996). A fluid limit model criterion for instability of multiclass queueing networks. *Ann. Appl. Probab.* **6**, 751-757.

[DaHV99] Dai, J.G., Hasenbein, J.J., and Vande Vate, J.H. (1999). Stability of a three-station fluid network. *Queueing Systems* **33**, 293-325.

[DaHV04] Dai, J.G., Hasenbein, J.J., and Vande Vate, J.H. (2004). Stability and instability of a two-station queueing network. *Ann. Appl. Probab.* **14**, 326-377.

[DaM95] Dai, J.G. and Meyn, S.P. (1995). Stability and convergence of moments for multiclass queueing networks via fluid limit models. *IEEE Trans. Automat. Control* **40**, 1889-1904.

[DaV96] Dai, J.G. and Vande Vate, J.H. (1996). Virtual stations and the capacity of two-station queueing networks. Preprint.

[DaV00] Dai, J.G. and Vande Vate J.H. (2000). The stability of two-station multitype fluid networks. *Operations Research* **48**, 721-744.

[DaWa93] Dai, J.G. and Wang, Y. (1993). Nonexistence of Brownian models of certain multiclass queueing networks. Queueing Systems **13**, 41-46.

[DaWe96] Dai, J.G. and Weiss, G. (1996). Stability and instability of fluid models for re-entrant lines. *Math. of Oper. Res.* **21**, 115-134.

[Da84] Davis, M.H.A. (1984). Piecewise deterministic Markov processes: a general class of nondiffusion stochastic models. *J. Roy. Statist. Soc. Ser. B* **46**, 353-388.

[Da93] Davis, M.H.A. (1993). *Markov Models and Optimization*. Chapman & Hall, London.

[Do53] Doob, J.L. (1953). *Stochastic Processes.* Wiley, New York.

[DoM94] Down, D. and Meyn, S.P. (1994). Piecewise linear test functions for stability of queueing networks. *Proc. 33rd Conference on Decision and Control*, 2069-2074.

[Du96] Dumas, V. (1996). Approches fluides pour la stabilité et l'instabilité de réseaux de files d'attente stochastiques à plusieurs classes de clients. Doctoral thesis, Ecole Polytechnique.

[Du97] Dumas, V. (1997). A multiclass network with non-linear, non-convex, non-monotonic stability conditions. *Queueing Systems* **25**, 1-43.

[Dur96] Durrett, R. (1996). *Probability: Theory and Examples*, second edition. Duxbury Press, Belmont.

[Fi89] Filonov, Y.P. (1989). A criterion for the ergodicity of discrete homogeneous Markov chains. (Russian). *Ukrain. Mat. Zh.* **41**, 1421-1422; translation in *Ukrainian Math. J.* **41**, 1223-1225.

[Fo91] Foss, S.G. (1991). Ergodicity of queueing networks. *Sib. Math. J.* **32**, 183-202.

[FoK04] Foss, S.G. and Konstantopoulos, T. (2004). An overview of some stochastic stability methods. *J. Oper. Res. Soc. Japan* **47**, 275-303.

[FoK99] Foss, S.G. and Kovalevskii, A. (1999). A stability criterion via fluid limits and its application to a polling system. *Queueing Systems* **32**, 131-168.

[Fo53] Foster, F.G. (1953). On the stochastic matrices associated with certain queueing processes. *Ann. Math. Stat.* **24**, 355-360.

[GaH05] Gamarnik, D. and Hasenbein, J.J. (2005). Instability in stochastic and fluid queueing networks. *Ann. Appl. Probab.* **15**, 1652-1690.

[Ge79] Getoor, R. (1979). Transience and recurrence of Markov processes. In *Séminaire de Probabilités XIV* (J. Azéma, M. Yor editors), 397-409. Springer, Berlin.

[Ha56] Harris, T.E. (1956). The existence of stationary measures for certain Markov processes. *Proc. 3rd Berkeley Symp.*, Vol. II, 113-124.

[Ha97] Hasenbein, J.J. (1997). Necessary conditions for global stability of multiclass queueing networks. *Oper. Res. Lett.* **21**, 87-94.

[Hu94] Humes, C. (1994). A regular stabilization technique: Kumar-Seidman revisited. *IEEE Trans. Automat. Control* **39**, 191-196.

[Ja54] Jackson, R.R.P. (1954). Queueing systems with phase-type service. *Operat. Res. Quart.* **5**, 109-120.

[Ja63] Jackson, R.R.P. (1963). Jobshop-like queueing systems. *Mgmt. Sci.* **10**, 131-142.

[KaM94] Kaspi, H. and Mandelbaum, A. (1994). On Harris recurrence in continuous time. *Math. Oper. Res.* **19**, 211-222.

[Ke75] Kelly, F.P. (1975). Networks of queues with customers of different types. *J. Appl. Probab.* **12**, 542-554.

[Ke76] Kelly, F.P. (1976). Networks of queues. *Adv. Appl. Prob.* **8**, 416-432.

[Ke79] Kelly, F.P. (1979). *Reversibility and Stochastic Networks.* Wiley, New York.

[Kn84] Knight, F.B. (1984). On the ray topology. *Séminaire de Probabilités (Strasbourg)* **18**, 56-69. Springer, Berlin.

[Ku93] Kumar, P.R. (1993). Re-entrant lines. *Queueing Systems* **13**, 87-110.

[KuS90] Kumar, P.R. and Seidman, T.I. (1990). Dynamic instabilities and stabilization methods in distributed real-time scheduling of manufacturing systems. *IEEE Trans. Automat. Control* **35**, 289-298.

[LuK91] Lu, S.H. and Kumar, P.R. (1991). Distributed scheduling based on due dates and buffer priorities. *IEEE Trans. Automat. Control* **36**, 1406-1416.

[MaM81] Malyshev, V.A. and Menshikov, M.V. (1981). Ergodicity, continuity, and analyticity of countable Markov chains. *Transactions of the Moscow Mathematical Society* **1**, 1-47.

[Me95] Meyn, S.P. (1995). Transience of multiclass queueing networks via fluid limit models. *Ann. Appl. Probab.* **5**, 946-957.

[MeD94] Meyn, S.P. and Down, D. (1994). Stability of generalized Jackson networks. *Ann. Appl. Probab.* **4**, 124-148.

[MeT93a] Meyn, S.P. and Tweedie, R.L. (1993). Generalized resolvents and Harris recurrence of Markov processes. *Contemporary Mathematics* **149**, 227-250.

[MeT93b] Meyn, S.P. and Tweedie, R.L. (1993). Stability of Markov processes II: continuous-time processes and sampled chains. *Adv. Appl. Prob.* **25**, 487-517.

[MeT93c] Meyn, S.P. and Tweedie, R.L. (1993). Stability of Markov processes III: Foster-Lyapunov criteria for continuous-time processes. *Adv. Appl. Prob.* **25**, 518-548.

[MeT93d] Meyn, S.P. and Tweedie, R.L. (1993). *Markov Chains and Stochastic Stability*. Springer, London.

[MeT94] Meyn, S.P. and Tweedie, R.L. (1994). State-dependent criteria for convergence of Markov chains. *Ann. Appl. Probab.* **4**, 149-168.

[Mu72] Muntz, R.R. (1972). *Poisson Departure Processes and Queueing Networks*, IBM Research Report RC 4145. A shortened version appeared in *Proceedings of the Seventh Annual Conference on Information Science and Systems*, Princeton, 1973, 435-440.

[Ne82] Newell, G.F. (1982). *Applications of Queueing Theory*. Chapman and Hall.

[Nu78] Nummelin, E. (1978). A splitting technique for Harris recurrent Markov chains. *Z. Wahrscheinlichkeitsth. verw. Geb.* **43**, 309-318.

[Nu84] Nummelin, E. (1984). *General Irreducible Markov Chains and Non-Negative Operators*. Cambridge University Press, Cambridge.

[Or71] Orey, S. (1971). *Lecture Notes on Limit Theorems for Markov Chain Transition Probabilities*. Van Nostrand Reinhold, London.

[PeK89] Perkins, J.R. and Kumar, P.R. (1989). Stable distributed real-time scheduling of flexible manufacturing/assembly/disassembly systems. *IEEE Trans. Automat. Control* **34**, 139-148.

[PuR00] Puhalskii, A. and Rybko, A. (2000). Nonergodicity of a queueing network under nonstability of its fluid model. *Probl. Inf. Transm.* **36**, 23-41.

[Re57] Reich, E. (1957). Waiting times when queues are in tandem. *Ann. Math. Statist.* **28**, 768-773.

[Re92] Resnick, S.I. (1992). *Adventures in Stochastic Processes*. Birkhäuser, Boston.

[RyS92] Rybko, A.N. and Stolyar, A.L. (1992). Ergodity of stochastic processes that describe functioning of open queueing networks. *Problems Inform. Transmission* **28**, 3-26 (in Russian).

[Se94] Seidman, T.I. (1994). "First come, first served" can be unstable! *IEEE Trans. Automat. Control* **39**, 2166-2170.

[Se99] Serfozo, R. (1999). *Introduction to Stochastic Networks*. Springer, New York.

[Sh88] Sharpe, M. (1988). *General Theory of Markov Processes*. Academic Press, San Diego.

[Si90] Sigman, K. (1990). The stability of open queueing networks. *Stoch. Proc. Applns.* **35**, 11-25.

[St95] Stolyar, A.L. (1995). On the stability of multiclass queueing networks: a relaxed sufficient condition via limiting fluid processes. *Markov Process. Related Fields* **1**, 491-512.

[StR99] Stolyar, A.L. and Ramakrishnan, K.K. (1999). The stability of a flow merge point with non-interleaving cut-through scheduling disciplines. In *Proceedings of INFOCOM 99*.

[Wa82] Walrand, J. (1982). Discussant's report on "Networks of quasi-reversible nodes", by F.P. Kelly. *Applied Probability - Computer Science: The Interface*, Volume 1, 27-29. Birkhäuser, Boston.

[Wa83] Walrand, J. (1983). A probabilistic look at networks of quasi-reversible queues. *IEEE Trans. Info. Theory*, **IT-29**, 825-836.

[Wa88] Walrand, J. (1988). *An Introduction to Queueing Networks*. Prentice-Hall, Englewood Cliffs.

[Wi98] Williams, R.J. (1998). Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing Systems* **30**, 27-88.

Index

List of Participants

Lecturers

Maury BRAMSON	University of Minnesota, Minneapolis, USA
Alice GUIONNET	École Normale Supérieure de Lyon, France
Steffen LAURITZEN	University of Oxford, UK

Participants

Marie ALBENQUE	Université Denis Diderot, Paris, France
Louis-Pierre ARGUIN	Princeton University, USA
Sylvain ARLOT	Université Paris-Sud, Orsay, France
Claudio ASCI	Università La Sapienza, Roma, Italy
Jean-Yves AUDIBERT	École Nationale des Ponts et Chaussées, Marne-la-Vallée, France
Wlodzimierz BRYC	University of Cincinnati, USA
Thierry CABANAL-DUVILLARD	Université Paris 5, France
Alain CAMANES	Université de Nantes, France
Mireille CAPITAINE	Université Paul Sabatier, Toulouse, France
Muriel CASALIS	Université Paul Sabatier, Toulouse, France

François CHAPON	Université Pierre et Marie Curie, Paris, France
Adriana CLIMESCU-HAULICA	CNRS, Marseille, France
Marek CZYSTOLOWSKI	Wrocław University, Poland
Manon DEFOSSEUX	Université Pierre et Marie Curie, Paris, France
Catherine DONATI-MARTIN	Université Pierre et Marie Curie, Paris, France
Coralie EYRAUD-DUBOIS	Université Claude Bernard, Lyon, France
Delphine FERAL	Université Paul Sabatier, Toulouse, France
Mathieu GOURCY	Université Blaise Pascal, Clermont-Ferrand, France
Mihai GRADINARU	Université Henri Poincaré, Nancy, France
Benjamin GRAHAM	University of Cambridge, UK
Katrin HOFMANN-CREDNER	Ruhr Universität, Bochum, Germany
Manuela HUMMEL	LMU, München, Germany
Jérémie JAKUBOWICZ	École Normale Supérieure de Cachan, France
Abdeldjebbar KANDOUCI	Université de Rouen, France
Achim KLENKE	Universität Mainz, Germany
Krzysztof LATUSZYNSKI	Warsaw School od Economics, Poland
Liangzhen LEI	Université Blaise Pascal, Clermont-Ferrand, France
Manuel LLADSER	University of Colorado, Boulder, USA
Dhafer MALOUCHE	École Polytechnique de Tunisie, Tunisia
Hélène MASSAM	York University, Toronto, Canada
Robert PHILIPOWSKI	Universität Bonn, Germany
Jean PICARD	Université Blaise Pascal, Clermont-Ferrand, France

Júlia RÉFFY	Budapest University of Technology and Economics, Hungary
Anthony REVEILLAC	Université de La Rochelle, France
Alain ROUAULT	Université de Versailles, France
Markus RUSCHHAUPT	German Cancer Research Center, Heidelberg, Germany
Erwan SAINT LOUBERT BIÉ	Université Blaise Pascal, Clermont-Ferrand, France
Pauline SCULLI	London School of Economics, UK
Sylvie SEVESTRE-GHALILA	Université Paris 5, France
Frederic UTZET	Universitat Autonoma de Barcelona, Spain
Nicolas VERZELEN	Université Paris-Sud, Orsay, France
Yvon VIGNAUD	Centre de Physique Théorique, Marseille, France
Pompiliu Manuel ZAMFIR	Stanford University, USA

List of Short Lectures

Louis-Pierre Arguin	Spin glass systems and Ruelle's probability cascades
Sylvain Arlot	Model selection by resampling in statistical learning
Claudio Asci	Generalized Beta distributions and random continued fractions
Wlodek Bryc	Families of probability measures generated by the Cauchy kernel
Thierry Cabanal-Duvillard	Lévy unitary ensemble and free Poisson point processes
Manon Defosseux	Random words and the eigenvalues of the minors of random matrices
Delphine Féral	The largest eigenvalue of rank one deformation of large Wigner matrices
Mathieu Gourcy	A large deviation principle for 2D stochastic Navier-Stokes equations
Mihaï Gradinaru	On the stochastic heat equation
Katrin Hofmann-Credner	Limiting laws for non-classical random matrices
Jérémie Jakubowicz	Detecting segments in digital images
Krzysztof Łatuszyński	$(\varepsilon - \alpha)$-MCMC approximation under drift condition
Liangzhen Lei	Large deviations of kernel density estimator

Gérard Letac	Pavage par séparateurs minimaux d'un arbre de jonction et applications aux modèles graphiques
Hélène Massam	Discrete graphical models Markov w.r.t. an undirected graph: the conjugate prior and its normalizing constant
Manuel Lladser	Asymptotics for the coefficients of mixed-powers generating functions
Robert Philipowski	Approximation de l'équation des milieux poreux visqueuse par des équations différentielles stochastiques non linéaires
Anthony Réveillac	Stein estimation on the Wiener space
Alain Rouault	Asymptotic properties of determinants of some random matrices
Markus Ruschhaupt	Simulation of matrices with constraints by using moralised graphs
Pauline Sculli	Counterparty default risk in affine processes with jump decay
Frederic Utzet	On the orthogonal polynomials associated to a Lévy process
Yvon Vignaud	Rigid interfaces for some lattice models

Saint-Flour Probability Summer Schools

In order to facilitate research concerning previous schools we give here the names of the authors, the series*, and the number of the volume where their lectures can be found:

Summer School	Authors	Series*	Vol. Nr.
1971	Bretagnolle; Chatterji; Meyer	LNM	307
1973	Meyer; Priouret; Spitzer	LNM	390
1974	Fernique; Conze; Gani	LNM	480
1975	Badrikian; Kingman; Kuelbs	LNM	539
1976	Hoffmann-Jörgensen; Liggett; Neveu	LNM	598
1977	Dacunha-Castelle; Heyer; Roynette	LNM	678
1978	Azencott; Guivarc'h; Gundy	LNM	774
1979	Bickel; El Karoui; Yor	LNM	876
1980	Bismut; Gross; Krickeberg	LNM	929
1981	Fernique; Millar; Stroock; Weber	LNM	976
1982	Dudley; Kunita; Ledrappier	LNM	1097
1983	Aldous; Ibragimov; Jacod	LNM	1117
1984	Carmona; Kesten; Walsh	LNM	1180
1985/86/87	Diaconis; Elworthy; Föllmer; Nelson; Papanicolaou; Varadhan	LNM	1362
1986	Barndorff-Nielsen	LNS	50
1988	Ancona; Geman; Ikeda	LNM	1427
1989	Burkholder; Pardoux; Sznitman	LNM	1464
1990	Freidlin; Le Gall	LNM	1527
1991	Dawson; Maisonneuve; Spencer	LNM	1541
1992	Bakry; Gill; Molchanov	LNM	1581
1993	Biane; Durrett;	LNM	1608
1994	Dobrushin; Groeneboom; Ledoux	LNM	1648
1995	Barlow; Nualart	LNM	1690
1996	Giné; Grimmett; Saloff-Coste	LNM	1665
1997	Bertoin; Martinelli; Peres	LNM	1717
1998	Emery; Nemirovski; Voiculescu	LNM	1738
1999	Bolthausen; Perkins; van der Vaart	LNM	1781
2000	Albeverio; Schachermayer; Talagrand	LNM	1816
2001	Tavaré; Zeitouni	LNM	1837
	Catoni	LNM	1851
2002	Tsirelson; Werner	LNM	1840
	Pitman	LNM	1875
2003	Dembo; Funaki	LNM	1869
	Massart	LNM	1896
2004	Cerf	LNM	1878
	Slade	LNM	1879
	Lyons**, Caruana, Lévy	LNM	1908
2005	Doney	LNM	1897
	Evans	LNM	1920
	Villani	Forthcoming	GL 338
2006	Bramson	LNM	1950
	Guionnet	Forthcoming	
	Lauritzen	Forthcoming	

*Lecture Notes in Mathematics (LNM), Lecture Notes in Statistics (LNS), Grundlehren der mathematischen Wissenschaften (GL)

**The St. Flour lecturer was T.J. Lyons.

Lecture Notes in Mathematics

For information about earlier volumes
please contact your bookseller or Springer
LNM Online archive: springerlink.com

Vol. 1766: H. Hennion, L. Hervé, Limit Theorems for Markov Chains and Stochastic Properties of Dynamical Systems by Quasi-Compactness (2001)
Vol. 1767: J. Xiao, Holomorphic Q Classes (2001)
Vol. 1768: M. J. Pflaum, Analytic and Geometric Study of Stratified Spaces (2001)
Vol. 1769: M. Alberich-Carramiñana, Geometry of the Plane Cremona Maps (2002)
Vol. 1770: H. Gluesing-Luerssen, Linear Delay-Differential Systems with Commensurate Delays: An Algebraic Approach (2002)
Vol. 1771: M. Émery, M. Yor (Eds.), Séminaire de Probabilités 1967-1980. A Selection in Martingale Theory (2002)
Vol. 1772: F. Burstall, D. Ferus, K. Leschke, F. Pedit, U. Pinkall, Conformal Geometry of Surfaces in S^4 (2002)
Vol. 1773: Z. Arad, M. Muzychuk, Standard Integral Table Algebras Generated by a Non-real Element of Small Degree (2002)
Vol. 1774: V. Runde, Lectures on Amenability (2002)
Vol. 1775: W. H. Meeks, A. Ros, H. Rosenberg, The Global Theory of Minimal Surfaces in Flat Spaces. Martina Franca 1999. Editor: G. P. Pirola (2002)
Vol. 1776: K. Behrend, C. Gomez, V. Tarasov, G. Tian, Quantum Comohology. Cetraro 1997. Editors: P. de Bartolomeis, B. Dubrovin, C. Reina (2002)
Vol. 1777: E. García-Río, D. N. Kupeli, R. Vázquez-Lorenzo, Osserman Manifolds in Semi-Riemannian Geometry (2002)
Vol. 1778: H. Kiechle, Theory of K-Loops (2002)
Vol. 1779: I. Chueshov, Monotone Random Systems (2002)
Vol. 1780: J. H. Bruinier, Borcherds Products on O(2,1) and Chern Classes of Heegner Divisors (2002)
Vol. 1781: E. Bolthausen, E. Perkins, A. van der Vaart, Lectures on Probability Theory and Statistics. Ecole d' Eté de Probabilités de Saint-Flour XXIX-1999. Editor: P. Bernard (2002)
Vol. 1782: C.-H. Chu, A. T.-M. Lau, Harmonic Functions on Groups and Fourier Algebras (2002)
Vol. 1783: L. Grüne, Asymptotic Behavior of Dynamical and Control Systems under Perturbation and Discretization (2002)
Vol. 1784: L. H. Eliasson, S. B. Kuksin, S. Marmi, J.-C. Yoccoz, Dynamical Systems and Small Divisors. Cetraro, Italy 1998. Editors: S. Marmi, J.-C. Yoccoz (2002)
Vol. 1785: J. Arias de Reyna, Pointwise Convergence of Fourier Series (2002)
Vol. 1786: S. D. Cutkosky, Monomialization of Morphisms from 3-Folds to Surfaces (2002)
Vol. 1787: S. Caenepeel, G. Militaru, S. Zhu, Frobenius and Separable Functors for Generalized Module Categories and Nonlinear Equations (2002)
Vol. 1788: A. Vasil'ev, Moduli of Families of Curves for Conformal and Quasiconformal Mappings (2002)
Vol. 1789: Y. Sommerhäuser, Yetter-Drinfel'd Hopf algebras over groups of prime order (2002)
Vol. 1790: X. Zhan, Matrix Inequalities (2002)
Vol. 1791: M. Knebusch, D. Zhang, Manis Valuations and Prüfer Extensions I: A new Chapter in Commutative Algebra (2002)
Vol. 1792: D. D. Ang, R. Gorenflo, V. K. Le, D. D. Trong, Moment Theory and Some Inverse Problems in Potential Theory and Heat Conduction (2002)
Vol. 1793: J. Cortés Monforte, Geometric, Control and Numerical Aspects of Nonholonomic Systems (2002)
Vol. 1794: N. Pytheas Fogg, Substitution in Dynamics, Arithmetics and Combinatorics. Editors: V. Berthé, S. Ferenczi, C. Mauduit, A. Siegel (2002)
Vol. 1795: H. Li, Filtered-Graded Transfer in Using Noncommutative Gröbner Bases (2002)
Vol. 1796: J.M. Melenk, hp-Finite Element Methods for Singular Perturbations (2002)
Vol. 1797: B. Schmidt, Characters and Cyclotomic Fields in Finite Geometry (2002)
Vol. 1798: W.M. Oliva, Geometric Mechanics (2002)
Vol. 1799: H. Pajot, Analytic Capacity, Rectifiability, Menger Curvature and the Cauchy Integral (2002)
Vol. 1800: O. Gabber, L. Ramero, Almost Ring Theory (2003)
Vol. 1801: J. Azéma, M. Émery, M. Ledoux, M. Yor (Eds.), Séminaire de Probabilités XXXVI (2003)
Vol. 1802: V. Capasso, E. Merzbach, B. G. Ivanoff, M. Dozzi, R. Dalang, T. Mountford, Topics in Spatial Stochastic Processes. Martina Franca, Italy 2001. Editor: E. Merzbach (2003)
Vol. 1803: G. Dolzmann, Variational Methods for Crystalline Microstructure – Analysis and Computation (2003)
Vol. 1804: I. Cherednik, Ya. Markov, R. Howe, G. Lusztig, Iwahori-Hecke Algebras and their Representation Theory. Martina Franca, Italy 1999. Editors: V. Baldoni, D. Barbasch (2003)
Vol. 1805: F. Cao, Geometric Curve Evolution and Image Processing (2003)
Vol. 1806: H. Broer, I. Hoveijn. G. Lunther, G. Vegter, Bifurcations in Hamiltonian Systems. Computing Singularities by Gröbner Bases (2003)
Vol. 1807: V. D. Milman, G. Schechtman (Eds.), Geometric Aspects of Functional Analysis. Israel Seminar 2000-2002 (2003)
Vol. 1808: W. Schindler, Measures with Symmetry Properties (2003)
Vol. 1809: O. Steinbach, Stability Estimates for Hybrid Coupled Domain Decomposition Methods (2003)
Vol. 1810: J. Wengenroth, Derived Functors in Functional Analysis (2003)
Vol. 1811: J. Stevens, Deformations of Singularities (2003)

Vol. 1812: L. Ambrosio, K. Deckelnick, G. Dziuk, M. Mimura, V. A. Solonnikov, H. M. Soner, Mathematical Aspects of Evolving Interfaces. Madeira, Funchal, Portugal 2000. Editors: P. Colli, J. F. Rodrigues (2003)
Vol. 1813: L. Ambrosio, L. A. Caffarelli, Y. Brenier, G. Buttazzo, C. Villani, Optimal Transportation and its Applications. Martina Franca, Italy 2001. Editors: L. A. Caffarelli, S. Salsa (2003)
Vol. 1814: P. Bank, F. Baudoin, H. Föllmer, L.C.G. Rogers, M. Soner, N. Touzi, Paris-Princeton Lectures on Mathematical Finance 2002 (2003)
Vol. 1815: A. M. Vershik (Ed.), Asymptotic Combinatorics with Applications to Mathematical Physics. St. Petersburg, Russia 2001 (2003)
Vol. 1816: S. Albeverio, W. Schachermayer, M. Talagrand, Lectures on Probability Theory and Statistics. Ecole d'Eté de Probabilités de Saint-Flour XXX-2000. Editor: P. Bernard (2003)
Vol. 1817: E. Koelink, W. Van Assche (Eds.), Orthogonal Polynomials and Special Functions. Leuven 2002 (2003)
Vol. 1818: M. Bildhauer, Convex Variational Problems with Linear, nearly Linear and/or Anisotropic Growth Conditions (2003)
Vol. 1819: D. Masser, Yu. V. Nesterenko, H. P. Schlickewei, W. M. Schmidt, M. Waldschmidt, Diophantine Approximation. Cetraro, Italy 2000. Editors: F. Amoroso, U. Zannier (2003)
Vol. 1820: F. Hiai, H. Kosaki, Means of Hilbert Space Operators (2003)
Vol. 1821: S. Teufel, Adiabatic Perturbation Theory in Quantum Dynamics (2003)
Vol. 1822: S.-N. Chow, R. Conti, R. Johnson, J. Mallet-Paret, R. Nussbaum, Dynamical Systems. Cetraro, Italy 2000. Editors: J. W. Macki, P. Zecca (2003)
Vol. 1823: A. M. Anile, W. Allegretto, C. Ringhofer, Mathematical Problems in Semiconductor Physics. Cetraro, Italy 1998. Editor: A. M. Anile (2003)
Vol. 1824: J. A. Navarro González, J. B. Sancho de Salas, $\mathscr{C}^\infty$ – Differentiable Spaces (2003)
Vol. 1825: J. H. Bramble, A. Cohen, W. Dahmen, Multiscale Problems and Methods in Numerical Simulations, Martina Franca, Italy 2001. Editor: C. Canuto (2003)
Vol. 1826: K. Dohmen, Improved Bonferroni Inequalities via Abstract Tubes. Inequalities and Identities of Inclusion-Exclusion Type. VIII, 113 p, 2003.
Vol. 1827: K. M. Pilgrim, Combinations of Complex Dynamical Systems. IX, 118 p, 2003.
Vol. 1828: D. J. Green, Gröbner Bases and the Computation of Group Cohomology. XII, 138 p, 2003.
Vol. 1829: E. Altman, B. Gaujal, A. Hordijk, Discrete-Event Control of Stochastic Networks: Multimodularity and Regularity. XIV, 313 p, 2003.
Vol. 1830: M. I. Gil', Operator Functions and Localization of Spectra. XIV, 256 p, 2003.
Vol. 1831: A. Connes, J. Cuntz, E. Guentner, N. Higson, J. E. Kaminker, Noncommutative Geometry, Martina Franca, Italy 2002. Editors: S. Doplicher, L. Longo (2004)
Vol. 1832: J. Azéma, M. Émery, M. Ledoux, M. Yor (Eds.), Séminaire de Probabilités XXXVII (2003)
Vol. 1833: D.-Q. Jiang, M. Qian, M.-P. Qian, Mathematical Theory of Nonequilibrium Steady States. On the Frontier of Probability and Dynamical Systems. IX, 280 p, 2004.
Vol. 1834: Yo. Yomdin, G. Comte, Tame Geometry with Application in Smooth Analysis. VIII, 186 p, 2004.
Vol. 1835: O.T. Izhboldin, B. Kahn, N.A. Karpenko, A. Vishik, Geometric Methods in the Algebraic Theory of Quadratic Forms. Summer School, Lens, 2000. Editor: J.-P. Tignol (2004)
Vol. 1836: C. Năstăsescu, F. Van Oystaeyen, Methods of Graded Rings. XIII, 304 p, 2004.
Vol. 1837: S. Tavaré, O. Zeitouni, Lectures on Probability Theory and Statistics. Ecole d'Eté de Probabilités de Saint-Flour XXXI-2001. Editor: J. Picard (2004)
Vol. 1838: A.J. Ganesh, N.W. O'Connell, D.J. Wischik, Big Queues. XII, 254 p, 2004.
Vol. 1839: R. Gohm, Noncommutative Stationary Processes. VIII, 170 p, 2004.
Vol. 1840: B. Tsirelson, W. Werner, Lectures on Probability Theory and Statistics. Ecole d'Eté de Probabilités de Saint-Flour XXXII-2002. Editor: J. Picard (2004)
Vol. 1841: W. Reichel, Uniqueness Theorems for Variational Problems by the Method of Transformation Groups (2004)
Vol. 1842: T. Johnsen, A. L. Knutsen, K_3 Projective Models in Scrolls (2004)
Vol. 1843: B. Jefferies, Spectral Properties of Noncommuting Operators (2004)
Vol. 1844: K.F. Siburg, The Principle of Least Action in Geometry and Dynamics (2004)
Vol. 1845: Min Ho Lee, Mixed Automorphic Forms, Torus Bundles, and Jacobi Forms (2004)
Vol. 1846: H. Ammari, H. Kang, Reconstruction of Small Inhomogeneities from Boundary Measurements (2004)
Vol. 1847: T.R. Bielecki, T. Björk, M. Jeanblanc, M. Rutkowski, J.A. Scheinkman, W. Xiong, Paris-Princeton Lectures on Mathematical Finance 2003 (2004)
Vol. 1848: M. Abate, J. E. Fornaess, X. Huang, J. P. Rosay, A. Tumanov, Real Methods in Complex and CR Geometry, Martina Franca, Italy 2002. Editors: D. Zaitsev, G. Zampieri (2004)
Vol. 1849: Martin L. Brown, Heegner Modules and Elliptic Curves (2004)
Vol. 1850: V. D. Milman, G. Schechtman (Eds.), Geometric Aspects of Functional Analysis. Israel Seminar 2002-2003 (2004)
Vol. 1851: O. Catoni, Statistical Learning Theory and Stochastic Optimization (2004)
Vol. 1852: A.S. Kechris, B.D. Miller, Topics in Orbit Equivalence (2004)
Vol. 1853: Ch. Favre, M. Jonsson, The Valuative Tree (2004)
Vol. 1854: O. Saeki, Topology of Singular Fibers of Differential Maps (2004)
Vol. 1855: G. Da Prato, P.C. Kunstmann, I. Lasiecka, A. Lunardi, R. Schnaubelt, L. Weis, Functional Analytic Methods for Evolution Equations. Editors: M. Iannelli, R. Nagel, S. Piazzera (2004)
Vol. 1856: K. Back, T.R. Bielecki, C. Hipp, S. Peng, W. Schachermayer, Stochastic Methods in Finance, Bressanone/Brixen, Italy, 2003. Editors: M. Fritelli, W. Runggaldier (2004)
Vol. 1857: M. Émery, M. Ledoux, M. Yor (Eds.), Séminaire de Probabilités XXXVIII (2005)
Vol. 1858: A.S. Cherny, H.-J. Engelbert, Singular Stochastic Differential Equations (2005)
Vol. 1859: E. Letellier, Fourier Transforms of Invariant Functions on Finite Reductive Lie Algebras (2005)
Vol. 1860: A. Borisyuk, G.B. Ermentrout, A. Friedman, D. Terman, Tutorials in Mathematical Biosciences I. Mathematical Neurosciences (2005)

Vol. 1861: G. Benettin, J. Henrard, S. Kuksin, Hamiltonian Dynamics – Theory and Applications, Cetraro, Italy, 1999. Editor: A. Giorgilli (2005)
Vol. 1862: B. Helffer, F. Nier, Hypoelliptic Estimates and Spectral Theory for Fokker-Planck Operators and Witten Laplacians (2005)
Vol. 1863: H. Führ, Abstract Harmonic Analysis of Continuous Wavelet Transforms (2005)
Vol. 1864: K. Efstathiou, Metamorphoses of Hamiltonian Systems with Symmetries (2005)
Vol. 1865: D. Applebaum, B.V. R. Bhat, J. Kustermans, J. M. Lindsay, Quantum Independent Increment Processes I. From Classical Probability to Quantum Stochastic Calculus. Editors: M. Schürmann, U. Franz (2005)
Vol. 1866: O.E. Barndorff-Nielsen, U. Franz, R. Gohm, B. Kümmerer, S. Thorbjønsen, Quantum Independent Increment Processes II. Structure of Quantum Lévy Processes, Classical Probability, and Physics. Editors: M. Schürmann, U. Franz, (2005)
Vol. 1867: J. Sneyd (Ed.), Tutorials in Mathematical Biosciences II. Mathematical Modeling of Calcium Dynamics and Signal Transduction. (2005)
Vol. 1868: J. Jorgenson, S. Lang, $\mathrm{Pos}_n(\mathrm{R})$ and Eisenstein Series. (2005)
Vol. 1869: A. Dembo, T. Funaki, Lectures on Probability Theory and Statistics. Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003. Editor: J. Picard (2005)
Vol. 1870: V.I. Gurariy, W. Lusky, Geometry of Müntz Spaces and Related Questions. (2005)
Vol. 1871: P. Constantin, G. Gallavotti, A.V. Kazhikhov, Y. Meyer, S. Ukai, Mathematical Foundation of Turbulent Viscous Flows, Martina Franca, Italy, 2003. Editors: M. Cannone, T. Miyakawa (2006)
Vol. 1872: A. Friedman (Ed.), Tutorials in Mathematical Biosciences III. Cell Cycle, Proliferation, and Cancer (2006)
Vol. 1873: R. Mansuy, M. Yor, Random Times and Enlargements of Filtrations in a Brownian Setting (2006)
Vol. 1874: M. Yor, M. Émery (Eds.), In Memoriam Paul-André Meyer - Séminaire de Probabilités XXXIX (2006)
Vol. 1875: J. Pitman, Combinatorial Stochastic Processes. Ecole d'Eté de Probabilités de Saint-Flour XXXII-2002. Editor: J. Picard (2006)
Vol. 1876: H. Herrlich, Axiom of Choice (2006)
Vol. 1877: J. Steuding, Value Distributions of *L*-Functions (2007)
Vol. 1878: R. Cerf, The Wulff Crystal in Ising and Percolation Models, Ecole d'Eté de Probabilités de Saint-Flour XXXIV-2004. Editor: Jean Picard (2006)
Vol. 1879: G. Slade, The Lace Expansion and its Applications, Ecole d'Eté de Probabilités de Saint-Flour XXXIV-2004. Editor: Jean Picard (2006)
Vol. 1880: S. Attal, A. Joye, C.-A. Pillet, Open Quantum Systems I, The Hamiltonian Approach (2006)
Vol. 1881: S. Attal, A. Joye, C.-A. Pillet, Open Quantum Systems II, The Markovian Approach (2006)
Vol. 1882: S. Attal, A. Joye, C.-A. Pillet, Open Quantum Systems III, Recent Developments (2006)
Vol. 1883: W. Van Assche, F. Marcellàn (Eds.), Orthogonal Polynomials and Special Functions, Computation and Application (2006)
Vol. 1884: N. Hayashi, E.I. Kaikina, P.I. Naumkin, I.A. Shishmarev, Asymptotics for Dissipative Nonlinear Equations (2006)
Vol. 1885: A. Telcs, The Art of Random Walks (2006)
Vol. 1886: S. Takamura, Splitting Deformations of Degenerations of Complex Curves (2006)
Vol. 1887: K. Habermann, L. Habermann, Introduction to Symplectic Dirac Operators (2006)
Vol. 1888: J. van der Hoeven, Transseries and Real Differential Algebra (2006)
Vol. 1889: G. Osipenko, Dynamical Systems, Graphs, and Algorithms (2006)
Vol. 1890: M. Bunge, J. Funk, Singular Coverings of Toposes (2006)
Vol. 1891: J.B. Friedlander, D.R. Heath-Brown, H. Iwaniec, J. Kaczorowski, Analytic Number Theory, Cetraro, Italy, 2002. Editors: A. Perelli, C. Viola (2006)
Vol. 1892: A. Baddeley, I. Bárány, R. Schneider, W. Weil, Stochastic Geometry, Martina Franca, Italy, 2004. Editor: W. Weil (2007)
Vol. 1893: H. Hanßmann, Local and Semi-Local Bifurcations in Hamiltonian Dynamical Systems, Results and Examples (2007)
Vol. 1894: C.W. Groetsch, Stable Approximate Evaluation of Unbounded Operators (2007)
Vol. 1895: L. Molnár, Selected Preserver Problems on Algebraic Structures of Linear Operators and on Function Spaces (2007)
Vol. 1896: P. Massart, Concentration Inequalities and Model Selection, Ecole d'Été de Probabilités de Saint-Flour XXXIII-2003. Editor: J. Picard (2007)
Vol. 1897: R. Doney, Fluctuation Theory for Lévy Processes, Ecole d'Été de Probabilités de Saint-Flour XXXV-2005. Editor: J. Picard (2007)
Vol. 1898: H.R. Beyer, Beyond Partial Differential Equations, On linear and Quasi-Linear Abstract Hyperbolic Evolution Equations (2007)
Vol. 1899: Séminaire de Probabilités XL. Editors: C. Donati-Martin, M. Émery, A. Rouault, C. Stricker (2007)
Vol. 1900: E. Bolthausen, A. Bovier (Eds.), Spin Glasses (2007)
Vol. 1901: O. Wittenberg, Intersections de deux quadriques et pinceaux de courbes de genre 1, Intersections of Two Quadrics and Pencils of Curves of Genus 1 (2007)
Vol. 1902: A. Isaev, Lectures on the Automorphism Groups of Kobayashi-Hyperbolic Manifolds (2007)
Vol. 1903: G. Kresin, V. Maz'ya, Sharp Real-Part Theorems (2007)
Vol. 1904: P. Giesl, Construction of Global Lyapunov Functions Using Radial Basis Functions (2007)
Vol. 1905: C. Prévôt, M. Röckner, A Concise Course on Stochastic Partial Differential Equations (2007)
Vol. 1906: T. Schuster, The Method of Approximate Inverse: Theory and Applications (2007)
Vol. 1907: M. Rasmussen, Attractivity and Bifurcation for Nonautonomous Dynamical Systems (2007)
Vol. 1908: T.J. Lyons, M. Caruana, T. Lévy, Differential Equations Driven by Rough Paths, Ecole d'Été de Probabilités de Saint-Flour XXXIV-2004 (2007)
Vol. 1909: H. Akiyoshi, M. Sakuma, M. Wada, Y. Yamashita, Punctured Torus Groups and 2-Bridge Knot Groups (I) (2007)
Vol. 1910: V.D. Milman, G. Schechtman (Eds.), Geometric Aspects of Functional Analysis. Israel Seminar 2004-2005 (2007)
Vol. 1911: A. Bressan, D. Serre, M. Williams, K. Zumbrun, Hyperbolic Systems of Balance Laws. Cetraro, Italy 2003. Editor: P. Marcati (2007)

Vol. 1912: V. Berinde, Iterative Approximation of Fixed Points (2007)
Vol. 1913: J.E. Marsden, G. Misiołek, J.-P. Ortega, M. Perlmutter, T.S. Ratiu, Hamiltonian Reduction by Stages (2007)
Vol. 1914: G. Kutyniok, Affine Density in Wavelet Analysis (2007)
Vol. 1915: T. Bıyıkoğlu, J. Leydold, P.F. Stadler, Laplacian Eigenvectors of Graphs. Perron-Frobenius and Faber-Krahn Type Theorems (2007)
Vol. 1916: C. Villani, F. Rezakhanlou, Entropy Methods for the Boltzmann Equation. Editors: F. Golse, S. Olla (2008)
Vol. 1917: I. Veselić, Existence and Regularity Properties of the Integrated Density of States of Random Schrödinger (2008)
Vol. 1918: B. Roberts, R. Schmidt, Local Newforms for GSp(4) (2007)
Vol. 1919: R.A. Carmona, I. Ekeland, A. Kohatsu-Higa, J.-M. Lasry, P.-L. Lions, H. Pham, E. Taflin, Paris-Princeton Lectures on Mathematical Finance 2004. Editors: R.A. Carmona, E. Çinlar, I. Ekeland, E. Jouini, J.A. Scheinkman, N. Touzi (2007)
Vol. 1920: S.N. Evans, Probability and Real Trees. Ecole d'Été de Probabilités de Saint-Flour XXXV-2005 (2008)
Vol. 1921: J.P. Tian, Evolution Algebras and their Applications (2008)
Vol. 1922: A. Friedman (Ed.), Tutorials in Mathematical BioSciences IV. Evolution and Ecology (2008)
Vol. 1923: J.P.N. Bishwal, Parameter Estimation in Stochastic Differential Equations (2008)
Vol. 1924: M. Wilson, Littlewood-Paley Theory and Exponential-Square Integrability (2008)
Vol. 1925: M. du Sautoy, L. Woodward, Zeta Functions of Groups and Rings (2008)
Vol. 1926: L. Barreira, V. Claudia, Stability of Nonautonomous Differential Equations (2008)
Vol. 1927: L. Ambrosio, L. Caffarelli, M.G. Crandall, L.C. Evans, N. Fusco, Calculus of Variations and Non-Linear Partial Differential Equations. Cetraro, Italy 2005. Editors: B. Dacorogna, P. Marcellini (2008)
Vol. 1928: J. Jonsson, Simplicial Complexes of Graphs (2008)
Vol. 1929: Y. Mishura, Stochastic Calculus for Fractional Brownian Motion and Related Processes (2008)
Vol. 1930: J.M. Urbano, The Method of Intrinsic Scaling. A Systematic Approach to Regularity for Degenerate and Singular PDEs (2008)
Vol. 1931: M. Cowling, E. Frenkel, M. Kashiwara, A. Valette, D.A. Vogan, Jr., N.R. Wallach, Representation Theory and Complex Analysis. Venice, Italy 2004. Editors: E.C. Tarabusi, A. D'Agnolo, M. Picardello (2008)
Vol. 1932: A.A. Agrachev, A.S. Morse, E.D. Sontag, H.J. Sussmann, V.I. Utkin, Nonlinear and Optimal Control Theory. Cetraro, Italy 2004. Editors: P. Nistri, G. Stefani (2008)
Vol. 1933: M. Petkovic, Point Estimation of Root Finding Methods (2008)
Vol. 1934: C. Donati-Martin, M. Émery, A. Rouault, C. Stricker (Eds.), Séminaire de Probabilités XLI (2008)
Vol. 1935: A. Unterberger, Alternative Pseudodifferential Analysis (2008)
Vol. 1936: P. Magal, S. Ruan (Eds.), Structured Population Models in Biology and Epidemiology (2008)
Vol. 1937: G. Capriz, P. Giovine, P.M. Mariano (Eds.), Mathematical Models of Granular Matter (2008)
Vol. 1938: D. Auroux, F. Catanese, M. Manetti, P. Seidel, B. Siebert, I. Smith, G. Tian, Symplectic 4-Manifolds and Algebraic Surfaces. Cetraro, Italy 2003. Editors: F. Catanese, G. Tian (2008)
Vol. 1939: D. Boffi, F. Brezzi, L. Demkowicz, R.G. Durán, R.S. Falk, M. Fortin, Mixed Finite Elements, Compatibility Conditions, and Applications. Cetraro, Italy 2006. Editors: D. Boffi, L. Gastaldi (2008)
Vol. 1940: J. Banasiak, V. Capasso, M.A.J. Chaplain, M. Lachowicz, J. Miękisz, Multiscale Problems in the Life Sciences. From Microscopic to Macroscopic. Będlewo, Poland 2006. Editors: V. Capasso, M. Lachowicz (2008)
Vol. 1941: S.M.J. Haran, Arithmetical Investigations. Representation Theory, Orthogonal Polynomials, and Quantum Interpolations (2008)
Vol. 1942: S. Albeverio, F. Flandoli, Y.G. Sinai, SPDE in Hydrodynamic. Recent Progress and Prospects. Cetraro, Italy 2005. Editors: G. Da Prato, M. Röckner (2008)
Vol. 1943: L.L. Bonilla (Ed.), Inverse Problems and Imaging. Martina Franca, Italy 2002 (2008)
Vol. 1944: A. Di Bartolo, G. Falcone, P. Plaumann, K. Strambach, Algebraic Groups and Lie Groups with Few Factors (2008)
Vol. 1945: F. Brauer, P. van den Driessche, J. Wu (Eds.), Mathematical Epidemiology (2008)
Vol. 1946: G. Allaire, A. Arnold, P. Degond, T.Y. Hou, Quantum Transport. Modelling, Analysis and Asymptotics. Cetraro, Italy 2006. Editors: N.B. Abdallah, G. Frosali (2008)
Vol. 1947: D. Abramovich, M. Mariño, M. Thaddeus, R. Vakil, Enumerative Invariants in Algebraic Geometry and String Theory. Cetraro, Italy 2005. Editors: K. Behrend, M. Manetti (2008)
Vol. 1948: F. Cao, J-L. Lisani, J-M. Morel, P. Musé, F. Sur, A Theory of Shape Identification (2008)
Vol. 1949: H.G. Feichtinger, B. Helffer, M.P. Lamoureux, N. Lerner, J. Toft, Pseudo-differential Operators. Quantization and Signals. Cetraro, Italy 2006. Editors: L. Rodino, M.W. Wong (2008)
Vol. 1950: M. Bramson, Stability of Queueing Networks, Ecole d' Eté de Probabilités de Saint-Flour XXXVI-2006 (2008)

Recent Reprints and New Editions

Vol. 1702: J. Ma, J. Yong, Forward-Backward Stochastic Differential Equations and their Applications. 1999 – Corr. 3rd printing (2007)
Vol. 830: J.A. Green, Polynomial Representations of GL_n, with an Appendix on Schensted Correspondence and Littelmann Paths by K. Erdmann, J.A. Green and M. Schoker 1980 – 2nd corr. and augmented edition (2007)
Vol. 1693: S. Simons, From Hahn-Banach to Monotonicity (Minimax and Monotonicity 1998) – 2nd exp. edition (2008)
Vol. 470: R.E. Bowen, Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms. With a preface by D. Ruelle. Edited by J.-R. Chazottes. 1975 – 2nd rev. edition (2008)
Vol. 523: S.A. Albeverio, R.J. Høegh-Krohn, S. Mazzucchi, Mathematical Theory of Feynman Path Integral. 1976 – 2nd corr. and enlarged edition (2008)
Vol. 1764: A. Cannas da Silva, Lectures on Symplectic Geometry 2001 – Corr. 2nd printing (2008)

LECTURE NOTES IN MATHEMATICS

Editorial Policy (for the publication of monographs)

1. Lecture Notes aim to report new developments in all areas of mathematics and their applications - quickly, informally and at a high level. Mathematical texts analysing new developments in modelling and numerical simulation are welcome.

   Monograph manuscripts should be reasonably self-contained and rounded off. Thus they may, and often will, present not only results of the author but also related work by other people. They may be based on specialised lecture courses. Furthermore, the manuscripts should provide sufficient motivation, examples and applications. This clearly distinguishes Lecture Notes from journal articles or technical reports which normally are very concise. Articles intended for a journal but too long to be accepted by most journals, usually do not have this "lecture notes" character. For similar reasons it is unusual for doctoral theses to be accepted for the Lecture Notes series, though habilitation theses may be appropriate.

2. Manuscripts should be submitted (preferably in duplicate) either to Springer's mathematics editorial in Heidelberg, or to one of the series editors (with a copy to Springer). In general, manuscripts will be sent out to 2 external referees for evaluation. If a decision cannot yet be reached on the basis of the first 2 reports, further referees may be contacted: The author will be informed of this. A final decision to publish can be made only on the basis of the complete manuscript, however a refereeing process leading to a preliminary decision can be based on a pre-final or incomplete manuscript. The strict minimum amount of material that will be considered should include a detailed outline describing the planned contents of each chapter, a bibliography and several sample chapters.

   Authors should be aware that incomplete or insufficiently close to final manuscripts almost always result in longer refereeing times and nevertheless unclear referees' recommendations, making further refereeing of a final draft necessary.

   Authors should also be aware that parallel submission of their manuscript to another publisher while under consideration for LNM will in general lead to immediate rejection.

3. Manuscripts should in general be submitted in English. Final manuscripts should contain at least 100 pages of mathematical text and should always include

   – a table of contents;
   – an informative introduction, with adequate motivation and perhaps some historical remarks: it should be accessible to a reader not intimately familiar with the topic treated;
   – a subject index: as a rule this is genuinely helpful for the reader.

   For evaluation purposes, manuscripts may be submitted in print or electronic form (print form is still preferred by most referees), in the latter case preferably as pdf- or zipped ps-files. Lecture Notes volumes are, as a rule, printed digitally from the authors' files. To ensure best results, authors are asked to use the LaTeX2e style files available from Springer's web-server at:

   ftp://ftp.springer.de/pub/tex/latex/svmonot1/ (for monographs) and
   ftp://ftp.springer.de/pub/tex/latex/svmultt1/ (for summer schools/tutorials).

Additional technical instructions, if necessary, are available on request from: lnm@springer.com.

4. Careful preparation of the manuscripts will help keep production time short besides ensuring satisfactory appearance of the finished book in print and online. After acceptance of the manuscript authors will be asked to prepare the final LaTeX source files (and also the corresponding dvi-, pdf- or zipped ps-file) together with the final printout made from these files. The LaTeX source files are essential for producing the full-text online version of the book (see http://www.springerlink.com/content/110312/ for the existing online volumes of LNM).

   The actual production of a Lecture Notes volume takes approximately 12 weeks.

5. Authors receive a total of 50 free copies of their volume, but no royalties. They are entitled to a discount of 33.3% on the price of Springer books purchased for their personal use, if ordering directly from Springer.

6. Commitment to publish is made by letter of intent rather than by signing a formal contract. Springer-Verlag secures the copyright for each volume. Authors are free to reuse material contained in their LNM volumes in later publications: a brief written (or e-mail) request for formal permission is sufficient.

Addresses:

Professor J.-M. Morel, CMLA,
École Normale Supérieure de Cachan,
61 Avenue du Président Wilson, 94235 Cachan Cedex, France
E-mail: Jean-Michel.Morel@cmla.ens-cachan.fr

Professor F. Takens, Mathematisch Instituut,
Rijksuniversiteit Groningen, Postbus 800,
9700 AV Groningen, The Netherlands
E-mail: F.Takens@math.rug.nl

Professor B. Teissier, Institut Mathématique de Jussieu,
UMR 7586 du CNRS, Équipe "Géométrie et Dynamique",
175 rue du Chevaleret
75013 Paris, France
E-mail: teissier@math.jussieu.fr

For the "Mathematical Biosciences Subseries" of LNM:

Professor P.K. Maini, Center for Mathematical Biology,
Mathematical Institute, 24-29 St Giles,
Oxford OX1 3LP, UK
E-mail: maini@maths.ox.ac.uk

Springer, Mathematics Editorial I, Tiergartenstr. 17
69121 Heidelberg, Germany,
Tel.: +49 (6221) 487-8410
Fax: +49 (6221) 487-8355
E-mail: lnm@springer.com