Conservation Laws in Polling Systems

Behoudswetten in polling systemen

(met een samenvatting in het Nederlands)

Proefschrift ter verkrijging van de graad van doctor
aan de Rijksuniversiteit te Utrecht
op gezag van de Rector Magnificus, Prof. Dr. J.A. van Ginkel
ingevolge het besluit van het College van Dekanen
in het openbaar te verdedigen
op 18 januari des namiddags te 16.15 uur

door

Wilhelmus Petrus Groenendijk

geboren op 10 december 1961, te Utrecht

Promotoren: Prof. dr. ir. J.W. Cohen,
Faculteit der Wiskunde en Informatica

Prof. dr. ir. O.J. Boxma,
Faculteit der Economische Wetenschappen,
Katholieke Universiteit Brabant

Dankwoord

Het verschijnen van dit proefschrift biedt mij de welkome gelegenheid dank te zeggen aan allen die op enigerlei wijze aan de totstandkoming hiervan hebben bijgedragen.

Gaarne bedank ik Prof. dr. ir. O.J. Boxma voor de stimulerende, intensieve begeleiding die ik van hem mocht ontvangen. De samenwerking met hem heb ik als een bijzonder voorrecht ervaren. Prof. dr. ir. J.W. Cohen dank ik in het bijzonder voor de vele gesprekken, waarin hij steeds weer nieuwe gezichtspunten naar voren wist te brengen.

Mijn positie als wetenschappelijk medewerker op het Centrum voor Wiskunde en Informatica (CWI) te Amsterdam is voor mij een ideale plaats gebleken om het onderzoek voor dit proefschrift te verrichten. Ik ben mij ervan bewust dat dit boekje niet geschreven zou zijn zonder de grote mate van vrijheid die mij op het CWI is geboden.

Wat betreft de technische afwerking van het proefschrift bedank ik ten eerste R.T. Baanders voor de tekening van zijn hand op pagina 1 van dit proefschrift. Voorts bedank ik Wim Aspers: bij alle moeilijkheden rond de afdruk-apparatuur bleef hij steeds een rustig aanspreekpunt waar ik met mijn problemen terecht kon.

Mijn kamergenoot, Hans van den Berg, bedank ik voor de vele intensieve discussies die wij gehad hebben. Rob van den Berg bedank ik voor enkele adviezen omtrent een correcte formulering van Stelling 2.2 in Hoofdstuk 2.

I thank Dr. Hanoch Levy from Tel-Aviv University, Israel, for the opportunity to work with him. His enthusiasm and great expertise made this a very special experience. Special thanks go to Dr. Y.T. Wang for inviting me to work at AT&T Bell Laboratories in Holmdel, New Jersey (USA). The two months I have spent there are unforgettable.

Ten slotte bedank ik mijn vrouw Marja voor alle steun die ik door de jaren heen van haar heb mogen ontvangen.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION AND OVERVIEW

1.1 Background

This study is devoted to queueing systems in which a single server serves several classes of customers, visiting the queues of these customer classes one at a time. Such single-server multiple-queue models are generally referred to as *polling models*. The characteristics of many practical systems can be represented by polling models. To obtain insight into the quality and efficiency of service of such systems mathematical methods from queueing theory are used to study these (stochastic) models.

One of the earliest applications of polling models which appeared in the open literature is the patrolling repairman problem. A repairman consecutively inspects each of N machines to check whether a breakdown has occurred. If not, he moves to the next machine. If a breakdown has occurred, he repairs the machine. The repair takes a certain *repair time*. After this time, the repairman moves to the next machine.

Polling models were used in the 1960s to analyze traffic signal control. Since most of the progress in the analysis of polling systems has been made only recently, it would be interesting to employ today's knowledge on polling models to analyze models of congestion at traffic lights (cf. also Figure 1.1).

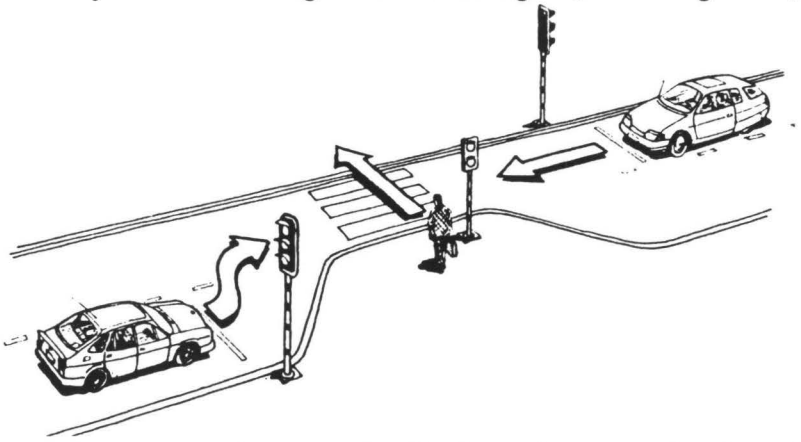

Figure 1.1 Traffic congestion on a signalized crossing: example of a polling system with three queues

Another example of where a polling model can be used is that of a bus serving a number of bus stops along a closed tour (cf. Figure 1.2). At each bus stop the bus picks up passengers who embark the bus one at a time and buy a ticket. If all passengers have entered, the bus drives on to the next stop. In the terminology of queueing theory the bus is the server, the bus stops are represented by the queues and the passengers by the customers. The server subsequently 'polls' each queue to see if there are any waiting customers, serving them before moving to the next queue.

These examples show that polling models may be encountered in very different situations.

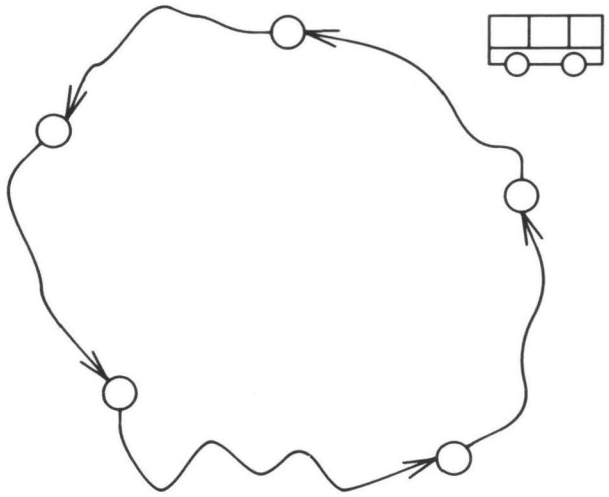

Figure 1.2 A bus serving a number of bus stops along a closed route

The advent of computers, computer-communication networks and digital communication opened up a rich area of new applications and strongly enhanced the interest in polling models. A first application in this area was in the analysis of a polling scheme for a computer with multidrop terminals. In such a system several terminals are connected to a central computer. The computer 'polls' the terminals in a cyclic order to check whether they have anything to send. A statistical multiplexer is another example: several streams of messages are merged into one single stream by giving each stream in turn permission to use the common channel for a certain time.

Nowadays, the prime application for polling models is in the performance evaluation of a variety of demand-based multiple-access schemes in computer and communication systems. Although the analysis in this study can be applied to any of the systems mentioned above, our main motivation has been the performance analysis of Local Area Networks (LAN's) employing some form of token passing. In a token-passing LAN, several stations (terminals,

file servers, hosts, gateways, etc.) are connected to a common transmission medium in a ring or bus topology. A special bit sequence called the *token* is passed from one station to the next; a station that 'possesses the token' is allowed to transmit messages. After completion of his transmission the station releases the token, giving the next station in turn an opportunity to transmit.

In the next section a description of the general queueing model is given and the mathematical problem is stated. Section 1.3 discusses some important assumptions that are assumed to be valid throughout this study. In Section 1.4 some related literature is discussed and the place of this study within the context of present-day literature is indicated. Finally, in Section 1.5 an overview is presented of results appearing in the next seven chapters.

1.2 Description of the queueing model and statement of the mathematical problem

1.2.1 Model description

In this subsection we give a brief discussion of the central model of this study. Details and variants will be presented in later chapters.

A single server S serves N classes of arrival streams of customers, or rather N queues $Q_1, \ldots, Q_N$ with infinite waiting rooms (cf. Figure 1.3). Customers arriving at Q_j are referred to as class-j customers. Arrival epochs of customers (or possibly customer batches) occur according to a Poisson process.

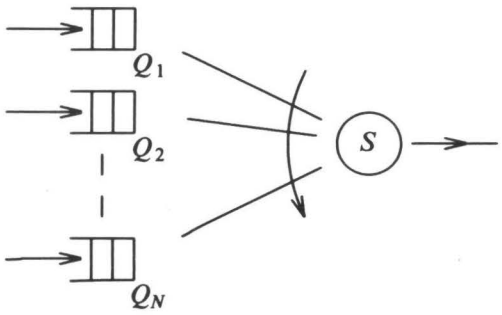

Figure 1.3 Queueing model of a polling system

The server visits the queues in a fixed cyclic order, i.e., in the order of their indices: $Q_1, Q_2, \ldots, Q_N, Q_1, Q_2, \ldots$. Service requirements of class-j customers are independent, identically distributed stochastic variables. Overhead is imposed on the system in the form of switch-over times between queues. The switch-over times of the server between successive queues are independent, identically distributed stochastic variables. The interarrival process, the service process and the switch-over time processes are assumed to be mutually independent processes.

1.2.2 Statement of the mathematical problem

Consider a queueing system in which a single server serves at constant rate. The scheduling discipline for the system is the procedure for deciding which customer should be in service at any time. Suppose the scheduling discipline is *work conserving*, i.e., it does not allow the server to be idle when work is present and does not alter the service requirements of customers or influence the arrival times of customers. Comparing the sample paths of the workload process for such a system under different scheduling disciplines leads us to the observation that the workload process is independent of the scheduling discipline. This is called the principle of 'work conservation'. The principle of work conservation has in the past proven to be very useful in the analysis of queueing systems with a non-FCFS (First Come First Served) service discipline and zero switch-over times. If the scheduling discipline for the system is work conserving this implies in particular that the amount of work present should not depend on the order of service - and hence should equal the amount of work in the 'corresponding' system with FCFS service discipline. If, moreover, the queueing discipline selects customers in a way that is independent of (any measure of) the required service time, then the distribution of the number of customers in the system is also independent of the order of service. But even if this is not the case, as in systems with different service requirements for different classes of customers, or in systems with different priorities for different classes of customers, the principle of work conservation yields an *exact* expression for an appropriate weighted sum of the mean queue lengths; hence (by using Little's formula) a weighted sum of mean waiting times of customers can be obtained: a *conservation law*. The conservation law implies that a certain weighted sum of mean waiting times will not change, no matter how sophisticated or elaborate we may choose our scheduling discipline. This justifies the word 'conservation'.

The conservation law is an important tool in the analysis of queueing systems with complicated scheduling mechanisms; often it provides the only exact information concerning mean waiting (sojourn) times that is available in the analysis of such systems. In particular it is often used for the derivation of bounds on waiting times for specific classes of customers and in devising approximations.

When overhead in the form of switch-over times between queues is introduced in the system, the principle of work conservation is violated in the sense that the server may be forced to work on other activities (switching) instead of serving the offered traffic. Comparing the sample paths of the workload for two such systems under different scheduling disciplines reveals that these sample paths will not in general be identical. Consequently, the amount of work in the system at some time t is not the same for different scheduling disciplines and the principle of work conservation is no longer satisfied; this fact clearly prohibits the formulation of a conservation law. However, it appears to be possible to formulate a so-called 'pseudoconservation law' for such systems. Like the 'ordinary' conservation law a pseudoconservation law is an exact expression for a weighted sum of the mean waiting times at the various

queues; however, this weighted sum now depends on the scheduling discipline. Its use is similar to that of the conservation law and it reduces to the conservation law when the switch-over times become zero. Therefore it seems appropriate to refer to such a law as to a '*pseudo*conservation law'. Denote the mean waiting time of class-i customers by $E\mathbf{W}_i$. The pseudoconservation law states that a weighted sum of the mean waiting times is equal to some function of the scheduling discipline and the system parameters:

$$\sum_{i=1}^{N} g_i E\mathbf{W}_i = f(\textit{scheduling discipline, system parameters}).$$

The discovery of the pseudoconservation law was a major step forward, since usually the analysis of polling systems is mathematically so complex that even mean waiting times are unknown. In view of this mathematical complexity the existence of this law was even more surprising.

Some special instances of the pseudoconservation law were obtained by Ferguson & Aminetzah [1985] and Watson [1985] as a by-product of their analysis. Unfortunately, the derivations of their results were lengthy and cumbersome and no satisfactory explanation for the occurrence of these relations was provided; as it turned out, more questions were raised than were answered.

In this study the following questions will be addressed:

- What is the reason for the occurrence of the pseudoconservation law? Does there exist an analogon of the principle of work conservation that holds for systems *with* switch-over times?
- Can a general framework be given which somehow generalizes and unifies existing results concerning this law?
- Does there exist a simpler derivation?
- Is there an interpretation for the various terms that occur in the pseudoconservation law?

It will be shown in this study that all these questions can be answered affirmatively. In addition, it will be shown how the pseudoconservation law can be used in deriving approximations for the mean waiting times at the various queues.

It is obvious that waiting times are important performance measures in networks of queues. Therefore, most of the analysis in the literature on polling models is concerned with waiting times. The majority of the studies considers cyclic polling, in which the server visits each station exactly once in each round. Various scheduling disciplines at the queues have been considered, ranging from *exhaustive service* (when the server visits a queue, he serves its customers until the queue has become empty) via *gated service* (when the server visits a queue he serves only those customers that were present upon his arrival

at the queue; customers arriving at the queue during his visit are served in the next round) to *1-limited service* (when the server visits a queue he serves only one customer, if present). It may be seen that, for systems with switch-over times, exhaustive service is very efficient. However, use of this policy can result in one heavily loaded queue monopolizing the server, degrading the performance at other queues and causing long waiting times there. This introduces the issue of *fairness*. 1-limited service is usually considered very fair w.r.t. different classes of customers, but not very efficient. The trade-off between efficiency and fairness is the main reason for introducing complicated scheduling disciplines.

1.3 Assumptions and conventions

Throughout this study the emphasis is on developing methodologies and methods which can be applied in the analysis of practical systems. Therefore, in a few places we have omitted details to enhance clarity of presentation. However, there are some points that require special care in their definition. This section is devoted to a detailed discussion of an assumption regarding the stochastic processes under consideration and of some of the conventions adopted in this study.

An important assumption is that we assume all considered systems to be in steady state. More precisely, we assume that the basic processes (i.e., the queue-length process, the workload process, etc.) are ergodic and stationary.[1]

For cyclic-service systems with at most two queues necessary and sufficient conditions for ergodicity of the basic processes are in most cases well known. For systems with more than two queues the formulation of these ergodicity conditions is in general a more complicated problem. For the cyclic-service systems under consideration we are usually only able to formulate *necessary* conditions for ergodicity. In Szpankowski and Rego [1988] a general method is presented for deducing the ergodicity conditions for (higher dimensional) queueing processes. Unfortunately, their method is rather complicated and, according to Nauta [1989] seems to contain heuristic elements. We refer to the study of Nauta [1989] for an overview of results in this area and a detailed investigation of ergodicity results for a general class of Markov chains with a two-dimensional state space.

The amount of work in the system at time t plays a key role in most of the analysis in the first few chapters. It is defined as the sum of the remaining service times of the customers in the system at time t. The terms "amount of work" and "workload" will be used interchangeably. Note that we assume the workload process to be ergodic and stationary for each system under consideration. On some occasions we will refer to the "amount of work at an arbitrary epoch". The random variable we are referring to in such cases is an independent copy[2] of the amount of work at some time t. Similarly, the

1. In physics such processes are often referred to as being in 'statistical equilibrium'.
2. By an independent copy of a random variable **X** we mean a random variable which 1) has the same distribution as **X** and 2) is independent of **X**.

"amount of work at an arbitrary epoch during a (switch-over) interval" refers to an independent copy of the amount of work at time t, conditioned on the fact that t lies in a (switch-over) interval.

It is possible to view switch-over times between queues as additional work introduced in the system by special classes of recurrent customers. By such a representation the discussion of waiting times and other operating characteristics can be made somewhat more uniform. However, in view of the application area of our results we prefer to have our nomenclature as close as possible to the colloquial term used in system engineering. The approach of viewing switch-over times as additional work is sometimes taken in the analysis of 'vacation' models (cf. Section 1.4). We can in that case apply the principle of work conservation directly. However, for systems with a non-FCFS service discipline or for polling systems with more than one queue, the relation between workload and waiting times of the 'regular' customers is not simple any more.

1.4 Discussion of related literature

In this section we discuss some literature related to this study and indicate the place of this study within the literature.

The term 'conservation law' was introduced by Kleinrock [1964,1965], who formulated and proved this relation for Poisson arrivals. Schrage [1970] showed that the conservation law also holds for general arrival processes. For some fundamental discussions of the conservation law and its implications, the reader is referred to the books of Gelenbe and Mitrani [1980] and Heyman and Sobel [1982]. For a survey with special emphasis on polling systems see Boxma [1989].

The first instances of the (pseudo)conservation law for polling systems with non-zero switch-over times were discovered by Ferguson and Aminetzah [1985] (for exhaustive and gated service) and by Watson [1985] (for exhaustive, gated and 1-limited service). The approach used by these authors was to derive these relations directly from a set of functional equations for the generating functions of the queue lengths. In the same year in which the results of Ferguson and Aminetzah, and Watson, appeared, Fuhrmann and Cooper [1985] published a paper in which they derived a (stochastic) decomposition for the *queue length* in a class of $M/G/1$ queues with so-called 'vacations'. In such models, the server occasionally 'takes a vacation'; during a vacation, no work is removed from the system. Vacation models are related to polling models, cf. also Section 2.6.

In this study a method is described that leads to a decomposition for the amount of *work* in polling models. The method was inspired by the above-mentioned result for vacation models by Fuhrmann and Cooper [1985]. A close examination of their arguments has led us to a surprising insight, viz. the technique of stochastic decomposition for the amount of work in polling systems. It is this decomposition which is one of the main results of this study. The decomposition result is subsequently used to give a simple but insightful derivation of a general pseudoconservation law, containing only one unknown

term that depends on the scheduling discipline. By specifying the scheduling discipline, we are then able to derive the pseudoconservation law explicitly for several special cases, amongst which are exhaustive, gated and 1-limited service. These results have appeared in Boxma and Groenendijk [1987,1989]. Since then, numerous papers have been published, deriving special cases of the pseudoconservation law using the decomposition result, or using the pseudoconservation law as a basis for approximations.

The validity of stochastic (work, queue length) decompositions in *vacation* models has been known for quite some time. We refer to Doshi [1986] for an extensive survey. There are also many recent generalizations, see Section 2.6 for a discussion of some of those results.

The literature on the analysis of polling systems is huge and still growing rapidly. For an extensive discussion of the many results that have recently been obtained, we refer to the book and survey paper of Takagi [1986,1988]. For a survey with special emphasis on the applications of polling systems see Levy and Sidi [1988].

1.5 Overview of results in the next chapters

In Chapter 2 the amount of work in the system is the primary quantity of interest. The principle of work conservation will be discussed and it will be shown that it has a *natural extension* to systems with switch-over times. In particular it will be shown that, under some fairly general conditions, a simple work decomposition is valid in the system with switch-over times: the amount of work in the system is distributed as the sum of two independent quantities, viz. (i) the amount of work in the corresponding system with identical traffic characteristics but *without* the switch-over times (hence *with* work conservation) and (ii) the amount of work in the original system at an arbitrary epoch at which the server is switching. Some generalizations of this result will be presented and the chapter is closed with a discussion of decomposition results in 'vacation' models.

In Chapter 3 it is shown how and under which conditions the principle of work conservation leads to the formulation of (Kleinrock's) conservation law for mean waiting times. The work decomposition is shown to give rise to the *pseudo*conservation law in a similar manner. For several scheduling disciplines an exact expression for this law is found.

The model formulation in Chapters 2 and 3 is continuous time. The generally time-synchronous configuration in many practical applications in communication networks would suggest a discrete-time formulation. Therefore in Chapter 4 the results of Chapters 2 and 3 are derived for the discrete-time model. The formulation and analysis of the discrete-time model is notationally more laborious and less easy to read. It is shown that via a limiting procedure the continuous-time pseudoconservation law can be derived from its discrete-time counterpart.

The results in Chapters 2 and 3 can be generalized to polling systems with a *polling table*. Such a table prescribes a fixed, not necessarily cyclic, order in which the queues are to be visited by the server. Each queue occurs at least

once in the table. The use of polling tables is of great interest for the operation of the system; in particular it provides possibilities for optimization of visiting schedules of the server at the various queues. The polling model with polling table was (re)introduced by Baker and Rubin [1987], see also Eisenberg [1972]. In Chapter 5 it is shown that by associating a unique *pseudostation* to each entry in the polling table, the same methods of analysis as in the (strictly) cyclic-service system can be applied. An extension of the work decomposition result enables the derivation of a pseudoconservation law for this system. The chapter is closed by some comparisons of the amount of work in the system under several important polling policies.

In Chapter 6 some two-queue models with 1-limited service at one queue are analyzed exactly. Generally speaking, the analysis of cyclic-service systems with exhaustive or gated type service is complex but tractable; limited-type service on the other hand gives rise to intricate mathematical problems, which only have been solved for some special models with not more than two queues. The analysis of the two-queue model with 1-limited service at both queues in this case leads to the formulation of a Riemann-Hilbert boundary value problem and illustrates the difficulties one may encounter in a detailed analysis; it also indicates the boundaries of mathematical tractability for slightly more general models. The model with exhaustive service at the other queue appears to be much simpler. The Laplace-Stieltjes Transforms of the marginal waiting-time distributions at both queues are derived in an explicit form for that case. The exact results that are obtained for two-queue models give insight into the behavior of more general polling systems and are useful for testing the accuracy of approximations.

A pseudoconservation law provides an excellent basis for the construction of approximations for mean waiting times of customers. It can also be used to test the quality of suggested mean waiting time approximations or to test the quality of simulation results. In Chapter 7 some approaches for approximation methods for mean waiting times based on the pseudoconservation law are developed. A simple procedure is discussed, yielding a closed-form expression for the individual mean waiting times. The method is shown to be exact in several special cases and is supported by extensive numerical testing. A refinement of the approximation scheme that is more accurate under heavy load conditions is also discussed.

Finally, in Chapter 8 it is investigated how the obtained results can be used in the analysis of Local Area Networks with 'multi-access based' protocols. This is illustrated by means of the token-passing protocol. The polling model as discussed in this study is shown to be a natural model for such networks. Some attention is furthermore given to the queueing analysis of several token ring local area networks interconnected via bridges. A simple procedure is presented to approximate the end-to-end delay of a message that traverses the network.

The results of Chapters 2 and 3 are based on Boxma and Groenendijk [1987], of Chapter 4 on Boxma and Groenendijk [1989], and of Chapter 5 on

Boxma, Groenendijk and Weststrate [1988]. Section 1 of Chapter 6 is based on Boxma and Groenendijk [1988b] and part of Section 2 on Groenendijk [1989]. Chapter 7 is based on Groenendijk [1988a,1989] and partly on Groenendijk and Levy [1989].

Throughout, symbols indicating random variables are denoted by capitals and printed bold. Sections, formulas, theorems, tables, figures, etc. are referred to by their number prefixed by a numeral indicating the chapter, e.g., Theorem 2.3 refers to the third theorem of Chapter 2. References to the literature are always presented as the name(s) of the author(s) followed by the year of publication. In case of more than one publication by the same author(s) in one year, a letter is added to distinguish between these various publications.

Chapter 2

WORK DECOMPOSITION: AN EXTENSION OF THE PRINCIPLE OF WORK CONSERVATION

2.1 Introduction

Consider a queueing system in which a single server serves at constant rate. Suppose the scheduling discipline, i.e., the procedure for deciding which customer(s) should be in service at any time, has the properties that it does not allow the server to be idle when work is present and does not affect the amount of service given to a customer or the arrival time of any customer. Comparing the sample paths of the workload process for such a system under different scheduling disciplines leads us to the observation that the workload process is independent of the scheduling discipline. This is called the 'principle of work conservation'; a scheduling discipline which possesses the above properties is said to be 'work conserving' (cf. Definition 2.1 below). The principle of work conservation has in the past proven to be very useful. It enables one to reduce the analysis of the workload process of queueing systems with highly complicated scheduling disciplines to the analysis of the workload process of queueing systems with a relatively simple scheduling discipline, such as the First Come First Served (FCFS) or the Last Come First Served (LCFS) discipline. In particular, for the cyclic-service system with zero switch-over times and a work conserving scheduling discipline the principle of work conservation implies that the steady-state amount of work in the system is the same as in a 'corresponding' system with, for instance, a FCFS discipline.

When the switch-over times of the cyclic-service system are non-zero the principle of work conservation is no longer valid, since now the server may be idle (switching) when there is work in the system. However, when the arrival process of customers is a Poisson process there appears to exist a natural modification of the principle of work conservation for (cyclic-service) systems with switch-over times, based on a *decomposition* of the amount of work in the system. The result, to be proved in this chapter, states that under certain conditions the following relation holds in the cyclic-service model with switch-over times:

$$\mathbf{V}_C \overset{D}{=} \mathbf{V}_{M/G/1} + \mathbf{Y},$$

with $\stackrel{D}{=}$ denoting equality in distribution and

$\mathbf{V}_C$:= amount of work in the cyclic-service system,

$\mathbf{V}_{M/G/1}$:= amount of work in a 'corresponding' $M/G/1$ system,

$\mathbf{Y}$:= amount of work in the cyclic-service system at an arbitrary epoch in a switching period.

$\mathbf{V}_{M/G/1}$ and $\mathbf{Y}$ are independent.

The work conserving property of the scheduling discipline will play a crucial role in our analysis. Following Heyman and Sobel [1982], we define it more precisely as:

DEFINITION 2.1
A scheduling discipline is *work conserving* when it does not allow the server to be idle when there is at least one customer in the system and does not alter the service requirements of customers nor the arrival times of customers.

REMARK 2.1
As some simple examples, the service disciplines FCFS, LCFS, Service In Random Order (SIRO) and Processor Sharing (PS) are work conserving. A Preemptive-Repeat service discipline is not work conserving. Also, a scheduling discipline that causes customers to balk, renege or defect, can never be work conserving.

In the next section a detailed model description is presented. In Section 2.3 the principle of work conservation will be formulated for cyclic-service systems with *zero* switch-over times. In Section 2.4 *non-zero* switch-over times will be considered and our main theorem for the work decomposition in such systems will be derived. Section 2.5 considers results for interrupted processes of a more general type than those with switch-over times only. Finally, in Section 2.6, we look at decomposition results in *vacation* models: a class of queueing models strongly related to cyclic-service models.

2.2 MODEL DESCRIPTION

The model under consideration consists of N queues, $Q_1, \ldots, Q_N$, each of which has infinite waiting room. The queues are served by a single server S who visits the queues in a fixed cyclic order: $Q_1,Q_2,\ldots,Q_N,Q_1,Q_2,\ldots$ (cf. Figure 2.1). Without loss of generality the server is assumed to serve at unit rate. The arrival process of customers at the queues is a compound Poisson process with a correlation structure, as recently introduced in Levy and Sidi [1988]. This arrival process is defined below.

DEFINITION 2.2
Arrival epochs occur according to a Poisson process with rate λ. At each arrival epoch, batches of size $\mathbf{K}=(\mathbf{K}_1, \ldots, \mathbf{K}_N)$ of customers for queues 1,...,N

arrive with some not specified joint batch size distribution. The vector $\mathbf{K}$ is assumed to have the same joint distribution at each arrival epoch. Furthermore it is assumed that $\mathbf{K}$ is independent of the number of customers arriving at previous or future arrival epochs. Let,

$$k_i := E\mathbf{K}_i, \quad k_i^{(2)} := E\mathbf{K}_i^2, \quad k_{i,i} := E\mathbf{K}_i^2 - E\mathbf{K}_i, \quad k_{i,j} := E\mathbf{K}_i K_j, \quad i \neq j. \tag{2.1}$$

The arrival rate of customers at queue i (class-i customers) is denoted by $\lambda_i := \lambda k_i$.

The service times of class-i customers are independent, identically distributed stochastic variables; their distribution $B_i(\cdot)$ has first moment β_i and second moment $\beta_i^{(2)}$. The offered traffic at Q_i, ρ_i, is defined as

$$\rho_i := \lambda_i \beta_i, \quad i = 1,...,N. \tag{2.2}$$

The total offered traffic ρ is defined as

$$\rho := \sum_{i=1}^{N} \rho_i. \tag{2.3}$$

The switch-over times of the server between the ith and $(i+1)$th queue are independent, identically distributed stochastic variables with first moment s_i and second moment $s_i^{(2)}$. The first moment of the total switch-over time during a cycle of the server, s, is given by:

$$s := \sum_{i=1}^{N} s_i, \tag{2.4}$$

its second moment is denoted by $s^{(2)}$. It is assumed that the interarrival process, the service process and the switch-over processes are mutually independent.

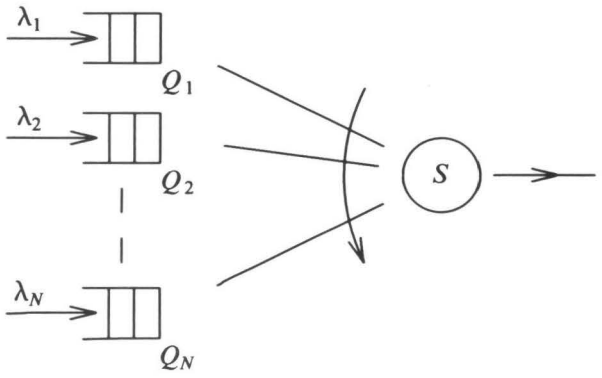

Figure 2.1 Queueing model of the cyclic-service system

Since in cyclic-service systems the visiting-order of the queues is fixed, the (global) scheduling discipline is entirely determined by the (local) scheduling disciplines at the individual queues. The scheduling discipline at a particular queue consists of two components, viz. the visit discipline and the service discipline at that queue:

1. visit discipline: Controls the number of customers served (or the amount of time spent) at a queue by the server during one uninterrupted visit of the server at this queue. Examples are: exhaustive service (the server serves a queue until its empty) or 1-limited service (the server serves only one customer, if present, and then moves to the next queue).

2. service discipline: Controls the order of service within a queue. Examples are: FCFS, LCFS, SIRO, PS or LCFS-Preemptive-Repeat.

2.3 All switch-over times equal to zero: work conservation

For a particular realization of the queueing process under a scheduling discipline D, the amount of work in the system at time t, $V_D(t)$, is defined as the sum of the remaining service times of the customers in the system at time t. When the scheduling discipline is work conserving, a typical realization of $V_D(t)$ is given in Figure 2.2.

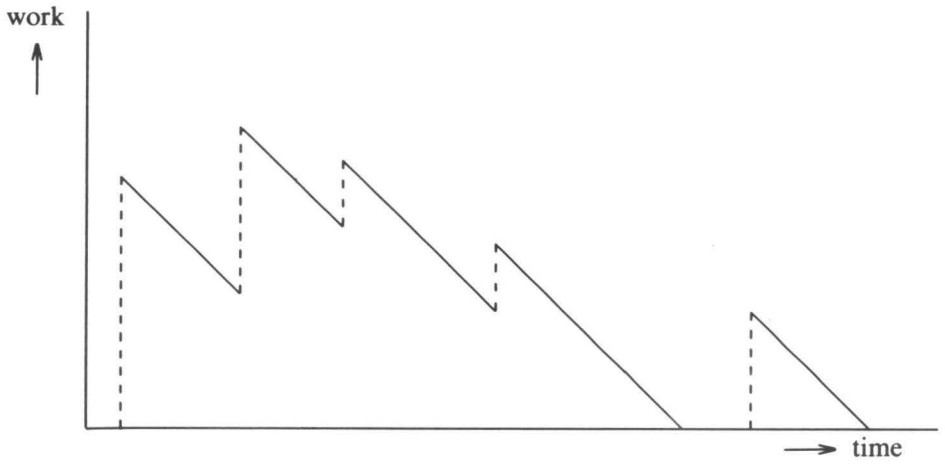

Figure 2.2
Amount of work in the system when the scheduling discipline is work conserving

At arrival epochs of customers, $V_D(t)$ jumps upwards by an amount equal to the required service time of an arriving batch of customers; it decreases linearly with slope -1 as long as the server is serving a (any) customer.

If the scheduling discipline is work conserving, then the only way for the scheduling discipline to influence $\mathbf{V}_D$ is to force the server to be idle when work is in the system, or by making customers leave before they have completed their service. Since this is not allowed due to the assumption that the scheduling discipline is work conserving, $\mathbf{V}_D$ must be independent of D. This is known as the 'principle of work conservation' (see also Gelenbe and Mitrani [1980, Ch. 6], Heyman and Sobel [1982, Ch. 11]). We formulate this principle in the following theorem.

THEOREM 2.1 The Principle of Work Conservation
Consider the cyclic-service system as described in Section 2.2. Suppose all switch-over times are identically zero. Then the steady-state workload in the system is the same for any work conserving scheduling discipline.

Theorem 2.1 implies the following weaker statement, which is of great practical importance in the analysis of cyclic-service systems without switch-over times.

COROLLARY 2.1
Denote by $\mathbf{V}_{FCFS}$ *the steady-state workload in the system under a FCFS scheduling discipline. The steady-state workload in the cyclic-service system with zero switch-over times is equal to* $\mathbf{V}_{FCFS}$ *for any work conserving scheduling discipline D. Hence, for any work conserving scheduling discipline D,*

$$\mathbf{V}_D \overset{D}{=} \mathbf{V}_{FCFS};$$

in which $\overset{D}{=}$ *denotes equality in distribution.*

The principle of work conservation is in a sense a similar property as local and global stochastic balance, Little's theorem and 'Poisson arrivals see time averages' (PASTA, cf. Wolff [1982]). In Heyman and Sobel [1982, p. 383] such properties are referred to as 'system properties'. System properties are shared by a large number of specific models; they can be used in structured models to obtain more specific conclusions. For example, under certain assumptions the mean workload of a particular class of customers can be expressed in the mean number of those customers and then, via Little's theorem, in their mean sojourn time. Thus the principle of work conservation may lead to a so-called *conservation law*: a linear relation between the mean waiting (or sojourn) times in a single-server multi-class system.

2.4 NON-ZERO SWITCH-OVER TIMES: WORK DECOMPOSITION

In the sequel, non-zero switch-over times are incorporated in the systems under consideration. Since now the server may be idle (switching) even though there is work in the system, the principle of work conservation is no longer valid. Denote the scheduling discipline of the system by D. Figure 2.3 gives a typical realization of $\mathbf{V}_D(t)$, the amount of work in the system at time t. Note that during switching periods the only changes in workload that occur are

caused by arrivals.

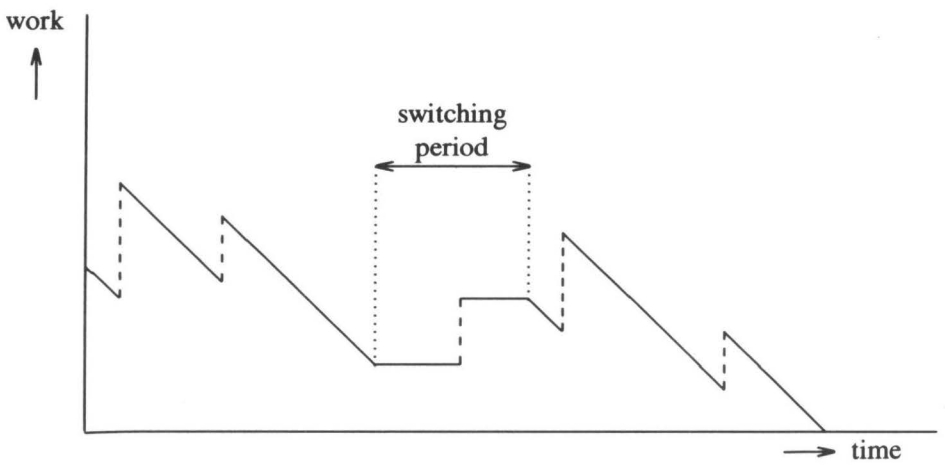

Figure 2.3
Amount of work in the system with non-zero switch-over times

The following theorem presents a natural modification of the work conservation principle. Before stating the theorem, we first introduce the notion of the '*corresponding*' $M/G/1$ queue: this is an $M/G/1$ queue in which the arrival stream is identical to the total arrival stream in the cyclic-service system (cf. Definition 2.2) and in which the service times are identical to the sum of the service times in the arriving batches in the cyclic-service system. Note that we therewith establish a *stochastic coupling* between the 'corresponding' $M/G/1$ queue and the cyclic-service system.

Consider a single-server cyclic-service system with switch-over times as described in Section 2.2. Suppose the scheduling discipline is work conserving. We define:

$\mathbf{V}_C$: the amount of work in the cyclic-service system,

$\mathbf{V}_{M/G/1}$: the amount of work in the 'corresponding' $M/G/1$ system,

$\mathbf{Y}$: the amount of work in the cyclic-service system at an arbitrary epoch in a switching interval.

We now formulate the main theorem of this chapter:

THEOREM 2.2 Work Decomposition
The amount of work in the cyclic-service system is distributed as the sum of the amount of work in the 'corresponding' M/G/1 system and the amount of work in the cyclic-service system at an arbitrary epoch in a switching interval:

$$\mathbf{V}_C \overset{D}{=} \mathbf{V}_{M/G/1} + \mathbf{Y}. \tag{2.5}$$

Furthermore, $\mathbf{V}_{M/G/1}$ *and* $\mathbf{Y}$ *are independent.*

PROOF:
In the cyclic-service system the server S is in one of two possible states: S is either serving or switching. As the system is assumed to be in steady state and an amount of work ρ per time unit is offered to the server, it is readily verified that

$$Pr\{S \text{ is serving}\} = \rho,$$

$$Pr\{S \text{ is switching}\} = 1 - \rho.$$

Hence we obtain (with (A) denoting the indicator function of the event A),

$$\begin{aligned} E[e^{-\omega \mathbf{V}_c}] &= E[e^{-\omega \mathbf{V}_c}(S \text{ is serving})] + E[e^{-\omega \mathbf{V}_c}(S \text{ is switching})] \\ &= \rho E[e^{-\omega \mathbf{V}_c} \mid S \text{ is serving}] + (1-\rho)E[e^{-\omega \mathbf{V}_c} \mid S \text{ is switching}] \\ &= \rho E[e^{-\omega \mathbf{V}_c} \mid S \text{ is serving}] + (1-\rho)E[e^{-\omega \mathbf{Y}}], \quad \text{Re } \omega \geqslant 0. \end{aligned} \tag{2.6}$$

We now need the following lemma.

LEMMA 2.1
The amount of work in the cyclic-service system at an arbitrary epoch in a service interval is distributed as the sum of two independent quantities, viz., the amount of work in the 'corresponding' M/G/1 queue at an arbitrary epoch in a service interval and the amount of work in the cyclic-service system at an arbitrary epoch in a switching interval. So,

$$E[e^{-\omega \mathbf{V}_c} \mid S \text{ is serving}] =$$

$$E[e^{-\omega \mathbf{V}_{M/G/1}} \mid \text{server in } M/G/1 \text{ is serving}]E[e^{-\omega \mathbf{Y}}], \quad \text{Re } \omega \geqslant 0. \tag{2.7}$$

Note that the distribution of $\mathbf{V}_{M/G/1}$ does not depend on the service discipline in the $M/G/1$ queue as long as no work is created or destroyed within the

system; this is the principle of work conservation, cf. Theorem 2.1.
From (2.6) and (2.7):

$$E[e^{-\omega \mathbf{V}_c}] = E[e^{-\omega \mathbf{Y}}]\left[1-\rho+\rho E[e^{-\omega \mathbf{V}_{M/G/1}} \mid \textit{server in } M/G/1 \textit{ is serving}]\right]$$

$$= E[e^{-\omega \mathbf{Y}}]E[e^{-\omega \mathbf{V}_{M/G/1}}], \quad \text{Re } \omega \geqslant 0.$$

Hence we have proved Theorem 2.2 once we have proved Lemma 2.1.

In the proof of Lemma 2.1 we shall need the concepts of 'ancestral line' and 'offspring' of a customer (cf. Fuhrmann and Cooper [1985]). Let K_A be a customer who arrives during a switching interval. The customers who arrive during the service of K_A are called the first generation offspring of K_A. The customers who arrive during the service of customers of the first generation offspring are called the second generation offspring of K_A, etc. The set of all customers who belong to the offspring of K_A, including K_A, is called the ancestral line of K_A and K_A is called the ancestor of all customers in this ancestral line.

PROOF OF LEMMA 2.1:
Adapting an idea of Fuhrmann and Cooper [1985], we consider an $M/G/1$ system with a Last Come First Served (LCFS) service discipline and with identically the same traffic process offered as the cyclic-service system, in which the server takes vacations *exactly* during the switching periods of the cyclic-service system (a switching *period* may consist of several consecutive switching *intervals,* e.g., switching intervals from Q_i to Q_{i+1} and from Q_{i+1} to Q_{i+2}). The LCFS discipline is assumed to be nonpreemptive, with one exception: if a service is interrupted by a vacation, forced upon the LCFS system by the cyclic-service system, and if during this vacation new customers arrive, then the interrupted service is resumed when all new customers (and offspring of these customers) have left.

Consider the cyclic-service system at an arbitrary service epoch. Obviously, the amount of work in the cyclic-service system and in the corresponding LCFS system with vacations are identical at any time and therefore we can (and we shall) from now on concentrate on the amount of work in the LCFS system at an arbitrary service epoch.

Let K denote the customer who is presently in service in the LCFS system. His ancestor is called K_A. Note that K could be K_A himself. By definition, K_A has arrived during a switching period (or, here: a vacation). Because of the 'Poisson arrivals see time averages' property, the amount of work found by K_A's batch upon arrival, $\mathbf{Y}_{K_A}$, is distributed like $\mathbf{Y}$. Note that, because of the LCFS service discipline, $\mathbf{Y}_{K_A}$ will still be present when K is in service.

We claim that the rest of the work, present at an arbitrary epoch at which K is being served, is distributed as the amount of work in an ordinary $M/G/1$

system with batch arrivals at a service epoch (the service discipline in this $M/G/1$ system may be FCFS or LCFS or any other work-conserving discipline). The motivations for this claim are that, even though it is possible that other customers have arrived after K_A in the same switching period (vacation), they do not belong to his ancestral line, they are served before K_A and so are their offspring - so they are of no interest to us.

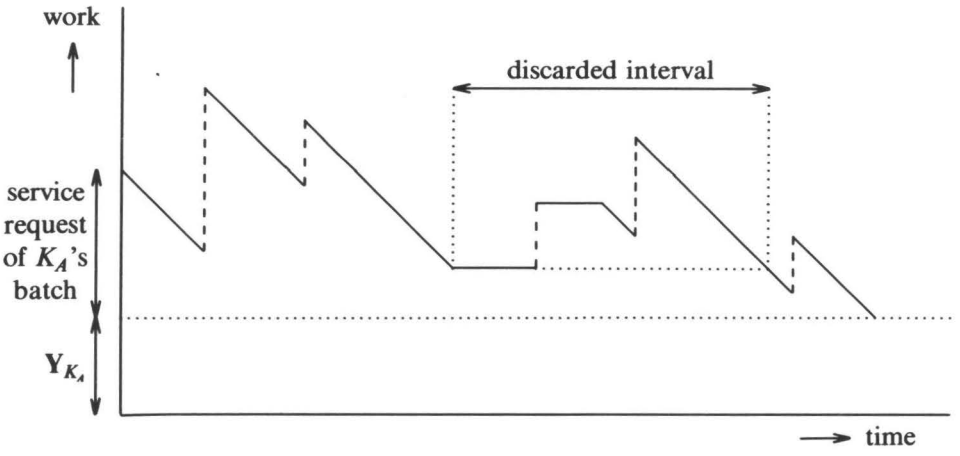

Figure 2.4
Amount of work in the LCFS system during service of K_A's ancestral line

Now consider the epoch at which the service of K_A starts (see Figure 2.4; cf. also Figure 2.3). Apart from Y_{K_A} no further work is present; and we ignore Y_{K_A}. The residual amount of work now evolves just as in an ordinary $M/G/1$ system with batch arrivals and non-preemptive LCFS (or any other work-conserving service discipline) with one exception: during the vacation periods, forced upon the LCFS system by the cyclic-service system, the work remains constant or may increase because of new arrivals. But these new arrivals, and their offspring, are served first (and do not belong to the ancestral line of K_A) and finally the work level is back again at the level immediately before the vacation started. Note that, due to the memoryless property, the arrival process also starts afresh and that, once more, only Y_{K_A} and work required by the offspring of K_A's batch is present.

This reasoning shows that, at an arbitrary service epoch of K, the amount of work present is composed of two independent parts: an amount of work Y_{K_A} that is distributed like Y and an amount of work that is distributed like the amount of work in an $M/G/1$ queue at an arbitrary service epoch.
This proves Lemma 2.1 and hence Theorem 2.2 is proved. □

REMARK 2.2
Our approach in the proof of Theorem 2.2, as first presented in Boxma and Groenendijk [1987], has been to look at 'serving' and 'switching' periods separately and subsequently to remove the conditioning. It was brought to our attention by S.W. Fuhrmann that a slightly more direct proof is possible by considering the arrival epoch of K's batch instead of an arbitrary epoch during K's service. This approach is taken in Boxma [1989].

REMARK 2.3
In the proof of Lemma 2.1 the same line of reasoning is used as in the proof of Proposition 5 of Fuhrmann and Cooper [1985]; but their reasoning is held for *customers* at *departure epochs* instead of for *steady-state workload*. In their case, this leads to a similar relation as (2.5) for *queue lengths*, for a class of vacation systems (cf. Section 2.6). Our cyclic-service model does not fall into this class, because Assumption 3 of Fuhrmann and Cooper [1985] is not fulfilled. It is easy to see that, when workload is considered instead of queue lengths, their Assumptions 2 and 3 may be replaced by the assumption that the service discipline is work conserving.

REMARK 2.4
The proof of Theorem 2.2, albeit insightful, is rather intuitive and is based on probabilistic arguments. B.T. Doshi kindly showed us a different, more analytic, proof of the decomposition result. That proof is based on a level crossing argument and uses regenerative processes; we present it below.

Let λ denote the rate of the Poisson arrival process of batches. Let $B(\cdot)$ denote the distribution of the amount of work brought in by an arbitrary arriving batch of customers. Denote its first moment by β and its Laplace-Stieltjes transform by $\beta(\cdot)$. The traffic intensity equals $\rho := \lambda\beta$. Let $V(\cdot)$ and $Y(\cdot)$ denote the distributions of $\mathbf{V}_C$ and $\mathbf{Y}$ in the cyclic-service system with switch-over times. Assume that their densities exist; denote them by $f(\cdot)$ and $g(\cdot)$ and denote the Laplace transforms of these densities by $\phi(\cdot)$ and $\eta(\cdot)$. Equating the down-crossing and up-crossing rates of level $x>0$ yields:

$$f(x)-(1-\rho)g(x) = \lambda\int_{0-}^{x}(1-B(x-y))f(y)dy.$$

Combining this relation with

$$f(0) = (1-\rho)g(0),$$

and taking Laplace transforms leads to:

$$\phi(s) = (1-\rho)\eta(s) + \lambda\phi(s)\frac{1-\beta(s)}{s}.$$

Hence

$$\phi(s) = \frac{(1-\rho)s}{s-\lambda+\lambda\beta(s)}\eta(s),$$

which proves the decomposition into two independent components.

2.5 More general interruption processes

An interesting question is whether the work decomposition (2.5) also holds for interruption structures that are more general than switch-over times between queues. This question is extensively addressed in Boxma [1989]. In Boxma's model, the server is in one of three possible states: free, interrupted or serving. Interruptions may occur in various forms:

- the server takes a vacation;
- the server requires switch-over times between classes, or between customers, or even between service intervals of one and the same customer;
- the server experiences a breakdown.

Accordingly, the process of service interruptions is a stochastic process which may be intricately interwoven with the arrival and service processes and the scheduling discipline. Let $\mathbf{V}_D^I(t)$ denote the amount of work in the system at time t for a scheduling discipline D and interruption process I. Introduce the following assumption:

Assumption 2.1

1. The stochastic process $\mathbf{V}_D^I(t)$, $t \geqslant 0$ possesses an equilibrium distribution.
2. The scheduling discipline is work conserving (cf. Definition 2.1).
3. The interruption process does not affect the amount of service time given to a customer or the arrival time of any customer.
4. The arrival process is the Poisson process described in Definition 2.2.

Consider a single-server multi-class system satisfying Assumption 2.1. The server is said to be *non-serving* when he is either free or interrupted. We define:

$\mathbf{V}_D$: the amount of work in this system,

$\mathbf{V}_{M/G/1}$: the amount of work in the 'corresponding' $M/G/1$ system,

$\mathbf{Y}$: the amount of work in the original system at an arbitrary epoch in a non-serving interval.

Boxma [1989] provides the following generalization of Theorem 2.2:

THEOREM 2.3
The amount of work in the system at an arbitrary epoch is distributed as the sum of the amount of work at an arbitrary epoch in the corresponding $M/G/1$ system and the amount of work present in the original system at an arbitrary epoch in a non-serving interval:

$$\mathbf{V}_D^I \overset{D}{=} \mathbf{V}_{M/G/1} + \mathbf{Y}. \tag{2.8}$$

Furthermore, $\mathbf{V}_{M/G/1}$ and $\mathbf{Y}$ are independent.

PROOF:
The proof of this theorem is analogous to that of Theorem 2.2, it will therefore be omitted here. □

Now consider a model with interruptions as in Boxma [1989], but with a renewal process as arrival process. Denote by Q_B the 'corresponding' queueing model *without* interruptions and by Q_I the model *with* interruptions. Denote by:

$F_B(\cdot)$: distribution of the amount of work in Q_B at an arbitrary epoch;

$F_I(\cdot)$: distribution of the amount of work in Q_I at an arbitrary epoch;

$G(\cdot)$: distribution of the amount of work at an arbitrary moment during a vacation;

$H_B(\cdot)$: distribution of the amount of work in Q_B at an arrival epoch;

$H_I(\cdot)$: distribution of the amount of work in Q_I at an arrival epoch;

assume that their densities exist and denote them by $f_B(\cdot)$, $f_I(\cdot)$, $g(\cdot)$, $h_B(\cdot)$ and $h_I(\cdot)$ respectively. Let ρ denote the steady-state probability that the server is serving. All notation of the previous sections remains valid.

A study of the sample paths of Q_B and Q_I reveals the validity of the following Volterra integral equations:

$$f_B(x) = \lambda \int_0^x h_B(y)\{1-B(x-y)\}dy + \lambda H_B(0)\{1-B(x)\}, \quad F_B(0) = 1-\rho, \tag{2.9}$$

$$f_I(x) - g(x)(1-\rho) = \lambda \int_0^x h_I(y)\{1-B(x-y)\}dy + \lambda H_I(0)\{1-B(x)\}. \tag{2.10}$$

Let $\tilde{F}_I := F_B \star G$ and $\tilde{H}_I := H_B \star G$ ("$\star$" denoting the convolution operator). Denote their densities by $\tilde{f}_I$ and $\tilde{h}_I$ respectively. Then the following lemma (Doshi [personal communication]) holds:

LEMMA 2.2

$$\tilde{f}_I(x) - g(x)(1-\rho) = \lambda \int_0^x \tilde{h}_I(y)\{1-B(x-y)\}dy + \lambda \tilde{H}_I(0)\{1-B(x)\}.$$

PROOF:
By direct substitution of $\tilde{f}_I$ and $\tilde{h}_I$ into Equation (2.10). □

Note that the fact that the densities of $\tilde{F}_I$ and $\tilde{H}_I$ satisfy Equation (2.10) is *necessary* but by no means *sufficient* to prove that they are the distributions for the amount of work at an arbitrary epoch and an arrival epoch respectively in the model with interruptions.

2.6 STOCHASTIC DECOMPOSITIONS IN VACATION MODELS

A class of models strongly related to cyclic-service models is the class of *vacation models*. In such models the server occasionally takes a vacation; during a vacation no work is removed from the system. We shall only consider models with so-called multiple vacations: the server keeps on taking vacations until on return from a vacation at least one customer is present. The relation with cyclic-service systems is that when in the cyclic-service system the server leaves a queue to possibly visit some other queues, from the point of view of this particular queue the server is temporarily not available ("takes a vacation"). In most vacation models however, the vacation sequence is assumed to be independent of the interarrival and service-time process. This clearly is not true for cyclic-service systems with more than one class of customers (more than one queue) because here the length of each vacation depends on the arrival and service processes at the other queues. Also, the distribution of the vacations is in general unknown in our model. Consequently, the results obtained from vacation models can not be applied directly to cyclic-service models. Note that when there is only one class of customers in the cyclic-service system the switch-over time takes the role of a vacation and the two models coincide.

Because of their wide applicability, vacation models have been extensively studied. Indeed they reflect the not uncommon behavior of a server who is occasionally occupied with background tasks. In particular, results from vacation models are often used to obtain either iterative procedures or approximations for cyclic-service systems.

The occurrence of stochastic decompositions in vacation models has in the past been noted by several researchers. The results that have been obtained depend heavily on the type of vacations that are allowed in the system. The most general results are obtained for models in which the server takes a vacation only when the queue is empty (*exhaustive vacation model*). When the server is allowed to take a vacation while there are still customers in the system (*nonexhaustive vacation model*), the problem becomes much more difficult. We shall briefly review the main results that have been obtained; the paper of

Doshi [1986] provides an extensive survey on decomposition results for queueing systems with vacations.

For the *exhaustive vacation model* with Poisson arrivals, Levy and Yechiali [1975], using an embedded Markov chain approach, show that the number of customers in the system at a customer departure epoch is distributed as the sum of two independent random variables, one of which is the number of customers at a customer departure epoch in the 'corresponding' $M/G/1$ queue. As noted by Fuhrmann [1984] the only effect of the vacations is on the number of customers present when the busy period starts; in Fuhrmann [1984] this fact is exploited to give a different, more probabilistic argument for the decomposition. For the $GI/G/1$ exhaustive vacation model, Gelenbe and Iasnogorodski [1980] present a purely analytical proof of a waiting-time (work) decomposition.

REMARK 2.5
As customary in the analysis of exhaustive vacation models, the service discipline is assumed to be FCFS. Furthermore, vacations are considered as additional work in the system. This allows one to interpret the amount of work in the system as (virtual) waiting time.

The work decomposition result for the exhaustive $GI/G/1$ vacation model has later been derived by Doshi [1985] using sample path comparisons. This result has been further generalized by Keilson and Servi [1986]; using an analytic treatment, they show the work decomposition for the Bernoulli $GI/G/1$ vacation model. The work decomposition result for the exhaustive vacation model has recently yet again been generalized by Lucantoni, Meier-Hellstern and Neuts [1988], who derived the decomposition for a class of non-renewal arrival processes called Markovian Arrival Process (MAP). A MAP is a point process, which is in general non-renewal, and which includes the Markov-modulated Poisson process, the PH-renewal process and superpositions of such processes as particular cases. Finally, the $G/G/1$ case was proved in Doshi [1988].

REMARK 2.6
All decomposition results in $GI/G/1$ vacation models concern amount of work (or waiting time); it appears that in these systems the amount of work is a more natural quantity than the queue length (cf. also Remark 2.3). Moreover, the relation between the queue length and the amount of work in the system is less simple here than it is for queues with Poisson arrivals.

For the *nonexhaustive vacation model* with Poisson arrivals, the first to notice the decomposition for the steady-state queue length was Gaver [1962]. It was subsequently observed in different settings by Cooper [1970], Scholl and Kleinrock [1983] and Ali and Neuts [1984]. Ott [1984] considered a single server queueing system with two independent input streams, one being of 'M/G' type and the other being a much more general process which need not be Markovian. For this system he proved a decomposition of the amount of work

similar to our Theorem 2.2. One of the most general and fundamental results for the (queue-length) decomposition in non-exhaustive *M/G/1* vacation models is presented in Fuhrmann and Cooper [1985]. In that paper, they consider an '*M/G/1* queue with generalized vacations': a queueing model in which:

a. Customers arrive to the system according to a Poisson process and have generally distributed service times. The service times of different customers are independent of each other and are independent of the arrival process. In addition, each service time is independent of the sequence of vacation periods that precede that service time.

b. All customers arriving to the system are eventually served. Thus, the system has an infinite queueing capacity and $\rho<1$. Moreover, customers do not balk, defect, or renege from the system.

c. Customers are served in an order that is independent of their service times.

d. Service is nonpreemptive. That is, once selected for service, a customer is served to completion uninterruptedly.

e. The rules that govern when the server begins and ends vacations do not anticipate future jumps of the Poisson arrival process. Here the notion of 'anticipate' is as defined in Wolff [1982].

Define $\chi(\cdot)$ to be the probability generating function (p.g.f.) for the number of customers in the system given that the server is on vacation. Furthermore let $\psi(\cdot)$ be the p.g.f. for the stationary distribution of the number of customers that an arbitrary departing customer leaves behind in the vacation system and let $\pi(\cdot)$ be the p.g.f. for the stationary distribution of the number of customers that an arbitrary departing customer leaves behind in the 'corresponding' *M/G/1* queue. Then the following theorem (Proposition 3 of Fuhrmann and Cooper [1985]) holds, for the proof of which we refer to that study.

THEOREM 2.4
Consider an M/G/1 queue with generalized vacations satisfying (a)-(e) above. The functions $\psi(\cdot)$, $\chi(\cdot)$ *and* $\pi(\cdot)$ *are related by*

$$\psi(z) = \chi(z)\pi(z), \quad |z| \leq 1.$$

Shanthikumar [1988] shows a queue-length decomposition at departure epochs of customers when the arrival rate during vacations may depend on the queue length and when some forms of reneging and balking are allowed during vacations. Harris and Marchal [1988] allow a state-dependent vacation at every service completion epoch. Each service time is extended to include a (possibly zero length) state-dependent vacation after each service and the length of the

vacation may depend on the number of customers in the system at the start of the vacation. The results are derived in terms of the queue-length distribution at customer departure epochs.

The decomposition results for the non-exhaustive vacation model are only known for Poisson arrivals. Whether these results may be extended to more general arrival processes remains an important open question. The only generalization that is known at present is that the ASTA property (cf. Melamed and Whitt [1988]) implies a similar work decomposition as for Poisson arrivals. This result is described in Rosberg and Gail [1989]; however, it is not clear from their paper just how general they allow the vacation sequence to be.

Chapter 3

A PSEUDOCONSERVATION LAW FOR CYCLIC-SERVICE SYSTEMS WITH SWITCH-OVER TIMES

3.1 Introduction

In the previous chapter we have seen that the steady-state workload in a system without switch-over times is the same for any work-conserving scheduling discipline; this is the principle of work conservation. In systems where the principle of work conservation is valid, the amount of work present should not depend on the order of service - and hence should equal the amount of work in the 'corresponding' system with FCFS service discipline. If, moreover, the scheduling discipline selects customers in a way that is independent of (any measure of) the required service time, then the distribution of the number of customers in the system is also independent of the order of service. But even if this is not the case, as in systems with different service requirements for different classes of customers, the principle of work conservation yields a useful expression for a weighted sum of the mean queue lengths. Hence (by using Little's formula) a weighted sum of mean waiting times can be obtained. Such an expression is called a *conservation law*; it was first formulated by Kleinrock [1965]. The conservation law implies that a certain weighted sum of mean waiting times will not change, no matter how sophisticated or elaborate we may choose our scheduling discipline (this justifies the use of the word 'conservation'). So if we decide to give preferential treatment to one class of customers this is afforded at the expense of others. This phenomenon is frequently encountered in physical systems; for example, consider a simple balance as in Figure 3.1. In order to keep the balance horizontally we need to ensure that $M_1D_1+M_2D_2+M_3D_3=C$, where C is some fixed constant. If we would move weight M_1 to the left (decrease D_1), we would have to increase D_2 or D_3 in order to keep the balance straight. For an extensive discussion on conservation laws in queueing systems see Gelenbe and Mitrani [1980, Section 6.2]. For an overview see Boxma [1989] and references contained therein.

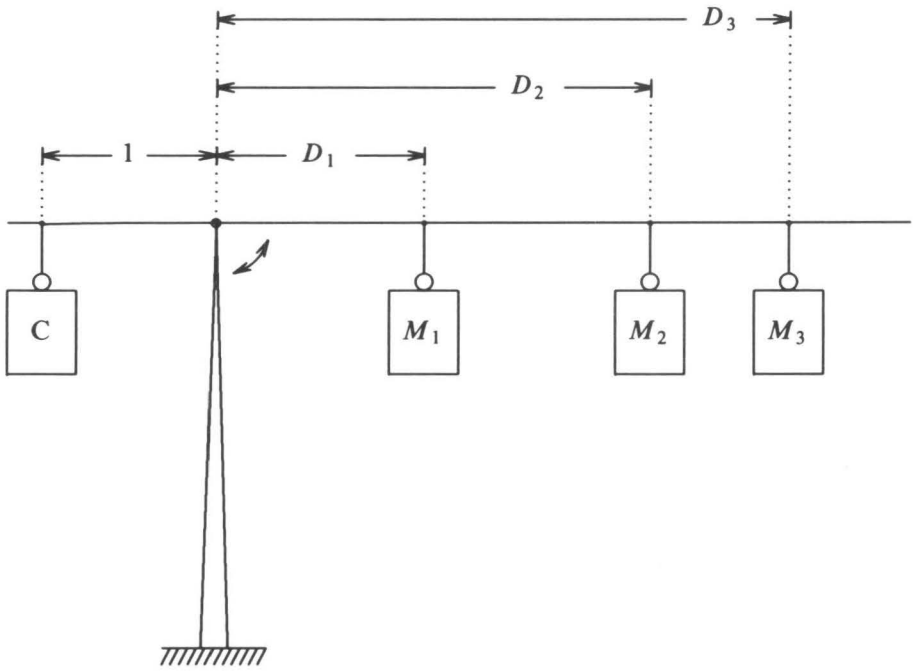

Figure 3.1 Example of conservation in a physical system: weights hanging on a balance

When switch-over times are introduced in the system the principle of work conservation is no longer valid, since the server may now be idle (switching) in the face of a non-empty queue. Comparing the sample paths of the workload of two such systems under different scheduling disciplines reveals that these sample paths will not in general be identical. Consequently, the amount of work in the system at some time t is not the same for different scheduling disciplines. This fact clearly prohibits the formulation of conservation laws in these systems. However, it will be shown that the work decomposition of Theorem 2.2 under rather restrictive assumptions concerning the scheduling discipline enables us to derive an exact expression for mean waiting times of customers in the cyclic-service system with switch-over times. Although it is no longer based on the principle of work conservation and the resulting expression is not invariant for the scheduling discipline, its use is similar to that of the conservation law. Furthermore, it reduces to the conservation law when the switch-over times become zero. We shall refer to this law as a 'pseudoconservation law'. The term 'pseudoconservation law' was for the first time used in Boxma and Groenendijk [1987]; it seems to be a commonly accepted term now.

Pseudoconservation laws for cyclic-service systems have been discovered by Ferguson and Aminetzah [1985] and Watson [1985]. Unfortunately, their

derivation was lengthy and cumbersome, and no satisfactory explanation for the occurrence of these laws was provided. Yet, their discovery was a major step forward and attracted considerable attention, since usually the analysis of cyclic-service systems is mathematically very complex. In some special cases the mean waiting times at the various queues can be found as the solution of a - usually huge - set of linear equations; but for most cases it is still unknown how to compute them.

The goal of this chapter is to generalize and unify the known pseudoconservation laws and to explain why they should hold. We start by showing that the principle of work conservation under some restrictions leads to the formulation of Kleinrock's conservation law. After having stated some general results for future reference, we proceed by deriving the main result of this chapter: the pseudoconservation law for cyclic-service systems with switch-over times. Finally, specifying the scheduling discipline, we explicitly calculate the pseudoconservation law for a number of important cases.

3.2 Kleinrock's conservation law

In this section zero switch-over times are assumed. The arrival process is the compound Poisson process with a correlation structure, as introduced in Definition 2.2. For this arrival process denote by $E\mathbf{W}_i$ the mean waiting time of a customer at an arbitrary position in a class-i batch. The waiting time of a customer is defined as the time from his arrival to the system until his departure from the system minus his required service time. Assume that $\rho<1$. This ensures that the joint stationary distribution of the waiting times exist. Note that we do not consider the waiting time of a batch but rather of an arbitrary customer in a batch. In Chapter 8 we shall occasionally consider the sojourn time of the whole batch (i.e., the waiting time of the last customer plus the last customer's service time).

We again need the assumption that the scheduling discipline is work conserving; it will enable us to use the results of the previous chapter. To be able to derive a general relation between the number of customers in the system and the amount of work we need some much stronger assumptions for the scheduling discipline. Following Gelenbe and Mitrani [1980] we assume that the scheduling discipline is *nonanticipating*:

Definition 3.1

A scheduling discipline is *nonanticipating* if it uses only information about the current state and the past of the queueing process in making scheduling decisions; thus it is possible to discriminate among customers on the basis of their expected remaining service times (since their classes and attained service are known) but not on the basis of exact remaining service times.

As some examples: FCFS, LCFS and SIRO are nonanticipating disciplines. The Shortest Remaining Processing Time (SRPT) discipline clearly is not nonanticipating. The 'nonanticipatory assumption' ensures that the mean service time of a class-i customer, given that the customer is in the system, is the

same as an unconditional mean class-i service time. For example, for the SRPT discipline the service time of a waiting customer tends to be longer than an arbitrary service time.

Our third and last assumption is that the scheduling discipline is nonpreemptive (i.e., once a customer has entered service he is served uninterruptedly until completion); however, when the required service times at Q_i are exponentially distributed the scheduling discipline is allowed to preempt class-i customers. For customer classes with non-exponential service times this assumption excludes service disciplines as Processor Sharing (PS) and LCFS-Preemptive-Resume. For such classes it also excludes visit disciplines like *time-limited* service where the server spends at most a time T visiting a particular queue before moving to the next queue.

Let us at this point formally restate the above assumptions:

ASSUMPTION 3.1
The scheduling discipline is assumed to be:

a. work conserving (cf. Definition 2.1),
b. nonanticipating (cf. Definition 3.1),
c. nonpreemptive for those classes of customers that have non-exponential service-time distributions.

REMARK 3.1
In the literature the case of exponentially distributed service times is usually treated separately. If it is assumed that the service-time distributions are general for all classes then the scheduling discipline is assumed to be nonpreemptive. If the service-time distributions for all classes are assumed to be exponential this assumption is omitted. The current formulation has the advantage that it allows for general service-time distributions for some classes and at the same time preemptive scheduling disciplines combined with exponential service times for other classes.

The following theorem presents (a slightly generalized version of) Kleinrock's conservation law for the mean waiting times:

THEOREM 3.1 Kleinrock's Conservation Law
For any scheduling discipline that satisfies Assumption 3.1,

$$\sum_{i=1}^{N} \rho_i E\mathbf{W}_i = \rho \frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\rho)} + \frac{\lambda \sum_{i=1}^{N} \sum_{j=1}^{N} \beta_i \beta_j k_{i,j}}{2(1-\rho)}, \tag{3.1}$$

with $E\mathbf{W}_i$ as defined above and the parameters as defined in Section 2.2.

PROOF:
Consider an $M/G/1$ queue in which the arrival stream is identical to the total

arrival stream in the cyclic-service system (cf. Definition 2.2) and in which the service times are distributed as the sum of the service times in an arbitrary batch in the cyclic-service system. As in section 2.4 this queue is subsequently denoted as the '*corresponding*' $M/G/1$ queue. It follows from Theorem 2.1 that the mean amount of work in the cyclic-service system $E\mathbf{V}_C$ and the mean amount of work in the 'corresponding' $M/G/1$ queue $E\mathbf{V}_{M/G/1}$ are equal:

$$E\mathbf{V}_C = E\mathbf{V}_{M/G/1}. \tag{3.2}$$

Denote by $\mathbf{V}_C^i$ the steady-state workload in the cyclic-service system due to class-i customers. For the mean amount of work in the 'corresponding' $M/G/1$ queue we have (cf. Levy and Sidi [1988]):

$$E\mathbf{V}_{M/G/1} = \frac{\sum_{i=1}^{N}\lambda_i\beta_i^{(2)}}{2(1-\rho)} + \frac{\lambda\sum_{i=1}^{N}\sum_{j=1}^{N}\beta_i\beta_j k_{i,j}}{2(1-\rho)}. \tag{3.3}$$

Denote by $E\mathbf{X}_i$ the mean number of class-i customers waiting at an arbitrary epoch. When the scheduling discipline at Q_i is nonpreemptive we can write the mean amount of class-i work in the cyclic-service system as:

$$\begin{aligned} E\mathbf{V}_C^i &= \beta_i E\mathbf{X}_i + \rho_i\frac{\beta_i^{(2)}}{2\beta_i} \\ &= \rho_i E\mathbf{W}_i + \rho_i\frac{\beta_i^{(2)}}{2\beta_i}. \end{aligned} \tag{3.4}$$

The first equality follows by noting that at an arbitrary epoch a class-i customer is being served with probability ρ_i, while his residual service time has mean $\beta_i^{(2)}/\beta_i$. The second equality follows from applying Little's formula.

Now suppose the scheduling discipline at Q_i is preemptive. Then according to Assumption 3.1 the service times of class-i customers have a negative exponential distribution. The memoryless property of the exponential distribution plus the assumption that the scheduling discipline is nonanticipating imply that the average remaining service time of any class-i customer in the system is β_i, regardless of how much service that customer has already received. It follows that for the mean amount of class-i work in the cyclic-service system:

$$E\mathbf{V}_C^i = \beta_i E\mathbf{X}_i. \tag{3.5}$$

Denote by $E\mathbf{S}_i$ the mean sojourn time of a typical customer at an arbitrary position in a typical class-i batch. Application of Little's formula yields,

$$E\mathbf{X}_i = \lambda_i E\mathbf{S}_i = \lambda_i(E\mathbf{W}_i + \beta_i) = \lambda_i E\mathbf{W}_i + \rho_i. \tag{3.6}$$

Hence,

$$E\mathbf{V}_C^i = \rho_i E\mathbf{W}_i + \rho_i \beta_i$$

$$= \rho_i E\mathbf{W}_i + \rho_i \frac{\beta_i^{(2)}}{2\beta_i}, \tag{3.7}$$

which is the same as (3.4).

It follows from (3.4) and (3.7) that the mean amount of work in the cyclic-service system due to all classes is equal to

$$E\mathbf{V}_C = \sum_{i=1}^{N} E\mathbf{V}_C^i = \sum_{i=1}^{N} \rho_i E\mathbf{W}_i + \sum_{i=1}^{N} \rho_i \frac{\beta_i^{(2)}}{2\beta_i}. \tag{3.8}$$

Equating (3.3) and (3.8), we obtain for any scheduling discipline satisfying Assumption 3.1 the following invariant expression for a weighted sum of mean waiting times:

$$\sum_{i=1}^{N} \rho_i E\mathbf{W}_i = \frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\rho)} + \frac{\lambda \sum_{i=1}^{N} \sum_{j=1}^{N} \beta_i \beta_j k_{i,j}}{2(1-\rho)} - \sum_{i=1}^{N} \rho_i \frac{\beta_i^{(2)}}{2\beta_i}$$

$$= \rho \frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\rho)} + \frac{\lambda \sum_{i=1}^{N} \sum_{j=1}^{N} \beta_i \beta_j k_{i,j}}{2(1-\rho)}. \tag{3.9}$$

□

The first statement and proof of a theorem of this type have been given for the *M/G/1* queue with single Poisson arrivals by L. Kleinrock; see Kleinrock [1965,1976 Sec. 3.4]. A proof for the *G/G/1* queue can be found in Schrage [1970]; note that for the latter queue no expression for the mean amount of work $E\mathbf{V}_{G/G/1}$ is known in general.

The conservation law puts a linear equality constraint on the set of mean waiting times $E\mathbf{W}_i$. This implies that any modification of the queueing discipline that reduces one of the $E\mathbf{W}_i$ will force an increase in one of the other $E\mathbf{W}_i$; however, this need not be an "even trade", since the weighting factors ρ_i are in general distinct.

Conservation laws are especially useful in the analysis of systems with a complicated scheduling discipline. For such systems the conservation law often provides the only exact information available about mean waiting times.

3.3 The pseudoconservation law

Before we start the derivation of the pseudoconservation law for the cyclic-service system *with* switch-over times, we state a few general known results for future reference.

For any cyclic service system we can define the cycle time $\mathbf{C}_i$ for Q_i as the time between two successive arrivals of the server S at Q_i. It is easily seen that $E\mathbf{C}_i$ is independent of i, so that we can drop the subscript i and write $E\mathbf{C}$ instead. Note that $\rho E\mathbf{C} = E\mathbf{C} - s$: the amount of work *arriving* during a cycle must on the average be the same as the amount of work *departing* during this cycle. This is a balancing argument. It follows, that

$$E\mathbf{C} = \frac{s}{1-\rho}. \tag{3.10}$$

Furthermore we can define the visit time $\mathbf{V}_i$ of S to Q_i as the time between the arrival of S at Q_i and his subsequent departure from that queue. Balancing the flow of class-i customers in and out of the system during a cycle shows that,

$$\lambda_i E\mathbf{C} = \frac{E\mathbf{V}_i}{\beta_i}, \tag{3.11}$$

and hence, from (3.10),

$$E\mathbf{V}_i = \frac{\rho_i s}{1-\rho}. \tag{3.12}$$

The intervisit time $\mathbf{I}_i$ for Q_i is defined as:

$$\mathbf{I}_i := \mathbf{C}_i - \mathbf{V}_i. \tag{3.13}$$

Finally, a remark about the conditions for ergodicity of cyclic-service systems. Clearly, $\rho<1$ is a necessary condition. However, for each scheduling discipline that we consider, additional conditions may be required. These will be discussed in detail in Section 3.4.

We now turn to the derivation of the general pseudoconservation law. Consider the single-server cyclic-service system described in Section 2.2. As usual the system is assumed to be in steady state. Suppose the scheduling discipline satisfies Assumption 3.1. From the work decomposition (2.5) we have (cf. also (3.3)):

$$\begin{aligned} E\mathbf{V}_C &= E\mathbf{V}_{M/G/1} + E\mathbf{Y} \\ &= \frac{\sum_{i=1}^{N}\lambda_i\beta_i^{(2)}}{2(1-\rho)} + \frac{\lambda\sum_{i=1}^{N}\sum_{j=1}^{N}\beta_i\beta_j k_{i,j}}{2(1-\rho)} + E\mathbf{Y}. \end{aligned} \tag{3.14}$$

As in Section 3.2, we have on the other hand:

$$E\mathbf{V}_C = \sum_{i=1}^{N} \rho_i E\mathbf{W}_i + \frac{1}{2}\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}. \tag{3.15}$$

From (3.14) and (3.15),

$$\sum_{i=1}^{N} \rho_i E\mathbf{W}_i = \rho \frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\rho)} + \frac{\lambda \sum_{i=1}^{N} \sum_{j=1}^{N} \beta_i \beta_j k_{i,j}}{2(1-\rho)} + E\mathbf{Y}. \tag{3.16}$$

To obtain an expression for this weighted sum of mean waiting times, it remains to determine $E\mathbf{Y}$, the amount of work in the cyclic-service system at an arbitrary epoch in a switching interval. Denote by $\mathbf{Y}_i$ the amount of work in the cyclic-service system at an arbitrary switching epoch during a switch-over from Q_i to Q_{i+1}; then it is easily seen that

$$E\mathbf{Y} = \sum_{i=1}^{N} \frac{s_i}{s} E\mathbf{Y}_i. \tag{3.17}$$

$E\mathbf{Y}_i$ is composed of three terms:

1. $E\mathbf{M}_i^{(1)}$: the mean amount of work in Q_i at a departure epoch of the server S from Q_i,
2. $E\mathbf{M}_i^{(2)}$: the mean amount of work in the rest of the system at a departure epoch of S from Q_i,
3. $\rho \frac{s_i^{(2)}}{2s_i}$: the mean amount of work that arrived at the system during the past part of the switching interval under consideration.

Hence we have

$$E\mathbf{Y}_i = E\mathbf{M}_i^{(1)} + E\mathbf{M}_i^{(2)} + \rho \frac{s_i^{(2)}}{2s_i}. \tag{3.18}$$

It will turn out that $E\mathbf{M}_i^{(1)}$ is the only term in the right-hand side of (3.18) which depends on the scheduling discipline at Q_i; it can only be determined when the visit discipline at Q_i is specified. Hence we shall first consider $E\mathbf{M}_i^{(2)}$, the total amount of work in $Q_{i+1},\ldots,Q_N,Q_1,\ldots,Q_{i-1}$ at a departure epoch of S from Q_i. By noting that the mean visit time at Q_h is given by $\rho_h s/(1-\rho)$ (cf. (3.12)), we obtain the following relation:

$$E\mathbf{M}_i^{(2)} = \rho_{i-1}(s_{i-1} + \frac{\rho_i s}{1-\rho}) + \rho_{i-2}(s_{i-2} + \frac{\rho_{i-1} s}{1-\rho} + s_{i-1} + \frac{\rho_i s}{1-\rho})$$

$$+ \cdots + \rho_{i+1}(s_{i+1} + \frac{\rho_{i+2}s}{1-\rho} + s_{i+2} + \frac{\rho_{i+3}s}{1-\rho} + \ldots + s_{i-1} + \frac{\rho_i s}{1-\rho})$$

$$+ \sum_{j \neq i} E\mathbf{M}_j^{(1)}, \tag{3.19}$$

and

$$\sum_{i=1}^{N} \frac{s_i}{s} E\mathbf{M}_i^{(2)} = \frac{\rho}{s} \sum\sum_{h<k} s_h s_k + \frac{s}{1-\rho} \sum\sum_{h<k} \rho_h \rho_k + \sum_{i=1}^{N} \frac{s_i}{s} \sum_{j \neq i} E\mathbf{M}_j^{(1)}. \tag{3.20}$$

Hence

$$E\mathbf{Y} = \sum_{i=1}^{N} \frac{s_i}{s} E\mathbf{Y}_i = \sum_{j=1}^{N} E\mathbf{M}_j^{(1)} + \frac{\rho}{s}[\sum\sum_{h<k} s_h s_k + \frac{1}{2}\sum_{i=1}^{N} s_i^{(2)}] + \frac{s}{1-\rho} \sum\sum_{h<k} \rho_h \rho_k$$

$$= \sum_{j=1}^{N} E\mathbf{M}_j^{(1)} + \rho \frac{s^{(2)}}{2s} + \frac{s}{2(1-\rho)}[\rho^2 - \sum_{i=1}^{N} \rho_i^2]. \tag{3.21}$$

Consider a single-server cyclic-service system with non-zero switch-over times as described in Section 2.2. Recall that for $j = 1,2,\ldots,N$, $E\mathbf{M}_j^{(1)}$ denotes the mean amount of work in Q_j at a departure epoch of the server from Q_j. Combining (3.16) and (3.21) we may now formulate our main result of this chapter.

THEOREM 3.2 The Pseudoconservation Law
For any scheduling discipline satisfying Assumption 3.1:

$$\sum_{i=1}^{N} \rho_i E\mathbf{W}_i = \rho \frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\rho)} + \frac{\lambda \sum_{i=1}^{N} \sum_{j=1}^{N} \beta_i \beta_j k_{i,j}}{2(1-\rho)} + \rho \frac{s^{(2)}}{2s}$$

$$+ \frac{s}{2(1-\rho)}[\rho^2 - \sum_{i=1}^{N} \rho_i^2] + \sum_{j=1}^{N} E\mathbf{M}_j^{(1)}. \tag{3.22}$$

Some comments about the meaning of the terms in the right-hand side of (3.22) are in order. The first two terms represent the mean amount of work waiting in the corresponding cyclic-service system *without* switch-over times (cf. (3.3)). The third, fourth and fifth terms reflect the influence of the presence of switch-over times. In fact they constitute the mean amount of work present at a switching epoch. The third term represents the mean amount of work that

arrived at all queues *during the switching intervals* after the last visit of S to those queues. Note that $s^{(2)}/2s$ represents the mean total past switching time from the departure of S from an arbitrary queue to the present random switching epoch. This interpretation explains why only s and $s^{(2)}$ occur in (3.22), and no moments of individual switch-over times. The fourth term reflects the interaction between queues; it represents the mean amount of work that arrived at queues after the last visit of S, during the subsequent service periods of other queues. Its most natural representation is perhaps

$$\sum_{k<h}\sum \rho_k E\mathbf{V}_h.$$

Finally $\sum_{j=1}^{N} E\mathbf{M}_j^{(1)}$ represents the mean amount of work that arrived at queues during the last service periods of those queues, but that was not handled by S at those service periods. $E\mathbf{M}_j^{(1)}$ depends on the visit discipline at Q_j; hence it can only be determined when the visit discipline at Q_j is specified. Finally, note that the order in which S visits the queues has no effect on the pseudoconservation law; however, the individual waiting times are in general affected if this order is changed.

In the next section we shall describe some of the various possibilities for the visit disciplines at the queues. For these cases we shall indicate any additional ergodicity conditions and explicitly calculate the pseudoconservation law.

3.4 Determination of $E\mathbf{M}_j^{(1)}$ for some special cases.

In this section we consider various possibilities for the visit disciplines at the queues; these disciplines differ in the number of customers that may be served in a queue during a visit of server S to that queue. Assume that S visits Q_i. When Q_i is empty S immediately begins to switch to Q_{i+1} (we disregard variants in which S does not switch if none of the queues contains customers). Otherwise S acts as follows, depending on the visit discipline at Q_i:

1) *Exhaustive service*: S serves class-i customers until Q_i is empty.

2) *Gated service*: S serves exactly those class-i customers present upon his arrival at Q_i (a 'gate' closes upon his arrival).

3) *1-Limited service*: S serves exactly one class-i customer.

4) *Semi-exhaustive service*: S continues serving class-i customers until the number present is one less than the number present upon his arrival.

5) *Binomial gated*: when S finds N_i customers present upon his arrival at Q_i, he serves a number of customers that is binomially distributed with parameters N_i and p_i, $0<p_i\leqslant 1$. Note that $p_i=1$ corresponds to gated service.

6) *Binomial exhaustive*: when S finds N_i customers present upon his arrival at Q_i, he sets aside a number of customers that is binomially distributed

with parameters N_i and p_i, $0 \leqslant p_i < 1$, and he serves the other customers and those arriving during their service, etc. Note that $p_i = 0$ corresponds to exhaustive service.

7) *Bernoulli*: after each service which does not leave Q_i empty, S serves another customer with probability p_i and moves to the next queue with probability $1-p_i$. This discipline has been introduced by Keilson and Servi [1986]. Note that for $p_i = 0$ and $p_i = 1$ the Bernoulli discipline reduces to the 1-limited and exhaustive disciplines respectively.

For an extensive discussion of the exhaustive, gated and 1-limited disciplines and detailed references see Takagi [1986].

For systems with an exhaustive or gated visit discipline the exact mean waiting times can be numerically calculated. The first complete solution of the problem, for single Poisson arrivals, was obtained by Eisenberg [1972]. In the case of exhaustive service at all queues, the mean waiting time $E\mathbf{W}_i$ can be expressed in the mean residual intervisit time (cf. (3.13)) for Q_i:

$$E\mathbf{W}_i = \frac{\lambda_i \beta_i^{(2)}}{2(1-\rho_i)} + \frac{E\mathbf{I}_i^2}{2E\mathbf{I}_i};$$

$E\mathbf{I}_i$ is given by $(1-\rho_i)s/(1-\rho)$ (cf. (3.12) and (3.13)). Ferguson and Aminetzah [1985] show that all $E\mathbf{I}_i^2$ can be computed by solving $N(N-1)$ linear equations (thus significantly improving upon earlier results that required $O(N^3)$ equations). To achieve this they study the *terminal service time* of Q_i, defined as the visit time of the server at Q_i plus the switch-over time from Q_{i-1} to Q_i (for gated service: the switch-over time from Q_i to Q_{i+1}).

The analysis for gated service is very similar. The mean waiting time at Q_i can be expressed in the mean residual cycle time at Q_i:

$$E\mathbf{W}_i = (1+\rho_i)\frac{E\mathbf{C}_i^2}{2E\mathbf{C}_i};$$

note that $E\mathbf{C}_i = s/(1-\rho)$. Ferguson and Aminetzah [1985] show that the $E\mathbf{C}_i^2$ can be computed by solving N^2 linear equations (again exploiting properties of the terminal service times).

The approach sketched above obviously breaks down in the case of zero switch-over times. It is however possible to derive the results for this case by suitably taking limits in the results for the model with non-zero switch-over times. The correct procedure for this is described in Levy and Kleinrock [1987].

The most economic approach thus far for systems with exhaustive or gated service at all queues is indicated by Sarkar and Zangwill [1987]; they derive the mean waiting times for each queue using only $O(N)$ equations. Although their algorithm requires $O(N^3)$ arithmetic operations to compute the coefficients for the set of equations, it represents a significant improvement

over the previous methods.

For 1-limited service the exact mean waiting times can only be solved when $N=1$ or $N=2$, the latter case requiring the solution of a Riemann-Hilbert boundary value problem (cf. Chapter 6). For $N>2$ this model has thus far defeated any attempts for an exact analysis of even the marginal mean waiting times.

The semi-exhaustive discipline has been introduced by Takagi [1985], who studies it in the case where all arrival rates, service-time and switch-over time distributions are the same for all queues. An exact analysis for the two-queue case has recently been given by Cohen [1988a]; again the method of formulating the problem as a boundary value problem has been employed.

The binomial gated discipline, which has been introduced and analyzed by Levy [1988], allows assigning priorities to the queues of a cyclic-service system by choosing the probabilities p_i. The binomial exhaustive discipline is in a sense dual to this discipline. The computing requirements of the binomial gated and binomial exhaustive discipline are similar to those of gated and exhaustive service. However, it is not clear whether the approach of Sarkar and Zangwill [1987] can be applied to these disciplines to reduce the number of equations from $O(N^2)$ to $O(N)$.

Finally, for the Bernoulli discipline the only exact solution that is known is for $N=1$. For $N=2$ queues the problem can in principle be formulated as a boundary value problem. However, the kernel of the functional equation occurring in the formulation of the boundary value problem is of a somewhat unusual type. The problem seems to have a similar structure as a model discussed in Nauta [1989]; it is conjectured that similar methods can be employed to solve it. This is however still an open problem. Boxma [1986] contains a concise survey on detailed mathematical studies of two-queue models. In Blanc [1988] an iterative numerical technique is developed for the evaluation of queue-length distributions. Blanc applies this procedure to the analysis of polling systems with Bernoulli visit disciplines and is able to numerically analyze systems with up to four queues. His technique is based on power series expansions of the state probabilities as functions of the load of the system.

We shall allow mixed visit disciplines (e.g., semi-exhaustive at Q_1, exhaustive at Q_2 and Q_4, 1-limited at Q_3, gated at Q_5 and binomial gated at $Q_6, \ldots, Q_N$). The order of service within each queue (the service discipline) will not be specified; the only restriction we impose is that the scheduling discipline satisfies Assumption 3.1.

REMARK 3.2

Allowing mixed strategies is not only of theoretical interest, but may also be of practical importance. For example, the cyclic-service model may model a local area network in which several rings are connected to each other by bridges. In such a system the queues which represent the bridges should have higher priority than the other queues at the ring. The visit discipline at the ordinary queues usually is 1-limited, but at the "bridge queue" one may consider

another visit discipline to model the preferential treatment received by these queues. Topics like these will be addressed in Chapter 8, where we consider the use of analytic models in the analysis of practical systems.

Additional ergodicity conditions

For some of the above visit disciplines we need additional conditions to ensure ergodicity of the system. Recall that in all cases $\rho<1$ is a necessary condition. For the exhaustive and gated type visit disciplines (1,2,5 and 6) this condition is also sufficient. For 1-limited service it can be seen that

$$\frac{\lambda_i s}{1-\rho} < 1, \tag{3.23}$$

is an additional necessary condition for the stability of Q_i, $i=1,...,N$; indeed, for every $i=1,...,N$ the mean number of class-i arrivals during a cycle should be less than one.

Similarly, for the semi-exhaustive case we have the following additional necessary conditions:

$$\lambda_i E\mathbf{I}_i = \frac{\lambda_i s(1-\rho_i)}{1-\rho} < 1, \quad i=1,...,N. \tag{3.24}$$

This reflects the fact that for semi-exhaustive service the mean number of class-i arrivals during the intervisit time $\mathbf{I}_i$ should be less than one, since during visit times the number of class-i customers is at most reduced by one.

Finally, for the Bernoulli discipline the mean number of class-i arrivals should be less then the mean number of class-i customers served under saturation. Hence, it is required that

$$\frac{\lambda_i s}{1-\rho} < \frac{1}{1-p_i}. \tag{3.25}$$

For the mixed strategies that we allow, the conditions (3.23), (3.24) and (3.25) should be added to the stability condition $\rho<1$ for those queues at which we have a 1-limited, semi-exhaustive or Bernoulli discipline. We shall assume throughout that the cyclic-service system under consideration is in equilibrium. The conditions (3.23), (3.24) and (3.25) are *necessary* conditions. Szpankowski and Rego [1988] pose that (3.23) is also a *sufficient* condition. However, according to Nauta [1989] their proof is incomplete.

We shall now turn to the determination of $E\mathbf{M}_j^{(1)}$ for each of the above listed visit disciplines.

1. Exhaustive:

$$E\mathbf{M}_j^{(1)} = 0; \tag{3.26}$$

this is a direct consequence of the definitions of $E\mathbf{M}_j^{(1)}$ and of the exhaustive discipline.

2. Gated:

$$E\mathbf{M}_j^{(1)} = \rho_j E\mathbf{V}_j = \rho_j \frac{\rho_j s}{1-\rho} = \rho_j^2 \frac{s}{1-\rho}; \tag{3.27}$$

the mean amount of work left behind by the server S at his departure from Q_j is for gated service equal to the mean amount of work that arrived during the visit period of S at Q_j.

3. 1-Limited:
At a departure epoch of S from Q_j S has just completed one service with probability $\lambda_j s/(1-\rho)$ and no service with probability $1-\lambda_j s/(1-\rho)$. Hence, with $\mathbf{T}_j$ the amount of work left behind in Q_j at the departure epoch of a *customer* from Q_j,

$$E\mathbf{M}_j^{(1)} = \frac{\lambda_j s}{1-\rho} E\mathbf{T}_j. \tag{3.28}$$

Note that $E\mathbf{T}_j$ consists of the mean amount of class-j work that has arrived during the departing customer's sojourn time plus the mean amount of work in his batch that is served after him. This latter term is given by $(k_{j,j}/2k_j)\beta_j$ (note that $k_{j,j}/2k_j$ represents the average number of customers *behind* an arbitrary customer in a class-j batch). Hence,

$$E\mathbf{T}_j = \rho_j(E\mathbf{W}_j + \beta_j) + \frac{k_{j,j}}{2k_j}\beta_j. \tag{3.29}$$

From (3.28) and (3.29):

$$E\mathbf{M}_j^{(1)} = \rho_j \frac{\lambda_j s}{1-\rho} E\mathbf{W}_j + \rho_j^2 \frac{s}{1-\rho} + \frac{\lambda s}{2(1-\rho)} k_{j,j}\beta_j. \tag{3.30}$$

4. Semi-Exhaustive:
Again, with the above definition of $\mathbf{T}_j$,

$$E\mathbf{T}_j = \rho_j E\mathbf{W}_j + \rho_j\beta_j + \frac{k_{j,j}}{2k_j}\beta_j.$$

Denote by $\mathbf{U}_j$ the number of customers in Q_j at an arrival epoch of S at Q_j. Due to the structure of the semi-exhaustive discipline we can also write

$$E\mathbf{T}_j = \beta_j E[\mathbf{U}_j - 1 \,|\, \mathbf{U}_j \geqslant 1] + \beta_j\left[\frac{\lambda_j^2\beta_j^{(2)}}{2(1-\rho_j)} + \rho_j + \frac{k_{j,j}}{2k_j}\frac{1}{1-\rho_j}\right], \tag{3.31}$$

(note that the second term in the right-hand side represents the amount of work left behind by a departing customer in an $M/G/1$ queue with batch arrivals, arrival rate λ_j and service time distribution $B_j(\cdot)$). Subsequently express $E\mathbf{M}_j^{(1)}$ in the first term in the right-hand side of (3.31):

$$E\mathbf{M}_j^{(1)} = \beta_j E[max(0,\mathbf{U}_j - 1)] = \beta_j E[\mathbf{U}_j - 1 \mid \mathbf{U}_j \geqslant 1] Pr\{\mathbf{U}_j \geqslant 1\}. \tag{3.32}$$

Because the mean visit time of S at Q_j during a cycle when positive equals the mean busy period of an $M/G/1$ system with arrival rate λ_j and service time distribution $B_j(\cdot)$, we have

$$E\mathbf{V}_j = \frac{\rho_j s}{1-\rho} = Pr\{\mathbf{U}_j \geqslant 1\} \frac{\beta_j}{1-\rho_j}, \tag{3.33}$$

so

$$Pr\{\mathbf{U}_j \geqslant 1\} = \frac{\lambda_j s(1-\rho_j)}{1-\rho}. \tag{3.34}$$

Combining (3.29), (3.31), (3.32) and (3.34),

$$\rho_j E\mathbf{W}_j + \rho_j \beta_j + \frac{k_{j,j}}{2k_j}\beta_j =$$

$$\frac{E\mathbf{M}_j^{(1)}}{\lambda_j s(1-\rho_j)/(1-\rho)} + \beta_j\left[\frac{\lambda_j^2 \beta_j^{(2)}}{2(1-\rho_j)} + \rho_j + \frac{k_{j,j}}{2k_j}\frac{1}{1-\rho_j}\right]; \tag{3.35}$$

and so we have

$$E\mathbf{M}_j^{(1)} = \rho_j \frac{\lambda_j s(1-\rho_j)}{1-\rho} E\mathbf{W}_j - \frac{\lambda_j^2 s}{2(1-\rho)}\rho_j \beta_j^{(2)} - \frac{\lambda s}{2(1-\rho)}\rho_j \beta_j k_{j,j}. \tag{3.36}$$

5. Binomial Gated:
Denote by $\mathbf{N}_j$ the number of customers in Q_j found by the server upon his arrival at Q_j. Note the obvious relation between the mean visit time at Q_j and $E\mathbf{N}_j$:

$$E\mathbf{V}_j = E\mathbf{N}_j p_j \beta_j. \tag{3.37}$$

So from (3.12) and (3.37) we immediately obtain,

$$E\mathbf{N}_j = \frac{\rho_j s}{1-\rho}\frac{1}{p_j \beta_j}. \tag{3.38}$$

The mean amount of work left behind at Q_j when the server departs from that

queue can be easily calculated from $E\mathbf{N}_j$:

$$E\mathbf{M}_j^{(1)} = E\mathbf{N}_j(1-p_j)\beta_j + E\mathbf{N}_j p_j \rho_j \beta_j$$

$$= (\rho_j + \frac{1-p_j}{p_j})\frac{\rho_j s}{1-\rho}. \tag{3.39}$$

6. Binomial Exhaustive:
Using similar arguments as for binomial gated, we find,

$$E\mathbf{V}_j = E\mathbf{N}_j(1-p_j)\frac{\beta_j}{1-\rho_j}, \tag{3.40}$$

hence,

$$E\mathbf{N}_j = \frac{\rho_j(1-\rho_j)s}{1-\rho}\frac{1}{(1-p_j)\beta_j}. \tag{3.41}$$

Finally,

$$E\mathbf{M}_j^{(1)} = E\mathbf{N}_j p_j \beta_j = \frac{\rho_j(1-\rho_j)s}{1-\rho}\frac{p_j}{1-p_j}. \tag{3.42}$$

7. Bernoulli:

$$E\mathbf{M}_j^{(1)} = (1-p_j)[\rho_j\frac{\lambda_j s}{1-\rho}E\mathbf{W}_j + \rho_j^2\frac{s}{1-\rho} + \frac{\lambda s}{2(1-\rho)}\beta_j k_{j,j}]. \tag{3.43}$$

Tedijanto [1988] has derived Equation (3.43) for single arrivals; the extension to correlated batch arrivals is due to Boxma [1989].

Denote by

g: the group of Gated queues,
se: the group of Semi Exhaustive queues,
bg: the group of Binomial Gated queues,
be: the group of Binomial Exhaustive queues,
b: the group of Bernoulli queues.

Combining (3.22) and the above expressions for $E\mathbf{M}_j^{(1)}$ allows us to express the pseudoconservation law for the visit disciplines listed above as follows:

$$\sum_{i\in b}\rho_i[1-(1-p_i)\frac{\lambda_i s}{1-\rho}]E\mathbf{W}_i + \sum_{i\in g,bg,be}\rho_i E\mathbf{W}_i + \sum_{i\in se}\rho_i[1-\frac{\lambda_i s(1-\rho_i)}{1-\rho}]E\mathbf{W}_i =$$

$$\rho\sum_{i=1}^{N}\frac{\lambda_i\beta_i^{(2)}}{2(1-\rho)} + \frac{\lambda\sum_{i=1}^{N}\sum_{j=1}^{N}\beta_i\beta_j k_{i,j}}{2(1-\rho)} - \sum_{i\in se}\frac{\lambda_i^2\beta_i^{(2)}\rho_i s}{2(1-\rho)} + \rho\frac{s^{(2)}}{2s} +$$

$$\frac{s}{2(1-\rho)}[\rho^2-\sum_{i=1}^{N}\rho_i^2]+\frac{s}{1-\rho}\sum_{i\in g}\rho_i^2+\frac{s}{1-\rho}\sum_{i\in bg}\rho_i(\frac{1-p_i}{p_i})+$$

$$\frac{s}{1-\rho}\sum_{i\in be}\frac{p_i}{1-p_i}\rho_i(1-\rho_i)+\frac{s}{1-\rho}\sum_{i\in b}(1-p_i)[\rho_i^2+\rho_i\frac{k_{i,i}}{2k_i}]. \tag{3.44}$$

Note that, by substituting $p_i=0$ or $p_i=1$ in the expressions for the Bernoulli discipline, we obtain the results for the 1-limited and exhaustive disciplines respectively.

REMARK 3.3
The case of $N=1$ queue yields some (mostly well known) expressions for mean waiting times in $M/G/1$ queues with some type of server vacations.

Some comments on the pseudoconservation law
For cyclic-service systems with switch-over times and single Poisson arrivals a pseudoconservation law has first been obtained by Ferguson and Aminetzah [1985] for the case of exhaustive service at all queues and gated service at all queues, and by Watson [1985] for the same two cases and also for 1-limited service at all queues. A pseudoconservation law for semi-exhaustive service at all queues was obtained in Boxma [1986]. These four pseudoconservation laws resulted from a direct manipulation of the functional equations for the queue lengths. Their derivation was lengthy and required much algebra. In particular this derivation did not elucidate why such simple expressions should exist for weighted sums of mean waiting times. Furthermore, the meaning of the various terms in the right-hand sides of the pseudoconservation laws remained obscure.

The general pseudoconservation law as given by Formula (3.22) has for the first time been published in Boxma and Groenendijk [1987]. In that paper an explicit expression for the pseudoconservation law has been given for a mixture of exhaustive, gated, 1-limited and semi-exhaustive service at the queues. A discrete-time version of the pseudoconservation law with an extension to batch arrivals has been published in Boxma and Groenendijk [1988a] (cf. also Chapter 4 of this study). Levy and Sidi [1988] have further generalized these results to the case of the correlated batch Poisson arrivals introduced in Definition 2.2.

The pseudoconservation laws for binomial gated, binomial exhaustive, or Bernoulli service at all queues were derived using Formula (3.22). A pseudoconservation law for binomial gated was obtained in Levy [1988]. This result inspired the derivation of a similar law for binomial exhaustive service, see Groenendijk [1988b]. Tedijanto [1988] obtained a pseudoconservation law for Bernoulli service at all queues. This latter result is also obtained as a special case in Boxma and Weststrate [1989].

It follows from (3.22) and the subsequent discussion that in a cyclic-service system with a mixture of various visit disciplines for each queue one only has

to determine its corresponding $E\mathbf{M}_j^{(1)}$. However, although fairly straightforward for the above listed visit disciplines, determination of $E\mathbf{M}_j^{(1)}$ may in some cases be far from simple. Consider the G-limited and E-limited visit disciplines: S serves a queue according to the gated or the exhaustive visit discipline, with the restriction that he serves at most, say, k customers. $k=1$ reduces to 1-limited service, whereas $k=\infty$ reduces to gated respectively exhaustive service. Everitt [1986b,1989] has derived a pseudoconservation law for G-limited and E-limited service respectively, but his formulas still contain the unknown second moment of the number of customers served in the queue during a visit of S. An exact expression for this term is yet unknown.

3.5. Applications of the pseudoconservation law

In this chapter we have used the decomposition result of Chapter 2 to obtain a general pseudoconservation law for cyclic-service systems with non-zero switch-over times. These results form a natural extension of Kleinrock's conservation law, see Kleinrock [1965,1976]. Below we shall mention a number of situations in which pseudoconservation laws can be applied.

The pseudoconservation law immediately yields the exact mean waiting times of customers in several important cases. As an example, when all mean waiting times in the system are identical the pseudoconservation law gives the mean waiting time at each queue exactly. This greatly simplifies the analysis of completely symmetric systems. The conservation laws also yields exact results for systems with only one queue, thereby solving several types of vacation models.

A prime application for pseudoconservation laws is in obtaining or testing approximations for individual mean waiting times. Indeed pseudoconservation laws like (3.44) have become indispensable tools in the analysis of cyclic-service systems. All recent approximation methods for analyzing cyclic-service are based on the pseudoconservation law. An approximation that satisfies the pseudoconservation law has the desirable property of producing exact results when the mean waiting times of customers at the various queues are identical. This happens for example in a completely symmetric system, in which all queues have identical parameters.

In Chapter 7 we shall use the pseudoconservation law as a basis to derive approximations for mean waiting times. Such approximations are badly needed in analytically untractable cases (as in the cases of, e.g., 1-limited, semi-exhaustive, or Bernoulli service) but also in analytically tractable cases; the latter because when the number of queues is large the numerical computation of the exact formulas can become very cumbersome.

An application that is frequently overlooked in the literature is in testing the quality of simulation results for cyclic-service systems. The results of a simulation of a cyclic-service system for which a pseudoconservation law holds can be tested as follows: by substituting the simulation results for the mean waiting times in the left-hand side of the pseudoconservation law, the relative difference between this value and the right-hand side of this law as computed from the system parameters can be used as a 'measure stick' for testing the

accuracy of the simulation. A similar procedure can be applied to check the correctness of computations when exact results are available.

Pseudoconservation laws can also be used to compare either the expected amount of work in the system or the expected delay in completely symmetric systems for various visit disciplines. Some comparisons of the latter type have been carried out by Takagi [1986]. We briefly describe his considerations. Denote by $E\mathbf{W}_E$, $E\mathbf{W}_G$, $E\mathbf{W}_{SE}$ and $E\mathbf{W}_{1L}$ the mean waiting times in a completely symmetric system with switch-over times and respectively exhaustive, gated, semi-exhaustive or gated service at all queues. Then (3.44) immediately implies that,

$$E\mathbf{W}_E \leqslant E\mathbf{W}_G \leqslant E\mathbf{W}_{1L},$$

and,

$$E\mathbf{W}_E \leqslant E\mathbf{W}_{SE} \leqslant E\mathbf{W}_{1L}.$$

There is no strict ordering between the gated and semi-exhaustive service discipline.

Levy, Sidi and Boxma [1988] extend these results and make comparisons with respect to the total amount of unfinished work actually found in the system by a typical customer upon his arrival at the system. In particular they compare the exhaustive visit discipline with any type of *non-exhaustive* visit discipline: any type of visit discipline that does not enforce buffer exhaustion at every visit of the server in each of the queues. Let $\mathbf{U}_E(t)$ and $\mathbf{U}_{NE}(t)$ be the total amount of unfinished work at time t in the exhaustive and the non-exhaustive systems, respectively. It may be proven that, for every $t \geqslant 0$, $\mathbf{U}_E(t) \leqslant \mathbf{U}_{NE}(t)$. Similarly, a comparison is made between the gated visit discipline and the G-limited visit discipline (cf. Section 3.4). Let $\mathbf{U}_G(t)$ and $\mathbf{U}_{GL}(t)$ be the total amount of unfinished work at time t in the gated and G-limited systems, respectively. Then, for every $t \geqslant 0$, $\mathbf{U}_G(t) \leqslant \mathbf{U}_{GL}(t)$. Comparison results like these are very important for studying the trade-off between 'efficiency' and 'fairness' (cf. Chapter 1).

Finally, pseudoconservation laws are very useful in studying asymptotics, yielding information about what happens when the number of queues becomes very large or when the offered traffic at a particular queue approaches its stability limit. Watson [1985] studies the asymptotic behavior of a system with 1-limited service at all queues of which one approaches its stability limit. He shows that when one of the queues, say Q_i, is heavily loaded, the system behaves like the same system with node Q_i discarded and the effect of Q_i taken into account by replacing the switch-over time from Q_{i-1} to Q_{i+1} by a random variable whose distribution is the convolution of:

1) the distribution of the switch-over time from Q_{i-1} to Q_i,
2) the distribution of a service time in Q_i,
3) the distribution of the switch-over time from Q_i to Q_{i+1}.

These results can easily be generalized to the case of a heavily loaded 1-limited queue in a cyclic-service system with mixed visit disciplines. We can also consider other visit disciplines than 1-limited service, as long as they are of 'limited type': they do not monopolize the system when saturated. The semi-exhaustive discipline and the Bernoulli discipline for $p<1$ satisfy this constraint.

Assume that Q_i is saturated. When Q_i has a semi-exhaustive visit discipline, the distribution of the switch-over time from Q_{i-1} to Q_{i+1} in the reduced system becomes the convolution of

1) the distribution of the switch-over time from Q_{i-1} to Q_i,
2) the distribution of the busy period in an $M/G/1$ queue with batch arrivals, arrival rate λ_i and service time distribution $B_i(\cdot)$,
3) the distribution of the switch-over time from Q_i to Q_{i+1}.

When the visit discipline at Q_i is Bernoulli, the switch-over time from Q_{i-1} to Q_{i+1} in the reduced system becomes the convolution of

1) the distribution of the switch-over time from Q_{i-1} to Q_i,
2) $(1-p_i)\sum_{k=1}^{\infty} p_i^{k-1} B_i^{k\star}(\cdot)$; $B_i^{k\star}(\cdot)$ denoting the k-fold convolution of $B_i(\cdot)$,
3) the distribution of the switch-over time from Q_i to Q_{i+1}.

The above ideas can be used in devising approximations as follows. Assume that a particular queue, say Q_i, is heavily loaded. Suppose the visit discipline at Q_i is 1-limited. The idea is then to remove this queue from the system and to replace it by a switch-over time, which is chosen such that the mean waiting times at the other queues remain approximately the same. Usually the approach is to set the mean switch-over time equal to the expected visit time of the removed queue and the second moment of the switch-over time equal to the square of the first moment. The motivation for removing heavily loaded queues from the system and subsequently approximating the reduced system is that the reduced system is more symmetric and - hopefully - easier to approximate. A more extensive discussion of such an 'elimination procedure' can be found at the end of Chapter 7.

Chapter 4

WORK DECOMPOSITION AND PSEUDOCONSERVATION LAW FOR DISCRETE-TIME CYCLIC-SERVICE SYSTEMS

4.1 Introduction

All results in the previous chapters are for continuous-time systems. However, the analysis in this chapter will be carried out in discrete time. The motivation is that discrete-time stochastic processes are more appropriate for modeling the generally time-synchronized configuration of present-day communication networks, and therefore are important in practical applications. Furthermore, we feel that a discrete-time approach to polling systems may often be slightly easier than a continuous-time approach, in particular in the important variants with time restrictions on visits and cycles. Discrete-time polling systems have been studied by Konheim and Meister [1974], Swartz [1980], Rubin and DeMoraes [1983] and Takagi [1986]; however, the bulk of the literature in this area is devoted to continuous-time systems. It is noted that continuous-time results are easily obtained from their discrete-time counterparts via a limiting procedure. The terminology that is used throughout this chapter stems from applications in communication networks.

The organization of this chapter is as follows. In Section 4.2 we consider cyclic-service systems without switch-over times. For such systems the principle of work conservation (cf. Section 2.3) obviously holds. This principle naturally leads to a discrete-time version of Kleinrock's conservation law for mean waiting times. The extension of the work conservation principle to the case *with* switch-over times is made in Section 4.3. The discrete-time pseudoconservation law for mean waiting times is derived in Section 4.4. In Section 4.5, the relation between the obtained discrete-time results and results for the continuous-time case is exposed. We close this introductory section by presenting a more detailed model description and some basic results.

MODEL DESCRIPTION

We consider a discrete-time queueing system with N stations (queues) $Q_1,...,Q_N$, where each station has an infinite buffer capacity to store waiting messages (customers). Each message consists of a number of packets, which are assumed to be of fixed length. Time is slotted with slot size equal to the transmission time of one packet (the service time of a packet). We shall call the time interval $[j,j+1[$ the jth time slot.

Message arrival process

Denote by $\mathbf{X}_i(j)$ the number of messages arriving at station i in the jth time slot. The message arrival process at each station is assumed to be independent of those at the other stations. The N stochastic processes $\{\mathbf{X}_i(j), j=1,2,3,...\}$, $i=1,...,N$ are assumed to be mutually independent. For fixed i, the $\mathbf{X}_i(j)$, $j=1,2,...$ are assumed to be independent, identically distributed random variables. Put:

$$A_i(z) := E[z^{\mathbf{X}_i(j)}], \quad |z| \leqslant 1; \quad \lambda_i := E[\mathbf{X}_i(j)]; \quad \lambda_i^{(2)} := E[\mathbf{X}_i^2(j)]. \tag{4.1}$$

Note that we can view the arrival process at Q_i as a Bernoulli arrival process with batch arrivals:

$$A_i(z) = A_i(0) + [1-A_i(0)]G_i(z),$$

with $G_i(z)$ denoting the z-transform of the size of a type-i batch. The interarrival times of batches at Q_i have the (memoryless) geometric distribution.

For future use we define here

$$\lambda := \sum_{i=1}^{N} \lambda_i, \quad \lambda^{(2)} := E[(\sum_{i=1}^{N} \mathbf{X}_i(j))^2]. \tag{4.2}$$

REMARK 4.1

In this description of the arrival process, we have not incorporated a similar correlation structure as in the arrival process of Definition 2.2. Extending the results for such a mechanism is, however, straightforward.

Denote by $\mathbf{B}_i$ the number of packets in a message at station i. It is assumed that $\mathbf{B}_i \in \{1,2,...\}$. Its z-transform, first and second moment are given by:

$$B_i(z) := E[z^{\mathbf{B}_i}], \quad |z| \leqslant 1; \quad \beta_i := E[\mathbf{B}_i]; \quad \beta_i^{(2)} := E[\mathbf{B}_i^2]. \tag{4.3}$$

The $\mathbf{B}_i$, $i=1,...,N$ are assumed to be mutually independent and independent of $\mathbf{X}_i(j)$, $i=1,...,N$, $j=1,2,3,...$. Further we introduce:

$$\beta := \sum_{i=1}^{N} \frac{\lambda_i}{\lambda} \beta_i, \quad \beta^{(2)} := \sum_{i=1}^{N} \frac{\lambda_i}{\lambda} \beta_i^{(2)}. \tag{4.4}$$

The offered traffic at the ith station ρ_i is defined by

$$\rho_i := \lambda_i \beta_i, \quad i = 1,2,...,N, \tag{4.5}$$

and the total offered traffic ρ is defined by

$$\rho := \sum_{i=1}^{N} \rho_i. \tag{4.6}$$

A single server S visits the N stations in the order of their indices. The scheduling discipline for the system is assumed to be work conserving (cf. Definition 2.1).

Switching process
A switch-over time is needed to switch from one station to the next. The switch-over times of the server between the ith and the $(i+1)$th station (measured in time slots) are independent, identically distributed random variables with first moment s_i and second moment $s_i^{(2)}$. The first moment of the total switch-over time during a cycle of the server, s, is given by:

$$s := \sum_{i=1}^{N} s_i, \tag{4.7}$$

its second moment is indicated by $s^{(2)}$.

As usual, in the sequel it will be assumed that the system is in steady state. Clearly, $\rho < 1$ is a necessary condition for ergodicity. However, some scheduling disciplines may require additional conditions.

For future use we introduce some notation and concepts.

$\mathbf{X}_i$: the number of type-i messages in the system at an arbitrary epoch;
$\mathbf{X}_i^w$: the number of *waiting* type-i messages in the system at an arbitrary epoch;
$\mathbf{W}_i$: the waiting time of a type-i message.

The waiting time of a message is defined as follows: the number of time slots counted from the time slot following the one in which the message arrived, until the time slot at the end of which the message departs from the system, minus the number of packets in the message. Note that transmission of each packet requires one time slot.

REMARK 4.2
It should be noted that, as customary in discrete-time queueing literature, an *arbitrary epoch* is by convention the instant just after the beginning of an arbitrary time slot (cf. Hunter [1983]).

Below we state a few general results for future reference. As in the continuous-time system, we can define the cycle time $\mathbf{C}_i$ for Q_i as the time between two successive arrivals of S at Q_i. Again, it is easily seen that the *mean* cycle time for Q_i, $E\mathbf{C}_i$, is independent of i; we shall denote it by $E\mathbf{C}$. The visit time $\mathbf{V}_i$ of S for Q_i is the time between the arrival of S at Q_i and his subsequent departure from that queue. As in Section 3.3, it may be seen that,

$$E\mathbf{C} = \frac{s}{1-\rho}, \tag{4.8}$$

and,

$$E\mathbf{V}_i = \frac{\rho_i s}{1-\rho}. \tag{4.9}$$

The intervisit time, $\mathbf{I}_i$, for Q_i is defined as:

$$\mathbf{I}_i := \mathbf{C}_i - \mathbf{V}_i. \tag{4.10}$$

4.2. Conservation law for the discrete-time Geom/G/1 model

In this section the switch-over times are taken to be zero; hence the server works whenever there is work in the system, and is idle when there is no work in the system. Therefore the principle of work conservation holds: the total amount of work, $\mathbf{V}_C$, in the cyclic-service system does not depend on the order of service, and should hence equal the amount of work in a '*corresponding*' FCFS Geom/G/1 queueing system. This observation, under some additional assumptions, will allow us to derive a conservation law for mean waiting times in the cyclic-service system without switch-over times.

We first introduce the notion of the 'corresponding' Geom/G/1 queueing model. This is a discrete-time queueing model, consisting of one server and one queue with batch arrivals. The arrival process is constructed as follows: the arrival streams at all N queues of the cyclic-service model are aggregated into a single (*vector*) arrival stream. The batch of all the messages arriving in a time slot is called a *train*. In any time slot, no train arrives with probability $\prod_1^N A_i(0)$ and a train does arrive with probability $1-\prod_1^N A_i(0)$. An arbitrarily chosen message in this train poses a service request whose z-transform is the mixture $\sum_1^N \frac{\lambda_i}{\lambda} B_i(z)$.

The principle of work conservation now states that $\mathbf{V}_C$ equals the amount of work in the corresponding Geom/G/1 system, $\mathbf{V}_{Geom/G/1}$. Therefore, $\mathbf{V}_C$ also equals $\mathbf{V}_{Geom/G/1}$ in distribution:

$$\mathbf{V}_C \overset{D}{=} \mathbf{V}_{Geom/G/1}. \tag{4.11}$$

In the remainder of this chapter, it will be assumed that the scheduling discipline satisfies Assumption 3.1, i.e., it is:

1) work conserving,
2) nonanticipating,
3) nonpreemptive; however, when $B_i(z) = (1-p_i)z/(1-p_iz)$ (shifted geometric), this restriction may be omitted for type-i messages.

According to Kobayashi and Konheim [1977], the mean number of messages in the corresponding system at an arbitrary epoch is given by:

$$E\mathbf{X}_{Geom/G/1} = \frac{\lambda^2\beta^{(2)}}{2(1-\rho)} + \frac{(\lambda^{(2)}-\lambda^2-\lambda)\beta}{2(1-\rho)} + \rho. \tag{4.12}$$

Note that the second term in the right-hand side disappears when the arrival process is Poisson. The mean number of messages in service is ρ; the residual service time of the message in service is $\frac{\beta^{(2)}}{2\beta} + \frac{1}{2}$. Hence

$$E\mathbf{V}_{Geom/G/1} = [\frac{\lambda^2\beta^{(2)}}{2(1-\rho)} + \frac{(\lambda^{(2)}-\lambda^2-\lambda)\beta}{2(1-\rho)}]\beta + \rho[\frac{\beta^{(2)}}{2\beta} + \tfrac{1}{2}]. \tag{4.13}$$

REMARK 4.3
It should be observed that in the renewal process in discrete time with interevent-time distribution having first moment β and second moment $\beta^{(2)}$, the mean residual life time is $\beta^{(2)}/2\beta + 1/2$ and the mean past life time is $\beta^{(2)}/2\beta - 1/2$ (cf. Hunter [1983]).

Using similar arguments as in Section 3.2, we can write $E\mathbf{V}_C$ as (cf. the definitions above Remark 4.2):

$$E\mathbf{V}_c = \sum_{i=1}^{N}\beta_i E\mathbf{X}_i^w + \sum_{i=1}^{N}\rho_i[\frac{\beta_i^{(2)}}{2\beta_i} + \tfrac{1}{2}] \tag{4.14}$$

$$= \sum_{i=1}^{N}\rho_i E\mathbf{W}_i + \sum_{i=1}^{N}\rho_i[\frac{\beta_i^{(2)}}{2\beta_i} + \tfrac{1}{2}].$$

The second equality is based on Little's formula.

From (4.11), (4.13) and (4.14) we obtain the following expression for the weighted sum of the mean message waiting times:

$$\sum_{i=1}^{N}\rho_i E\mathbf{W}_i = \frac{\lambda\beta^{(2)}}{2(1-\rho)}\rho + \frac{(\lambda^{(2)}-\lambda^2-\lambda)\beta}{2\lambda(1-\rho)}\rho. \tag{4.15}$$

We shall call Equation (4.15) the *Geom/G/1 conservation law*. We have found

no explicit references to this relation in the queueing literature.

4.3. The stochastic decomposition result

In the sequel switch-over times are incorporated in the systems under consideration. Because now the server may be idle (switching) although there is work in the system, Kleinrock's principle of work conservation is no longer valid. However, Theorem 4.1 below presents a natural modification of this work conservation principle. Theorem 4.1 is the discrete-time analogue of Theorem 2.2.

Consider a single-server cyclic-service system as described in Section 4.1. Suppose the scheduling discipline is work conserving. We define:

$\mathbf{V}_C$: the amount of work in the cyclic-service system,

$\mathbf{V}_{Geom/G/1}$: the amount of work in the 'corresponding' Geom/G/1 system,

$\mathbf{Y}$: the amount of work in the cyclic-service system at an arbitrary epoch in a switching interval.

Note that an arbitrary epoch is considered to be 'in' a switching interval if it marks the beginning of a switching time slot; the 'corresponding' Geom/G/1 system is the system (without switch-over times) introduced in the preceding section.

Theorem 4.1

The amount of work in the cyclic-service system is distributed as the sum of the amount of work in the 'corresponding' Geom/G/1 system and the amount of work in the cyclic-service system at an arbitrary epoch in a switching interval:

$$\mathbf{V}_C \overset{D}{=} \mathbf{V}_{Geom/G/1} + \mathbf{Y}, \tag{4.16}$$

where $\overset{D}{=}$ *stands for equality in distribution. Furthermore,* $\mathbf{V}_{Geom/G/1}$ *and* $\mathbf{Y}$ *are independent.*

Proof:

The proof is similar to that of Theorem 2.2 for the continuous-time case. It is based on the following observations:

(1) $\mathbf{V}_{Geom/G/1}$ does not change when the service discipline is LCFS nonpreemptive instead of FCFS;

(2) $\mathbf{V}_C$ also does not change when, instead of cyclic service, the following service discipline is enforced: all arriving trains are served LCFS and service is interrupted only at switching epochs of the cyclic-service system;

(3) It now suffices to prove that $\mathbf{V}_C^{LCFS} \overset{D}{=} \mathbf{V}_{Geom/G/1}^{LCFS} + \mathbf{Y}$. The validity of this decomposition is a consequence of the LCFS discipline. Consider a train T that arrives during a switch-over period. It has to wait until trains

that arrived after T, in the same switch-over period, have been served (and also trains arriving during their service, etc.). When, finally, T is taken into service, the only work present is the work that T found upon his arrival. This latter quantity is distributed like **Y**. This statement is implied by a discrete-time equivalent of the PASTA-property (Wolff [1982]), which we should like to call the BASTA-property (Bernoulli Arrivals See Time Averages). Note that because the input of trains to the system is Bernoulli (and due to the memoryless property of the underlying geometric distribution), the distribution of the amount of work at an arbitrary epoch is equal to the distribution of the amount of work immediately before an arrival epoch of a train (cf. also Halfin [1983]). T initiates a busy period, which evolves exactly like a busy period in the 'corresponding' Geom/G/1 system. So during the busy period initiated by T, the amount of work present in the system is distributed as the sum of $\mathbf{V}^{LCFS}_{Geom/G/1}$ and **Y**. We refer to the proof of Theorem 2.2 for further details.

4.4. The pseudoconservation law

As a consequence of Theorem 4.1 we have:

$$E\mathbf{V}_C = E\mathbf{V}_{Geom/G/1} + E\mathbf{Y}, \tag{4.17}$$

and hence, cf. (4.13) and (4.14),

$$\sum_{i=1}^{N} \rho_i E\mathbf{W}_i = \frac{\lambda\beta^{(2)}}{2(1-\rho)}\rho + \frac{(\lambda^{(2)}-\lambda^2-\lambda)\beta}{2\lambda(1-\rho)}\rho + E\mathbf{Y}. \tag{4.18}$$

We now derive an expression for $E\mathbf{Y}$, thus obtaining a general pseudoconservation law for the weighted sum of the mean waiting times at the various queues.

Let $E\mathbf{Y}_i$ denote the mean amount of work in the cyclic-service system at an arbitrary switching epoch during a switch-over from Q_i to Q_{i+1}. Obviously, $E\mathbf{Y} = \sum_{i=1}^{N} \frac{s_i}{s} E\mathbf{Y}_i$. As in the continuous-time case, $E\mathbf{Y}_i$ is composed of three terms:

1. $E\mathbf{M}_i^{(1)}$: the mean amount of work in Q_i at a departure epoch of the server from Q_i.
2. $E\mathbf{M}_i^{(2)}$: the mean amount of work in the rest of the system at a departure epoch of S from Q_i.
3. $\rho\{\frac{s_i^{(2)}}{2s_i} - \frac{1}{2}\}$: the mean amount of work that arrived in the system during the past part of the switching interval under consideration (cf. also Remark 4.3).

A similar derivation as in the continuous-time case yields (cf. also (3.20)):

$$\sum_{i=1}^{N}\frac{s_i}{s}E\mathbf{M}_i^{(2)} = \frac{\rho}{s}\sum\sum_{h<k}s_h s_k + \frac{s}{1-\rho}\sum\sum_{h<k}\rho_h\rho_k + \sum_{i=1}^{N}\frac{s_i}{s}\sum_{j\neq i}E\mathbf{M}_j^{(1)}, \quad (4.19)$$

and hence for $E\mathbf{Y}$:

$$E\mathbf{Y} = \rho\left(\frac{s^{(2)}}{2s} - \frac{1}{2}\right) + \frac{s}{2(1-\rho)}\left(\rho^2 - \sum_{i=1}^{N}\rho_i^2\right) + \sum_{j=1}^{N}E\mathbf{M}_j^{(1)}. \quad (4.20)$$

Consider a discrete-time cyclic-service system with switch-over times as described in Section 4.1. Recall that for $i=1,2,...,N$, $E\mathbf{M}_i^{(1)}$ denotes the mean amount of work in Q_i at a departure epoch of the server from this queue. Combining (4.18) and (4.20), we arrive at the main result of this chapter.

THEOREM 4.2 The Pseudoconservation Law for the Discrete-Time System
For any scheduling discipline satisfying Assumption 3.1:

$$\sum_{i=1}^{N}\rho_i E\mathbf{W}_i = \frac{\lambda\beta^{(2)}}{2(1-\rho)}\rho + \frac{(\lambda^{(2)}-\lambda^2-\lambda)\beta}{2\lambda(1-\rho)}\rho + \rho\frac{s^{(2)}}{2s} - \frac{1}{2}\rho +$$

$$\frac{s}{2(1-\rho)}\left(\rho^2 - \sum_{i=1}^{N}\rho_i^2\right) + \sum_{i=1}^{N}E\mathbf{M}_i^{(1)}. \quad (4.21)$$

Note that the form of Formula (4.21) is still independent of the visit disciplines at the various queues; only the $E\mathbf{M}_i^{(1)}$ depend on the choice of the visit disciplines. The terms in the right-hand side of (4.21) have a similar interpretation as in Section 3.3.

We shall now calculate $E\mathbf{Y}$ explicitly for a mixture of the exhaustive, gated, 1-limited and semi-exhaustive visit disciplines at the queues (cf. Section 3.4, visit disciplines (1) - (4); replace "customers" by "messages"). We have restricted ourself here to the main four visit disciplines exhaustive, gated, 1-limited and semi-exhaustive; they are most frequently encountered and thus sufficiently illustrate the argumentation used. The ergodicity conditions for the system are similar to the conditions in the continuous-time case. $\rho<1$ clearly is a necessary condition for ergodicity. For 1-limited service at Q_i, (3.23) is an additional condition for the stability of Q_i. For semi-exhaustive service at Q_i, (3.24) is an additional condition for the stability of Q_i.

The $E\mathbf{M}_i^{(1)}$ are again readily found for an exhaustive or gated visit discipline at Q_i:
Q_i exhaustive:

$$E\mathbf{M}_i^{(1)} = 0; \quad (4.22)$$

Q_i gated (cf. (4.9)):

$$E\mathbf{M}_i^{(1)} = \rho_i E\mathbf{V}_i = \rho_i^2 \frac{s}{1-\rho}. \tag{4.23}$$

Now suppose the visit discipline at Q_i is 1-limited. Denote by $\mathbf{T}_i$ the amount of work left behind at a departure epoch of a type-i message. Similar considerations as in Section 3.4 lead to (cf. also (3.28)):

$$E\mathbf{M}_i^{(1)} = \frac{\lambda_i s}{1-\rho} E\mathbf{T}_i. \tag{4.24}$$

To determine $E\mathbf{T}_i$, we calculate the mean number of packets left behind by a departing type-i message. Let $W_i(z)$ be the z-transform for the waiting time of an arbitrarily chosen type-i message (the tagged message); $E\mathbf{W}_i = W_i^{(1)}(1)$. Note that the messages left behind at station Q_i when the service of the tagged message has been completed are those which arrived during the sojourn time of the tagged message, and those which arrived in the same time slot as the tagged message but were placed behind the tagged message (the sojourn time is counted from the beginning of the time slot next to the one in which the arrival took place). The z-transform $Q_i(z)$ for the number of messages who arrived during the sojourn time of the tagged message is given by:

$$Q_i(z) = W_i(A_i(z))B_i(A_i(z)). \tag{4.25}$$

Next we introduce the z-transform $\tilde{Q}_i(z)$ for the number of messages that have arrived in the same time slot as the tagged message, but were placed behind the tagged message. An expression for $Q_i(z)$ can be obtained from discrete-time renewal theory: it is actually the equivalent of the backward recurrence time transform in continuous-time renewal theory. Hence we have,

$$\tilde{Q}_i(z) = \frac{1-A_i(z)}{\lambda_i(1-z)}. \tag{4.26}$$

The numbers of messages just mentioned are not independent, but we can still determine the first moment of their sum, i.e.,

$$Q_i^{(1)}(1) + \tilde{Q}_i^{(1)}(1), \tag{4.27}$$

where

$$Q_i^{(1)}(1) = \lambda_i(E\mathbf{W}_i + \beta_i), \tag{4.28}$$

$$\tilde{Q}_i^{(1)}(1) = \frac{\lambda_i^{(2)} - \lambda_i}{2\lambda_i}. \tag{4.29}$$

And so $E\mathbf{T}_i$, the mean amount of work left behind in Q_i at a departure epoch of a type-i message, equals:

$$E\mathbf{T}_i = \rho_i E\mathbf{W}_i + \rho_i \beta_i + \frac{\lambda_i^{(2)} - \lambda_i}{2\lambda_i} \beta_i. \tag{4.30}$$

From (4.24) and (4.30) we obtain:
For Q_i 1-limited (cf. also (3.30)):

$$E\mathbf{M}_i^{(1)} = \frac{\lambda_i s}{1-\rho} \rho_i E\mathbf{W}_i + \rho_i^2 \frac{s}{1-\rho} + \frac{\rho_i s}{1-\rho} \frac{\lambda_i^{(2)} - \lambda_i}{2\lambda_i}. \tag{4.31}$$

Finally, we consider semi-exhaustive service. With the above definition of $E\mathbf{T}_i$, (4.30) again holds. Denote by $\mathbf{U}_i$ the number of messages in Q_i at an arrival epoch of S at Q_i. Due to the structure of the SE discipline we can also write

$$E\mathbf{T}_i = \beta_i E[\mathbf{U}_i - 1 \mid \mathbf{U}_i \geqslant 1] + [\frac{\lambda_i^2 \beta_i^{(2)}}{2(1-\rho_i)} + \frac{(\lambda_i^{(2)} - \lambda_i^2 - \lambda_i)\beta_i}{2(1-\rho_i)} +$$

$$\rho_i + \frac{\lambda_i^{(2)} - \lambda_i}{2\lambda_i}]\beta_i. \tag{4.32}$$

Note that the second term in the right-hand side represents the mean amount of work left behind by a departing message in the Geom/G/1 queue with $A_i(z)$ and $B_i(z)$ respectively the z-transform of the number of message arrivals per time slot and the number of packets per message; the first three terms between square brackets represent the mean number of messages that have arrived during the sojourn time of the departing message (cf. (4.12) and Little's formula), and the fourth term is the mean number of messages that have arrived in the same time slot as this message, but were placed behind it, cf. (4.29). Subsequently express $E\mathbf{M}_i^{(1)}$ in the first term in the right-hand side of (4.32):

$$E\mathbf{M}_i^{(1)} = \beta_i E[max(0, \mathbf{U}_i - 1)] = \beta_i E[\mathbf{U}_i - 1 \mid \mathbf{U}_i \geqslant 1] Pr\{\mathbf{U}_i \geqslant 1\}. \tag{4.33}$$

Because the mean visit time of S at Q_i during a cycle, when positive, equals $\beta_i / (1-\rho_i)$ (the mean busy period of the Geom/G/1 system with mean number of arrivals per time slot λ_i and mean number of packets per message β_i), we have

$$E\mathbf{V}_i = \frac{\rho_i s}{1-\rho} = Pr\{\mathbf{U}_i \geqslant 1\} \frac{\beta_i}{1-\rho_i}, \tag{4.34}$$

so

$$Pr\{\mathbf{U}_i \geqslant 1\} = \frac{\lambda_i s(1-\rho_i)}{1-\rho}. \tag{4.35}$$

Combining (4.30), (4.32), (4.33) and (4.35) yields,

$$\rho_i E\mathbf{W}_i + \rho_i \beta_i = \frac{E\mathbf{M}_i^{(1)}}{\lambda_i s \frac{1-\rho_i}{1-\rho}} + [\frac{\lambda_i^2 \beta_i^{(2)}}{2(1-\rho_i)} + \frac{\lambda_i^{(2)} - \lambda_i^2 - \lambda_i}{2\lambda_i(1-\rho_i)}\rho_i + \rho_i]\beta_i, \tag{4.36}$$

and so we have (cf. also (3.36)):

$$E\mathbf{M}_i^{(1)} = \rho_i \frac{\lambda_i s(1-\rho_i)}{1-\rho} E\mathbf{W}_i - \frac{\lambda_i s(1-\rho_i)}{1-\rho}[\frac{\lambda_i \beta_i^{(2)}}{2(1-\rho_i)}\rho_i +$$

$$\frac{(\lambda_i^{(2)} - \lambda_i^2 - \lambda_i)}{2\lambda_i(1-\rho_i)}\rho_i \beta_i]. \tag{4.37}$$

Denote by

e: the group of Exhaustive queues,
g: the group of Gated queues,
1l: the group of 1-Limited queues,
se: the group of Semi Exhaustive queues.

Combining (4.21) and the above expressions for $E\mathbf{M}_i^{(1)}$, allows us to formulate the discrete-time pseudoconservation law for a mixture of the exhaustive, gated, 1-limited and semi-exhaustive visit disciplines:

$$\sum_{i \in e,g} \rho_i E\mathbf{W}_i + \sum_{i \in 1l} \rho_i [1 - \frac{\lambda_i s}{1-\rho}] E\mathbf{W}_i + \sum_{i \in se} \rho_i [1 - \frac{\lambda_i s(1-\rho_i)}{1-\rho}] E\mathbf{W}_i =$$

$$\frac{\lambda \beta^{(2)}}{2(1-\rho)}\rho + \frac{(\lambda^{(2)} - \lambda^2 - \lambda)\beta}{2\lambda(1-\rho)}\rho + \rho\frac{s^{(2)}}{2s} - \frac{1}{2}\rho + \frac{s}{2(1-\rho)}[\rho^2 - \sum_{\forall i}\rho_i^2] +$$

$$\frac{s}{1-\rho}\sum_{i \in g,\, 1l} \rho_i^2 - \frac{s}{2(1-\rho)}\sum_{i \in se} \lambda_i^2 \beta_i^{(2)} \rho_i + \frac{s}{1-\rho}\sum_{i \in 1l} \frac{\lambda_i^{(2)} - \lambda_i}{2\lambda_i}\rho_i -$$

$$\frac{s}{2(1-\rho)}\sum_{i \in se} (\lambda_i^{(2)} - \lambda_i^2 - \lambda_i)\beta_i \rho_i. \tag{4.38}$$

REMARK 4.4
The case with N = 1 yields expressions for mean waiting times in Geom/G/1 queues with some form of server vacations. In the completely symmetric case with all queues having identical characteristics and the same exhaustive (gated, 1-limited) visit discipline, Formula (4.38) reduces to Formula (3.63b) (resp. (5.23), (6.60)) of Takagi [1986].

REMARK 4.5
If we assume Poisson arrivals in (4.38) (and hence take $\lambda_i^{(2)} = \lambda_i^2 + \lambda_i$) we obtain the following relation for the weighted sum of the mean waiting times.

$$\sum_{i\in e,g} \rho_i E\mathbf{W}_i + \sum_{i\in 1l} \rho_i[1-\frac{\lambda_i s}{1-\rho}]E\mathbf{W}_i + \sum_{i\in se} \rho_i[1-\frac{\lambda_i s(1-\rho_i)}{1-\rho}]E\mathbf{W}_i =$$

$$\frac{\lambda\beta^{(2)}}{2(1-\rho)}\rho + \rho\frac{s^{(2)}}{2s} - \frac{1}{2}\rho + \frac{s}{2(1-\rho)}[\rho^2 - \sum_{\forall i}\rho_i^2] + \frac{s}{1-\rho}\sum_{i\in g,\,1l}\rho_i^2 -$$

$$\frac{s}{2(1-\rho)}\sum_{i\in se}\lambda_i^2\beta_i^{(2)}\rho_i + \frac{s}{2(1-\rho)}\sum_{i\in 1l}\lambda_i\rho_i. \tag{4.39}$$

4.5. RELATION TO THE CONTINUOUS-TIME CASE

In this chapter we have so far expressed all quantities involved, including waiting times, in time slots with the slot length equal to the time unit. If instead we assume a time slot to be of length Δ we are able, by taking the limit $\Delta \to 0$, to pass the results over to continuous time.

First we express the arrival process in messages per time unit. Recall that the z-transform of the number of message arrivals at Q_i in a time slot is given by $A_i(z)$, with first and second moment λ_i and $\lambda_i^{(2)}$ respectively. Denote by $\tilde{A}_i(z)$ the number of message arrivals at Q_i per time unit. Then

$$\tilde{A}_i(z) = [A_i(z)]^{1/\Delta}, \tag{4.40}$$

($1/\Delta$ is the number of time slots per time unit). From (4.40) we find:

$$\tilde{\lambda}_i = \frac{\lambda_i}{\Delta}, \qquad \tilde{\lambda}_i^{(2)} = \frac{\lambda_i^{(2)}}{\Delta} + \frac{1}{\Delta}(\frac{1}{\Delta}-1)\lambda_i^2. \tag{4.41}$$

For the service (switching) process let $\tilde{\beta}_i$, $\tilde{\beta}_i^{(2)}$ ($\tilde{s}_i$, $\tilde{s}_i^{(2)}$) denote the first and second moment respectively of the service (switching) time expressed in time units. It may be easily seen that,

$$\tilde{\beta}_i = \beta_i\Delta, \qquad \tilde{\beta}_i^{(2)} = \beta_i^{(2)}\Delta^2;$$

$$\tilde{s}_i = s_i\Delta, \qquad \tilde{s}_i^{(2)} = s_i^{(2)}\Delta^2. \tag{4.42}$$

Similarly, cf. (4.2), (4.4),

$$\tilde{\lambda} := \sum_{i=1}^{N}\tilde{\lambda}_i = \frac{\lambda}{\Delta}, \qquad \tilde{\lambda}^{(2)} := \tilde{\lambda}^2 + \sum_{i=1}^{N}(\tilde{\lambda}_i^{(2)} - \tilde{\lambda}_i^2), \tag{4.43}$$

hence,

$$\tilde{\lambda}^{(2)} - \tilde{\lambda}^2 - \tilde{\lambda} = \frac{1}{\Delta}(\lambda^{(2)} - \lambda^2 - \lambda);$$

furthermore,

$$\tilde{\beta} := \sum_{i=1}^{N}\frac{\tilde{\lambda}_i}{\tilde{\lambda}}\tilde{\beta}_i = \beta\Delta, \qquad \tilde{\beta}^{(2)} := \sum_{i=1}^{N}\frac{\tilde{\lambda}_i}{\tilde{\lambda}}\tilde{\beta}_i^{(2)} = \beta^{(2)}\Delta^2. \tag{4.44}$$

For the mean waiting time in time units, $E\tilde{\mathbf{W}}_i$, we have:

$$E\tilde{\mathbf{W}}_i = E\mathbf{W}_i\Delta. \tag{4.45}$$

Of course $\rho_i = \lambda_i\beta_i = \tilde{\lambda}_i\tilde{\beta}_i$. We can now express (4.38) in time units. With the slot length equal to Δ we obtain from (4.38) and (4.41)-(4.45):

$$\sum_{i\in e,g}\rho_i E\tilde{\mathbf{W}}_i\frac{1}{\Delta} + \sum_{i\in 1l}\rho_i[1-\frac{\tilde{\lambda}_i\tilde{s}}{1-\rho}]E\tilde{\mathbf{W}}_i\frac{1}{\Delta} + \sum_{i\in se}\rho_i[1-\frac{\tilde{\lambda}_i\tilde{s}(1-\rho_i)}{1-\rho}]E\tilde{\mathbf{W}}_i\frac{1}{\Delta} =$$

$$\frac{\tilde{\lambda}\tilde{\beta}^{(2)}}{2(1-\rho)}\rho\frac{1}{\Delta} + \frac{(\tilde{\lambda}^{(2)}-\tilde{\lambda}^2-\tilde{\lambda})\tilde{\beta}}{2\tilde{\lambda}(1-\rho)}\rho\frac{1}{\Delta} + \rho\frac{\tilde{s}^{(2)}}{2\tilde{s}}\frac{1}{\Delta} - \tfrac{1}{2}\rho +$$

$$\frac{\tilde{s}}{2(1-\rho)}[\rho^2 - \sum_{\forall i}\rho_i^2]\frac{1}{\Delta} + \frac{\tilde{s}}{1-\rho}\sum_{i\in g,\,1l}\rho_i^2\frac{1}{\Delta} - \frac{\tilde{s}}{2(1-\rho)}\sum_{i\in se}\tilde{\lambda}_i^2\tilde{\beta}_i^{(2)}\rho_i\frac{1}{\Delta} +$$

$$\frac{\tilde{s}}{1-\rho}\sum_{i\in 1l}\frac{\tilde{\lambda}_i^{(2)} - (1-\Delta)\tilde{\lambda}_i^2 - \tilde{\lambda}_i}{2\tilde{\lambda}_i}\rho_i\frac{1}{\Delta} - \frac{\tilde{s}}{2(1-\rho)}\sum_{i\in se}(\tilde{\lambda}_i^{(2)} - \tilde{\lambda}_i^2 - \tilde{\lambda}_i)\tilde{\beta}_i\rho_i\frac{1}{\Delta}. \tag{4.46}$$

In (4.46) we can take the limit for $\Delta \to 0$ by multiplying the left- and right-hand side with Δ and substituting $\Delta = 0$. If we do so, we obtain

$$\sum_{i\in e,g}\rho_i E\tilde{\mathbf{W}}_i + \sum_{i\in 1l}\rho_i[1-\frac{\tilde{\lambda}_i\tilde{s}}{1-\rho}]E\tilde{\mathbf{W}}_i + \sum_{i\in se}\rho_i[1-\frac{\tilde{\lambda}_i\tilde{s}(1-\rho_i)}{1-\rho}]E\tilde{\mathbf{W}}_i =$$

$$\frac{\tilde{\lambda}\tilde{\beta}^{(2)}}{2(1-\rho)}\rho + \frac{(\tilde{\lambda}^{(2)}-\tilde{\lambda}^2-\tilde{\lambda})\tilde{\beta}}{2\tilde{\lambda}(1-\rho)}\rho + \rho\frac{\tilde{s}^{(2)}}{2\tilde{s}} + \frac{\tilde{s}}{2(1-\rho)}[\rho^2 - \sum_{\forall i}\rho_i^2] +$$

$$\frac{\tilde{s}}{1-\rho}\sum_{i\in g,\,1l}\rho_i^2 - \frac{\tilde{s}}{2(1-\rho)}\sum_{i\in se}\tilde{\lambda}_i^2\tilde{\beta}_i^{(2)}\rho_i + \frac{\tilde{s}}{(1-\rho)}\sum_{i\in 1l}\frac{\tilde{\lambda}_i^{(2)}-\tilde{\lambda}_i^2-\tilde{\lambda}_i}{2\tilde{\lambda}_i}\rho_i -$$

$$\frac{\tilde{s}}{2(1-\rho)}\sum_{i\in se}(\tilde{\lambda}_i^{(2)}-\tilde{\lambda}_i^2-\tilde{\lambda}_i)\tilde{\beta}_i\rho_i. \tag{4.47}$$

At this point some remarks are in order. To obtain Formula (4.47) it is not necessary to specify precisely how the above limit $\Delta \to 0$ is taken. However, as we will explain in the following, the structure of the resulting arrival process does depend on it. Let us take a closer look at the arrival process. As has been noted in Section 4.1, the message arrival process at Q_i is a Bernoulli process with batch arrivals. We have a Bernoulli arrival process in the sense that

Pr{ type-i batch arrives in a time slot } $= 1 - A_i(0)$,

Pr{ type-i batch does not arrive in a time slot } $= A_i(0)$.

With respect to the batch arrivals let $G_i(z)$ denote the z-transform of the size of a type-i batch. Then we can write $A_i(z)$ as:

$$A_i(z) = A_i(0) + [1-A_i(0)]G_i(z), \tag{4.48}$$

and hence, with (4.40):

$$\tilde{A}_i(z) = (A_i(0) + [1-A_i(0)]G_i(z))^{1/\Delta} \tag{4.49}$$

$$= (1 - [\frac{1-A_i(0)}{\Delta} - \frac{1-A_i(0)}{\Delta}G_i(z)]\Delta)^{1/\Delta}.$$

Let,

$$\gamma_i := \frac{1-A_i(0)}{\Delta};$$

γ_i denotes the arrival intensity of type-i batches. Note that γ_i is also equal to $\lambda_i / G_i'(1)$. Now if in (4.49) we let $\Delta \to 0$ in such a way that γ_i remains constant, the z-transform for the number of message arrivals per time unit at Q_i becomes:

$$\tilde{A}_i(z) = e^{\gamma_i(G_i(z)-1)}, \tag{4.50}$$

which is the z-transform of a compound Poisson process. If we take $G_i(z) = z$

(single arrivals) we obtain the z-transform of the 'ordinary' Poisson process; in this case $\tilde{\lambda}_i^{(2)} = \tilde{\lambda}_i^2 + \tilde{\lambda}_i$, and (4.47) reduces to the pseudoconservation law for single Poisson arrivals in continuous time, a special case of Formula (3.44):

$$\sum_{i\in e,g} \rho_i E\tilde{\mathbf{W}}_i + \sum_{i\in 1l} \rho_i[1-\frac{\tilde{\lambda}_i\tilde{s}}{1-\rho}]E\tilde{\mathbf{W}}_i + \sum_{i\in se} \rho_i[1-\frac{\tilde{\lambda}_i\tilde{s}(1-\rho_i)}{1-\rho}]E\tilde{\mathbf{W}}_i =$$

$$\frac{\tilde{\lambda}\tilde{\beta}^{(2)}}{2(1-\rho)}\rho + \rho\frac{\tilde{s}^{(2)}}{2\tilde{s}} + \frac{\tilde{s}}{2(1-\rho)}[\rho^2 - \sum_{\forall i}\rho_i^2] + \frac{\tilde{s}}{1-\rho}\sum_{i\in g,\,1l}\rho_i^2 -$$

$$\frac{\tilde{s}}{2(1-\rho)}\sum_{i\in se}\tilde{\lambda}_i^2\tilde{\beta}_i^{(2)}\rho_i. \tag{4.51}$$

Chapter 5

CYCLIC-SERVICE SYSTEMS WITH A POLLING TABLE

5.1 Introduction

A large number of studies has been devoted to the queueing analysis of polling systems. However, the vast majority of these studies considers polling systems in which the server attends to the queues in a *fixed strictly cyclic* order. This extensive research on cyclic-service systems has been useful for performance evaluation. However, it has not yet led to a clear ability to control the systems under consideration and to affect their design. Recent advances of computer and communication technology make the use of more sophisticated scheduling and visit disciplines feasible and thus open possibilities for (optimal) control of complex networks. The control of such a system may be effectuated by introducing a *polling table* prescribing the order in which the queues are to be visited by the server.

This chapter is devoted to cyclic-service systems with a polling table. An extension of the work decomposition property of Chapter 2 leads to the main result of this chapter, the pseudoconservation law for mean waiting times. Although an analysis of optimal control and optimization of polling systems will not be carried out in this study, this result may serve as a starting point.

Remark 5.1

Very few studies have appeared which consider optimization of polling systems. Levy [1988] studies a cyclic-service system with a binomial-gated service discipline (cf. Section 3.4) at the queues. This leads to a mathematically tractable model in which the choice of binomial probabilities of numbers of customers served at the queues changes the quality of service. Browne and Yechiali [1989a,b] present a semi-dynamic polling policy with a finite horizon of one cycle: after each visit of the server to a queue the next queue to be visited is chosen so as to minimize some objective function. They assume global availability of local information and so their results can not be applied to distributed systems. Baker and Rubin [1987] derive the mean waiting times in a multi-queue single-server system with the server visiting the N queues according to a polling table of length $M \geqslant N$ and serves each queue exhaustively. Stations are given higher priority by listing them more often in the polling table. Eisenberg [1971] first derived a complete solution for an exhaustive polling

system with a general polling table. However, his method required the solution of a set of M^3 quite complex simultaneous equations in order to derive the mean waiting times. The approach of Baker and Rubin results in a set of $M(M-1)$ simultaneous equations, which can be solved recursively by adding M equations.

The rest of this chapter is organized as follows. Section 5.2 contains a model description and some preparatory results concerning cycle times and visit times; it also presents the extension of the work decomposition result of Chapter 2 to single-server, multi-queue systems with a polling table. Section 5.3 is devoted to the derivation of the pseudoconservation law. Finally, the special case of polling in a so-called star network is discussed in detail in Section 5.4; the mean workload in the star system is compared with the mean workload in a corresponding network with strictly cyclic service order. As in the previous chapter our model formulation is in discrete-time. Continuous-time results are easily obtained via a limiting procedure.

5.2. Model description and preliminary results

As in the previous chapter, we consider a discrete-time queueing system with N stations (queues) $Q_1,\ldots,Q_N$; each station has an infinite buffer capacity to store waiting messages (customers). Each message consists of a number of packets; packets are assumed to be of fixed length. Time is slotted with slot size equal to the transmission time of one packet (the service time of a packet). As before, we shall call the time interval $[j,j+1[$ the jth time slot.

Message arrival process

Let $\mathbf{X}_n(j)$ denote the number of messages arriving at station n in the jth time slot. The $\mathbf{X}_n(j)$, $j=1,2,\ldots$ are assumed to be independent, identically distributed random variables. Put

$$A_n(z) := E[z^{\mathbf{X}_n(j)}], \quad |z| \leqslant 1; \quad \lambda_n := E[\mathbf{X}_n(j)]; \quad \lambda_n^{(2)} := E[\mathbf{X}_n^2(j)], \tag{5.1}$$

and

$$\lambda := \sum_{n=1}^{N} \lambda_n, \quad \lambda^{(2)} := E[(\sum_{n=1}^{N} \mathbf{X}_n(j))^2]. \tag{5.2}$$

The message arrival process at each station is assumed to be independent of those at other stations.

Service process

Denote by $\mathbf{B}_n$ the number of packets included in a message at station n. It is assumed that $\mathbf{B}_n \in 1,2,\ldots$. We define

$$B_n(z) := E[z^{\mathbf{B}_n}], \quad |z| \leqslant 1; \quad \beta_n := E[\mathbf{B}_n]; \quad \beta_n^{(2)} := E[\mathbf{B}_n^2], \tag{5.3}$$

and

$$\beta := \sum_{n=1}^{N} \frac{\lambda_n}{\lambda} \beta_n, \quad \beta^{(2)} := \sum_{n=1}^{N} \frac{\lambda_n}{\lambda} \beta_n^{(2)}. \tag{5.4}$$

The $\mathbf{B}_n$, $n = 1,...,N$ are assumed to be independent of $\mathbf{X}_n(j)$, $n = 1,...,N$, $j = 1,2,3,...$.

The offered traffic at the nth station ρ_n is defined by

$$\rho_n := \lambda_n \beta_n, \quad n = 1,2,...,N, \tag{5.5}$$

and the total offered traffic ρ by

$$\rho := \sum_{n=1}^{N} \rho_n. \tag{5.6}$$

Polling strategy
The N stations are served by a single server S. The order in which S visits the stations is specified by a *polling table* $T = \{T(m), m = 1,...,M\}$. The first entry in the polling table, $T(1)$, is the index of the first station polled in a cycle, $T(2)$ the index of the second, etc. After station $T(M)$ is polled, the next cycle starts with $T(1)$. The figure below pictures the relation between the order in which the queues are to be visited by the server and the polling table T.

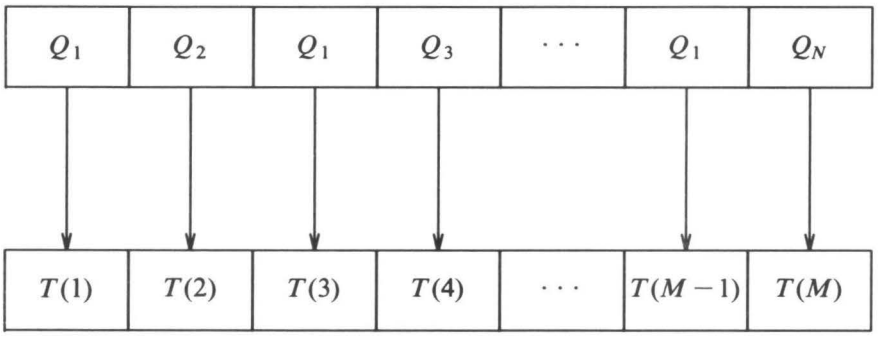

Figure 5.1. Relation of the visiting order of the queues and the polling table

Following the approach of Baker and Rubin [1987], a unique *pseudostation* will be associated with each entry in the polling table; as a result of this, the M pseudostations are visited in a strictly cyclic order. Denote by PS_m the pseudostation associated with the mth entry in the polling table; its

corresponding station has index $T(m)$ (as much as possible, we reserve n as an index for stations and m as an index for pseudostations). For simplicity of notation, all references to station indices and pseudostation indices are implicitly assumed to be modulo N and M respectively. We shall say that 'PS_i is connected with PS_j', if $T(i)=T(j)$, that is, if PS_i and PS_j correspond to the same station.

On a few occasions, we shall have need to specify the exact position of work. Suppose PS_j is the first pseudostation after PS_i that is connected with PS_i. By convention, the work in PS_i is shifted to PS_j immediately before S arrives at PS_j.

Visit discipline

For the visit disciplines at the pseudostations we consider three possibilities. Assume that S visits PS_m. When PS_m (and hence $Q_{T(m)}$) is empty, S immediately begins to switch to PS_{m+1}. Otherwise, depending on the visit discipline at PS_m:

1. Exhaustive service (E): messages are transmitted from PS_m until PS_m is empty;
2. Gated service (G): only messages present in PS_m upon arrival of S at PS_m are transmitted;
3. 1-Limited service ($1-L$): exactly one message is transmitted from PS_m.

To carry out an exact analysis, we impose the restriction that stations with a 1-limited visit discipline are served only once during a cycle.

In the sequel we will allow mixed visit disciplines (e.g., exhaustive at PS_1, 1-limited at PS_2 and PS_4, gated at PS_3, etc.). It is assumed that pseudostations corresponding to the same station have the same visit discipline, but this assumption is not essential for the analysis.

REMARK 5.2

We have restricted ourself here to the three main visit disciplines in polling systems. For the other visit disciplines mentioned in Section 3.4 the derivation is similar and therefore omitted. The derivation of the pseudoconservation law in Section 5.3 will, however, illustrate clearly the essential arguments and indicate how they should be adapted.

Switching process

A switch-over time is needed to switch from one pseudostation to the next. The switch-over times of the server between the mth and the $(m+1)$th pseudostation (measured in time slots) are independent, identically distributed random variables with first moment s_m and second moment $s_m^{(2)}$. The first moment s of the total switch-over time during a cycle of the server is given by

$$s := \sum_{m=1}^{M} s_m, \tag{5.7}$$

its second moment is indicated by $s^{(2)}$. The message arrival process, the service process and the switching processes are assumed to be mutually independent processes.

For ease of notation we define a 'cyclic sum' as

$$\sum_{m=i}^{j\bullet} x_m := \begin{cases} \sum_{m=i}^{j} x_m, & \text{if } i \leqslant j \\ \sum_{m=i}^{M} x_m + \sum_{m=1}^{j} x_m, & \text{if } i > j \end{cases} \tag{5.8a}$$

and, analogously, a 'cyclic product' as

$$\prod_{m=i}^{j\bullet} x_m := \begin{cases} \prod_{m=i}^{j} x_m, & \text{if } i \leqslant j \\ \prod_{m=i}^{M} x_m \times \prod_{m=1}^{j} x_m, & \text{if } i > j \end{cases} \tag{5.8b}$$

Preparatory results
We conclude this section by stating a few results for future reference.

Ergodicity conditions
The ergodicity conditions are the same as for the strictly cyclic service system. Hence, a necessary condition for ergodicity of the system is $\rho<1$. When the visit discipline at each queue is either exhaustive or gated, this condition is also sufficient. However, for each queue Q_n with a 1-limited visit discipline an additional ergodicity condition is needed, viz.,

$$\rho + \lambda_n s < 1, \tag{5.9}$$

see also Chapter 3. As usual it will be assumed that the ergodicity conditions are satisfied and that the system is in steady state.

Cycle- and visit-time results
For the (strictly) cyclic service system with the M pseudostations we can define the cycle time $\mathbf{C}_m$ for PS_m as the time between two successive arrivals of the server at PS_m. It is easily seen that $E\mathbf{C}_m$ is independent of m and from a similar balancing argument as in Section 3.3 it follows that the mean cycle time equals $E\mathbf{C}$ with

$$E\mathbf{C} = \frac{s}{1-\rho}. \tag{5.10}$$

Furthermore we define the visit time $\mathbf{V}_m$ of the server for PS_m as the time between the arrival of the server at PS_m and his subsequent departure from PS_m. The mean visit times play an important role in the waiting-time analysis in the next section. To calculate them, we must first introduce the binary $M \times M$ matrix $H = (h_{ij})$ (this matrix is the transposed of H in Baker and Rubin [1987]), where

$$h_{ij} := \min\{1, \prod_{m=i}^{j-1}{}^{\bullet} |T(j) - T(m)|\}. \tag{5.11}$$

Note that, for $i \neq j$, h_{ij} equals 1 iff $PS_i, \ldots, PS_{j-1}$ are not connected with the same pseudostation as PS_j.

EXAMPLE 5.1
Suppose $N = 3$ and $T = [1,2,1,3]$. Then

$$H = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

EXAMPLE 5.2
In the strictly cyclic case $h_{ij} = 1$ for all i,j, $i \neq j$.

To calculate the mean visit times for a pseudostation we distinguish between the cases that PS_m has a 1-limited, exhaustive or gated visit discipline. Our analysis follows Baker and Rubin [1987].

1. *PS_m has a 1-limited visit discipline.* By assumption PS_m is only visited once per cycle. Hence, balancing the flow of customers at PS_m in and out of the system during a cycle of the server shows that

$$\lambda_{T(m)} E\mathbf{C} = \frac{E\mathbf{V}_m}{\beta_{T(m)}}, \tag{5.12}$$

and so from (5.10):

$$E\mathbf{V}_m = \rho_{T(m)} \frac{s}{1-\rho}. \tag{5.13}$$

2. *PS_m has an exhaustive discipline.* The fact that stations with an exhaustive discipline may be served more than once during a cycle, complicates the determination of the mean visit times for their corresponding pseudostations. Note however, that we can write $E\mathbf{V}_m$ as the mean number of

messages found by the server upon his arrival at PS_m, multiplied by the mean length of a busy period started by one message. When the server arrives at PS_m, messages have accumulated during the interval

$$\sum_{i=m+1}^{m-1}{}^{\bullet} h_{im}(s_{i-1}+E\mathbf{V}_i)+s_{m-1}. \tag{5.14}$$

(5.14) represents the mean time (measured in time slots) between the arrival of S at PS_m and the departure of S from the last pseudostation before PS_m which is connected with PS_m. From (5.14) and the fact that the mean length of a busy period started by one message is $\beta_{T(m)}/(1-\rho_{T(m)})$, we obtain for the mean visit time at PS_m (note that $h_{mm}=0$ by definition):

$$E\mathbf{V}_m = \lambda_{T(m)}[\sum_{i=1}^{M} h_{im}(s_{i-1}+E\mathbf{V}_i)+s_{m-1}]\frac{\beta_{T(m)}}{1-\rho_{T(m)}}. \tag{5.15}$$

3. *PS_m has a gated service discipline.* In this case, we can write $E\mathbf{V}_m$ as the mean number of messages found by the server upon his arrival at PS_m, multiplied by the mean transmission time of a message at PS_m. When the server arrives at PS_m, messages have accumulated during the interval

$$\sum_{i=m+1}^{m-1}{}^{\bullet} h_{im}(E\mathbf{V}_{i-1}+s_{i-1})+E\mathbf{V}_{m-1}+s_{m-1}. \tag{5.16}$$

So for the mean visit times at the gated pseudostations we obtain:

$$E\mathbf{V}_m = \lambda_{T(m)}[\sum_{i=1}^{M} h_{im}(E\mathbf{V}_{i-1}+s_{i-1})+E\mathbf{V}_{m-1}+s_{m-1}]\beta_{T(m)}. \tag{5.17}$$

Balancing the flow of messages at Q_n in and out of the system during a cycle shows that

$$\sum_{\{m\,|\,T(m)=n\}} \frac{E\mathbf{V}_m}{\rho_{T(m)}} = E\mathbf{C} = \frac{s}{1-\rho}. \tag{5.18}$$

Note that (5.15), (5.17) and (5.18) represent a set of $M-N$ simultaneous linear equations for the mean visit times of the server at the pseudostations with a gated or exhaustive visit discipline.

Work decomposition

Suppose for the moment that there are no switch-over times. Then the principle of work conservation implies that the amount of work in the polling system equals the amount of work in a 'corresponding' Geom/G/1 queueing system

with batch arrivals; this system is the discrete-time queueing model for one queue and one server with a Bernoulli arrival process of batches introduced in Chapter 4.

As before, an *arbitrary epoch* is supposed to be the instant just after the beginning of an arbitrary time slot. Define $\mathbf{V}_{Geom/G/1}$ as the amount of work at an arbitrary epoch in this corresponding Geom/G/1 system with batch arrivals. The mean number of messages in this system at an arbitrary epoch is given by (cf. (4.12))

$$E\mathbf{X}_{Geom/G/1} = \frac{\lambda^2\beta^{(2)}}{2(1-\rho)} + \frac{(\lambda^{(2)}-\lambda^2-\lambda)\beta}{2(1-\rho)} + \rho. \tag{5.19}$$

Note that the second term in the right-hand side disappears when the number of messages arriving per slot has a Poisson distribution. The mean number of messages in service is ρ; the mean residual service time of the message in service is $\beta^{(2)}/2\beta + 1/2$ (cf. Remark 4.3). Hence, (cf. also (4.13))

$$E\mathbf{V}_{Geom/G/1} = \left[\frac{\lambda^2\beta^{(2)}}{2(1-\rho)} + \frac{(\lambda^{(2)}-\lambda^2-\lambda)\beta}{2(1-\rho)}\right]\beta + \rho\left[\frac{\beta^{(2)}}{2\beta} + \frac{1}{2}\right]. \tag{5.20}$$

Since there are switch-over times incorporated in the original system, the server may be idle (switching) although there is work in the system. Hence the principle of work conservation can not be applied. However, for the service system with a polling table, just as for systems with a strictly cyclic polling strategy (cf. Chapter 2), there appears to exist a natural extension of the work conservation principle. The amount of work in the system can be decomposed into the amount of work in the corresponding Geom/G/1 queue and an extra quantity. This decomposition is presented in Theorem 5.1 below.

Consider a single-server multi-queue system with a polling table as described in the beginning of this section. We define:

$\mathbf{V}_P$: the amount of work in the system with polling table,

$\mathbf{V}_{Geom/G/1}$: the amount of work in the 'corresponding' Geom/G/1 system,

$\mathbf{Y}$: the amount of work in the system with polling table at an arbitrary epoch in a switching interval.

Note that an arbitrary epoch is considered to be 'in' a switching interval if it marks the beginning of a switching slot.

THEOREM 5.1

The amount of work in the system with polling table is distributed as the sum of the amount of work in the 'corresponding' Geom/G/1 system with batch arrivals and the amount of work in the system with polling table at an arbitrary epoch in a switching interval:

$$\mathbf{V}_P \overset{D}{=} \mathbf{V}_{Geom/G/1} + \mathbf{Y}. \tag{5.21}$$

Furthermore, $\mathbf{V}_{Geom/G/1}$ *and* $\mathbf{Y}$ *are independent.*

PROOF:
The proof is analogous to that of Theorem 2.2 and is therefore omitted here. □

5.3. THE PSEUDOCONSERVATION LAW
As a consequence of Theorem 5.1:

$$E\mathbf{V}_P = E\mathbf{V}_{Geom/G/1} + E\mathbf{Y}, \tag{5.22}$$

and hence, cf. (5.20):

$$E\mathbf{V}_P = \frac{\lambda\beta^{(2)}}{2(1-\rho)}\rho + \frac{(\lambda^{(2)}-\lambda^2-\lambda)\beta}{2\lambda(1-\rho)}\rho + \rho\{\frac{\beta^{(2)}}{2\beta} + \frac{1}{2}\} + E\mathbf{Y}. \tag{5.23}$$

On the other hand, we can write $E\mathbf{V}_P$ as:

$$\begin{aligned} E\mathbf{V}_P &= \sum_{n=1}^{N}\beta_n E\mathbf{X}_n^w + \sum_{n=1}^{N}\rho_n\{\frac{\beta_n^{(2)}}{2\beta_n} + \frac{1}{2}\} \\ &= \sum_{n=1}^{N}\rho_n E\mathbf{W}_n + \sum_{n=1}^{N}\rho_n\{\frac{\beta_n^{(2)}}{2\beta_n} + \frac{1}{2}\}, \end{aligned} \tag{5.24}$$

where $\mathbf{X}_n^w$ denotes the number of waiting type-n messages at an arbitrary epoch, and $\mathbf{W}_n$ the waiting time of a type-n message; the waiting time is counted from the beginning of the time slot following the one in which the message arrived. The second equality is based on Little's formula. From (5.23) and (5.24) we obtain the following expression for a weighted sum of the mean message waiting times:

$$\sum_{n=1}^{N}\rho_n E\mathbf{W}_n = \frac{\lambda\beta^{(2)}}{2(1-\rho)}\rho + \frac{(\lambda^{(2)}-\lambda^2-\lambda)\beta}{2\lambda(1-\rho)}\rho + E\mathbf{Y}. \tag{5.25}$$

Note that in the derivation of (5.25) the notion of pseudostations has played no role. Only the last term in (5.25) depends on this notion. To obtain an expression for the weighted sum of the mean message waiting times at the various queues, we now derive an expression for $E\mathbf{Y}$, the mean amount of work in the modified system of pseudostations at an arbitrary epoch in a switching interval. Let $\mathbf{Y}_m$ denote the amount of work in the cyclic-service system at an arbitrary switching epoch during a switch-over from PS_m to PS_{m+1}. Obviously,

$$E\mathbf{Y} = \sum_{m=1}^{M} \frac{s_m}{s} E\mathbf{Y}_m. \tag{5.26}$$

As in Chapters 3 and 4, $E\mathbf{Y}_m$ is composed of three terms:

1. $E\mathbf{M}_m^{(1)}$: the mean amount of work in PS_m at a departure epoch of the server from PS_m.
2. $E\mathbf{M}_m^{(2)}$: the mean amount of work in the rest of the system at a departure epoch of S from PS_m.
3. $\rho\{\frac{s_m^{(2)}}{2s_m} - \frac{1}{2}\}$: the mean amount of work that arrived in the system during the past part of the switching interval under consideration.

It will turn out that only $E\mathbf{M}_m^{(1)}$ depends on the choice of the visit disciplines at the various (pseudo)stations. To calculate $E\mathbf{M}_m^{(2)}$, we must introduce the $M \times M$ (0-1) matrix $Z = (z_{ij})$, where

$$z_{ij} := \min\{1, \prod_{k=i+1}^{j\bullet} |T(i) - T(k)|\}. \tag{5.27}$$

Note that for $i \neq j$, $z_{ij} = 1$ iff $PS_{i+1}, \ldots, PS_j$ are not connected with PS_i. The differences in the matrices Z and H (cf. (5.11)) result from the fact that in the matrix Z one is looking 'forward' in a cycle, whereas in the matrix H one is looking 'backward'.

EXAMPLE 5.3
Suppose $N = 3$ and $T = [1,2,1,3]$. Then

$$Z = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

EXAMPLE 5.4
In the strictly cyclic case, $z_{ij} = 1$ for all i,j, $i \neq j$.

We shall now consider $E\mathbf{M}_m^{(2)}$, the mean amount of work in $PS_{m-1}, \ldots, PS_1, PS_M, \ldots, PS_{m+1}$ at a departure epoch of the server from PS_m. PS_k can make two contributions to $E\mathbf{M}_m^{(2)}$, viz. (i) the mean amount of work $E\mathbf{M}_k^{(1)}$ left behind in PS_k by S, and (ii) the mean amount of work that has arrived in $Q_{T(k)}$ during the switch-over times from PS_k to PS_m and the visit times of $PS_{k+1}, \ldots, PS_m$. Both contributions disappear when any of the pseudostations $PS_{k+1}, \ldots, PS_m$ is connected with PS_k, i.e. when $z_{km} = 0$ (but

not when PS_k is connected with, say, PS_{m+1}; cf. the convention introduced in Section 5.2).

Hence we have:

$$E\mathbf{M}_m^{(2)} = \sum_{k \neq m} z_{km} E\mathbf{M}_k^{(1)} + \sum_{k \neq m} z_{km} \rho_{T(k)} \sum_{j=k}^{m-1\bullet} (s_j + E\mathbf{V}_{j+1}). \tag{5.28}$$

Still leaving $E\mathbf{M}_m^{(1)}$ unspecified, we obtain the following expression for $E\mathbf{Y}$ from (5.26) and (5.28) (note that $z_{mm}=0$ by definition):

$$E\mathbf{Y} = \sum_{m=1}^{M} \frac{s_m}{s} [E\mathbf{M}_m^{(1)} + \sum_{k=1}^{M} z_{km} E\mathbf{M}_k^{(1)}] + \sum_{k=1}^{M} \rho_{T(k)} \sum_{m=1}^{M} \frac{s_m}{s} z_{km} \sum_{j=k}^{m-1\bullet} (s_j + E\mathbf{V}_{j+1}) +$$

$$\rho \sum_{m=1}^{M} \frac{s_m^{(2)}}{2s} - \frac{1}{2}\rho. \tag{5.29}$$

Finally, from (5.25) and (5.29):

$$\sum_{n=1}^{N} \rho_n E\mathbf{W}_n = \frac{\lambda\beta^{(2)}}{2(1-\rho)}\rho + \frac{(\lambda^{(2)} - \lambda^2 - \lambda)\beta}{2\lambda(1-\rho)}\rho + \sum_{m=1}^{M} \frac{s_m}{s} [E\mathbf{M}_m^{(1)} + \sum_{k=1}^{M} z_{km} E\mathbf{M}_k^{(1)}]$$

$$+ \sum_{k=1}^{M} \rho_{T(k)} \sum_{m=1}^{M} \frac{s_m}{s} z_{km} \sum_{j=k}^{m-1\bullet} (s_j + E\mathbf{V}_{j+1}) + \rho \sum_{m=1}^{M} \frac{s_m^{(2)}}{2s} - \frac{1}{2}\rho. \tag{5.30}$$

Note that the form of Formula (5.30) is still independent of the visit disciplines at the various pseudostations. Only the $E\mathbf{M}_m^{(1)}$ and $E\mathbf{V}_m$ depend on the choice of the visit disciplines.

The $E\mathbf{M}_m^{(1)}$ are readily found for an exhaustive, gated or 1-limited discipline at PS_m:

if PS_m has an exhaustive visit discipline, then

$$E\mathbf{M}_m^{(1)} = 0. \tag{5.31}$$

if PS_m has a gated visit discipline, then

$$E\mathbf{M}_m^{(1)} = \rho_{T(m)} E\mathbf{V}_m, \tag{5.32}$$

with $E\mathbf{V}_m$ as determined by (5.17).

if PS_m has a 1-limited visit discipline, then a similar derivation as in Chapter 4 leads to

$$E\mathbf{M}_m^{(1)} = \frac{\lambda_{T(m)} s}{1-\rho} \rho_{T(m)} E\mathbf{W}_{T(m)} + \rho_{T(m)}^2 \frac{s}{1-\rho} + \frac{\rho_{T(m)} s}{1-\rho} \frac{\lambda_{T(m)}^{(2)} - \lambda_{T(m)}}{2\lambda_{T(m)}}. \tag{5.33}$$

Substitution of (5.31), (5.32) and (5.33) into (5.30) gives our main result, which is formulated in Theorem 5.2 below. Denote by

e: the group of Exhaustive stations,
g: the group of Gated stations,
$\tilde{g}$ the group of Gated *pseudo*stations,
1l: the group of 1-Limited stations.

Consider a single-server system with a polling table and mixed visit disciplines as described in Section 5.2.

THEOREM 5.2 The Pseudoconservation Law for Polling Tables

$$\sum_{n\in e,g} \rho_n E\mathbf{W}_n + \sum_{n\in 1l} \rho_n[1-\frac{\lambda_n s}{1-\rho}]E\mathbf{W}_n = \frac{\lambda\beta^{(2)}}{2(1-\rho)}\rho + \frac{(\lambda^{(2)}-\lambda^2-\lambda)\beta}{2\lambda(1-\rho)}\rho +$$

$$\rho\sum_{m=1}^{M}\frac{s_m^{(2)}}{2s} - \frac{1}{2}\rho + \frac{s}{1-\rho}\sum_{n\in 1l}(\rho_n^2 + \rho_n\frac{\lambda_n^{(2)}-\lambda_n}{2\lambda_n}) + \sum_{m\in\tilde{g}}\rho_{T(m)}\frac{s_m}{s}E\mathbf{V}_m +$$

$$\sum_{k=1}^{M}\rho_{T(k)}\sum_{m=1}^{M}\frac{s_m}{s}z_{km}\sum_{j=k}^{m-1\bullet}(s_j+E\mathbf{V}_{j+1}) + \sum_{j\in\tilde{g}}\rho_{T(j)}E\mathbf{V}_j\sum_{m=1}^{M}\frac{s_m}{s}z_{jm}. \tag{5.34}$$

Note that for this complex polling system the right-hand side of (5.34) can be easily evaluated for each given set of parameter values, polling order and visit disciplines.

SPECIAL CASES

(i) When each station is polled only once during a cycle, Theorem 5.2 becomes a special case of Theorem 4.2.

(ii) If the numbers of message arrivals at the queues per time slot are independent Poisson distributed random variables, then $\lambda_n^{(2)}=\lambda_n^2+\lambda_n$ and $\lambda^{(2)}=\lambda^2+\lambda$; this leads to minor simplifications in the right-hand side of (5.34).

(iii) If messages arrive in batches with γ_n the arrival rate of a batch at Q_n and $G_n(z)$ the generating function of the batch size, and if the numbers of batch arrivals at the queues are independent Poisson distributed random variables, then $\lambda_n^{(2)}-\lambda_n^2-\lambda_n=\gamma_n G_n^{(2)}(1)$ and $\lambda^{(2)}-\lambda^2-\lambda=\sum_{n=1}^{N}\gamma_n G_n^{(2)}(1)$.

REMARK 5.3

Giannakouros and Laloux [1988] also derive a pseudoconservation law for a polling system with polling table. However, they do not specify the $E\mathbf{M}_m^{(1)}$ terms for particular visit disciplines. In their model, for a station occurring more than once in the table, the work in each of its pseudostations can only arrive during part of a cycle, and is not shifted to other connected

pseudostations (cf. the convention introduced in Section 5.2).

THE CONTINUOUS-TIME CASE

In this chapter we have expressed all quantities involved, including waiting times, in time slots with the slot length equal to the time unit. If instead we assume a time slot to be of length Δ we are able, by taking the limit $\Delta \to 0$, to pass the results over to continuous time.

We first translate (5.34), using a tilde to indicate that a quantity is expressed in time units instead of time slots. Introduce

$$\tilde{\lambda}_n = \lambda_n/\Delta, \quad \tilde{\lambda} = \lambda/\Delta, \quad \tilde{\beta}_n = \beta_n\Delta, \quad \tilde{\beta} = \beta\Delta,$$

$$\tilde{s}_n = s_n\Delta, \quad \tilde{s} = s\Delta, \quad E\tilde{\mathbf{W}}_n = E\mathbf{W}_n\Delta;$$

furthermore, cf. Boxma and Groenendijk [1988a],

$$\tilde{\lambda}_n^{(2)} = \frac{\lambda_n^{(2)}}{\Delta} + \frac{1}{\Delta}\left(\frac{1}{\Delta} - 1\right)\lambda_n^2, \quad \tilde{\lambda}^{(2)} - \tilde{\lambda}^2 = \sum_{n=1}^{N}(\lambda_n^{(2)} - \lambda_n^2)/\Delta,$$

$$\tilde{\beta}_n^{(2)} = \beta_n^{(2)}\Delta^2, \quad \tilde{\beta}^{(2)} = \beta^{(2)}\Delta^2, \quad \tilde{s}_n^{(2)} = s_n^{(2)}\Delta^2.$$

Of course $\rho_n = \lambda_n\beta_n = \tilde{\lambda}_n\tilde{\beta}_n$. Formula (5.34) now becomes:

$$\sum_{n\in e,g} \rho_n E\tilde{\mathbf{W}}_n \frac{1}{\Delta} + \sum_{n\in 1l} \rho_n\left[1 - \frac{\tilde{\lambda}_n\tilde{s}}{1-\rho}\right]E\tilde{\mathbf{W}}_n \frac{1}{\Delta} = \frac{\tilde{\lambda}\tilde{\beta}^{(2)}}{2(1-\rho)}\rho\frac{1}{\Delta} +$$

$$\frac{(\tilde{\lambda}^{(2)} - \tilde{\lambda}^2 - \tilde{\lambda})\tilde{\beta}}{2\tilde{\lambda}(1-\rho)}\rho\frac{1}{\Delta} + \rho\frac{\sum_{m=1}^{M}\tilde{s}_m^{(2)}}{2\tilde{s}}\frac{1}{\Delta} - \frac{1}{2}\rho + \sum_{m\in\tilde{g}} \rho_{T(m)}\frac{\tilde{s}_m}{\tilde{s}}E\tilde{\mathbf{V}}_m\frac{1}{\Delta} +$$

$$\frac{\tilde{s}}{1-\rho}\sum_{n\in 1l}\left(\rho_n^2 + \rho_n\frac{\tilde{\lambda}_n^{(2)} - (1-\Delta)\tilde{\lambda}_n^2 - \tilde{\lambda}_n}{2\tilde{\lambda}_n}\right)\frac{1}{\Delta} +$$

$$\sum_{k=1}^{M}\rho_{T(k)}\sum_{m=1}^{M}\frac{\tilde{s}_m}{\tilde{s}}z_{km}\sum_{j=k}^{m-1\bullet}(\tilde{s}_j + E\tilde{\mathbf{V}}_{j+1})\frac{1}{\Delta} + \sum_{j\in\tilde{g}}\rho_{T(j)}E\tilde{\mathbf{V}}_j\sum_{m=1}^{M}\frac{\tilde{s}_m}{\tilde{s}}z_{jm}\frac{1}{\Delta}. \tag{5.35}$$

In (5.35) we can take the limit for $\Delta \to 0$ by multiplying the left- and right-

hand side with Δ and substituting $\Delta = 0$. If we do so, we obtain

$$\sum_{n\in e,g} \rho_n E\tilde{\mathbf{W}}_n + \sum_{n\in 1l} \rho_n [1-\frac{\tilde{\lambda}_n \tilde{s}}{1-\rho}] E\tilde{\mathbf{W}}_n = \frac{\tilde{\lambda}\tilde{\beta}^{(2)}}{2(1-\rho)}\rho + \frac{(\tilde{\lambda}^{(2)}-\tilde{\lambda}^2-\tilde{\lambda})\tilde{\beta}}{2\tilde{\lambda}(1-\rho)}\rho +$$

$$\rho\frac{\sum_{m=1}^{M}\tilde{s}_m^{(2)}}{2\tilde{s}} + \frac{\tilde{s}}{(1-\rho)}\sum_{n\in 1l}(\rho_n^2 + \rho_n\frac{\tilde{\lambda}_n^{(2)}-\tilde{\lambda}_n^2-\tilde{\lambda}_n}{2\tilde{\lambda}_n}) + \sum_{m\in\tilde{g}} \rho_{T(m)}\frac{\tilde{s}_m}{\tilde{s}} E\tilde{\mathbf{V}}_m +$$

$$\sum_{k=1}^{M}\rho_{T(k)}\sum_{m=1}^{M}\frac{\tilde{s}_m}{\tilde{s}} z_{km}\sum_{j=k}^{m-1}{}^{\bullet}(\tilde{s}_j + E\tilde{\mathbf{V}}_{j+1}) + \sum_{j\in\tilde{g}}\rho_{T(j)}E\tilde{\mathbf{V}}_j\sum_{m=1}^{M}\frac{\tilde{s}_m}{\tilde{s}} z_{jm}. \tag{5.36}$$

Formula (5.36) is a pseudoconservation law of the same type as in Chapter 4.

5.4. Example: the star network

In this section we evaluate the pseudoconservation law (5.34) for a network with a star configuration, and we make a comparison with the network with corresponding stations and strictly cyclic service order. A polling network with a star configuration represents, e.g., a computer with multidrop terminals in which the computer, after polling a terminal, transmits its outbound traffic and then polls the next terminal. Two cases are considered. In Case A the central station Q_1 receives exhaustive service, whereas in Case B it receives gated service; in both cases $Q_2, \ldots, Q_N$, $N \geqslant 2$, receive 1-limited service. It will be shown below that in both cases the mean workload in the star system is smaller than the mean workload in the corresponding cyclic-service system.

The polling table is: $T=[1,2,1,3,\ldots,1,N]$. There are $M=2(N-1)$ pseudo-stations. Denote by

$\tilde{e}=\{1,3,\ldots,2N-3\}$: *the group of exhaustive pseudostations (Case A),*

$\tilde{g}=\{1,3,\ldots,2N-3\}$: *the group of gated pseudostations (Case B),*

$\tilde{1}\tilde{l}=\{2,4,\ldots,2N-2\}$: *the group of 1-limited pseudostations.*

Definition (5.27) of the matrix Z implies that

when $i\in\tilde{e}$ $(\tilde{g})$: $z_{i,i+1}=1$; $z_{ij}=0$ otherwise;

when $i\in\tilde{1}\tilde{l}$: $z_{ij}\equiv 1$, $i\neq j$.

Case A: Q_1 exhaustive service

Introducing

$$C := \frac{\lambda\beta^{(2)}}{2(1-\rho)}\rho + \frac{(\lambda^{(2)}-\lambda^2-\lambda)\beta}{2\lambda(1-\rho)}\rho - \frac{1}{2}\rho +$$

$$\frac{s}{1-\rho}\sum_{n\in 1l}(\rho_n^2 + \rho_n\frac{\lambda_n^{(2)}-\lambda_n}{2\lambda_n}), \tag{5.37}$$

and substituting the z_{ij} values calculated above into (5.34) gives the pseudoconservation law for a star network in discrete time ($\mathbf{W}_n^{star}$ denotes the waiting time at Q_n in the star network), viz.

$$\rho_1 E\mathbf{W}_1^{star} + \sum_{n\in 1l}\rho_n[1-\frac{\lambda_n s}{1-\rho}]E\mathbf{W}_n^{star} = C + \rho\sum_{m=1}^{M}\frac{s_m^{(2)}}{2s} +$$

$$\rho_1\sum_{k\in\tilde{e}}\frac{s_{k+1}}{s}(s_k + E\mathbf{V}_{k+1}) + \sum_{k\in\tilde{1}l}\rho_{T(k)}\sum_{m\neq k}\frac{s_m}{s}\sum_{j=k}^{m-1\bullet}(s_j + E\mathbf{V}_{j+1}). \tag{5.38}$$

The mean visit times $E\mathbf{V}_k$ can be specified using (5.13) and (5.15):

$$k\in\tilde{1}l: \quad E\mathbf{V}_k = \rho_{T(k)}\frac{s}{(1-\rho)}; \tag{5.39}$$

$$k\in\tilde{e}: \quad E\mathbf{V}_k = \frac{\rho_1}{1-\rho_1}[s_{k-1} + s_{k-2} + E\mathbf{V}_{k-1}]$$

$$= \frac{\rho_1}{1-\rho_1}[s_{k-1} + s_{k-2} + \rho_{T(k-1)}\frac{s}{1-\rho}]. \tag{5.40}$$

Note that the sum of the visit times in (5.40) satisfies (5.18). To simplify the following calculations, it is assumed that all M switch-over times are equal to the constant r. Substitution of (5.39) and (5.40) into (5.38) yields, after a tedious but straightforward calculation:

$$\rho_1 E\mathbf{W}_1^{star} + \sum_{n\in 1l}\rho_n[1-\frac{\lambda_n Mr}{1-\rho}]E\mathbf{W}_n^{star} =$$

$$C + \rho_1 r + \frac{1}{2}(\rho-\rho_1)Mr + \rho_1\frac{\rho-\rho_1}{1-\rho}r + \frac{1}{2}\rho_1\frac{\rho-\rho_1}{1-\rho_1}Mr +$$

$$\frac{\rho_1}{(1-\rho)(1-\rho_1)}(M-1)r\sum_{n=2}^{N}\rho_n^2 + \frac{\rho_1}{(1-\rho)(1-\rho_1)}(M-2)r\sum_{j=2}^{N-1}\sum_{i=j+1}^{N}\rho_i\rho_j +$$

$$\frac{1}{1-\rho} Mr \sum_{j=2}^{N-1} \sum_{i=j+1}^{N} \rho_i \rho_j. \tag{5.41}$$

We now compare the star network with the 'corresponding' strictly cyclic service system; this is a system with N queues $Q_1, \ldots, Q_N$ with cyclic service in this order, where Q_1 receives exhaustive service and $Q_2, \ldots, Q_N$ receive 1-limited service, and where each queue has exactly the same traffic characteristics as its counterpart in the star network. Furthermore, the total switch-over times in both systems correspond; so in the strictly cyclic system, $s = Mr$ and $s^{(2)} = M^2 r^2$.

A comparison between the mean workloads in both models amounts to a comparison between the expressions in the right-hand sides of the pseudoconservation laws for both models. The pseudoconservation law for the 'corresponding' strictly cyclic service system reads in this case (cf. Chapter 4; $\mathbf{W}_n^{cycl}$ denotes the waiting time at Q_n in the cyclic network):

$$\rho_1 E\mathbf{W}_1^{cycl} + \sum_{n \in 1l} \rho_n [1 - \frac{\lambda_n Mr}{1-\rho}] E\mathbf{W}_n^{cycl} =$$

$$C + \tfrac{1}{2}\rho Mr + \frac{Mr}{2(1-\rho)} [\rho^2 - \sum_{n=1}^{N} \rho_n^2]. \tag{5.42}$$

Subtract the right-hand side of (5.41) from the right-hand side of (5.42), and call the difference $Diff_E$. Note that, with an obvious notation,

$$Diff_E = E\mathbf{V}^{cycl} - E\mathbf{V}^{star} = E\mathbf{Y}^{cycl} - E\mathbf{Y}^{star}.$$

From (5.41) and (5.42),

$$Diff_E = \rho_1 (N-2) r [1 + \frac{\rho - \rho_1}{1-\rho_1}] + \frac{\rho_1}{(1-\rho)(1-\rho_1)} (2N-2) r \sum_{j=2}^{N-1} \sum_{i=j+1}^{N} \rho_i \rho_j$$

$$\geqslant 0. \tag{5.43}$$

This result leads to the following observations:

- The mean workload in the star system is at most equal to the mean workload in the corresponding cyclic system, and the difference increases roughly linearly in N.
- $Diff_E = 0$ when $N = 2$, and when $\rho_1 = 0$; indeed, in those cases the two systems coincide.
- $Diff_E$ approaches zero when $r \to 0$.

- $Diff_E$ depends on λ_n and β_n only via their product ρ_n.
- When $\rho=\rho_1$, the star and cyclic systems reduce to vacation queues with vacation $2r$ respectively Mr. A standard decomposition result for vacation queues (cf. Fuhrmann and Cooper [1985]) learns that the workload difference equals $\rho_1 \frac{1}{2} Mr - \rho_1 \frac{1}{2}(2r)$; this is confirmed by (5.43).

In our model of star polling, Q_1 is served exhaustively N times during a cycle. Generally speaking, exhaustive service minimizes workload in polling systems with switch-over times, cf. Takagi [1986]; therefore it is not surprising that $Diff_E \geqslant 0$, and that the difference increases roughly linearly in N.

REMARK 5.4
As in Manfield [1985], the mean waiting time for the exhaustive station can be explicitly calculated. Assume that all 1-limited stations in the network have the same traffic characteristics and all switch-over times are equal to r; then

$$E\mathbf{W}_1^{star} = \frac{\lambda\beta^{(2)}}{2(1-\rho_1)} + \frac{\lambda_1^{(2)}-\lambda_1^2-\lambda_1}{2\lambda_1(1-\rho_1)}\beta_1 + \frac{(\rho-\rho_1)r}{1-\rho_1} + (N-1)\frac{1-\rho_1}{1-\rho}r + \frac{1}{2}.$$

We can substitute this result into (5.41); since in this case all mean waiting times at the 1-limited queues are equal, we obtain the exact mean waiting time at the 1-limited queues.

Case B: Q_1 gated service
In this case the pseudoconservation law reduces to (cf. (5.38)):

$$\rho_1 E\mathbf{W}_1^{star} + \sum_{n\in 1l} \rho_n [1-\frac{\lambda_n s}{1-\rho}] E\mathbf{W}_n^{star} =$$

$$C + \rho \sum_{m=1}^{M} \frac{s_m^{(2)}}{2s} + \rho_1 \sum_{k\in\tilde{g}}^{\bullet} \frac{s_{k+1}}{s}(s_k + E\mathbf{V}_{k+1}) +$$

$$\sum_{k\in 1l} \rho_{T(k)} \sum_{m\neq k} \frac{s_m}{s} \sum_{j=k}^{m-1}{}^{\bullet} (s_j + E\mathbf{V}_{j+1}) + \rho_1 \sum_{j\in\tilde{g}} E\mathbf{V}_j \sum_{m=1}^{M} \frac{s_m}{s} z_{jm} + \rho_1 \sum_{m\in\tilde{g}} \frac{s_m}{s} E\mathbf{V}_m. \quad (5.44)$$

The mean visit times for the 1-limited pseudostations are again given by (5.39). It follows from (5.17) that

$$k\in\tilde{g}: \quad E\mathbf{V}_k = \rho_1 [s_{k-1} + E\mathbf{V}_{k-1} + \sum_{j=1}^{M} h_{jk}(s_{j-1} + E\mathbf{V}_{j-1})] =$$

$$\rho_1 [s_{k-2} + s_{k-1} + E\mathbf{V}_{k-2} + E\mathbf{V}_{k-1}] =$$

$$\rho_1[s_{k-2}+s_{k-1}+\rho_{T(k-1)}\frac{s}{1-\rho}] + \rho_1 E\mathbf{V}_{k-2} =$$

$$\rho_1[s_{k-2}+s_{k-1}+\rho_{T(k-1)}\frac{s}{1-\rho}] + \rho_1^2[s_{k-4}+s_{k-3}+\rho_{T(k-3)}\frac{s}{1-\rho}] +$$

$$\cdots + \rho_1^{\frac{1}{2}M}[s_k+s_{k+1}+\rho_{T(k+1)}\frac{s}{1-\rho}] + \rho_1^{\frac{1}{2}M} E\mathbf{V}_k, \tag{5.45}$$

leading to an explicit expression for $E\mathbf{V}_k$.
To simplify the calculations it is again assumed that all M switch-over times are equal to the constant r. Formula (5.45) now reduces to:

$$k \in \tilde{g}: \quad E\mathbf{V}_k = \frac{\rho_1}{1-\rho_1}2r + \frac{\rho_1}{1-\rho}R_k Mr, \tag{5.46}$$

with

$$R_k := \frac{1}{1-\rho_1^{\frac{1}{2}M}}[\rho_{T(k-1)}+\rho_1\rho_{T(k-3)}+\rho_1^2\rho_{T(k-5)}+\cdots+\rho_1^{\frac{1}{2}M-1}\rho_{T(k+1)}].$$

Substitution of (5.39) and (5.46) into (5.44) yields:

$$\rho_1 E\mathbf{W}_1^{star} + \sum_{n\in 1l}\rho_n[1-\frac{\lambda_n Mr}{1-\rho}]E\mathbf{W}_n^{star} = C + \rho_1 r + \tfrac{1}{2}(\rho-\rho_1)Mr +$$

$$\rho_1\frac{\rho-\rho_1}{1-\rho}r + 2\frac{\rho_1^2}{1-\rho}r + \tfrac{1}{2}\rho_1\frac{\rho-\rho_1}{1-\rho_1}Mr + \frac{1}{1-\rho}Mr\sum_{j=2}^{N-1}\sum_{i=j+1}^{N}\rho_i\rho_j +$$

$$\frac{\rho_1}{1-\rho}r\sum_{k\in\tilde{1l}}\rho_{T(k)}[(M-1)R_{k+1}+(M-3)R_{k+3}+\cdots+R_{k-1}]. \tag{5.47}$$

Again we make a comparison with the corresponding strictly cyclic service system. The pseudoconservation law reads in this case (cf. Chapter 4, and compare with (5.42)):

$$\rho_1 E\mathbf{W}_1^{cycl} + \sum_{n\in 1l}\rho_n[1-\frac{\lambda_n Mr}{1-\rho}]E\mathbf{W}_n^{cycl} =$$

$$C + \tfrac{1}{2}\rho Mr + \frac{Mr}{2(1-\rho)}[\rho^2 - \sum_{n=1}^{N}\rho_n^2] + \frac{\rho_1^2}{1-\rho}Mr. \tag{5.48}$$

Subtract the right-hand side of (5.47) from the right-hand side of (5.48), and call the difference $Diff_G$. Then

$$Diff_G = \rho_1(N-2)r[1+\frac{\rho-\rho_1}{1-\rho_1}] + \frac{\rho_1(\rho-\rho_1)^2}{(1-\rho)(1-\rho_1)}(2N-3)r + \frac{\rho_1^2}{1-\rho}2(N-2)r$$

$$- \frac{\rho_1}{1-\rho}r\sum_{k\in\tilde{1}l}\rho_{T(k)}[(M-1)R_{k+1}+(M-3)R_{k+3}+\cdots+R_{k-1}]. \qquad (5.49)$$

This result leads to similar observations as the ones for Case A (cf. below (5.43)). To see that $Diff_G \geqslant 0$, consider the last term, LT, in the right-hand side of (5.49). The coefficient of $\rho^2_{T(k)}$ in LT is:

$$\frac{\rho_1}{1-\rho}r\frac{1}{1-\rho_1^{\frac{1}{2}M}}[(M-1)+(M-3)\rho_1+\cdots+3\rho_1^{\frac{1}{2}M-2}+\rho_1^{\frac{1}{2}M-1}] \leqslant$$

$$\frac{\rho_1}{1-\rho}r\frac{1}{1-\rho_1^{\frac{1}{2}M}}(M-1)[1+\rho_1+\ldots+\rho_1^{\frac{1}{2}M-2}+\rho_1^{\frac{1}{2}M-1}] =$$

$$\frac{\rho_1}{(1-\rho)(1-\rho_1)}(M-1)r.$$

The coefficient of $\rho_{T(k)}\rho_{T(k-2)}$ in LT is:

$$\frac{\rho_1}{1-\rho}r\frac{1}{1-\rho_1^{\frac{1}{2}M}}([(M-1)\rho_1+(M-3)\rho_1^2+\cdots+3\rho_1^{\frac{1}{2}M-1}+1] +$$

$$[(M-1)\rho_1^{\frac{1}{2}M-1}+(M-3)+\ldots+3\rho_1^{\frac{1}{2}M-3}+\rho_1^{\frac{1}{2}M-2}]) \leqslant$$

$$\frac{\rho_1}{(1-\rho)(1-\rho_1)}2(M-1)r.$$

Similarly for the other products. Summing all the upper bounds yields:

$$LT \leqslant \frac{\rho_1}{(1-\rho)(1-\rho_1)}[\rho_{T(2)}+\ldots+\rho_{T(M)}]^2(M-1)r =$$

$$\frac{\rho_1}{(1-\rho)(1-\rho_1)}(\rho-\rho_1)^2(M-1)r.$$

Hence

$$Diff_G \geqslant \rho_1(N-2)r[1+\frac{\rho-\rho_1}{1-\rho_1}] + \frac{\rho_1^2}{1-\rho}2(N-2)r \geqslant 0.$$

REMARK 5.5
Translation of the results of this section to the continuous-time case is almost immediate. In particular, the expressions for $Diff_E$ and $Diff_G$ are not affected.

REMARK 5.6
In the symmetric case $\rho_2 = \cdots = \rho_N = (\rho-\rho_1)/(N-1)$, the formulas (5.43) for $Diff_E$ and (5.49) for $Diff_G$ become very simple:

$$Diff_E = (N-2)r\rho_1\frac{1-\rho_1}{1-\rho}, \tag{5.50}$$

and

$$Diff_G = (N-2)r\rho_1\frac{1+\rho_1}{1-\rho}. \tag{5.51}$$

We have also evaluated (5.34) for a network with scan polling (polling table $T=[1,2,...,N-1,N,N,N-1,...,2,1]$, cf. also Coffman and Hofri [1982], Takagi and Murata [1986]), and we have again compared the result with the network with corresponding stations and strictly cyclic service. Because of the complexity of the calculations, we have restricted ourself to the case of exhaustive service at all stations, constant switch-over times r between all pseudostations in the scan network, and equal traffic intensities at all stations: $\rho_1 = \cdots = \rho_N = \rho/N$. If the switch-over times in the cyclic system equal $2r$ (so that the mean cycle times in both systems are the same), then

$$Diff = E\mathbf{V}^{cycl} - E\mathbf{V}^{scan} = \frac{\rho r}{2} + \frac{\rho r(N-1)}{6(1-\rho)}\frac{2N-3\rho-1}{N-\rho} > 0; \tag{5.52}$$

this is not surprising, as the queues in the scan system are visited twice as often as in the cyclic system. However, it seems more realistic to choose the switch-over times in the cyclic system equal to r, just as in the scan system; then

$$Diff = E\mathbf{V}^{cycl} - E\mathbf{V}^{scan} = -\frac{\rho r(N-1)}{6(1-\rho)}\frac{N+1}{N-\rho} \leqslant 0. \tag{5.53}$$

Again, we might have expected this, because of the inefficient visiting pattern of scan polling.

Chapter 6

EXACT RESULTS FOR SOME TWO-QUEUE MODELS WITH 1-LIMITED SERVICE AT ONE QUEUE

6.1 Introduction

In this chapter some two-queue models with 1-limited service at one of the queues are analyzed exactly. Polling models consisting of two queues have received considerable attention in the literature. For an extensive survey with special emphasis on the mathematical techniques, cf. Boxma [1986]. The study of polling models with two queues is important, since they represent the simplest possible polling models which at the same time reflect the interaction between queues and yet can still be analyzed exactly in several cases. Furthermore, the results may yield considerable insight into the behavior of more general cyclic-service systems and may be of help in devising and testing approximations. Below, a brief overview is presented on some of the available literature on two-queue polling models. We restrict ourself to the exhaustive, gated, 1-limited and semi-exhaustive visit disciplines.

Exhaustive service at both queues.

This model is usually referred to as the *alternating-priority model*. For the case of *zero* switch-over times, cf. for instance Takács [1968]; for the case with *nonzero* switch-over times, cf. Sykes [1970]. In Eisenberg [1971] a similar model is studied; when both queues are empty, in Eisenberg's model the server remains at the queue last served, whereas in Sykes's model the server keeps switching. In Hofri [1986] an interesting extension to the alternating-priority model is presented in which the server when ready at one queue only switches to the other queue if the queue length there exceeds a certain threshold value. This latter model opens up certain possibilities for optimization.

Gated service at both queues.

The analysis of the system with gated service at both queues is very similar to that of the alternating-priority model. The gated service system with N asymmetric queues *without* switch-over times is considered in Cooper and Murray [1969] and Cooper [1970]. For the case *with* switch-over times, cf. for instance Ferguson and Aminetzah [1985].

Exhaustive service at one queue, gated at the other.
A system with a mixture of exhaustive and gated visit disciplines is analyzed in Ozawa [1987] for two queues, and by Takagi [1989] for a general number of queues.

1-limited service at both queues.
This model is often referred to as the *alternating-service model.* The model with two queues and no switch-over times has first been tackled in an important study of Eisenberg [1972]. Eisenberg transformed the problem of determining the joint queue-length distribution at the two queues into the problem of solving a singular Fredholm integral equation. As an example of the study of 2-dimensional queueing systems, Cohen and Boxma [1981,1983 Sect. III.2 and IV.1] applied methods as developed in Cohen and Boxma [1983] to provide a detailed analysis of the alternating-service model *without* switch-over times. In Boxma [1985] the analysis was extended to the case *with* switch-over times of the server between queues under the restriction that both queues have identical characteristics.

In Section 6.2 of this chapter the asymmetric alternating-service model with nonzero switch-over times is studied. This analysis has appeared before in Boxma and Groenendijk [1988b]. It is shown, that the joint stationary queue-length distribution at the instants at which the server becomes available to a queue can be determined via transformation to a Riemann boundary value problem. The latter problem can be completely solved for general service- and switch-over time distributions.

Exhaustive service at one queue, 1-limited service at the other.
The model with exhaustive service at one queue and 1-limited service at the other queue turns out to be considerably simpler than the alternating-service model. The model was first solved by Skinner [1967], assuming a constant switch-over time from the exhaustive service queue to the limited service queue. The case of generally distributed switch-over times is treated in Section 6.3 of this chapter; it was first published in Groenendijk [1989].

Semi-exhaustive service at both queues.
The model with semi-exhaustive service at both queues and nonzero switch-over times was first studied by Takagi [1985]. For a symmetric system, Takagi obtained the first moment of the waiting times at the queues. The model was subsequently analyzed by Cohen [1988a] for the asymmetric case, following a similar approach as for the alternating-service model and obtaining the joint stationary queue-length distributions.

Several interesting variants of the model with two parallel queues have been considered, most of which fall outside the scope of this monograph. We shall mention only a few. Single-server priority models with two classes of customers form such an example. These models have been extensively studied in the past, cf. Jaiswal [1968] and Cohen [1982] and references contained therein. Murata and Takagi [1987] present one of the few priority studies in which

nonzero switch-over times are assumed. Another example is the 'priority for the longer queue' model: after each service completion, the server chooses the first customer from the longer of two queues. This model is studied in Cohen [1987].

6.2. Two queues with alternating service and switch-over times

In this section we give an exact analysis of a system of two queues, attended by a single server who alternately serves one customer of each queue (if not empty). The server incurs a non-zero switch-over time when switch-over between the queues. This system is in the sequel referred to as the 'alternating-service model'.

The organization of this section is as follows. In Section 6.2.1 the model is reformulated. Section 6.2.2 contains the main analysis. The joint stationary queue-length distribution at the instants at which the server becomes available to serve a queue is determined as the solution of a homogeneous Riemann boundary value problem on the unit circle. This boundary value problem (BVP) can be formulated as follows:

Suppose $G(t)$ is a function on the unit circle C which does not vanish on C and which satisfies a so-called Hölder condition on C (this latter condition implies that we can find two positive constants A and μ with $0<\mu\leqslant 1$ such that for every two points t_1 and t_2 on C, $|G(t_1)-G(t_2)| \leqslant A|t_1-t_2|^{\mu}$). It is required to determine two functions $F_1(z)$ and $F_2(z)$, such that:

- $F_1(z)$ is regular *inside* C and continuous on C,
- $F_2(z)$ is regular *outside* C (including $z=\infty$) and continuous on C,
- $F_1(t) = G(t)F_2(t)$ for $t \in C$.

See Cohen and Boxma [1983], Cohen [1988b] for a short exposition and see Gakhov [1966] for a detailed discussion.

Once the joint queue-length distribution is known, one can easily derive expressions for various important performance measures like waiting times of customers and cycle times of the server (the time between two successive arrivals of the server at a particular queue). Waiting times are studied in Section 6.2.3, with particular attention for the mean waiting times; cycle times are studied in Section 6.2.4, with particular attention to second moments of the cycle times (first moments of cycle times are trivially determined). Section 6.2.5 is devoted to a numerical evaluation of some important performance measures of the alternating-service model. It is shown that the boundary value problem formulation leads to formulas which can be numerically evaluated in a straightforward manner. A profound study of the numerical results leads to new insight into the behavior of the alternating-service model and of more general cyclic-service models. For instance, while first moments of cycle times do not depend on the number of the queue at which the cycle starts (see Section 3.3), it is seen from the numerical results that second moments differ only slightly. This observation supports the assumption of equality of the second moments of the cycle times in the derivation of the approximation discussed in

6.2.1. Model description

A single server S serves two queues Q_1, Q_2 (with infinite buffer capacities) in cyclic order. The arrival process of customers at Q_i is a Poisson process with rate λ_i, $i=1,2$. The service times at Q_i are independent, identically distributed stochastic variables with distribution $B_i(\cdot)$, with first moment β_i, second moment $\beta_i^{(2)}$ and LST (Laplace-Stieltjes Transform) $\beta_i(\cdot)$. The various arrival and service processes are independent.

The utilization ρ_i at Q_i is defined by

$$\rho_i := \lambda_i \beta_i, \quad i=1,2, \tag{6.1}$$

the total utilization ρ of the server is defined by

$$\rho := \rho_1 + \rho_2. \tag{6.2}$$

The visit discipline at both queues is 1-limited: the server serves at most one customer at a queue before switching to the next queue. The successive switch-over times from Q_i to $Q_{(i+1) \bmod 2}$ are independent, identically distributed stochastic variables, also independent of the service times, with distribution $S_i(\cdot)$. Their first moment, second moment and LST are respectively denoted by s_i, $s_i^{(2)}$ and $\sigma_i(\cdot)$.

Let $\mathbf{C}_i$ denote the time between two successive arrivals of S at Q_i, the cycle time for Q_i. Clearly each cycle consists of two switches and at most one service at each of the two queues. The first and second moments of the total switch-over time during one cycle are denoted by, respectively,

$$s := s_1 + s_2, \tag{6.3}$$

and

$$s^{(2)} := s_1^{(2)} + 2s_1 s_2 + s_2^{(2)}. \tag{6.4}$$

As in Section 3.3 it is easily seen that the *mean cycle time* $E\mathbf{C}_i$ is independent of i and is given by

$$E\mathbf{C}_i = \frac{s}{1-\rho}. \tag{6.5}$$

Note that $E\mathbf{C}_i = 0$ when $s_1 = s_2 = 0$. In that case, when the system becomes empty, the server cycles infinitely often; each of such cycles has length zero. In the sequel we assume that

$$\rho + \max(\lambda_1 s, \lambda_2 s) < 1; \tag{6.6}$$

as usual the system is assumed to be in steady state.

6.2.2. Formulation and solution of the boundary value problem

Let $\mathbf{X}_i^{(1)}$, $i=1,2$, denote the number of customers at Q_i at those instants at which S arrives at Q_1; similarly for $\mathbf{X}_i^{(2)}$. Let

$$F_j(z_1,z_2) := E[z_1^{\mathbf{X}_1^{(j)}} z_2^{\mathbf{X}_2^{(j)}}], \quad |z_1| \leqslant 1, \ |z_2| \leqslant 1, \ j=1,2. \tag{6.7}$$

Our goal in this subsection is to determine $F_1(z_1,z_2)$ and $F_2(z_1,z_2)$. This goal will be accomplished by formulating and solving a so-called Riemann boundary value problem (cf. Cohen and Boxma [1983]). Once $F_i(z_1,z_2)$ is determined, the LST of the waiting-time distribution at Q_i and that of the distribution of the cycle time $\mathbf{C}_i$ can be obtained. In Sections 6.2.3 and 6.2.4 we will demonstrate this. In particular expressions for the mean waiting times and second moments of cycle times will be given.

The vector $(\mathbf{X}_1^{(j)},\mathbf{X}_2^{(j)})$, $j=1,2,\ldots$ of queue lengths at Q_1 and Q_2 at successive arrival epochs of server S at those queues is a vector Markov chain; this is a direct consequence of the various assumptions. Standard arguments (see, e.g., Boxma [1985]) lead to the following recurrence relations for the generating functions $F_1(z_1,z_2)$ and $F_2(z_1,z_2)$: for $|z_1| \leqslant 1$, $|z_2| \leqslant 1$,

$$F_2(z_1,z_2) = \{[F_1(z_1,z_2) - F_1(0,z_2)]\, z_1^{-1} \beta_1(x) + F_1(0,z_2)\}\, \sigma_1(x), \tag{6.8}$$

$$F_1(z_1,z_2) = \{[F_2(z_1,z_2) - F_2(z_1,0)]\, z_2^{-1} \beta_2(x) + F_2(z_1,0)\}\, \sigma_2(x), \tag{6.9}$$

with for $|z_1| \leqslant 1$, $|z_2| \leqslant 1$,

$$x := \lambda_1(1-z_1) + \lambda_2(1-z_2). \tag{6.10}$$

Note that $F_1(1,1) = F_2(1,1) = 1$. Substitution of (6.8) into (6.9) yields: for $|z_1| \leqslant 1$, $|z_2| \leqslant 1$,

$$K(z_1,z_2)F_1(z_1,z_2) = F_1(0,z_2)\{\beta_2(x)\, \sigma_1(x)\, \sigma_2(x)\, (z_1 - \beta_1(x))\} + F_2(z_1,0)\{z_1\, \sigma_2(x)\, (z_2 - \beta_2(x))\}, \tag{6.11}$$

and analogously,

$$K(z_1,z_2)F_2(z_1,z_2) = F_2(z_1,0)\{\beta_1(x)\, \sigma_1(x)\, \sigma_2(x)\, (z_2 - \beta_2(x))\} + F_1(0,z_2)\{z_2\, \sigma_1(x)\, (z_1 - \beta_1(x))\}; \tag{6.12}$$

here $K(z_1,z_2)$ is the *kernel of the functional equation,* defined as

$$K(z_1,z_2) := z_1z_2 - \beta_1(x)\,\beta_2(x)\,\sigma_1(x)\,\sigma_2(x), \quad |z_1| \leqslant 1,\ |z_2| \leqslant 1. \tag{6.13}$$

Relation (6.11) is the starting point for our analysis. The main idea is similar to that in Cohen and Boxma [1981,1983] for the model without switch-over times and in Boxma [1985] for the model with switch-over times and with both queues having identical characteristics ('the symmetric model').

The determination of $F_1(z_1,z_2)$ from (6.11) will be reduced to the solution of a boundary value problem. The analysis consists of four steps.

Step 1: the set-up

According to its definition as a generating function, $F_1(z_1,z_2)$ should be regular for $|z_1| < 1$, continuous for $|z_1| \leqslant 1$, for every fixed z_2 with $|z_2| \leqslant 1$; and similarly with z_1 and z_2 interchanged. Hence every zero (z_1,z_2) of the kernel $K(z_1,z_2)$ in (6.11) should be a zero of the right-hand side of (6.11). This condition should lead to the determination of the yet unknown functions $F_1(0,z_2)$ and $F_2(z_1,0)$ in the right-hand side of (6.11), and hence to that of $F_1(z_1,z_2)$.

Step 2: analysis of the kernel

$K(z_1,z_2)$ is a so-called Poisson kernel (cf. Ch. II.4 of Cohen and Boxma [1983]). It has the same structure as the Poisson kernel defined in (2.4) on p. 274 of Cohen and Boxma [1983], where the alternating-service model *without* switch-over times is studied; we, therefore, proceed as in Cohen and Boxma [1983] (cf. also Boxma [1985]). First introduce

$$\beta(x) := \beta_1(x)\,\beta_2(x)\,\sigma_1(x)\,\sigma_2(x),$$

$$\lambda := \lambda_1 + \lambda_2,$$

$$r_1 := \lambda_1/\lambda, \qquad r_2 := \lambda_2/\lambda.$$

Without loss of generality, it will henceforth be assumed that

$$r_1 \geqslant r_2.$$

By introducing

$$w_1 := 2r_1z_1, \qquad w_2 := 2r_2z_2,$$

we can rewrite $4r_1r_2K(z_1,z_2)$ as

$$w_1w_2 - 4r_1r_2\beta(\lambda(1-(w_1+w_2)/2)).$$

The symmetry of this expression suggests to look for pairs of zeros of the kernel that are each other's complex conjugates: $(w_1,w_2)=(w,\bar{w})$. These pairs of

zeros turn out to supply all the information we need. The following should hold for w:

$$|w|^2 = 4r_1r_2\beta(\lambda(1-\operatorname{Re} w)).$$

Write

$$w = e^{i\phi}2\sqrt{r_1r_2}\sqrt{\beta(\lambda(1-\operatorname{Re} w))}, \quad 0\leqslant\phi\leqslant 2\pi.$$

Consider the function

$$\delta - 2\sqrt{r_1r_2}\cos(\phi)\sqrt{\beta(\lambda(1-\delta))}, \quad 0\leqslant\phi\leqslant 2\pi, \ \operatorname{Re}\delta\leqslant 1. \tag{6.14}$$

By applying Rouché's theorem, cf. Cohen and Boxma [1983], it is directly seen that (6.14) has exactly one zero for each $\phi\in(0,2\pi]$. Defining $\delta(\phi)$ to be this unique zero it is seen that when ϕ once traverses the trajectory $(0,2\pi]$,

$$w = w(\phi) := \delta(\phi)(1+i\tan\phi)$$

once encircles a simple, smooth contour F that is contained in the unit circle. F is an egg-shaped contour. Using the notation L^+ (L^-) for the interior (exterior) of a contour L, we have $0\in F^+$. Let $w\in F$. Put $z_1 = w/2r_1$, $z_2 = \overline{w}/2r_2$; then it is seen by direct substitution that (z_1,z_2) is a zero tuple of the kernel. Note that it cannot be guaranteed that $|z_2|\leqslant 1$; for a discussion see below.

Step 3: formulation of a Riemann boundary value problem

The choice of zero-pairs $(z_1,z_2) = (w/2r_1,\overline{w}/2r_2)$ of the kernel leads, in a natural way, to the formulation of a Riemann BVP. In the formulation and solution of the BVP a few technical difficulties will arise; these are mainly related to the position of the point $2r_2$ with respect to the contour F. Depending on the choice of parameters, this point can be inside, on or outside the contour. For the sake of clarity, we restrict ourself here to the case $2r_2\in F^+$; see Remark 6.2 for a short discussion of the case $2r_2\in F$ (which occurs, e.g., for $r_2=1/2$) and the (relatively rare) case $2r_2\in F^-$.

As stated in the beginning of Section 6.2 the Riemann BVP basically amounts to finding two functions, one regular inside a certain smooth contour and the other one regular outside that contour, such that a certain linear relation exists between these functions on the contour. The first part of Step 3 concerns the formulation of that linear relation between the two functions on the contour. The right-hand side of (6.11) should be zero for all those $w\in F$, for which $(w/2r_1,\overline{w}/2r_2)$ forms a pair of zeros of $K(z_1,z_2)$ inside the product of unit circles. Now $|w/2r_1|\leqslant 1$ always holds for $w\in F$, but the possibility that $|\overline{w}/2r_2|>1$ cannot be excluded. Fortunately, in the latter case analytic continuation can be used (cf. below (6.27)) to show that, for *all* $w\in F$, the

right-hand side of (6.11) should be zero. So for *all* $w \in F$, the following linear relation should exist between $F_1(0,\overline{w}/2r_2)$ and $F_2(w/2r_1,0)$:

$$F_1(0,\overline{w}/2r_2) \times$$

$$\left[\beta_2(\lambda(1-\mathrm{Re}\ w))\ \sigma_1(\lambda(1-\mathrm{Re}\ w))\ \sigma_2(\lambda(1-\mathrm{Re}\ w))\ \{\frac{w}{2r_1}-\beta_1(\lambda(1-\mathrm{Re}\ w))\}\right] +$$

$$F_2(w/2r_1,0)\ \left[\frac{w}{2r_1}\sigma_2(\lambda(1-\mathrm{Re}\ w))\ \{\frac{\overline{w}}{2r_2}-\beta_2(\lambda(1-\mathrm{Re}\ w))\}\right] = 0. \tag{6.15}$$

Hence

$$F_2(w/2r_1,0) = G(w)\ F_1(0,\overline{w}/2r_2), \quad w \in F, \tag{6.16}$$

with

$$G(w) := -\beta_2(\lambda(1-\mathrm{Re}\ w))\ \sigma_1(\lambda(1-\mathrm{Re}\ w))\ \frac{1}{w/2r_1}\times$$

$$\frac{w/2r_1-\beta_1(\lambda(1-\mathrm{Re}\ w))}{\overline{w}/2r_2-\beta_2(\lambda(1-\mathrm{Re}\ w))}, \quad w \in F. \tag{6.17}$$

In the standard formulation of the Riemann BVP, the involved smooth contour is the unit circle. A conformal mapping of F^+ onto the interior C^+ of the unit circle C will lead us to a standard Riemann BVP. The second part of Step 3 concerns this conformal mapping:

$$z = f(w):\ F^+ \to C^+, \tag{6.18}$$

and its inverse, the conformal mapping

$$w = f_0(z):\ C^+ \to F^+. \tag{6.19}$$

One can write (cf. Gaier [1964], Section 2.1; see also Section I.4.4 of Cohen and Boxma [1983]):

$$f_0(z) = z\ \exp[\frac{1}{2\pi}\int_0^{2\pi} \log\{\frac{\delta(\theta(\omega))}{\cos(\theta(\omega))}\}\ \frac{e^{i\omega}+z}{e^{i\omega}-z}\ d\omega], \quad |z|<1, \tag{6.20}$$

with the angular deformation $\theta(\cdot)$ being uniquely determined as the continuous solution of the *Theodorsen* integral equation

$$\theta(\phi) = \phi - \frac{1}{2\pi}\int_0^{2\pi} \log\{\frac{\delta(\theta(\omega))}{\cos(\theta(\omega))}\} \operatorname{cotan}\{\frac{1}{2}(\omega-\phi)\}\, d\omega, \quad 0\leqslant\phi\leqslant 2\pi; \tag{6.21}$$

$\theta(\phi)$ is a strictly increasing and continuous function of ϕ, and $\theta(\phi)=2\pi-\theta(2\pi-\phi)$. According to the corresponding boundaries theorem (Cohen and Boxma [1983], p. 66), $f_0(z)$ maps C onto F and is continuous in $C^+ \cup C$.

Application of the conformal mapping $f_0(\cdot)$ transforms (6.16) into:

$$F_2(f_0(z)/2r_1,0) = G(f_0(z))\, F_1(0,f_0(1/z)/2r_2), \quad z\in C. \tag{6.22}$$

Introducing the functions

$$\hat{F}_2(z) := F_2(f_0(z)/2r_1,0), \quad z\in C^+ \cup C, \tag{6.23}$$

$$\hat{F}_1(z) := F_1(0,f_0(1/z)/2r_2), \quad z\in C\cup C^-, \tag{6.24}$$

$$H(z) := G(f_0(z)) = -\beta_2(\lambda(1-\operatorname{Re} f_0(z)))\,\sigma_1(\lambda(1-\operatorname{Re} f_0(z)))\,\frac{1}{f_0(z)/2r_1}\times$$

$$\frac{f_0(z)/2r_1-\beta_1(\lambda(1-\operatorname{Re} f_0(z)))}{f_0(1/z)/2r_2-\beta_2(\lambda(1-\operatorname{Re} f_0(z)))}, \quad z\in C, \tag{6.25}$$

(6.22) can be rewritten as:

$$\hat{F}_2(z) = H(z)\,\hat{F}_1(z), \quad z\in C. \tag{6.26}$$

We have now arrived at a standard, homogeneous, Riemann BVP on the unit circle:
Determine two functions $\hat{F}_1(z)$ and $\hat{F}_2(z)$, such that

- (6.26) holds, with $H(\cdot)$ satisfying a Hölder condition on C and $0<|H(z)|<\infty$, $z\in C$;
- $\hat{F}_1(z)$ is regular for $z\in C^-$, continuous for $z\in C\cup C^-$;
- $\hat{F}_2(z)$ is regular for $z\in C^+$, continuous for $z\in C^+\cup C$;
- $\hat{F}_1(z)\to A$ for $|z|\to\infty$, with A a constant.

We shall consider each of the four requirements in the BVP formulation in turn.

That $H(z)$ satisfies a Hölder condition on C can be easily verified, and will not be further discussed. We shall present a proof that $H(z)\neq 0$ for $z\in C$ (the proof that $|H(z)|<\infty$ is left to the reader). We prove the equivalent

statement that $G(w) \neq 0$ for $w \in F$. The two points of F on the real axis are the only candidate zeros of $G(w)$, $w \in F$. It is soon clear that we can concentrate on $w = \delta(0) \in F$, and that it remains to show that $\delta(0)/2r_1 - \beta_1(\lambda(1-\delta(0))) \neq 0$. The assumption that $2r_2 \in F^+$ implies that $\delta(0)/2r_2 > \beta_2(\lambda(1-\delta(0)))$. The definition of $\delta(0)$, see (6.14), implies that

$$\delta(0) = 2\sqrt{r_1 r_2}\sqrt{\beta(\lambda(1-\delta(0)))} < 2\sqrt{r_1 r_2}\sqrt{\beta_1(\lambda(1-\delta(0)))}\sqrt{\beta_2(\lambda(1-\delta(0)))}.$$

Hence

$$\delta(0)/2r_1 \geqslant \beta_1(\lambda(1-\delta(0))) \quad \text{and} \quad \delta(0)/2r_2 \geqslant \beta_2(\lambda(1-\delta(0)))$$

are not simultaneously possible. In view of the above,

$$\delta(0)/2r_1 < \beta_1(\lambda(1-\delta(0))). \tag{6.27}$$

It remains to show that $\hat{F}_1(z)$ is regular for $z \in C^-$, continuous for $z \in C \cup C^-$, and $\hat{F}_2(z)$ is regular for $z \in C^+$, continuous for $z \in C^+ \cup C$. We show, equivalently, that $F_2(w/2r_1, 0)$ and $F_1(0, w/2r_2)$ are regular in F^+ and continuous in $F^+ \cup F$. Since by assumption $2r_1 \geqslant 1$, while F is contained in the unit circle, it immediately follows that $F_2(w/2r_1, 0)$ is regular in F^+, and continuous in $F^+ \cup F$. It is somewhat more difficult to show that $F_1(0, w/2r_2)$ is also regular in F^+ and continuous in $F^+ \cup F$. First note that $\delta(0) = \max|w|$, $w \in F^+ \cup F$. Subsequently note that $F_1(0, \delta(0)/2r_2)$ is finite, because $F_2(\delta(0)/2r_1, 0)/G(\delta(0))$ is finite. These observations, combined with the fact that the coefficients in the series expansion of $F_1(0, w/2r_2)$ are nonnegative, lead to the stated regularity and continuity properties of $F_1(0, w/2r_2)$.

Finally by requiring that $\hat{F}_1(z) \to A$ for $z \to \infty$ and A some constant, the existence of a unique solution of the boundary value problem is ensured.

REMARK 6.1
In the model without switch-over times the BVP is a Dirichlet problem; in the symmetric model with switch-over times it is a Riemann-Hilbert problem; and in the present more general model it can be formulated as a Riemann problem. In fact in the latter two cases both a Riemann-Hilbert and a Riemann problem formulation are possible; the Riemann formulation seems to be somewhat more natural here.

Step 4: solution of the Riemann boundary value problem
A crucial role in the solution of the Riemann BVP is played by the *index*, χ, of the function $H(\cdot)$ on C. This index is by definition:

$$\chi := ind_{z \in C}\, H(z) = \frac{1}{2\pi} \int_{z \in C} d\{arg\, H(z)\}$$

$$= ind_{w\in F}\, G(w) = \frac{1}{2\pi} \int_{w\in F} d\{arg\ G(w)\}. \tag{6.28}$$

LEMMA 6.1
$\chi = 0$ *for* $2r_2 \in F^+$.

PROOF:
From (6.17),

$$\chi = -ind_{w\in F} \frac{w}{2r_1} + ind_{w\in F}\,[w/2r_1 - \beta_1(\lambda(1-\mathrm{Re}\ w))] -$$

$$ind_{w\in F}\,[\bar{w}/2r_2 - \beta_2(\lambda(1-\mathrm{Re}\ w))] =$$

$$-1 + ind_{w\in F}\,[w/2r_1 - \beta_1(\lambda(1-\mathrm{Re}\ w))] +$$

$$ind_{w\in F}\,[w/2r_2 - \beta_2(\lambda(1-\mathrm{Re}\ w))]. \tag{6.29}$$

The fact that $2r_2 \in F^+$ implies that $w/2r_2 > \beta_2(\lambda(1-\mathrm{Re}\ w))$ for $w = \delta(0) \in F$. It now readily follows that

$$ind_{w\in F}\,[w/2r_2 - \beta_2(\lambda(1-\mathrm{Re}\ w))] = 1. \tag{6.30}$$

Similarly, from (6.27)

$$ind_{w\in F}\,[w/2r_1 - \beta_1(\lambda(1-\mathrm{Re}\ w))] = 0. \tag{6.31}$$

The lemma follows from (6.29), (6.30) and (6.31). □

The homogeneous Riemann BVP formulated below (6.26), with index 0, has the following unique solution (cf. Cohen and Boxma [1983], Section I.2.3):

$$\hat{F}_1(z) = A\ \exp[\frac{1}{2\pi i} \int_{t\in C} \frac{\log H(t)}{t-z}\, dt], \quad z \in C^-, \tag{6.32}$$

$$\hat{F}_2(z) = A\ \exp[\frac{1}{2\pi i} \int_{t\in C} \frac{\log H(t)}{t-z}\, dt], \quad z \in C^+, \tag{6.33}$$

with $A = F_1(0,0)$ (let $z \to \infty$ in (6.32)) yet to be determined by the norming

condition $F_1(1,1) = F_2(1,1) = 1$.
Formulas (6.32) and (6.33) lead, in combination with (6.23), (6.24) and the definition (6.19) of the conformal mapping $f(\cdot)$, to our main result:

THEOREM 6.1 Solution of the Riemann Boundary Value Problem
Under the conditions as formulated below (6.26) and under the assumption that $2r_2 \in F^+$, *the following relations hold:*

$$F_1(0,w/2r_2) = A \exp[\frac{1}{2\pi i} \int_{t \in C} \frac{\log H(t)}{t - 1/f(w)}\, dt], \quad w \in F^+, \tag{6.34}$$

$$F_2(w/2r_1,0) = A \exp[\frac{1}{2\pi i} \int_{t \in C} \frac{\log H(t)}{t - f(w)}\, dt], \quad w \in F^+. \tag{6.35}$$

It remains to determine the constant $A = F_1(0,0)$. Substitution of $w = 2r_2$ in (6.34) yields a linear relation between $F_1(0,1)$ and $F_1(0,0)$. $F_1(0,1)$ can be determined in various ways. For example, substituting $z_2 = 1$ in (6.11) and subsequently letting $z_1 \to 1$ gives one linear relation between $F_1(0,1)$ and $F_2(1,0)$; applying a similar procedure to (6.12) gives a second linear relation between those quantities. Solution of the two equations yields:

$$F_1(0,1) = 1 - \frac{\lambda_1 s}{1-\rho}, \tag{6.36}$$

$$F_2(1,0) = 1 - \frac{\lambda_2 s}{1-\rho}. \tag{6.37}$$

The following observation also immediately leads to (6.36) (and similarly (6.37)): $1 - F_1(0,1)$ is the probability that server S finds Q_1 not empty upon his arrival. Therefore it equals the fraction of times that S serves a customer in Q_1 during his visit. By a balance argument, this fraction also equals the mean number of arrivals at Q_1 during a cycle of the server. According to (6.5), the mean cycle time of the server equals $s/(1-\rho)$; hence the mean number of arrivals at Q_1 during a cycle of the server equals $\lambda_1 s/(1-\rho)$.

The above implies that the constant A is given by:

$$A = F_1(0,0) = (1 - \frac{\lambda_1 s}{1-\rho}) \exp[-\frac{1}{2\pi i} \int_{t \in C} \frac{\log H(t)}{t - 1/f(2r_2)}\, dt]. \tag{6.38}$$

REMARK 6.2
Above the restrictive assumption $2r_2 \in F^+$ has been made. However, the cases $2r_2 \in F$ and $2r_2 \in F^-$ can also occur. For example, if $r_1 = r_2 = \frac{1}{2}$, then $2r_2 = 1 = \delta(0) \in F$; examples in which $2r_2 \in F^-$ can also be constructed,

although one has to be careful not to violate the ergodicity condition (cf. Cohen and Boxma [1983], pp. 360-361). We briefly consider Steps 3 and 4 above for these two cases.

(i) $2r_2 \in F$.
The Riemann BVP formulation and the proof that the index $\chi=0$ proceed as before. The special case $r_1=r_2=\frac{1}{2}$ requires some extra attention. Now $\delta(0)=1$; both the numerator and denominator of the right-hand side of (6.17) become zero, but the zeros cancel and again $G(\delta(0))\neq 0$. Furthermore, (6.29) reduces to $\chi=-1+\frac{1}{2}+\frac{1}{2}=0$.

For all cases in which $2r_2 \in F$, the solution of the Riemann BVP proceeds as before, and Theorem 6.1 still holds. A minor difficulty is that $F_1(0,1)$ cannot be obtained from (6.34) by substitution of $w=2r_2$. But application of the so-called Plemelj-Sokhotski formula (cf. Cohen and Boxma [1983], Formula (I.1.6.4)) to (6.34) leads to an expression for $F_1(0,w/2r_2)$, $w \in F$. In the resulting expression, the substitution $w=2r_2$ can be made.

(ii) $2r_2 \in F^-$.
The formulation of the Riemann BVP proceeds as before. This time verification of the regularity of $F_1(0,w/2r_2)$ in F^+ is trivial. Again the index $\chi=0$, but verification is less straightforward. Theorem 6.1 still holds, but a major difficulty now is that $F_1(0,1)$ cannot be obtained by substitution of $w=2r_2$ in (6.34). We refer to Cohen and Boxma [1983] and Blanc [1982] for more extensive discussions of these difficulties.

Expressions for the generating functions $F_1(z_1,z_2)$ and $F_2(z_1,z_2)$ follow from (6.11), (6.12) and Theorem 6.1. In the next two sections we use the results obtained about queue-length generating functions to derive information about the distributions of waiting times and cycle times, and in particular about their moments.

6.2.3. Waiting times

In this section we shall derive an expression for $E\mathbf{W}_2$, the mean waiting time at Q_2. $E\mathbf{W}_2$ will be expressed in some given model parameters and in the function $d/dz\ F_1(0,z)$, evaluated at $z=1$. The latter function is obtained from (6.34) after differentiation with respect to w and substitution of $w=2r_2$. $E\mathbf{W}_1$ cannot be obtained from (6.35) in a similar way; the fact that $2r_1 \in F^-$ poses a problem. There are several ways to overcome this problem; however, discussing them is not within the scope of this study. We refer to Cohen & Boxma [1983] and Blanc [1982] for a more extensive discussion.

Anyway, once we have calculated $E\mathbf{W}_2$, $E\mathbf{W}_1$ can be obtained directly from the functional equation, or from the pseudoconservation law derived in Chapter 3. For two queues, each with 1-limited service, this law reduces to:

$$\rho_1[1-\frac{\lambda_1 s}{1-\rho}]E\mathbf{W}_1 + \rho_2[1-\frac{\lambda_2 s}{1-\rho}]E\mathbf{W}_2 =$$

$$\rho\frac{\lambda_1\beta_1^{(2)}+\lambda_2\beta_2^{(2)}}{2(1-\rho)} + \rho\frac{s^{(2)}}{2s} + \frac{s}{2(1-\rho)}[\rho^2+\rho_1^2+\rho_2^2]. \tag{6.39}$$

$E\mathbf{W}_2$ is obtained in the following way. By a standard M/G/1-type argument (cf. Watson [1985], Takagi [1986]) we can write:

$$E\{e^{-\lambda_2(1-z)\mathbf{W}_2}\} = \frac{F_2(1,z) - F_2(1,0)}{z(1-F_2(1,0))}. \tag{6.40}$$

Indeed, the customers present in Q_2 at the start of a non-empty service period at that queue, excluding the customer about to be served, are just the customers who had arrived during the waiting time of that customer. Note that (6.40) completely determines the waiting-time distribution at Q_2; a similar relation holds for the transform of the waiting-time distribution at Q_1. From (6.40) and (6.37),

$$E\mathbf{W}_2 = \frac{1-\rho}{\lambda_2^2 s}\{\frac{d}{dz}F_2(1,z)\}_{z=1} - \frac{1}{\lambda_2}. \tag{6.41}$$

The derivative occurring in (6.41) follows from (6.12) after a tedious but straightforward calculation. Denote by β and $\beta^{(2)}$ the first and second moments of the *sum* of a service time at Q_1, a switch-over time from Q_1 to Q_2, a service time at Q_2 and a switch-over time from Q_2 to Q_1. Then

$$\{\frac{d}{dz}F_2(1,z)\}_{z=1} = (1-\frac{\lambda_2 s}{1-\rho})\left[\frac{\lambda_2(\beta_1+s)(1-\lambda_2\beta_2)}{1-\lambda_2\beta} - \right.$$

$$\left.\frac{\frac{1}{2}\lambda_2^2\beta_2^{(2)}}{1-\lambda_2\beta} + (1-\lambda_2\beta_2)\frac{\frac{1}{2}\lambda_2^2\beta^{(2)}}{(1-\lambda_2\beta)^2}\right] -$$

$$\{\frac{d}{dz}F_1(0,z)\}_{z=1}\frac{\lambda_2\beta_1}{1-\lambda_2\beta} -$$

$$(1-\frac{\lambda_1 s}{1-\rho})\left[\frac{\lambda_2\beta_1(1+\lambda_2 s_1)+\frac{1}{2}\lambda_2^2\beta_1^{(2)}}{1-\lambda_2\beta} + \frac{\frac{1}{2}\lambda_2^3\beta_1\beta^{(2)}}{(1-\lambda_2\beta)^2}\right]. \tag{6.42}$$

It remains to determine the derivative occurring in the right-hand side of (6.42). From (6.34) and Cohen and Boxma [1983, p.38],

$$\{\frac{d}{dz}F_1(0,z)\}_{z=1} = 2r_2\{\frac{d}{dw}F_1(0,w/2r_2)\}_{w=2r_2} =$$

$$= 2r_2\, F_1(0,1)\, \frac{-f^{(1)}(2r_2)}{(f(2r_2))^2}\, \frac{1}{2\pi i} \int_{t \in C} \frac{\log H(t)}{(t-1/f(2r_2))^2}\, dt$$

$$= 2r_2\, F_1(0,1)\, \frac{-f^{(1)}(2r_2)}{(f(2r_2))^2}\, \frac{1}{2\pi i} \int_{t \in C} \frac{H^{(1)}(t)/H(t)}{(t-1/f(2r_2))}\, dt. \quad (6.43)$$

$E\mathbf{W}_2$ follows from (6.41), (6.42), (6.43) and (6.36). As indicated above, $E\mathbf{W}_1$ subsequently follows from (6.39).

6.2.4. Cycle times

In Section 6.2.1, the cycle time $\mathbf{C}_i$ for Q_i has been defined as the time between two consecutive arrivals of S at Q_i. Both from a theoretical and a practical point of view, cycle times are important quantities in cyclic-service systems. Mean cycle times are easily calculated (cf. (6.5)), but in cyclic-service systems with 1-limited service hardly any other exact cycle-time results are known. Only for the special case of two completely symmetric queues an exact formula for the LST of the cycle-time distribution has been obtained by Boxma [1985]. In the present section we extend this result to the asymmetric case. We are thus able to compare $E\mathbf{C}_1^2$ and $E\mathbf{C}_2^2$, and also to determine

$$E\mathbf{C}_{b,i} := E[\mathbf{C}_i \mid A_i], \quad (6.44)$$

with A_i the indicator function of the event 'the cycle contains a service at Q_i'. This quantity plays an important role in several mean waiting-time approximations, cf. Boxma and Meister [1986,1987], Fuhrmann and Wang [1988] and Kühn [1979]. Generally speaking exact cycle-time formulas for the two-queue case give more insight into the accuracy of general approximations for cycle-time distributions, as were proposed by Hashida and Ohara [1972] and Kühn [1979].

We now derive an exact expression for the LST of the distribution of $\mathbf{C}_1$; the analogous result for $\mathbf{C}_2$ is obtained by interchanging all indices. Starting-point of the analysis is the relation

$$E[e^{-\omega \mathbf{C}_1}] = F_1(0,0)\, E[e^{-\omega \mathbf{C}_1} \mid \mathbf{X}_1^{(1)}=0, \mathbf{X}_2^{(1)}=0] +$$

$$[F_1(0,1)-F_1(0,0)]\, E[e^{-\omega \mathbf{C}_1} \mid \mathbf{X}_1^{(1)}=0, \mathbf{X}_2^{(1)}>0] +$$

$$[F_1(1,0)-F_1(0,0)]\, E[e^{-\omega \mathbf{C}_1} \mid \mathbf{X}_1^{(1)}>0, \mathbf{X}_2^{(1)}=0] +$$

$$[1-F_1(0,1)-F_1(1,0)+F_1(0,0)]\, E[e^{-\omega \mathbf{C}_1} \mid \mathbf{X}_1^{(1)}>0, \mathbf{X}_2^{(1)}>0]$$

$$= F_1(0,0)\,\sigma_2(\omega)[\sigma_1(\omega+\lambda_2)+\{\sigma_1(\omega)-\sigma_1(\omega+\lambda_2)\}\beta_2(\omega)] +$$

$$[F_1(0,1)-F_1(0,0)]\,\beta_2(\omega)\sigma_1(\omega)\sigma_2(\omega) +$$

$$[F_1(1,0)-F_1(0,0)]\,\sigma_2(\omega)\times$$

$$[\beta_1(\omega+\lambda_2)\sigma_1(\omega+\lambda_2)+\{\beta_1(\omega)\sigma_1(\omega)-\beta_1(\omega+\lambda_2)\sigma_1(\omega+\lambda_2)\}\beta_2(\omega)] +$$

$$[1-F_1(0,1)-F_1(1,0)+F_1(0,0)]\,\beta_1(\omega)\sigma_1(\omega)\beta_2(\omega)\sigma_2(\omega). \qquad (6.45)$$

$F_1(0,1)$ is given by (6.36). Hence $E[e^{-\omega \mathbf{C}_1}]$ can be expressed in $F_1(0,0)$ and $F_1(1,0)$. Substitution of $z_1=1$, $z_2=0$ into (6.11) leads to a linear relation between those two terms:

$$F_1(1,0) = F_1(0,0)\frac{\beta_1(\lambda_2)-1}{\beta_1(\lambda_2)} + \frac{1-\lambda_2 s/(1-\rho)}{\beta_1(\lambda_2)\sigma_1(\lambda_2)}. \qquad (6.46)$$

Differentiation of the expressions in (6.45) with respect to ω and substitution of $F_1(1,0)$ into $F_1(0,0)$ using (6.46) leads to cycle-time moments. A simple calculation yields the mean cycle time given in (6.5); a lengthy calculation yields

$$E\mathbf{C}_1^2 = s^{(2)} + \sum_{i=1}^{2}\lambda_i\beta_i^{(2)}\frac{s}{1-\rho} + 2\beta_1(\beta_2+s)\frac{\lambda_1 s}{1-\rho} +$$

$$2\beta_2 s - 2\beta_2 s_2(1-\frac{\lambda_2 s}{1-\rho}) + 2\beta_2(1-\frac{\lambda_2 s}{1-\rho})[\frac{\beta_1'(\lambda_2)}{\beta_1(\lambda_2)} + \frac{\sigma_1'(\lambda_2)}{\sigma_1(\lambda_2)}] -$$

$$2\beta_2 F_1(0,0)\sigma_1(\lambda_2)\frac{\beta_1'(\lambda_2)}{\beta_1(\lambda_2)}. \qquad (6.47)$$

Note that if the switch-over time from Q_1 to Q_2 is a constant (s_1), then $E\mathbf{C}_1^2$ only depends on the individual mean switch-over times via the term involving $F_1(0,0)\sigma_1(\lambda_2)$ - apart from that term, only s and $s^{(2)}$ occur. We will return to this point in the Section 6.2.6.

We now turn to the cycle-time distribution of $\mathbf{C}_1$ under the condition A_1: "the cycle contains a service at Q_1". Analogously to the derivation in (6.45),

$$E[e^{-\omega \mathbf{C}_1} \mid A_1] = \frac{F_1(1,0)-F_1(0,0)}{1-F_1(0,1)}\sigma_2(\omega)\times$$

$$[\beta_1(\omega+\lambda_2)\sigma_1(\omega+\lambda_2)+\{\beta_1(\omega)\sigma_1(\omega)-\beta_1(\omega+\lambda_2)\sigma_1(\omega+\lambda_2)\}\beta_2(\omega)] +$$

$$\frac{1-F_1(0,1)-F_1(1,0)+F_1(0,0)}{1-F_1(0,1)}\beta_1(\omega)\sigma_1(\omega)\beta_2(\omega)\sigma_2(\omega). \tag{6.48}$$

A simple calculation leads to

$$E\mathbf{C}_{b,1} = E[\mathbf{C}_1 \mid A_1]$$

$$= \beta_1 + s + \beta_2[1-\frac{1-\rho}{\lambda_1 s}\{1-\frac{\lambda_2 s}{1-\rho}-F_1(0,0)\sigma_1(\lambda_2)\}]. \tag{6.49}$$

A similar expression, with all indices interchanged, holds for $E\mathbf{C}_{b,2}$. Note that the term between curly brackets represents the difference between the probability that S finds Q_2 empty and the probability that S finds first Q_1 and then Q_2 empty. Also observe that $\lambda_1 s/(1-\rho)$ is the probability that S does serve at Q_1. Hence the term between square brackets represents the conditional probability that S does serve at Q_2, under the condition A_1.

6.2.5. Numerical analysis

This section is devoted to a numerical evaluation of some important performance measures of the alternating-service model. Our reasons for including this section are twofold:

(i) We want to show that the BVP formulation leads to formulas which can be numerically evaluated in a straightforward manner;

(ii) We want to contribute to the insight into the behavior of the alternating-service model and, more generally, of cyclic-service models; in particular, the numerical results to be presented may be helpful for devising and testing approximations.

For the sake of (i), we now discuss the numerical evaluation of $F_i(0,0)$ and $E\mathbf{W}_i$; other performance measures are easily evaluated from these quantities. The numerical analysis basically consists of five steps. For details we refer to Ch. IV.1 of Cohen and Boxma [1983], in which numerical calculations of this kind have been extensively discussed.

Step 1: Solving Theodorsen's integral equation (cf. (6.21)).

Determine $\theta(\phi)$, iteratively, from (cf. Gaier [1964]):

$$\theta_0(\phi) = \phi, \quad 0 \leqslant \phi \leqslant 2\pi,$$

$$\theta_{n+1}(\phi) = \phi - \frac{1}{2\pi}\int_0^{2\pi} \log\{\frac{\delta(\theta_n(\omega))}{\cos(\theta_n(\omega))}\} \operatorname{cotan}\{\tfrac{1}{2}(\omega-\phi)\}\, d\omega, \quad 0\leqslant\phi\leqslant 2\pi, \tag{6.50}$$

where $\delta(\theta_n(\omega))$ is determined from (cf. (6.14)):

$$\delta(\theta_n(\omega)) = 2\sqrt{r_1 r_2}\cos(\theta_n(\omega))\sqrt{\beta(\lambda(1-\delta(\theta_n(\omega))))}, \tag{6.51}$$

using the Newton-Raphson root-finding procedure. In our calculations, the iteration has been continued until the differences between successive iterations of $\theta(\cdot)$ (in the supremum norm) were in absolute value less than 10^{-6}. This required between 6 and 14 iterations.

REMARK 6.3
Due to various symmetry properties we can restrict ourself in the computations, here and in the sequel, to $\phi\in[0,\pi]$. As various integrands that will have to be computed change more rapidly for ϕ close to 0 than for other values of ϕ, a finer subdivision has been chosen for the interval $[0,\pi/5]$ (20 points) than for the interval $[\pi/5,\pi]$ (40 points). All involved integrals have been evaluated using the repeated trapezoidal rule.

Step 2: Determination of the conformal mapping $f_0(e^{i\phi})$, $0\leqslant\phi\leqslant 2\pi$.

Applying the Plemelj-Sokhotski formula (cf. Cohen and Boxma [1983], Formula (I.1.6.4)) to (6.20) yields:

$$\begin{aligned} f_0(e^{i\phi}) &= e^{i\phi}\exp[\log\{\frac{\delta(\theta(\phi))}{\cos(\theta(\phi))}\} + \\ &\qquad \frac{1}{2\pi i}\int_0^{2\pi}\log\{\frac{\delta(\theta(\omega))}{\cos(\theta(\omega))}\}\operatorname{cotan}\{\tfrac{1}{2}(\omega-\phi)\}\, d\omega] \\ &= e^{i\theta(\phi)}\,\frac{\delta(\theta(\phi))}{\cos(\theta(\phi))} \\ &= \delta(\theta(\phi))[1+i\tan(\theta(\phi))], \quad 0\leqslant\phi\leqslant 2\pi; \end{aligned} \tag{6.52}$$

this result could also have been derived from the formula below (6.14).

Step 3: Determination of $f(2r_2)$ and $f^{(1)}(2r_2)$.

Using (6.20), $f(2r_2)$ is obtained as the solution, on $[0,1]$, of $f_0(z) = 2r_2$. Again we have used the Newton-Raphson root-finding procedure.
$f^{(1)}(2r_2)$ can be obtained in two ways:

(i) by numerical differentiation of $f_0(\cdot)$; note that

$$f^{(1)}(2r_2) = \frac{1}{f_0^{(1)}(2r_2)}; \tag{6.53}$$

(ii) by a numerical evaluation of the expression:

$$f_0^{(1)}(2r_2) = \frac{2r_2}{f(2r_2)} + 2r_2 \frac{1}{2\pi} \int_0^{2\pi} \log\{\frac{\delta(\theta(\omega))}{\cos(\theta(\omega))}\} \frac{2e^{i\omega}}{(e^{i\omega} - f(2r_2))^2}\, d\omega, \tag{6.54}$$

and substitution of the result in (6.53).

For a discussion of (ii) see Cohen and Boxma [1983], p. 351. We have used both (i) and (ii), but due to the fact that we have chosen a relatively fine subdivision we have found no significant differences.

Step 4: Calculation of $H(e^{i\phi})$, $0 \leqslant \phi \leqslant 2\pi$.

$H(e^{i\phi})$ is obtained from (6.25) by noting that $\text{Re}\, f_0(e^{i\phi}) = \delta(\theta(\phi))$:

$$H(e^{i\phi}) = -\beta_2(\lambda(1-\delta(\theta(\phi)))) \, \sigma_1(\lambda(1-\delta(\theta(\phi)))) \frac{2r_1}{f_0(e^{i\phi})} \times$$

$$\frac{f_0(e^{i\phi})/2r_1 - \beta_1(\lambda(1-\delta(\theta(\phi))))}{f_0(e^{-i\phi})/2r_2 - \beta_2(\lambda(1-\delta(\theta(\phi))))}, \quad 0 \leqslant \phi \leqslant 2\pi. \tag{6.55}$$

Step 5: Determination of $E\mathbf{W}_i$ and $F_i(0,0)$, $i = 1,2$.

Once we have calculated $\{\frac{d}{dz} F_1(0,z)\}_{z=1}$ from (6.43), we can obtain $E\mathbf{W}_2$ (cf. (6.41), (6.42) and (6.43)), and subsequently $E\mathbf{W}_1$ from (6.39).
$F_1(0,0)$ is easily calculated from (6.38); $F_2(0,0)$ is obtained from (cf. (6.35)):

$$F_2(0,0) = F_1(0,0) \exp[\frac{1}{2\pi} \int_0^{2\pi} \log H(e^{i\omega})\, d\omega]. \tag{6.56}$$

With the subdivision we have chosen, each row in the tables below takes about 15 sec. of CPU time on a Cyber 170 model 750, with very small memory requirements. Using fewer $\theta(\cdot)$ iterations and a less fine subdivision of the interval $[0,\pi]$ leads to a considerable reduction of CPU time, without sacrificing too much accuracy. The computer program was written in Pascal.

6.2.6. Discussion of the results

The numerical results are presented in Tables 6.1 and 6.2 at the end of this chapter. The performance measures under consideration are the mean waiting times $E\mathbf{W}_i$, the second moments of cycle times $E\mathbf{C}_i^2$, the conditional first

moments of cycle times $E\mathbf{C}_{b,i}$ (cf. (6.44) and (6.49)) and the empty-system probabilities at server-arrival epochs, $F_i(0,0)$. Table 6.1 studies the influence of the switch-over times on these performance measures. Table 6.1.a shows that the choice of switch-over time *distributions* has hardly any effect on $E\mathbf{C}_{b,i}$ and $F_i(0,0)$, and only has a considerable effect on $E\mathbf{W}_i$ and $E\mathbf{C}_i^2$ when mean service times are relatively small. Exactly the same statement can be made concerning the choice of s_1 and s_2, for given *total* mean switch-over time $s=s_1+s_2$. In Table 6.1.b results for $\beta_1=\beta_2=0.8$ are printed, exhibiting almost-insensitivity for the choice of s_1 and s_2. In the (non-printed) case with $\beta_1=\beta_2=0.2$ and all other parameter values as in Table 6.1.b, the largest difference (with respect to $E\mathbf{W}_i$ and $E\mathbf{C}_i^2$) due to changes in s_i is in the order of 25%. For deterministic switch-over time distributions, $E\mathbf{W}_i$, $E\mathbf{C}_i^2$ and $E\mathbf{C}_{b,i}$ appear to be completely independent of s_1 and s_2, given their sum s. The structure of (6.47) and (6.49) shows that the same must hold for $F_1(0,0)\,\sigma_1(\lambda_2) = Pr\{S \textit{ finds first } Q_1 \textit{ and then } Q_2 \textit{ empty}\}$. The robustness of the model for switch-over times is also being expressed by the pseudo-conservation law for mean waiting times mentioned in Section 6.2.3 (cf. (6.39) for the alternating-service model). In the pseudo-conservation law, the expression for a weighted sum of mean waiting times is seen to depend on the switch-over time distributions only through the mean s and the second moment $s^{(2)}$ of the *total* switch-over time - and the influence of the factor involving $s^{(2)}$ is usually small.

Table 6.2 presents mean waiting times and cycle-time moments for three different combinations of service-time distributions, viz.:
Case A: both service-time distributions are negative exponential;
Case B: both service-time distributions are hyperexponential distributions with squared coefficient of variation 4 ($H_2(4)$) and balanced means (cf. Tijms [1986]);
Case C: $B_1(\cdot)$ is a $H_2(4)$ distribution with balanced means, and $B_2(\cdot)$ is deterministic.
Out of a wide range of distributions and parameter values we have tried to make a representative choice. The observations from Table 6.1 allow us to restrict ourself to constant switch-over times, with $s_1=s_2$. In all cases considered $s=0.2$. We discuss all tabulated performance measures in turn.

(i) $E\mathbf{W}_i$
In Boxma and Meister [1986,1987] a mean waiting-time approximation has been proposed for a cyclic-service model with and without switch-over times, respectively. It is first argued that

$$E\mathbf{W}_i = \frac{E\mathbf{R}_i}{1-\lambda_i E\mathbf{C}_{b,i}}, \tag{6.57}$$

with $E\mathbf{R}_i$ the mean residual cycle time for Q_i. In fact, this is not an exact result. In the alternating-service model it appears to be quite close in most cases, but there are a few exceptions.

Taking $E\mathbf{R}_i = E\mathbf{C}_i^2 / 2E\mathbf{C}_i$ (acting as if the cycle-time process is a renewal process, cf. also Remark 7.1) and assuming that $E\mathbf{C}_{b,i}$ is independent of $E\mathbf{C}_i^2$, Formula (6.57) would imply that $E\mathbf{W}_i$ changes linearly with $E\mathbf{C}_i^2$ for fixed first moments of service times and switch-over times. The table entries for these two quantities suggest that this is indeed more or less the case.

In Boxma and Meister [1986,1987], two approximation assumptions are introduced to estimate the unknown $E\mathbf{R}_i$ and $E\mathbf{C}_{b,i}$, viz.:

Assumption 1: $E\mathbf{C}_{b,i} = E\tilde{\mathbf{C}}_{b,i} := (\beta_i + s)/(1-\rho+\rho_i)$
(this approximation is due to Kühn [1979]).

Assumption 2: $E\mathbf{R}_i$ is the same for all i.

Subsequently the pseudo-conservation law (cf. (6.39) for the alternating-service model) is used to estimate the one unknown $E\mathbf{R}_1$. Below we investigate the accuracy of these assumptions for the alternating-service model.

(ii) $E\mathbf{C}_i^2$
Again taking $E\mathbf{R}_i = E\mathbf{C}_i^2 / 2E\mathbf{C}_i$, Assumption 2 above would imply that all $E\mathbf{C}_i^2$ are the same. Indeed, in all considered cases, $E\mathbf{C}_1^2$ and $E\mathbf{C}_2^2$ differ less than 7% (and usually much less). Fuhrmann and Wang [1988] suggest another mean waiting-time approximation along similar lines as Boxma and Meister [1986], but they assume that

$$E\mathbf{C}_1^2 / E\mathbf{C}_2^2 \approx E\mathbf{C}_{b,2} / E\mathbf{C}_{b,1}; \tag{6.58}$$

our numerical results show that this assumption is not accurate for the alternating-service model. Still, Fuhrmann and Wang improve upon Boxma and Meister [1986] in case of heavy traffic. It is not yet fully clear whether (6.58) becomes more accurate when the number of queues is larger, or whether (6.58) counteracts an inaccuracy in (6.57) or in the approximation for $E\mathbf{C}_{b,i}$.

(iii) $E\mathbf{C}_{b,i}$ and $E\tilde{\mathbf{C}}_{b,i}$
In all considered cases, the approximation $E\mathbf{C}_{b,1} \approx E\tilde{\mathbf{C}}_{b,1}$ (see Assumption 1 above) is extremely accurate. The approximation $E\mathbf{C}_{b,2} \approx E\tilde{\mathbf{C}}_{b,2}$ is much less accurate: the flow-balancing argument on which the approximation is based, should not be applied to the situation of a rarely occurring cycle $\mathbf{C}_2$ with a - sometimes large - service time at Q_2. The approximation becomes useless in the cases marked with an asterisk, because $E\mathbf{C}_{b,2}$ exceeds the obvious upper bound $\beta_1 + \beta_2 + s$; in those cases we have printed the latter number. In Chapter 7 an iterative scheme for computing $E\mathbf{C}_{b,i}$ is proposed which is more accurate than the above approximation.

Finally we observe that $E\mathbf{C}_{b,i}$ is hardly dependent on the choice of the service-time distributions.

(iv) $F_i(0,0)$
$F_i(0,0)$, too, appears to be hardly dependent on the choice of the service-time

distributions.

6.3. Exact results for the two-queue E/1L model

In this section we consider the model consisting of two queues, Q_1 and Q_2, where the service strategy is exhaustive at Q_1 and 1-limited at Q_2. For a further description of the model and the model parameters we refer to the previous section and Chapter 4. Note that in case the switch-over times are zero the model coincides with the non-preemptive priority M/G/1 queue with two types of customers. As usual the system is assumed to be in steady state.

As a notational convention we shall write $\alpha\beta(s)$ for the product of the Laplace Stieltjes Transforms $\alpha(s)$ and $\beta(s)$. We further introduce:

$$r_i := \lambda_i / \lambda;$$

$$x := \lambda(1 - r_1 z_1 - r_2 z_2).$$

Let $F_j(\cdot,\cdot)$ denote the joint generating function of the queue-length distribution at arrival instants of the server at Q_j, $j = 1,2$. It is easily found, that

$$F_1(z_1,z_2) = \frac{\sigma_2\beta_2(x)}{z_2} F_2(z_1,z_2) + \sigma_2(x)\frac{z_2 - \beta_2(x)}{z_2} F_2(z_1,0); \quad (6.59)$$

$$F_2(z_1,z_2) = F_1(\gamma_1(\lambda_2(1-z_2)),z_2)\sigma_1(x), \quad (6.60)$$

where $\gamma_1(\cdot)$ is the LST of the length of the busy period at Q_1 starting with one customer present. For notational convenience we shall write

$$\delta_1(z_2) := \gamma_1(\lambda_2(1-z_2)),$$

and

$$\tilde{x} := \lambda(1 - r_1\delta_1(z_2) - r_2 z_2).$$

Taking $z_1 = \delta_1(z_2)$ in (6.59) yields:

$$F_1(\delta_1(z_2),z_2) = \frac{\sigma_2\beta_2(\tilde{x})}{z_2} F_2(\delta_1(z_2),z_2) + \sigma_2(\tilde{x})\frac{z_2 - \beta_2(\tilde{x})}{z_2} F_2(\delta_1(z_2),0). \quad (6.61)$$

From (6.60) and (6.61) we obtain

$$F_2(z_1,z_2) = \sigma_1(x)\frac{\sigma_2\beta_2(\tilde{x})}{z_2} F_2(\delta_1(z_2),z_2) +$$

$$\sigma_1(x)\sigma_2(\tilde{x})\frac{z_2-\beta_2(\tilde{x})}{z_2}F_2(\delta_1(z_2),0). \tag{6.62}$$

Again taking $z_1=\delta_1(z_2)$, now in (6.62), yields:

$$F_2(\delta_1(z_2),z_2) = \frac{\sigma_1\sigma_2\beta_2(\tilde{x})}{z_2}F_2(\delta_1(z_2),z_2) +$$

$$\sigma_1\sigma_2(\tilde{x})\frac{z_2-\beta_2(\tilde{x})}{z_2}F_2(\delta_1(z_2),0), \tag{6.63}$$

which may be written as

$$F_2(\delta_1(z_2),z_2) = \sigma_1\sigma_2(\tilde{x})\frac{z_2-\beta_2(\tilde{x})}{z_2-\sigma_1\sigma_2\beta_2(\tilde{x})}F_2(\delta_1(z_2),0). \tag{6.64}$$

Note that, cf. (6.60):

$$F_2(z_1,0) = F_1(\gamma_1(\lambda_2),0)\sigma_1(\lambda(1-r_1z_1)). \tag{6.65}$$

Of course we have, cf. for instance Section 6.2,

$$F_2(1,0) = \frac{1-\rho-\lambda_2 s}{1-\rho}. \tag{6.66}$$

Hence, from (6.65) and (6.66),

$$F_2(\delta_1(z_2),0) = \frac{1-\rho-\lambda_2 s}{1-\rho}\frac{\sigma_1(\lambda(1-r_1\delta_1(z_2)))}{\sigma_1(\lambda_2)}. \tag{6.67}$$

Finally, from (6.62), (6.64) and (6.67):

$$F_2(z_1,z_2) = \sigma_1(x)\sigma_1(\lambda(1-r_1\delta_1(z_2)))\sigma_2(\tilde{x})\times$$

$$\frac{z_2-\beta_2(\tilde{x})}{z_2-\sigma_1\sigma_2\beta_2(\tilde{x})}\frac{1-\rho-\lambda_2 s}{1-\rho}\frac{1}{\sigma_1(\lambda_2)}. \tag{6.68}$$

The generating functions of the queue lengths at polling instants and the Laplace Stieltjes Transforms of the waiting time are related as follows (cf. Watson [1985]):

$$E[e^{-\lambda_1(1-z_1)W_1}] = \frac{1-\lambda_1\beta_1}{\frac{d}{dz_1}F_1(z_1,1)|_{z_1=1}}\frac{1-F_1(z_1,1)}{\beta_1(\lambda_1(1-z_1))-z_1}, \tag{6.69}$$

$$E[e^{-\lambda_2(1-z_2)\mathbf{W}_2}] = \frac{F_2(1,z_2)-F_2(1,0)}{z(1-F_2(1,0))}.$$

A straightforward calculation now yields the Laplace Stieltjes Transforms of the waiting times at Q_1 and Q_2 respectively:

$$E[e^{-v\mathbf{W}_1}] = \frac{\sigma_1\sigma_2\beta_2(v)-1}{\lambda_1-v-\lambda_1\beta_1(v)}\frac{1-\rho}{s} +$$

$$\frac{1-\rho-\lambda_2 s}{s}\frac{\sigma_1(\lambda_2+v)\sigma_2(v)}{\sigma_1(\lambda_2)}\frac{1-\beta_2(v)}{\lambda_1-v-\lambda_1\beta_1(v)}.$$

$$E[e^{-v\mathbf{W}_2}] = \frac{1-\rho-\lambda_2 s}{s}\sigma_1(v)\frac{\sigma_1(\lambda(1-r_1\gamma_1(v)))}{\sigma_1(\lambda_2)}\sigma_2(\lambda(1-r_1\gamma_1(v)-r_2(1-\frac{v}{\lambda_2})))\times$$

$$\frac{\lambda_2-v-\lambda_2\beta_2(\lambda(1-r_1\gamma_1(v)-r_2(1-\frac{v}{\lambda_2})))}{\lambda_2-v-\lambda_2\sigma_1\sigma_2\beta_2(\lambda(1-r_1\gamma_1(v)-r_2(1-\frac{v}{\lambda_2})))}\frac{1}{\lambda_2-v} -$$

$$\frac{1}{\lambda_2-v}\frac{1-\rho-\lambda_2 s}{s}. \tag{6.70}$$

The mean waiting times are easily calculated from (6.70) as:

$$E\mathbf{W}_1 = \frac{\lambda_1\beta_1^{(2)}+\lambda_2\beta_2^{(2)}}{2(1-\rho_1)} + \frac{1-\rho}{1-\rho_1}(\frac{s^{(2)}}{2s}+\beta_2) + \frac{1-\rho-\lambda_2 s}{s(1-\rho_1)}(\frac{\sigma_1'(\lambda_2)}{\sigma_1(\lambda_2)}-s_2)\beta_2;$$

$$E\mathbf{W}_2 = \frac{\lambda_1\beta_1^{(2)}+\lambda_2\beta_2^{(2)}}{2(1-\rho_1)(1-\rho-\lambda_2 s)} + \frac{1-\rho}{1-\rho_1}\frac{1}{1-\rho-\lambda_2 s}(\frac{s^{(2)}}{2s}+\beta_2) -$$

$$\frac{\rho_1(1-\rho)}{\lambda_2 s(1-\rho_1)}(\frac{\sigma_1'(\lambda_2)}{\sigma_1(\lambda_2)}-s_2) - \frac{\rho}{\lambda_2}. \tag{6.71}$$

If we take $\sigma_1(v)=\exp[-s_1 v]$ (deterministic switch-over time from the exhaustive service queue to the 1-limited service queue), we obtain a special case of a model previously studied by Skinner [1967]. In this particular case

$$E\mathbf{W}_1 = (1-\rho-\lambda_2 s)E\mathbf{W}_2, \tag{6.72}$$

while in general we have

$$E\mathbf{W}_1 \geqslant (1-\rho-\lambda_2 s)E\mathbf{W}_2. \tag{6.73}$$

This latter inequality may be proven as follows. From (6.71) we obtain

$$E\mathbf{W}_1 - (1-\rho-\lambda_2 s)E\mathbf{W}_2 = \rho\frac{1-\rho-\lambda_2 s}{\lambda_2 s}[\frac{\sigma_1'(\lambda_2)}{\sigma_1(\lambda_2)}+s_1]; \tag{6.74}$$

hence it is sufficient to prove that, for all $\lambda_2 \geqslant 0$:

$$\frac{\sigma_1'(\lambda_2)}{\sigma_1(\lambda_2)} \geqslant -s_1 = \frac{\sigma_1'(0)}{\sigma_1(0)}. \tag{6.75}$$

For $\lambda_2=0$, (6.75) turns into an equality; this implies that we are done if $\sigma_1'(\lambda_2)/\sigma_1(\lambda_2)$ is non-decreasing in $\lambda_2>0$, or equivalently, if for $\lambda_2>0$:

$$\{\frac{d}{dy}\frac{\sigma_1'(y)}{\sigma_1(y)}\}_{y=\lambda_2} = \frac{\sigma_1''(\lambda_2)\sigma_1(\lambda_2)-[\sigma_1'(\lambda_2)]^2}{[\sigma_1'(\lambda_2)]^2} \geqslant 0. \tag{6.76}$$

Note that, using Cauchy-Schwarz's inequality:

$$\sigma_1''(\lambda_2)\sigma_1(\lambda_2) = \int_0^\infty t^2 e^{-\lambda_2 t} dS_1(t) \int_0^\infty e^{-\lambda_2 t} dS_1(t) \geqslant$$

$$\left[\int_0^\infty te^{-\lambda_2 t} dS_1(t)\right]^2 = [\sigma_1'(\lambda_2)]^2, \tag{6.77}$$

and the result follows.

REMARK 6.4

In the current section, we have calculated the mean waiting times at the two queues from the LST's of the marginal waiting time distributions. However, it is also possible to employ a more direct approach. Denote by $\mathbf{X}_1^W$ the number of waiting class-1 customers at an arbitrary epoch; denote by SO_1 a typical switch-over interval from Q_1 to Q_2. By looking at the amount of work found by an arriving type-1 customer, and conditioning on the state of the server at the customer's arrival epoch, we find:

$$E\mathbf{W}_1 = \beta_1 E\mathbf{X}_1^W + \rho_1\frac{\beta_1^{(2)}}{2\beta_1} + \rho_2(\frac{\beta_2^{(2)}}{2\beta_2}+s_2) +$$

$$(1-\rho)[\frac{s_2}{s}\frac{s_2^{(2)}}{2s_2}+\frac{s_1}{s}(\frac{s_1^{(2)}}{2s_1}+s_2+q_1\beta_2)], \tag{6.78}$$

where,

$q_1 :=$ Pr{ server finds Q_2 not empty | type$-$1 arrival in SO_1}.

Rearranging terms in (6.78) and applying Little's formula, we find:

$$EW_1 = \frac{\lambda_1\beta_1^{(2)}+\lambda_2\beta_2^{(2)}}{2(1-\rho_1)} + \frac{1-\rho}{1-\rho_1}\frac{s^{(2)}}{2s} + \frac{\rho_2 s_2}{1-\rho_1} + \frac{1-\rho}{1-\rho_1}q_1\beta_2\frac{s_1}{s}. \tag{6.79}$$

We shall determine q_1 using a probabilistic argument.

$q_1 = 1 -$ Pr{ server finds Q_2 empty | type$-$1 arrival in SO_1}

$= 1 -$ Pr{ Q_2 is empty when server starts SO_1}$\times$

Pr{ no type$-$1 arrivals during SO_1 | type$-$1 arrival in SO_1 }

$$= 1 - \frac{1-\rho-\lambda_2 s}{1-\rho}\frac{1}{\sigma_1(\lambda_2)} \times$$

Pr{ no type$-$2 arrivals during SO_1 | type$-$1 arrival in SO_1 }.

According to Feller [1966], the density function of the length of the switch-over interval in which an arrival takes place, is given by:

$$\frac{t}{s_1}\frac{dS_1(t)}{dt}.$$

Hence we have,

Pr{ no type$-$2 arrivals during SO_1 | type$-$1 arrival in SO_1 } =

$$\int_0^\infty te^{-\lambda_2 t}d\frac{S_1(t)}{s_1} = \frac{-\sigma_1'(\lambda_2)}{s_1}. \tag{6.80}$$

Combining (6.79) and the expression for q_1, we find the exact mean waiting time of class-1 customers, cf. also (6.71). Note that we can find EW_2 by substituting the expression for EW_1 in the pseudoconservation law.

TABLE 6.1 For a detailed discussion of these tables see p.101.

Mean waiting times and cycle-time moments for the alternating-service model; influence of the switch-over times

6.1.a The influence of the switch-over time distributions.
$B_i(\cdot)$ negative exponential, $i=1,2$; in the first four rows $\beta_1=\beta_2=0.2$, in the last four $\beta_1=\beta_2=0.8$.
$\lambda=1$, $r_1=0.7$, $s_1=s_2=0.1$.

$S_1(\cdot)$	$S_2(\cdot)$	$E\mathbf{W}_1$	$E\mathbf{W}_2$	$E\mathbf{C}_1^2$	$E\mathbf{C}_2^2$	$E\mathbf{C}_{b,1}$	$E\mathbf{C}_{b,2}$	$F_1(0,0)$	$F_2(0,0)$
det	det	0.236	0.192	0.082	0.082	0.425	0.461	0.796	0.829
det	exp	0.265	0.221	0.093	0.095	0.427	0.462	0.797	0.827
exp	det	0.268	0.216	0.094	0.092	0.426	0.465	0.796	0.830
exp	exp	0.297	0.244	0.104	0.105	0.427	0.466	0.797	0.829
det	det	16.566	2.158	2.100	2.007	1.317	1.745	0.286	0.300
det	exp	16.671	2.172	2.110	2.020	1.318	1.746	0.287	0.299
exp	det	16.674	2.169	2.113	2.017	1.317	1.746	0.286	0.300
exp	exp	16.779	2.183	2.124	2.031	1.318	1.747	0.287	0.300

6.1.b The influence of s_1 and s_2 for given $s = s_1 + s_2$.
$B_i(\cdot)$ negative exponential, $i = 1,2$; $\beta_1 = \beta_2 = 0.8$.
$\lambda = 1$, $r_1 = 0.7$.

$S_1(\cdot)$	$S_2(\cdot)$	s_1	s_2	EW_1	EW_2	EC_1^2	EC_2^2	$EC_{b,1}$	$EC_{b,2}$	$F_1(0,0)$	$F_2(0,0)$
det	det	0.05	0.15	16.566	2.158	2.100	2.007	1.317	1.745	0.282	0.310
det	det	0.1	0.1	16.566	2.158	2.100	2.007	1.317	1.745	0.286	0.300
det	det	0.15	0.05	16.566	2.158	2.100	2.007	1.317	1.745	0.290	0.289
det	exp	0.05	0.15	16.802	2.189	2.124	2.037	1.319	1.747	0.283	0.309
det	exp	0.1	0.1	16.671	2.172	2.110	2.020	1.318	1.746	0.287	0.299
det	exp	0.15	0.05	16.592	2.161	2.102	2.010	1.317	1.745	0.290	0.289
exp	det	0.05	0.15	16.593	2.161	2.103	2.010	1.317	1.745	0.282	0.310
exp	det	0.1	0.1	16.674	2.169	2.113	2.017	1.317	1.746	0.286	0.300
exp	det	0.15	0.05	16.808	2.183	2.130	2.031	1.318	1.748	0.290	0.290
exp	exp	0.05	0.15	16.829	2.192	2.127	2.040	1.319	1.747	0.283	0.310
exp	exp	0.1	0.1	16.779	2.183	2.124	2.031	1.318	1.747	0.287	0.300
exp	exp	0.15	0.05	16.834	2.187	2.133	2.034	1.318	1.748	0.291	0.290

For a detailed discussion of these tables see p.101.

TABLE 6.2 For a detailed discussion of these tables see p.101.

Mean waiting times and cycle-time moments for the alternating-service model

Case A: $B_i(\cdot)$ negative exponential, $i=1,2$;
$\lambda=1$, $s_1=s_2=0.1$ (constant switch-over times).

r_1	β_1	β_2	$E\mathbf{W}_1$	$E\mathbf{W}_2$	$E\mathbf{C}_1^2$	$E\mathbf{C}_2^2$	$E\mathbf{C}_{b,1}$	$E\tilde{\mathbf{C}}_{b,1}$	$E\mathbf{C}_{b,2}$	$E\tilde{\mathbf{C}}_{b,2}$	$F_1(0,0)$	$F_2(0,0)$
0.7	0.2	0.2	0.236	0.192	0.082	0.082	0.425	0.426	0.461	0.465	0.796	0.829
0.7	0.2	0.5	0.387	0.330	0.139	0.142	0.478	0.471	0.789	0.814	0.772	0.811
0.7	0.2	0.8	0.726	0.599	0.248	0.256	0.554	0.526	1.114	1.163	0.743	0.786
0.7	0.5	0.2	0.738	0.440	0.237	0.231	0.743	0.745	0.629	0.615	0.734	0.759
0.7	0.5	0.5	1.152	0.671	0.369	0.362	0.824	0.824	0.999	1.077	0.690	0.720
0.7	0.5	0.8	2.131	1.098	0.618	0.612	0.932	0.921	1.361	1.500*	0.630	0.663
0.7	0.8	0.2	2.680	0.977	0.707	0.686	1.062	1.064	0.943	0.909	0.606	0.623
0.7	0.8	0.5	5.088	1.406	1.141	1.096	1.175	1.176	1.351	1.500*	0.494	0.513
0.7	0.8	0.8	16.566	2.158	2.100	2.007	1.317	1.316	1.745	1.800*	0.286	0.300
0.9	0.2	0.2	0.256	0.170	0.081	0.081	0.408	0.408	0.477	0.488	0.767	0.831
0.9	0.2	0.5	0.308	0.208	0.099	0.099	0.422	0.421	0.803	0.854	0.758	0.825
0.9	0.2	0.8	0.400	0.269	0.127	0.129	0.438	0.435	1.121	1.200*	0.749	0.816
0.9	0.5	0.2	1.084	0.416	0.292	0.288	0.714	0.714	0.703	0.727	0.653	0.706
0.9	0.5	0.5	1.279	0.473	0.341	0.336	0.737	0.737	1.050	1.200*	0.633	0.687
0.9	0.5	0.8	1.586	0.557	0.413	0.406	0.762	0.761	1.384	1.500*	0.610	0.664
0.9	0.8	0.2	9.564	0.871	1.197	1.182	1.020	1.020	1.122	1.200*	0.304	0.328
0.9	0.8	0.5	16.161	0.959	1.460	1.420	1.053	1.053	1.459	1.500*	0.215	0.233
0.9	0.8	0.8	43.679	1.077	1.841	1.767	1.087	1.087	1.786	1.800*	0.099	0.108

TABLE 6.2 (CONT'D) For a detailed discussion of these tables see p.101.

Case B: $B_i(\cdot)$ hyperexponential (H_2) with squared coefficient of variation 4 and balanced means (cf. Tijms [1986]);
$\lambda = 1$, $s_1 = s_2 = 0.1$ (constant switch-over times).

r_1	β_1	β_2	$E\mathbf{W}_1$	$E\mathbf{W}_2$	EC_1^2	EC_2^2	$E\mathbf{C}_{b,1}$	$E\tilde{\mathbf{C}}_{b,1}$	$E\mathbf{C}_{b,2}$	$E\tilde{\mathbf{C}}_{b,2}$	$F_1(0,0)$	$F_2(0,0)$
0.7	0.2	0.2	0.330	0.267	0.114	0.113	0.426	0.426	0.463	0.465	0.796	0.830
0.7	0.2	0.5	0.684	0.547	0.230	0.230	0.481	0.471	0.789	0.814	0.773	0.811
0.7	0.2	0.8	1.546	1.122	0.469	0.470	0.559	0.526	1.112	1.163	0.744	0.785
0.7	0.5	0.2	1.389	0.832	0.435	0.425	0.744	0.745	0.642	0.615	0.735	0.762
0.7	0.5	0.5	2.325	1.273	0.689	0.667	0.828	0.824	1.006	1.077	0.692	0.722
0.7	0.5	0.8	4.660	2.118	1.192	1.154	0.937	0.921	1.363	1.500*	0.632	0.663
0.7	0.8	0.2	5.706	2.073	1.450	1.418	1.063	1.064	0.964	0.909	0.608	0.627
0.7	0.8	0.5	11.124	2.855	2.275	2.188	1.177	1.176	1.361	1.500*	0.496	0.516
0.7	0.8	0.8	37.323	4.258	4.130	3.932	1.319	1.316	1.748	1.800*	0.287	0.301
0.9	0.2	0.2	0.355	0.235	0.112	0.111	0.408	0.408	0.478	0.488	0.767	0.831
0.9	0.2	0.5	0.470	0.309	0.148	0.147	0.422	0.421	0.799	0.854	0.758	0.824
0.9	0.2	0.8	0.689	0.441	0.212	0.211	0.438	0.435	1.115	1.200*	0.749	0.815
0.9	0.5	0.2	2.084	0.795	0.558	0.548	0.714	0.714	0.712	0.727	0.653	0.707
0.9	0.5	0.5	2.509	0.902	0.659	0.637	0.737	0.737	1.050	1.200*	0.633	0.687
0.9	0.5	0.8	3.206	1.069	0.813	0.776	0.762	0.761	1.379	1.500*	0.610	0.664
0.9	0.8	0.2	20.659	1.861	2.564	2.521	1.020	1.020	1.127	1.200*	0.304	0.329
0.9	0.8	0.5	35.133	2.021	3.108	2.988	1.053	1.053	1.460	1.500*	0.215	0.233
0.9	0.8	0.8	95.841	2.248	3.908	3.687	1.087	1.087	1.786	1.800*	0.099	0.107

TABLE 6.2 (CONT'D) For a detailed discussion of these tables see p.101.

Case C: $B_1(\cdot)$ hyperexponential (H_2) with squared coefficient of variation 4 and balanced means (cf. Tijms [1986]);
$B_2(\cdot)$ deterministic;
$\lambda=1$, $s_1=s_2=0.1$ (constant switch-over times).

r_1	β_1	β_2	$E\mathbf{W}_1$	$E\mathbf{W}_2$	$E\mathbf{C}_1^2$	$E\mathbf{C}_2^2$	$E\mathbf{C}_{b,1}$	$E\tilde{\mathbf{C}}_{b,1}$	$E\mathbf{C}_{b,2}$	$E\tilde{\mathbf{C}}_{b,2}$	$F_1(0,0)$	$F_2(0,0)$
0.7	0.2	0.2	0.293	0.235	0.102	0.100	0.425	0.426	0.464	0.465	0.796	0.830
0.7	0.2	0.5	0.400	0.333	0.145	0.143	0.475	0.471	0.794	0.814	0.771	0.813
0.7	0.2	0.8	0.605	0.507	0.220	0.219	0.547	0.526	1.121	1.163	0.741	0.789
0.7	0.5	0.2	1.334	0.795	0.418	0.405	0.743	0.745	0.645	0.615	0.734	0.762
0.7	0.5	0.5	1.862	1.021	0.568	0.540	0.823	0.824	1.015	1.077	0.689	0.725
0.7	0.5	0.8	2.909	1.391	0.815	0.772	0.929	0.921	1.376	1.500*	0.628	0.667
0.7	0.8	0.2	5.602	2.030	1.425	1.388	1.063	1.064	0.966	0.909	0.607	0.628
0.7	0.8	0.5	9.975	2.566	2.067	1.970	1.175	1.176	1.366	1.500*	0.494	0.518
0.7	0.8	0.8	29.029	3.410	3.361	3.157	1.317	1.316	1.752	1.800*	0.286	0.303
0.9	0.2	0.2	0.342	0.226	0.108	0.107	0.408	0.408	0.482	0.488	0.767	0.832
0.9	0.2	0.5	0.382	0.252	0.122	0.121	0.422	0.421	0.812	0.854	0.758	0.826
0.9	0.2	0.8	0.444	0.293	0.143	0.141	0.437	0.435	1.135	1.200*	0.748	0.818
0.9	0.5	0.2	2.061	0.785	0.552	0.541	0.714	0.714	0.718	0.727	0.653	0.708
0.9	0.5	0.5	2.346	0.842	0.619	0.596	0.737	0.737	1.069	1.200*	0.633	0.689
0.9	0.5	0.8	2.728	0.915	0.703	0.666	0.762	0.761	1.407	1.500*	0.610	0.667
0.9	0.8	0.2	20.557	1.851	2.552	2.508	1.020	1.020	1.129	1.200*	0.304	0.329
0.9	0.8	0.5	34.081	1.961	3.021	2.901	1.053	1.053	1.465	1.500*	0.215	0.234
0.9	0.8	0.8	88.887	2.091	3.652	3.431	1.087	1.087	1.789	1.800*	0.099	0.108

Chapter 7

APPROXIMATIONS FOR MEAN WAITING TIMES IN POLLING SYSTEMS

7.1. Introduction

In this chapter we consider the application of the pseudoconservation law as a tool for the construction of approximations for mean waiting times. As discussed in Chapters 2 and 3 the mathematical complexity of a detailed analysis of polling models makes the need for such approximations evident. The model and notation under consideration are those of Section 2.2, but with *single* Poisson arrivals; the visit disciplines at the queues are either exhaustive, gated or 1-limited. The pseudoconservation law for this case reads (cf. (3.44)):

$$\sum_{i\in e}\rho_i E\mathbf{W}_i + \sum_{i\in g}\rho_i E\mathbf{W}_i + \sum_{i\in 1l}\rho_i\left(1-\frac{\lambda_i s}{1-\rho}\right)E\mathbf{W}_i =$$

$$\rho\sum_{i=1}^{N}\frac{\lambda_i\beta_i^{(2)}}{2(1-\rho)} + \rho\frac{s^{(2)}}{2s} + \frac{s}{2(1-\rho)}\left(\rho^2-\sum_{i\in e}\rho_i^2+\sum_{i\in g,\,1l}\rho_i^2\right) \tag{7.1}$$

Several mean waiting time approximations for this model have been suggested in the literature. We shall briefly discuss some of the more recent ones. Most of these recent approximations are based on the following idea, that has independently been developed in Everitt [1986a] for the two cases of exhaustive and gated service at all queues, and in Boxma and Meister [1986] for 1-limited service at all queues:

For each class of customers obtain a linear relation between the mean waiting time $E\mathbf{W}_i$ *and the mean residual cycle time* $E\mathbf{R}_i$*:* $E\mathbf{W}_i = A_i E\mathbf{R}_i + B_i$*, where* A_i *and* B_i *are known expressions. Now assume that the N mean residual cycle times are exactly the same:* $E\mathbf{R}_i \equiv E\mathbf{R}$*. As a result, the mean waiting times for all classes are expressed in a common unknown:* $E\mathbf{W}_i = A_i E\mathbf{R} + B_i$*,* $i=1,\ldots,N$*. Finally substitute these N expressions into (7.1) and solve for the one unknown* $E\mathbf{R}$*.*

We shall apply this idea to derive a mean waiting time approximation for

cyclic-service systems with a *mixture* of exhaustive, gated and 1-limited visit disciplines.

In Groenendijk [1988a] a refinement of this approximation is studied, based on a more detailed investigation of cycle times and taking into account information about previous cycles. The resulting approximation is more accurate but less transparent than the previous method. The ideas in Groenendijk [1988a] are partly based on those of Srinivasan [1988]. For the case of 1-limited service at all queues, Srinivasan had also improved upon Boxma and Meister [1986,1987] by studying in more detail (conditional) cycle times before eventually applying the pseudoconservation law.

Fuhrmann and Wang [1988] consider the difficult but important case of k-limited service (cf. Section 3.4). They derive heuristic mean waiting time approximations based on tight bounds (cf. Fuhrmann [1987]) for the pseudoconservation law. In Groenendijk and Levy [1989] an approximation procedure combining ideas of Boxma and Meister [1987] and Fuhrmann and Wang [1988] is proposed for a model with bulk arrivals and 1-limited service at all queues. This approximation is then applied to the performance analysis of Transaction Driven Computer Systems. Pang and Donaldson [1986] suggest a mean waiting-time approximation for discrete-time cyclic-service systems with gated service at all queues. They express the mean waiting time at Q_i in the second moment $v_{i,i}$ of the sum of Q_i's visit time and the subsequent switch-over time; next they obtain a linear relation between $v_{i+1,i+1}$ and $v_{i,i}$ for all i; and finally they solve for the $v_{i,i}$ by deriving an extra relation between $v_{1,1}, \ldots, v_{N,N}$. At this last stage the pseudoconservation law is elegantly brought into the picture. It turns out that their approximation is very accurate.

Eckberg and Meier-Hellstern [1988] consider a completely symmetric cyclic-service system with gated service at all queues and a general arrival process. They propose an approximation for the first two moments of the waiting time. They claim their approximation to become more accurate as the number of queues grows. Their arguments contain a number of heuristic elements, some of which seem hard to justify.

The organization of this chapter is as follows. In Sections 7.2.1 and 7.2.2 the basic mean waiting time approximation based on the pseudoconservation law is developed. A refinement of this approximation is briefly discussed in Section 7.2.3. In Section 7.2.4 an extension of the basic approximation to bulk arrivals is considered. Some special cases for which the approximation method yields exact results are discussed in Section 7.2.5. Section 7.3 contains an extensive investigation of the results of the approximation. The detailed numerical results may be found at the end of the chapter. As can be seen from the results in the tables the approximation yields very accurate results for low and moderate loads; for high load the accuracy decreases somewhat. Finally in Section 7.4 some concluding remarks are made.

7.2. Basic mean waiting time approximation

In this section we present a simple approximation for the mean waiting times at the individual queues. The approximation is based on the pseudoconservation law. Before presenting a detailed derivation we shall briefly outline the method.

7.2.1 *Outline of the method*

The approximation basically consists of two steps.

The first step is to express the mean waiting time at Q_i in $E\mathbf{R}_i$: the expected time until the next arrival of S at Q_i. Below it will be shown that
for Q_i gated:

$$E\mathbf{W}_i = (1+\rho_i)E\mathbf{R}_i; \tag{7.2}$$

for Q_i exhaustive:

$$E\mathbf{W}_i \approx (1-\rho_i)E\mathbf{R}_i, \tag{7.3}$$

and for Q_i 1-limited:

$$E\mathbf{W}_i \approx \frac{E\mathbf{R}_i}{1-\lambda_i E\mathbf{C}_{b,i}}, \tag{7.4}$$

where $E\mathbf{C}_{b,i}$ denotes the mean cycle time at Q_i, given the cycle contains a service at Q_i. An approximation for $E\mathbf{C}_{b,i}$ will be presented in the sequel.

The second step is to assume $E\mathbf{R}_i$ is the same for all i: $E\mathbf{R}_i \equiv E\mathbf{R}$ (cf. Remark 7.1 below). We then substitute the expressions for the mean waiting times into the pseudo-conservation law (7.1) and thus obtain the following approximation for the mean residual cycle time:

$$E\mathbf{R} \approx \left[\rho \sum_{i=1}^{N} \frac{\lambda_i \beta_i^{(2)}}{2(1-\rho)} + \rho \frac{s^{(2)}}{2s} + \frac{s}{2(1-\rho)} (\rho^2 - \sum_{i \in e} \rho_i^2 + \sum_{i \in g,\, 1l} \rho_i^2) \right] \times$$

$$\times \left[\sum_{i \in e} \rho_i (1-\rho_i) + \sum_{i \in g} \rho_i (1+\rho_i) + \sum_{i \in 1l} \rho_i \left(1 - \frac{\lambda_i s}{1-\rho}\right) \frac{1}{1-\lambda_i E\mathbf{C}_{b,i}} \right]^{-1}. \tag{7.5}$$

Substitution of $E\mathbf{R}$ in the above formulas (7.2)-(7.4) for $E\mathbf{W}_i$ yields the approximation for the mean waiting times at the various queues.

Remark 7.1

By using stochastic mean value theorems it may be easily proved that although successive cycle times for Q_i are dependent and hence do not form a renewal process,

$$E\mathbf{R}_i = \frac{E\mathbf{C}_i^2}{2E\mathbf{C}_i}; \tag{7.6}$$

for an alternative proof see Franken, König, Arndt and Schmidt [1982]. Note that this result resembles a well-known result for the forward recurrence time from the theory of regenerative processes which, however, is not applicable here.

Although *mean* cycle times are independent of i, the same obviously does not in general hold for the *second moments* of the cycle time. However, the differences between the $E\mathbf{C}_i^2$ are in general very small and for the purpose of our approximation we *assume* them all to be equal (cf. also the discussion at the end of this section).

7.2.2 *Development of the approximation method*

We shall show in detail how the approximation method is developed. First we derive the relations (7.2)-(7.4) between the mean waiting times and the residual cycle times for each visit discipline.

1. *Gated visit discipline at Q_i.*
   Consider a tagged customer arriving at Q_i. The mean waiting time of this customer consists of two components. First a mean residual type-i cycle time ($E\mathbf{R}_i$), because due to the gating mechanism a customer is never served in the cycle in which it arrived. Secondly, the mean time from the instant the server arrives at Q_i till the moment the tagged customer starts to receive service. This component consists of all the service times from the customers that arrived after the start of the previous visit period, but before the tagged customer arrived. Hence it is given by $\lambda_i E\mathbf{R}_i\beta_i$ $(=\rho_i E\mathbf{R}_i)$.

2. *Exhaustive visit discipline at Q_i.*
   We shall prove that

$$E\mathbf{W}_i = (1-\rho_i)E\tilde{\mathbf{R}}_i, \tag{7.7}$$

   where $E\tilde{\mathbf{R}}_i$ is the mean residual cycle time at Q_i that now corresponds to a type-i cycle, $\tilde{\mathbf{C}}_i$, being defined as the time between two successive *departures* of S from Q_i. After we have proved (7.7) we shall as an approximation *ignore* the (small) difference between $E\mathbf{R}_i$ and $E\tilde{\mathbf{R}}_i$, and so we arrive at relation (7.3).

   We elucidate (7.7) by providing two arguments, each having their own merits. The first argument consists of an algebraic proof. Denote the LST of the cycle time and intervisit time for Q_i by $\tilde{C}_i(\cdot)$, and $\tilde{I}_i(\cdot)$, respectively. Denote the LST of the length of a visit period at Q_i starting with one customer present by $\gamma_i(\cdot)$. As in Bux and Truong [1983], it may be proved that

$$\tilde{C}_i(s) = \tilde{I}_i(s+\lambda_i-\lambda_i\gamma_i(s)), \quad \text{Re } s \geqslant 0. \tag{7.8}$$

Hence we obtain the following result for the second moment of the inter-visit time for Q_i:

$$E\tilde{\mathbf{I}}_i^2 = (1-\rho_i)^2 E\tilde{\mathbf{C}}_i^2 - \frac{\lambda_i \beta_i^{(2)}}{1-\rho} s. \tag{7.9}$$

Furthermore we have, cf. Doshi [1986],

$$E\mathbf{W}_i = \frac{\lambda_i \beta_i^{(2)}}{2(1-\rho_i)} + \frac{E\tilde{\mathbf{I}}_i^2}{2E\tilde{\mathbf{I}}_i}, \tag{7.10}$$

Substituting (7.9) into (7.10) and using $E\tilde{\mathbf{R}}_i = E\tilde{\mathbf{C}}_i^2 / 2E\tilde{\mathbf{C}}_i$ yields (7.7). Note that the derivation above is independent of the visit disciplines at the other queues.

The second argument is due to Doshi [personal communication]. It is heuristic but it provides a useful insight. He observes that $E\tilde{\mathbf{R}}_i$ consists of two components: first $E\mathbf{W}_i$, the mean waiting time of the hypothetical customer whose arrival marks the beginning of the residual cycle, and secondly $\rho_i E\tilde{\mathbf{R}}_i$, the mean work arriving at Q_i during the residual cycle. As noted by Levy [personal communication], a minor variant of this argument is to write

$$E\tilde{\mathbf{R}}_i = E\mathbf{W}_i + \lambda_i E\mathbf{W}_i \frac{\beta_i}{1-\rho_i},$$

the last term in the right-hand side representing the mean number of arrivals at Q_i during $E\mathbf{W}_i$ times the mean length of the visit period at Q_i generated by one such arrival. Note that the hypothetical customer himself should not contribute to $E\tilde{\mathbf{R}}_i$!

3. *1-Limited visit discipline at Q_i.*
Denote by $\mathbf{X}_i$ the number of waiting customers at Q_i just before the arrival of a type-i customer, and by $\mathbf{C}_{b,i}$ the length of a cycle of the server starting with a service at Q_i. A customer arriving at Q_i first has to wait until the server returns to that queue. Subsequently he has to wait until all customers in front of him have been served. Therefore, the waiting time of this customer consists of two parts: a residual cycle $\mathbf{R}_i$ and just as many 'busy i-cycles' $\mathbf{C}_{b,i}$ as there are waiting customers in front of him:

$$E\mathbf{W}_i \approx E\mathbf{R}_i + E\mathbf{X}_i E\mathbf{C}_{b,i}. \tag{7.11}$$

This is actually an approximation, because the use of Wald's lemma, leading to the last term in the right-hand side of (7.11), is not justified. Using the fact that Poisson arrivals see time averages, and Little's formula, we

may write $E\mathbf{X}_i = \lambda_i E\mathbf{W}_i$. Substitution of this in (7.11) leads to (7.4).

For the approximation of $E\mathbf{C}_{b,i}$, note that it consists of a type-i service, and, possibly, services of customers at the other queues, plus the total switch-over time. It is now assumed that the number of type-j customers *arriving* during $E\mathbf{C}_{b,i}$, equals the number of type-j customers *departing* during $E\mathbf{C}_{b,i}$ (balance of flow). Hence we approximate the average number of type-j services during $E\mathbf{C}_{b,i}$ by $\lambda_j E\mathbf{C}_{b,i}$. The idea of this assumption is due to Kühn [1979]. However, for the case that Q_j has a 1-limited visit discipline, we bound $\lambda_j E\mathbf{C}_{b,i}$ by one since the number of type-i services during any cycle is at most one. So our approximation for $E\mathbf{C}_{b,i}$ will be calculated from the equation

$$E\mathbf{C}_{b,i} = \beta_i + s + \sum_{j \in e,g} \lambda_j E\mathbf{C}_{b,i} \beta_j + \sum_{j \in 1l} \min(1, \lambda_j E\mathbf{C}_{b,i}) \beta_j. \tag{7.12}$$

If $\lambda_j E\mathbf{C}_{b,i} \leqslant 1$ for all queues Q_j with 1-limited service, then (7.12) simplifies to

$$E\mathbf{C}_{b,i} = \frac{\beta_i + s}{1 - \rho + \rho_i}, \tag{7.13}$$

which is the approximation suggested by Boxma and Meister [1987]. In Chapter 6, $E\mathbf{C}_{b,i}$ has been calculated exactly for the two-queue model with 1-limited service at both queues.

In order to compute $E\mathbf{C}_{b,i}$ from (7.12) we propose to use the following iterative scheme. Let for $n = 1,2,...$:

$$x^{(n)} = \beta_i + s + \sum_{j \in e,g} \rho_j x^{(n-1)} + \sum_{\substack{j \in 1l \\ j \neq i}} \min(1, \lambda_j x^{(n-1)}) \beta_j. \tag{7.14}$$

For all starting values $x^{(0)} > 0$ and for all $n = 1,2,...$, it is easily proved by induction that

$$\left| x^{(n+1)} - x^{(n)} \right| < \left| x^{(n)} - x^{(n-1)} \right|; \tag{7.15}$$

hence, according to the fixed point theorem the recursion (7.14) has a unique fixed point x^*, which we choose as our approximation for $E\mathbf{C}_{b,i}$. Numerical experiences suggest that

$$x^{(0)} := \frac{\beta_i + s}{1 - \rho + \rho_i} \tag{7.16}$$

is a good starting point for the iteration.

In general, the assumption that $E\mathbf{R}_i \equiv E\mathbf{R}$ for all i seems to be fairly accurate. However, see Everitt [1986a] and Boxma and Groenendijk [1988b] for a

discussion of some factors which may influence accuracy. The conclusion seems to be that the assumption becomes less accurate as the system is more asymmetric. Furthermore, it appears that large variances of the switch-over times also have a negative effect on the accuracy of this particular assumption. See Chapter 6 for some exact results for second moments of cycle times in the two-queue case with 1-limited service at both queues. In Blanc [1988] a table is included which demonstrates the effect of the order in which the server attends the queues on a system with four 1-limited queues and zero switch-over times. If one queue, say Q_1, is more heavily loaded than the other queues, it appears that for these queues it is best to immediately follow Q_1 and it is worst to immediately precede Q_1. However, these differences are rather small (in Blanc's example with $\rho=0.9$ the largest difference was 4.3%). When all queues in the system have an *exhaustive* visit discipline we can observe the same effect. It is interesting to remark that when the visit discipline for all queues in the system is *gated* we observe the opposite effect: for the more lightly loaded queues it is best to immediately precede the heavily loaded queue and it is worst to immediately follow this queue. Again the differences are usually rather small - in the order of a few percent.

7.2.3 *Refinement of the approximation*

In this section we briefly formulate another algorithm for approximating the mean waiting times at the various queues. A more detailed study of the cycle times and in particular taking into account information about previous cycles leads to the formulation of an iteration scheme. As a result, the method is considerably more involved than the one described in the previous section and lacks the simplicity and transparency of that method. However, the scheme is easily programmed on any (personal) computer and requires little run time and memory.

The algorithm can be formulated as follows. In each step of the iteration:

- we express all mean waiting times in the mean residual cycle time $E\mathbf{R}_i$. For gated and exhaustive service we shall use the expressions given in (7.2) and (7.3). The expression we shall use for 1-limited service is of the form

$$E\mathbf{W}_i \approx \frac{1}{1-\lambda_i E\mathbf{C}_{b,i}}[E\mathbf{R}_i + H_i(E\mathbf{R}_i)]; \tag{7.17}$$

  $H_i(\cdot)$, which is a function of $E\mathbf{R}_i$, will be specified below.

- as before we assume that $E\mathbf{R}_i \equiv E\mathbf{R}$ for all i.

To start the iteration we take $H_i(E\mathbf{R}_i) =: H_i^{(0)} = 0$. Then (7.2), (7.3) and (7.17) represent a set of N linear relations between $E\mathbf{W}_i$ and $E\mathbf{R}_i$. Substituting these relations into the pseudoconservation law (7.1) and solving for the one unknown $E\mathbf{R}$, we obtain our first approximation for the mean residual cycle time, which we shall denote by $E\mathbf{R}^{(1)}$. Note that this initial step is just the

approximation of Section 7.2.2. In the second step of the iteration, we use $E\mathbf{R}^{(1)}$ to compute the $H_i(E\mathbf{R}_i^{(1)}) =: H_i^{(1)}$. So now we take for a queue Q_i with 1-limited service:

$$E\mathbf{W}_i \approx \frac{1}{1-\lambda_i E\mathbf{C}_{b,i}}[E\mathbf{R}_i + H_i^{(1)}]. \tag{7.18}$$

(7.2), (7.3) and (7.18) again represent a set of N linear relations between $E\mathbf{W}_i$ and $E\mathbf{R}_i$. Substituting these relations into the conservation law and solving for the mean residual cycle time gives us our second approximation for $E\mathbf{R}_i$, which we shall denote by $E\mathbf{R}^{(2)}$. We then compute $H_i^{(2)}$ from $E\mathbf{R}^{(2)}$. In the third step we use $H_i^{(2)}$ to compute $E\mathbf{R}^{(3)}$, etc... . Thus we switch back and forth between the computation of the mean residual cycle time and the terms H_i. The iteration is continued until the mean waiting times in subsequent steps do not significantly change any more. Unfortunately, convergence of this procedure is not guaranteed. We have had in fact a few cases in which the algorithm does not converge and alternates between two values for $E\mathbf{R}$. At this moment we have no explanation for this phenomenon; we have no insight whether it should be contributed to numerical difficulties or whether there are other causes.

Assume that Q_i has a 1-limited visit discipline. An approximation for $E\mathbf{C}_{b,i}$ has already been provided in Section 7.2.2. So the only unknown left in (7.17) is $H_i(E\mathbf{R}_i)$, which we shall next specify. From Groenendijk [1988a],

$$H_i(E\mathbf{R}_i) = \rho_i(1-p_i)[B_i^{(1)} - E\mathbf{C}_{b,i}] + (1-\rho_i-p_i)[B_i^{(2)} - E\mathbf{C}_{b,i}], \tag{7.19}$$

with

$$B_i^{(1)} = \beta_i + s + \sum_{j!=i} \min(1, \lambda_j[A_i^* - \frac{\beta_i^{(2)}}{\beta_i} + \beta_i])\beta_j, \tag{7.20}$$

$$B_i^{(2)} = \beta_i + s + \sum_{j\neq i} \min(1, \lambda_j[\frac{\lambda_i \beta_i^{(2)}}{2(1-\rho_i)} + \frac{E\mathbf{R}_i}{1-\rho_i} - \frac{\rho_i}{1-\rho_i} A_i^* + \beta_i])\beta_j, \tag{7.21}$$

and

$$p_i = \frac{1-\lambda_i E\mathbf{C}_{b,i}}{1 - \lambda_i E\mathbf{C}_{b,i} + \frac{\lambda_i}{2(1-\rho_i)}(E\mathbf{R}_i - \rho_i \frac{\beta_i^{(2)}}{2\beta_i} - \rho_i A_i^*) + \lambda_i B_i^{(2)}}. \tag{7.22}$$

We shall only briefly discuss the interpretation of the various terms occurring in (7.19). Call the cycle in which the tagged (type-i) customer arrived the A-cycle and the cycle following the A-cycle the B-cycle. Note that the B-cycle always contains a type-i service. $B_i^{(1)}$ is an approximation for the mean length

of the B-cycle, given the tagged customer arrived during the visit time of the server at Q_i in the A-cycle. $B_i^{(2)}$ is an approximation for the mean length of the B-cycle, given the tagged customer arrived during the intervisit time of the server with respect to Q_i in the A-cycle. A_i^* is an approximation for the mean length of the A-cycle given the tagged customer arrived during the visit time of the server at Q_i. Finally, p_i is an approximation for the steady-state probability that there are zero type-i customers in the system. Its derivation is based on a comparison with an M/G/1 model with exceptional service for the first customer in a busy period.

In (7.20), (7.21) and (7.22) A_i^* denotes the (unique) solution of

$$A_i = \frac{\beta_i^{(2)}}{\beta_i} + s + \sum_{j \neq i} \min(1, \lambda_j A_i)\beta_j. \tag{7.23}$$

(7.23) is easily solved using the same iteration scheme as for $E\mathbf{C}_{b,i}$, cf. (7.14). Again it may be proved that for any starting value $x^{(0)}>0$ (7.23) has a unique fixed point. A good starting value appears to be

$$x^{(0)} = \frac{\beta_i^{(2)}/\beta_i + s}{1-\rho+\rho_i}. \tag{7.24}$$

Some comments on the background of the algorithm are in order. A detailed derivation of the algorithm will not be discussed here; for this the reader is referred to Groenendijk [1988a].

The basic idea of the approximation is similar to that of the previous section and is based on the observation that the waiting time of an arbitrary type-i customer arriving to the system consists of two components. First he has to wait until the server returns to Q_i and subsequently he has to wait as many cycles as there are customers in front of him. In the previous section the second component has been approximated by the product of the expected number of waiting type-i customers and $E\mathbf{C}_{b,i}$ (cf. (7.11)). The current algorithm is based on a more detailed approximation of this second component; in particular it takes into account during which part of the cycle the tagged customer arrived. This more detailed study involves the approximation of several conditional probabilities and conditional expectations which have to be approximated; but the reward is an approximation that improves on previous approximations. The various approximations of conditional probabilities and conditional expectations may be rather inaccurate in some cases; however, even in such cases the conservation-law constraint leads to reasonable approximations for the mean waiting times.

7.2.4 *Extension to bulk arrivals*

Adapting the approximation method described in Section 7.2.2 to accommodate *bulk* Poisson arrivals is relatively straight-forward. Consider the cyclic-

service model as described before but now with the following arrival process. Assume that customers arrive at Q_i as an independent Poisson stream with rate λ_i and that each arrival to Q_i consists of a bulk of $\mathbf{L}_i$ customers: $\mathbf{L}_i$ is a generally distributed random variable with mean l_i and second moment $l_i^{(2)}$ and is assumed to be independent of other bulks at this queue or at other queues. Note that this arrival process differs from that of Definition 2.2 in that it does not incorporate correlation between bulks. Let $\rho_i := \lambda_i l_i \beta_i$ and denote by $E\mathbf{W}_i$ the mean waiting time of an arbitrary type-i customer in a bulk. Similarly as for (7.2), (7.3) and (7.4) it can easily be shown that
for Q_i gated:

$$E\mathbf{W}_i = (1+\rho_i)E\mathbf{R}_i + \frac{l_i^{(2)} - l_i}{2l_i}\beta_i; \tag{7.25}$$

for Q_i exhaustive:

$$E\mathbf{W}_i \approx (1-\rho_i)E\mathbf{R}_i + \frac{l_i^{(2)} - l_i}{2l_i}\beta_i; \tag{7.26}$$

and for Q_i 1-limited (cf. Groenendijk and Levy [1989]):

$$E\mathbf{W}_i \approx \frac{E\mathbf{R}_i}{1-\lambda_i l_i E\mathbf{C}_{b,i}} + \frac{\frac{l_i^{(2)} - l_i}{2l_i} E\mathbf{C}_{b,i}}{1-\lambda_i l_i E\mathbf{C}_{b,i}}. \tag{7.27}$$

Note that the pseudoconservation law in this case reads (cf. (3.44)):

$$\sum_{i\in e}\rho_i E\mathbf{W}_i + \sum_{i\in g}\rho_i E\mathbf{W}_i + \sum_{i\in 1l}\rho_i\left(1-\frac{\lambda_i l_i s}{1-\rho}\right)E\mathbf{W}_i = \rho\sum_{i=1}^{N}\frac{\lambda_i l_i \beta_i^{(2)}}{2(1-\rho)} +$$

$$\sum_{i=1}^{N}\frac{\lambda_i\beta_i(l_i^{(2)} - l_i)(\beta_i + s)}{2(1-\rho)} + \rho\frac{s^{(2)}}{2s} + \frac{s}{2(1-\rho)}\left(\rho^2 - \sum_{i\in e}\rho_i^2 + \sum_{i\in g,\,1l}\rho_i^2\right). \tag{7.28}$$

The same procedure as before can be applied to solve for the one unknown $E\mathbf{R}$. Having derived $E\mathbf{W}_i$ we now calculate the mean waiting time of the last customer in a bulk at Q_i, $E\mathbf{W}_i^L$. As in (7.25)-(7.27) we get:
for Q_i gated or exhaustive:

$$E\mathbf{W}_i^L = E\mathbf{W}_i + [l_i - 1 - \frac{l_i^{(2)} - l_i}{2l_i}]\beta_i; \tag{7.29}$$

for Q_i 1-limited:

$$E\mathbf{W}_i^L \approx E\mathbf{R}_i + [E\mathbf{X}_i + l_i - 1]E\mathbf{C}_{b,i}, \tag{7.30}$$

which yields

$$E\mathbf{W}_i^L \approx E\mathbf{W}_i + [l_i - 1 - \frac{l_i^{(2)} - l_i}{2l_i}]E\mathbf{C}_{b,i}. \tag{7.31}$$

Finally, a measure that may be of particular interest in systems with bulk arrivals is the mean *sojourn time* $E\mathbf{T}_i$ of a class-i bulk. It can obviously be approximated as

$$E\mathbf{T}_i \approx E\mathbf{W}_i^L + \beta_i. \tag{7.32}$$

REMARK 7.2
Note that when the distribution of $\mathbf{L}_i$ is the shifted-geometric distribution (i.e., $\Pr\{\mathbf{L}_i = 0\} = 0$, $\Pr\{\mathbf{L}_i = l\} = (1-p)p^{l-1}$ for $l = 1,2,3,...$) then we have $(l_i^{(2)} - l_i)/2l_i = l_i - 1$, which implies that $E\mathbf{W}_i^L \equiv E\mathbf{W}_i$.

REMARK 7.3
Generally speaking, increasing the bulk size introduces more dependency in the system and therefore causes higher mean waiting times.

7.2.5 *Special cases*

There are several special cases for which the approximation methods yields exact results. First of all, the approximation methods yield exact results when the system under consideration is completely symmetric, i.e., when all N queues have exactly the same parameters. Furthermore the results are exact when $N = 1$ (vacation models!). For the two-queue case with 1-limited service at one queue and exhaustive service at the other that was analyzed in Section 6.3, it was noted that (cf. (6.72)) if the switch-over time from Q_1 to Q_2 is deterministic then $E\mathbf{W}_1 = (1-\rho-\lambda_2 s)E\mathbf{W}_2$. It may be easily seen that this is the same ratio as obtained by combining (7.2) and (7.4); applying the procedure of Section 7.2.1 implies that the approximation for this case also yields exact results.

When the visit discipline is exhaustive at all queues or gated at all queues, the approximation reduces to an approximation by Everitt [1986a]. Although effective numerical procedures are available in both of these cases (cf. Ferguson and Aminetzah [1985], Sarkar and Zangwill [1987]), the simplicity of the approximations, while sacrificing some accuracy, offers valuable insight into the qualitative behavior of the mean waiting times over a reasonably wide range of parameter values. In the case that all queues have a 1-limited visit discipline, the approximation closely resembles the approximation of Boxma and Meister [1986].

7.3 Numerical results

To test the accuracy of the approximation method of Section 7.2.1 and the refinement discussed in Section 7.2.3 we investigated several cyclic-service models with mixed visit disciplines. We shall first give a description of the cases that are considered. The detailed results may be found in the tables at the end of this chapter. In the tables, when more than one number is given in the column "queue index" the mean waiting times in the columns right of it are averaged over the corresponding group of queues. The switch-over times are in all cases taken deterministic. From the exact results in Chapter 6 it may be seen that the influence of the type of the switch-over time distributions is usually small.

Case 1

3 queues; $\lambda_1=0.6$, $\lambda_2=\lambda_3=0.2$; $\beta_1=\beta_2=\beta_3$ (negative exponential);
$s_1=s_2=s_3=0.1$ (deterministic).
For $\rho=0.3$, 0.5 and 0.8 respectively we investigate in 6 sub-cases the influence of different visit disciplines at the various queues.

Case 2

3 queues; $\lambda_1=\lambda_2=\lambda_3=1/3$; $\beta_1=3\beta_2=3\beta_3$ (negative exponential);
$s_1=s_2=s_3=0.1$ (deterministic).
For $\rho=0.3$, 0.5 and 0.8 respectively we investigate in 6 sub-cases the influence of different visit disciplines at the various queues.

Case 3

12 queues; $\beta_1=\cdots=\beta_{12}=0.5$ (negative exponential);
$s_1=\cdots=s_{12}=0.05$ (deterministic); $\rho=0.5$.
The visit disciplines are exhaustive at Q_1, Q_2 and Q_3, gated at $Q_4,\ldots,Q_7$ and 1-limited at $Q_8,\ldots,Q_{12}$.
We investigate the influence of a high arrival rate at one of the queues for the various visit disciplines.
Case 3a: $\lambda_1=0.56$, $\lambda_2=\cdots=\lambda_{12}=0.04$.
Case 3b: $\lambda_4=0.56$, $\lambda_1=\cdots=\lambda_3=\lambda_5=\cdots=\lambda_{12}=0.04$.
Case 3c: $\lambda_8=0.56$, $\lambda_1=\cdots=\lambda_7=\lambda_9=\cdots=\lambda_{12}=0.04$.

Case 4

12 queues; $\lambda_1=\cdots=\lambda_{12}=1/12$; $s_1=\cdots=s_{12}=0.05$ (deterministic); $\rho=0.5$.
The visit disciplines are exhaustive at Q_1, Q_2 and Q_3, gated at $Q_4,\ldots,Q_7$ and 1-limited at $Q_8,\ldots,Q_{12}$.
We investigate the influence of a large service time at one of the queues for the various visit disciplines.
Case 4a: $\beta_1=3.36$, $\beta_2=\cdots=\beta_{12}=0.24$.
Case 4b: $\beta_4=3.36$, $\beta_1=\cdots=\beta_3=\beta_5=\cdots=\beta_{12}=0.24$.
Case 4c: $\beta_8=3.36$, $\beta_1=\cdots=\beta_7=\beta_9=\cdots=\beta_{12}=0.24$.

Case 5

12 queues; $s_1=\cdots=s_{12}=0.05$ (deterministic).

Q_1, Q_2 and Q_3 served exhaustively, $Q_4,\ldots,Q_{12}$ served 1-limited.
Several combinations of arrival rates and service times with varying degrees of asymmetry are tested for a system with exhaustive and 1-limited visit disciplines.
Case 5a: $\lambda_1 = \cdots = \lambda_{12} = 1/12$; $\beta_1 = \cdots = \beta_{12}$.
Case 5b: $\lambda_1 = \cdots = \lambda_3 = 4/15$, $\lambda_4 = \cdots = \lambda_{12} = 1/45$; $\beta_1 = \cdots = \beta_{12}$.
Case 5c: $\lambda_1 = \cdots = \lambda_3 = 1/15$, $\lambda_4 = \cdots = \lambda_{12} = 4/45$; $\beta_1 = \cdots = \beta_{12}$.
Case 5d: $\lambda_1 = \cdots = \lambda_{12} = 1/12$; $\beta_1 = \cdots = \beta_3$, $\beta_4 = \cdots = \beta_{12}$, $\beta_1 = 12\beta_4$.
Case 5e: $\lambda_1 = \cdots = \lambda_{12} = 1/12$; $\beta_1 = \cdots = \beta_3$, $\beta_4 = \cdots = \beta_{12}$, $4\beta_1 = 3\beta_4$.

Case 6
12 queues; $s_1 = \cdots = s_{12} = 0.05$ (deterministic).
Q_1, Q_2 and Q_3 served gated, $Q_4,\ldots,Q_{12}$ served 1-limited.
Several combinations of arrival rates and service times with varying degrees of asymmetry are tested for a system with gated and 1-limited visit disciplines.
Case 6a: $\lambda_1 = \cdots = \lambda_{12} = 1/12$; $\beta_1 = \cdots = \beta_{12}$.
Case 6b: $\lambda_1 = \cdots = \lambda_3 = 4/15$, $\lambda_4 = \cdots = \lambda_{12} = 1/45$; $\beta_1 = \cdots = \beta_{12}$.
Case 6c: $\lambda_1 = \cdots = \lambda_3 = 1/15$, $\lambda_4 = \cdots = \lambda_{12} = 4/45$; $\beta_1 = \cdots = \beta_{12}$.
Case 6d: $\lambda_1 = \cdots = \lambda_{12} = 1/12$; $\beta_1 = \cdots = \beta_3$, $\beta_4 = \cdots = \beta_{12}$, $\beta_1 = 12\beta_4$.
Case 6e: $\lambda_1 = \cdots = \lambda_{12} = 1/12$; $\beta_1 = \cdots = \beta_3$, $\beta_4 = \cdots = \beta_{12}$, $4\beta_1 = 3\beta_4$.

Case 7
2 queues, both served 1-limited; $s_1 = s_2 = 0.1$ (deterministic).
We investigate the influence of different arrival intensities and mean service times for $\rho = 0.2$, 0.5 and 0.9.
In Case 7a the service-time distribution at both queues is negative exponential, in Case 7b the service-time distribution at both queues is hyperexponential and in Case 7c the service-time distribution is hyperexponential at one queue and deterministic at the other.

We shall now discuss the results. In Case 1 to 6 the results of the approximation method are compared with simulation results and, for cases with a relatively high system load and/or much asymmetry, with the refined approximation discussed in Section 7.2.3. The results for Case 7 are compared with exact results from Chapter 6 and results from the approximation described in Fuhrmann and Wang [1988].

REMARK 7.4
The *accuracy of the simulation results* can be verified by substituting the mean waiting times as obtained from the simulation into the left-hand side of the pseudoconservation law and comparing the result to the right-hand side of the pseudoconservation law as computed from the system parameters. The relative error between these two measures appeared in all cases to be smaller than 1 percent.

The simulation program was written in Simula '67. Each simulation was run

with at least 1,000,000 customers.

We have deliberately refrained from adding the relative errors of the mean waiting times in the tables. The reason is that one should be very careful in interpreting the accuracy of the mean waiting times observed at small queues. For instance if the observed mean waiting times in a two-queue system are $E\mathbf{W}_1 = 50$ and $E\mathbf{W}_2 = 0.5$ and the mean waiting times as predicted by the approximation are 48 and 1 respectively, then the relative error at the small queue is 100 percent. Yet intuitively (of course it also depends on the application) we feel that the approximation is not so bad at all. In such cases we need another accuracy measure; for instance instead of defining the accuracy measure as $(E\mathbf{W}_i^{sim} - E\mathbf{W}_i^{app})/E\mathbf{W}_i^{sim}$ (the relative error) we could alternatively define it as $(E\mathbf{W}_i^{sim} - E\mathbf{W}_i^{app})/E\overline{\mathbf{W}}$, in which $E\overline{\mathbf{W}}$ is given by:

$$E\overline{\mathbf{W}} = \frac{1}{\Lambda} \sum_{i=1}^{N} \lambda_i E\mathbf{W}_i^{sim}; \tag{7.33}$$

$\Lambda = \sum_{i=1}^{N} \lambda_i$. In Whitt [1985] this issue is extensively discussed and an accuracy measure combining absolute and relative errors is proposed.

Table 7.1 and Table 7.2 demonstrate that in general the approximation is quite accurate at low and medium loads ($\rho = 0.3, 0.5$) and less accurate at high loads ($\rho = 0.8$). For low and medium loads the refined approximation of Section 7.2.3 obtains the same degree of accuracy as the basic approximation; as can be seen in the tables, the refined approximation improves upon the results for $\rho = 0.8$. When there are no 1-limited queues in the system the approximation is accurate even when the load is high. A relatively high *arrival rate* has a much stronger effect on the mean waiting times at a 1-limited queue than a relatively large *service time* at that queue. This is also reflected in the stability condition for such a queue: $\rho + \lambda_i s < 1$; in this condition λ_i occurs explicitly, whereas β_i influences this condition only through ρ. For queues with an exhaustive or gated visit discipline we observe the opposite effect: for the same ρ_i a relatively large *service time* gives rise to a higher mean waiting time at Q_i than a relatively large *arrival rate*. These effects are illustrated in Tables 7.1 to 7.6. From Tables 7.3,7.4,7.5 and 7.6 we may conclude that the approximation is better when arrival rates are asymmetric than when there is asymmetry w.r.t. the service times. This effect is most noticeable at high loads. Especially large relative errors (of about 50 %) occur under a load of $\rho = 0.8$ in Cases 5d and 6d, where a small group of queues with exhaustive and gated service respectively has large service times and a large group of queues with 1-limited service has small service times. The refined approximation improves considerably upon the results in these cases, reducing the errors in some extreme cases with a factor of two.

In the three cases of Table 7.7, the approximation is compared with exact results and with the Fuhrmann-Wang (F&W) approximation. The exact results follow from the analysis in Chapter 6. Note that for this case, with all queues having a 1-limited visit discipline, the approximation reduces to the Boxma-

Meister approximation, the only difference being the calculation of $E\mathbf{C}_{b,i}$. Interestingly, the results of this approximation are better than those of the F&W approximation in almost all investigated cases, even at high loads. Yet it is observed by Fuhrmann and Wang [1988] that their approximation at high load is in general better than the original Boxma-Meister approximation. This suggests that the approach of calculating $E\mathbf{C}_{b,i}$ via the set of equations (7.12) is a considerable improvement of the Boxma-Meister approximation.

REMARK 7.5

As in Boxma and Meister [1986] we can apply a modification procedure for the case that there is a heavily loaded 1-limited queue in the system. This procedure is based on the following idea. Remove the 1-limited queue(s) with a relatively high arrival rate from the system and enlarge the switch-over times to compensate for the service times at the removed queues. The resulting reduced system has a lower and more symmetric traffic load and can be more accurately approximated. Boxma and Meister recommend application of this 'elimination procedure' when i) $\rho \geqslant 0.7$, ii) the total switch-over time is not negligible, and iii) the arrival rate at a small group of queues is at least three times as high as at any of the other queues. As suggested in Groenendijk [1989] we can improve on their procedure by deriving the mean waiting time for the eliminated queue by a back-substitution of the mean waiting times obtained from the elimination procedure into the pseudoconservation law.

In Groenendijk and Levy [1989] a similar elimination procedure is derived and studied for the case of identically zero switch-over times and unit service times.

7.4 DISCUSSION AND CONCLUSIONS

In this section we shall discuss some of the properties of the approximation algorithm.

The approximation method provides a very simple formula for $E\mathbf{W}_i$ and thus gives much qualitative insight into the behavior of the systems under consideration. The approximation is reasonably accurate and useful for engineering purposes in a wide area of parameter values. In extreme situations (e.g., high system load in combination with a very asymmetric system) the detailed approximation discussed in section 7.2.3 can be used as a more accurate alternative.

Since the approximations are based on the conservation law, they are exact in the completely symmetric case (same traffic characteristics, switch-over time distributions and visit disciplines at all queues). This suggests that the quality of the approximations should increase with the symmetry of the system considered. This observation is confirmed by the results in the tables.

For systems with a mixture of only gated and exhaustive service at all queues the approximations are identical to the approximation described in Everitt [1986a], and thus very accurate over a wide range of parameter values.

Several sources contribute to the error in the approximation for the mean

waiting times. First the assumption that the mean residual cycle times are all equal. Then, for queues with an exhaustive or 1-limited visit discipline, the approximation for the relation between the mean waiting time and the mean residual cycle time (cf. (7.3) and (7.4)). In addition, for queues with a 1-limited visit discipline two more approximations are used: 1) the approximation in (7.11) assuming independence between the number of customers found by the tagged customer upon his arrival and the number of busy-i cycles this customer has to wait; 2) the approximation (7.12) for the length of a busy-i cycle. As it appears the use of the pseudoconservation law averages out the 'local' errors and leads to accurate approximations for the mean waiting times at the various queues. This 'robustness' property is an inherent property of approximations based on the pseudoconservation law; the pseudoconservation law is therefore an indispensable and very efficient tool for the construction of approximations.

TABLE 7.1
3 queues; $\lambda_1 = 0.6$, $\lambda_2 = \lambda_3 = 0.2$; $\beta_1 = \beta_2 = \beta_3$ (negative exponential); $s_1 = s_2 = s_3 = 0.1$ (deterministic); $\rho = 0.3$, 0.5, 0.8.

$\rho = 0.3$				
case	visit discipline	queue index	simul.	approx.
1a	exh.	1	0.286	0.288
	1-lim	2,3	0.419	0.417
1b	gated	1	0.381	0.383
	1-lim	2,3	0.393	0.386
1c	1-lim	1	0.535	0.534
	exh.	2	0.294	0.297
	1-lim	3	0.373	0.375
1d	1-lim	1	0.532	0.531
	gated	2	0.328	0.332
	1-lim	3	0.370	0.372
1e	exh.	1	0.291	0.291
	gated	2,3	0.379	0.376
1f	gated	1	0.392	0.393
	exh.	2,3	0.314	0.313

$\rho = 0.5$				
case	visit discipline	queue index	simul.	approx.
1a	exh.	1	0.58	0.59
	1-lim	2,3	1.17	1.15
1b	gated	1	0.86	0.89
	1-lim	2,3	1.00	0.94
1c	1-lim	1	1.57	1.55
	exh.	2	0.54	0.56
	1-lim	3	0.81	0.84
1d	1-lim	1	1.56	1.53
	gated	2	0.63	0.67
	1-lim	3	0.79	0.83
1e	exh.	1	0.61	0.62
	gated	2,3	0.98	0.97
1f	gated	1	0.95	0.96
	exh.	2,3	0.67	0.66

For a detailed discussion of these tables see p.124 ff.

$\rho=0.8$					
case	visit discipline	queue index	simul.	approx.	refined approx.
1a	exh.	1	1.64	1.91	1.50
	1-lim	2,3	9.96	9.42	10.29
1b	gated	1	2.99	3.88	3.42
	1-lim	2,3	8.78	6.74	7.73
1c	1-lim	1	57.83	57.49	58.56
	exh.	2	1.18	1.42	1.25
	1-lim	3	2.56	3.09	2.87
1d	1-lim	1	60.38	56.68	58.00
	gated	2	1.47	1.93	1.70
	1-lim	3	2.48	3.04	2.82
1e	exh.	1	2.51	2.53	2.53
	gated	2,3	5.71	5.65	5.65
1f	gated	1	4.94	4.98	4.98
	exh.	2,3	2.84	2.83	2.83

TABLE 7.2
3 queues; $\lambda_1=\lambda_2=\lambda_3=1/3$; $\beta_1=3\beta_2=3\beta_3$ (negative exponential); $s_1=s_2=s_3=0.1$ (deterministic); $\rho=0.3$, 0.5, 0.8.

$\rho=0.3$				
case	visit discipline	queue index	simul.	approx.
2a	exh.	1	0.321	0.324
	1-lim	2,3	0.507	0.500
2b	gated	1	0.420	0.426
	1-lim	2,3	0.468	0.457
2c	1-lim	1	0.515	0.515
	exh.	2	0.323	0.330
	1-lim	3	0.452	0.445
2d	1-lim	1	0.516	0.511
	gated	2	0.357	0.369
	1-lim	3	0.453	0.441
2e	exh.	1	0.326	0.328
	gated	2,3	0.426	0.424
2f	gated	1	0.439	0.438
	exh.	2,3	0.346	0.349

For a detailed discussion of these tables see p.124 ff.

$\rho=0.5$				
case	visit discipline	queue index	simul.	approx.
2a	exh.	1	0.68	0.71
	1-lim	2,3	1.59	1.53
2b	gated	1	0.97	1.06
	1-lim	2,3	1.39	1.22
2c	1-lim	1	1.44	1.47
	exh.	2	0.63	0.66
	1-lim	3	1.12	1.10
2d	1-lim	1	1.43	1.45
	gated	2	0.73	0.79
	1-lim	3	1.20	1.08
2e	exh.	1	0.73	0.75
	gated	2,3	1.20	1.17
2f	gated	1	1.16	1.14
	exh.	2,3	0.79	0.79

$\rho=0.8$					
case	visit discipline	queue index	simul.	approx.	refined approx.
2a	exh.	1	1.99	2.42	1.72
	1-lim	2,3	18.45	16.77	18.88
2b	gated	1	3.06	4.82	3.90
	1-lim	2,3	17.02	11.73	14.48
2c	1-lim	1	12.99	13.74	13.41
	exh.	2	1.49	1.70	1.57
	1-lim	3	10.33	7.28	8.54
2d	1-lim	1	12.53	13.53	13.24
	gated	2	1.85	2.31	2.13
	1-lim	3	10.04	7.16	8.41
2e	exh.	1	3.14	3.22	3.22
	gated	2,3	7.31	7.18	7.18
2f	gated	1	6.23	6.22	6.22
	exh.	2,3	3.50	3.53	3.53

For a detailed discussion of these tables see p.124 ff.

TABLE 7.3
12 queues; $\beta_1 = \cdots = \beta_{12} = 0.5$ (negative exponential);
$s_1 = \cdots = s_{12} = 0.05$ (deterministic); $\rho = 0.5$.
In Case 3a: $\lambda_1 = 0.56$, $\lambda_2 = \cdots = \lambda_{12} = 0.04$.
In Case 3b: $\lambda_4 = 0.56$, $\lambda_1 = \cdots = \lambda_3 = \lambda_5 = \cdots = \lambda_{12} = 0.04$.
In Case 3c: $\lambda_8 = 0.56$, $\lambda_1 = \cdots = \lambda_7 = \lambda_9 = \cdots = \lambda_{12} = 0.04$.

case	visit discipline	queue index	simul.	approx.	refined approx.
3a	exh.	1	0.84	0.85	0.85
	exh.	2,3	1.15	1.16	1.15
	gated	4,5,6,7	1.20	1.21	1.20
	1-lim	8,9,10,11,12	1.33	1.29	1.33
3b	exh.	1,2,3	1.00	1.01	1.00
	gated	4	1.31	1.32	1.31
	gated	5,6,7	1.05	1.05	1.04
	1-lim	8,9,10,11,12	1.15	1.12	1.14
3c	exh.	1,2,3	0.85	0.89	0.85
	gated	4,5,6,7	0.89	0.92	0.89
	1-lim	8	4.40	4.31	4.40
	1-lim	9,10,11,12	0.94	0.99	0.95

TABLE 7.4
12 queues; $\lambda_1 = \cdots = \lambda_{12} = 1/12$; $s_1 = \cdots = s_{12} = 0.05$ (deterministic);
$\rho = 0.5$.
In Case 4a: $\beta_1 = 3.36$, $\beta_2 = \cdots = \beta_{12} = 0.24$.
In Case 4b: $\beta_4 = 3.36$, $\beta_1 = \cdots = \beta_3 = \beta_5 = \cdots = \beta_{12} = 0.24$.
In Case 4c: $\beta_8 = 3.36$, $\beta_1 = \cdots = \beta_7 = \beta_9 = \cdots = \beta_{12} = 0.24$.

case	visit discipline	queue index	simul.	approx.	refined approx.
4a	exh.	1	1.91	2.11	1.98
	exh.	2,3	2.68	2.87	2.70
	gated	4,5,6,7	2.92	2.99	2.81
	1-lim	8,9,10,11,12	4.08	3.38	4.02
4b	exh.	1,2,3	2.13	2.26	2.17
	gated	4	2.81	2.95	2.84
	gated	5,6,7	2.03	2.35	2.26
	1-lim	8,9,10,11,12	3.21	2.66	3.11
4c	exh.	1,2,3	1.81	1.99	1.95
	gated	4,5,6,7	1.96	2.07	2.02
	1-lim	8	3.50	3.52	3.44
	1-lim	9,10,11,12	2.60	2.35	2.73

For a detailed discussion of these tables see p.124 ff.

TABLE 7.5
12 queues; $s_1 = \cdots = s_{12} = 0.05$ (deterministic).
Q_1, Q_2 and Q_3 served exhaustively, $Q_4, \cdots, Q_{12}$ served 1-limited.
In Case 5a:
$\lambda_1 = \cdots = \lambda_{12} = 1/12$; $\beta_1 = \cdots = \beta_{12}$.
In Case 5b:
$\lambda_1 = \cdots = \lambda_3 = 4/15$, $\lambda_4 = \cdots = \lambda_{12} = 1/45$; $\beta_1 = \cdots = \beta_{12}$.
In Case 5c:
$\lambda_1 = \cdots = \lambda_3 = 1/15$, $\lambda_4 = \cdots = \lambda_{12} = 4/45$; $\beta_1 = \cdots = \beta_{12}$.
In Case 5d:
$\lambda_1 = \cdots = \lambda_{12} = 1/12$; $\beta_1 = \cdots = \beta_3$, $\beta_4 = \cdots = \beta_{12}$, $3\beta_1 = 12\beta_4$.
In Case 5e:
$\lambda_1 = \cdots = \lambda_{12} = 1/12$; $\beta_1 = \cdots = \beta_3$, $\beta_4 = \cdots = \beta_{12}$, $4\beta_1 = 3\beta_4$.

$\rho = 0.3$				
case	visit discipline	queue index	simul.	approx.
5a	exh.	1,2,3	1.44	1.45
	1-lim.	4,...,12	2.00	2.00
5b	exh.	1,2,3	1.39	1.39
	1-lim.	4,...,12	1.63	1.62
5c	exh.	1,2,3	1.45	1.45
	1-lim.	4,...,12	2.04	2.04
5d	exh.	1,2,3	1.58	1.59
	1-lim.	4,...,12	2.30	2.26
5e	exh.	1,2,3	1.45	1.46
	1-lim.	4,...,12	2.00	2.00

$\rho = 0.5$					
case	visit discipline	queue index	simul.	approx.	refined approx.
5a	exh.	1,2,3	2.18	2.24	2.19
	1-lim.	4,...,12	3.75	3.73	3.75
5b	exh.	1,2,3	2.14	2.14	2.14
	1-lim.	4,...,12	2.80	2.76	2.79
5c	exh.	1,2,3	2.19	2.25	2.20
	1-lim.	4,...,12	3.86	3.84	3.86
5d	exh.	1,2,3	2.82	2.92	2.82
	1-lim.	4,...,12	5.66	5.07	5.64
5e	exh.	1,2,3	2.19	2.25	2.21
	1-lim.	4,...,12	3.74	3.73	3.75

For a detailed discussion of these tables see p.124 ff.

$\rho=0.8$					
case	visit discipline	queue index	simul.	approx.	refined approx.
5a	exh.	1,2,3	5.2	6.2	5.5
	1-lim.	4,...,12	46.1	44.1	45.3
5b	exh.	1,2,3	6.3	6.7	6.5
	1-lim.	4,...,12	13.4	11.8	12.9
5c	exh.	1,2,3	5.1	6.1	5.4
	1-lim.	4,...,12	59.9	59.7	60.8
5d	exh.	1,2,3	8.6	11.5	9.7
	1-lim.	4,...,12	136.7	79.4	115.1
5e	exh.	1,2,3	5.2	6.1	5.5
	1-lim.	4,...,12	44.4	44.0	44.8

TABLE 7.6
12 queues; $s_1 = \cdots = s_{12} = 0.05$ (deterministic).
Q_1, Q_2 and Q_3 served gated, Q_4, $\cdots$, Q_{12} served 1-limited.
In Case 6a:
$\lambda_1 = \cdots = \lambda_{12} = 1/12$; $\beta_1 = \cdots = \beta_{12}$.
In Case 6b:
$\lambda_1 = \cdots = \lambda_3 = 4/15$, $\lambda_4 = \cdots = \lambda_{12} = 1/45$; $\beta_1 = \cdots = \beta_{12}$.
In Case 6c:
$\lambda_1 = \cdots = \lambda_3 = 1/15$, $\lambda_4 = \cdots = \lambda_{12} = 4/45$; $\beta_1 = \cdots = \beta_{12}$.
In Case 6d:
$\lambda_1 = \cdots = \lambda_{12} = 1/12$; $\beta_1 = \cdots = \beta_3$, $\beta_4 = \cdots = \beta_{12}$, $3\beta_1 = 12\beta_4$.
In Case 6e:
$\lambda_1 = \cdots = \lambda_{12} = 1/12$; $\beta_1 = \cdots = \beta_3$, $\beta_4 = \cdots = \beta_{12}$, $4\beta_1 = 3\beta_4$.

$\rho=0.3$				
case	visit discipline	queue index	simul.	approx.
6a	gated	1,2,3	1.51	1.52
	1-lim.	4,...,12	2.00	1.99
6b	gated	1,2,3	1.61	1.61
	1-lim.	4,...,12	1.60	1.60
6c	gated	1,2,3	1.51	1.51
	1-lim.	4,...,12	2.05	2.04
6d	gated	1,2,3	1.81	1.82
	1-lim.	4,...,12	2.25	2.21
6e	gated	1,2,3	1.45	1.46
	1-lim.	4,...,12	2.00	2.00

For a detailed discussion of these tables see p.124 ff.

$\rho=0.5$					
case	visit discipline	queue index	simul.	approx.	refined approx.
6a	gated	1,2,3	2.36	2.43	2.38
	1-lim.	4,...,12	3.74	3.71	3.74
6b	gated	1,2,3	2.68	2.68	2.68
	1-lim.	4,...,12	2.68	2.64	2.66
6c	gated	1,2,3	2.33	2.40	2.34
	1-lim.	4,...,12	3.87	3.84	3.86
6d	gated	1,2,3	3.39	3.50	3.42
	1-lim.	4,...,12	5.31	4.65	5.13
6e	gated	1,2,3	2.33	2.40	2.35
	1-lim.	4,...,12	3.76	3.73	3.74

$\rho=0.8$					
case	visit discipline	queue index	simul.	approx.	refined approx.
6a	gated	1,2,3	5.9	7.0	6.2
	1-lim.	4,...,12	46.0	43.8	45.1
6b	gated	1,2,3	8.7	9.0	8.9
	1-lim.	4,...,12	11.8	10.3	11.0
6c	gated	1,2,3	5.7	6.8	6.0
	1-lim.	4,...,12	61.6	59.4	60.7
6d	gated	1,2,3	10.8	14.3	12.8
	1-lim.	4,...,12	135.8	64.1	93.9
6e	gated	1,2,3	5.8	6.8	6.1
	1-lim.	4,...,12	45.4	43.8	44.7

For a detailed discussion of these tables see p.124 ff.

TABLE 7.7
2 queues with 1-limited service;
Case 7a: $B_i(\cdot)$ negative exponential, $i=1,2$;
$\lambda=1$, $r_1=\lambda_1/\lambda$; $s_1=s_2=0.1$ (deterministic).

parameters			exact results		approx.		F&W approx.	
r_1	β_1	β_2	$E\mathbf{W}_1$	$E\mathbf{W}_2$	$E\mathbf{W}_1$	$E\mathbf{W}_2$	$E\mathbf{W}_1$	$E\mathbf{W}_2$
0.7	0.2	0.2	0.236	0.192	0.236	0.192	0.242	0.180
0.7	0.2	0.5	0.387	0.330	0.379	0.337	0.486	0.249
0.7	0.2	0.8	0.726	0.599	0.655	0.635	1.025	0.450
0.7	0.5	0.2	0.738	0.440	0.740	0.434	0.723	0.514
0.7	0.5	0.5	1.152	0.671	1.133	0.709	1.203	0.575
0.7	0.5	0.8	2.131	1.098	1.973	1.275	2.299	0.908
0.7	0.8	0.2	2.680	0.977	2.685	0.943	2.663	1.094
0.7	0.8	0.5	5.088	1.406	5.006	1.606	5.139	1.283
0.7	0.8	0.8	16.57	2.158	15.98	2.743	16.37	2.351
0.9	0.2	0.2	0.256	0.170	0.256	0.171	0.260	0.145
0.9	0.2	0.5	0.308	0.208	0.308	0.209	0.341	0.114
0.9	0.2	0.8	0.400	0.269	0.397	0.275	0.486	0.120
0.9	0.5	0.2	1.084	0.416	1.084	0.417	1.084	0.410
0.9	0.5	0.5	1.279	0.473	1.276	0.489	1.309	0.293
0.9	0.5	0.8	1.586	0.557	1.578	0.585	1.666	0.269
0.9	0.8	0.2	9.564	0.871	9.563	0.887	9.583	0.652
0.9	0.8	0.5	16.16	0.959	16.15	1.000	16.30	0.482
0.9	0.8	0.8	43.68	1.077	43.60	1.160	44.30	0.460

For a detailed discussion of these tables see p.124 ff.

TABLE 7.7 (CONT'D)

Case 7b: $B_i(\cdot)$ hyperexponential (H_2) with squared coefficient of variation 4 and balanced means (cf. Tijms [1986]);
$\lambda = 1$, $r_1 = \lambda_1 / \lambda$; $s_1 = s_2 = 0.1$ (constant switch-over times).

parameters			exact results		approx.		F&W approx.	
r_1	β_1	β_2	$E\mathbf{W}_1$	$E\mathbf{W}_2$	$E\mathbf{W}_1$	$E\mathbf{W}_2$	$E\mathbf{W}_1$	$E\mathbf{W}_2$
0.7	0.2	0.2	0.330	0.267	0.329	0.269	0.337	0.252
0.7	0.2	0.5	0.684	0.547	0.649	0.576	0.831	0.426
0.7	0.2	0.8	1.546	1.122	1.289	1.251	2.018	0.886
0.7	0.5	0.2	1.389	0.832	1.392	0.817	1.362	0.968
0.7	0.5	0.5	2.325	1.273	2.254	1.410	2.393	1.145
0.7	0.5	0.8	4.660	2.118	4.157	2.685	4.843	1.913
0.7	0.8	0.2	5.706	2.073	5.716	2.007	5.670	2.329
0.7	0.8	0.5	11.12	2.855	10.87	3.486	11.15	2.784
0.7	0.8	0.8	37.32	4.258	35.49	6.091	36.36	5.223
0.9	0.2	0.2	0.355	0.235	0.355	0.236	0.360	0.200
0.9	0.2	0.5	0.470	0.309	0.468	0.318	0.518	0.174
0.9	0.2	0.8	0.689	0.441	0.675	0.467	0.825	0.204
0.9	0.5	0.2	2.084	0.795	2.084	0.803	2.085	0.789
0.9	0.5	0.5	2.509	0.902	2.500	0.957	2.564	0.573
0.9	0.5	0.8	3.206	1.069	3.176	1.177	3.351	0.540
0.9	0.8	0.2	20.66	1.861	20.65	1.916	20.70	1.408
0.9	0.8	0.5	35.13	2.021	35.09	2.173	35.42	1.047
0.9	0.8	0.8	95.84	2.248	95.56	2.533	97.09	0.999

For a detailed discussion of these tables see p.124 ff.

TABLE 7.7 (CONT'D)

Case 7c: $B_1(\cdot)$ hyperexponential (H_2) with squared coefficient of variation 4 and balanced means (cf. Tijms [1986]);
$B_2(\cdot)$ deterministic;
$\lambda = 1$, $r_1 = \lambda_1 / \lambda$; $s_1 = s_2 = 0.1$ (constant switch-over times).

parameters			exact results		approx.		F&W approx.	
r_1	β_1	β_2	$E\mathbf{W}_1$	$E\mathbf{W}_2$	$E\mathbf{W}_1$	$E\mathbf{W}_2$	$E\mathbf{W}_1$	$E\mathbf{W}_2$
0.7	0.2	0.2	0.293	0.235	0.292	0.238	0.299	0.223
0.7	0.2	0.5	0.400	0.333	0.387	0.344	0.496	0.254
0.7	0.2	0.8	0.605	0.507	0.551	0.534	0.862	0.379
0.7	0.5	0.2	1.334	0.795	1.337	0.785	1.307	0.929
0.7	0.5	0.5	1.862	1.021	1.805	1.130	1.917	0.917
0.7	0.5	0.8	2.909	1.391	2.634	1.701	3.068	1.212
0.7	0.8	0.2	5.602	2.030	5.610	1.970	5.565	2.286
0.7	0.8	0.5	9.975	2.566	9.745	3.127	10.00	2.497
0.7	0.8	0.8	29.03	3.410	27.69	4.752	28.36	4.074
0.9	0.2	0.2	0.342	0.226	0.342	0.227	0.347	0.193
0.9	0.2	0.5	0.382	0.252	0.380	0.258	0.422	0.141
0.9	0.2	0.8	0.444	0.293	0.438	0.303	0.536	0.132
0.9	0.5	0.2	2.061	0.785	2.060	0.794	2.061	0.780
0.9	0.5	0.5	2.346	0.842	2.337	0.895	2.397	0.536
0.9	0.5	0.8	2.728	0.915	2.704	1.003	2.854	0.460
0.9	0.8	0.2	20.56	1.851	20.55	1.907	20.59	1.401
0.9	0.8	0.5	34.08	1.961	34.04	2.108	34.36	1.015
0.9	0.8	0.8	88.89	2.091	88.63	2.350	90.05	0.927

For a detailed discussion of these tables see p.124 ff.

CHAPTER 8

PERFORMANCE MODELING AND ANALYSIS OF TOKEN-PASSING LOCAL AREA NETWORKS

8.1 INTRODUCTION

In this chapter we shall apply the results obtained in the previous chapters to the analysis of message delay in local area networks (LAN's) employing some form of token passing. This chapter is only meant to illustrate the applicability of the methods derived in this monograph and not to present an exhaustive analysis of LAN performance. In particular many problems are left open; the methods presented in this chapter only aim to serve as a stepping stone upon which to base further studies. See also the discussion at the end of this chapter. The technical aspects of the access mechanism in the various token-passing networks will be presented in some detail to highlight the differences between the polling model and the actual systems.

A LAN is a network that has a well-defined topology and medium access control connecting computers and communication equipment within the bounds of a single office building, building complex, or campus. It is characterized by the following elements (Kümmerle and Reiser [1987]):

- A transmission medium shared among the participating stations (workstations, hosts, servers, etc.), and providing a broadcast capability.
- A distributed protocol, the Medium Access Control (MAC) protocol which controls access to the medium and provides recovery mechanisms where necessary.
- A set of cooperating adapters (also called LAN interfaces) through which stations attach to the network, and which execute the MAC protocol and interface with the attaching stations.

The geographical constraints together with advanced transmission technologies enable information transfer rates of many millions of bits per second within the local network.

The pioneering LAN for connecting computers and peripheral devices was Ethernet, the product of a joint venture of the Digital Equipment Corporation and Xerox corporation. Ethernet had its origins in the ALOHA system (Abramson [1970]), which was used to connect central computers to terminals

throughout the Hawaiian Islands, over the radio airways. The 'ether' of Ethernet is coaxial cable, configured as a baseband bus. Stations are 'tapped' onto the cable. Access to the media is gained through a preemptive protocol of first listening for media activity and then broadcasting. If a 'collision' with someone else's transmission occurs, each backs off a random amount of time and tries again. This access technique is formally called Carrier Sense Multiple Access with Collision Detection (CSMA/CD). For a more detailed description of Ethernet see Metcalfe and Boggs [1976].

CSMA/CD is a so-called contention-based protocol. In contrast, the token-passing scheme is an example of a contention-free demand-based multiple-access protocol. A LAN employing the token-passing scheme is usually configured in a ring or a bus topology. Depending on the topology such networks are referred to as token rings and token buses respectively. A *token ring* consists of a set of stations connected to a transmission medium in a ring topology. A special bit pattern called the *token* is passed sequentially from one active station to the next. The station that receives the token gains the right to transmit a frame consisting of a header, user-supplied data and a trailer. After completion of his transmission this station releases the token again, giving the next station downstream the opportunity to transmit. The token is embedded in the header of the frame. This technique is called *implicit* token passing, in contrast to *explicit* token passing, where passing the token from one station to the next requires the transmission of an explicit-token frame. A *token bus* consists of a set of stations connected to each other in a bus topology. The intention behind this technique has been to combine the attractive features of a bus topology (e.g., use of broadband transmission) with those of a contention-free medium access protocol. In a token bus, as the token is passed a logical ring is formed. Since the bus topology does not impose any sequential ordering of the stations, the logical ring is defined by a sequence of station addresses. Conceptually, token passing on buses and rings is very similar and the same type of performance model can be used to describe the two techniques. However, the model parameters are rather different; in a token bus, explicit token passing is employed. Hence passing the token from a station to its successor requires the transmission of an explicit-token frame of 152 bits long, whereas in a token ring the time to pass the token only consists of the signal propagation delay between the stations and the latency within a station. This latter delay can be kept as small as only one bit time. Note that a token bus is more flexible w.r.t. the polling order of the stations. Hence the queueing model with polling table described in Chapter 5 may be particularly appropriate for modeling a token bus rather than a token ring.

Local area network standards apply to the data link and physical layer of the OSI Reference Model. As shown in Figure 8.1, the OSI data link layer is split into two sublayers, the medium independent *Logical Link Control* (LLC) sublayer and the medium dependent *Medium Access Control* (MAC) sublayer. Together they perform the duties of the data link sublayer, namely frame exchange between peer data link layer entities. Frame exchange encompasses

frame delimiting, frame transmission, address recognition, error checking, association with a data link user, and, if desired, link flow control and assurance of delivery without loss, duplication, or misordering. Medium access control performs the delimiting, transmission, address recognition, and error checking; logical link control performs the remainder.

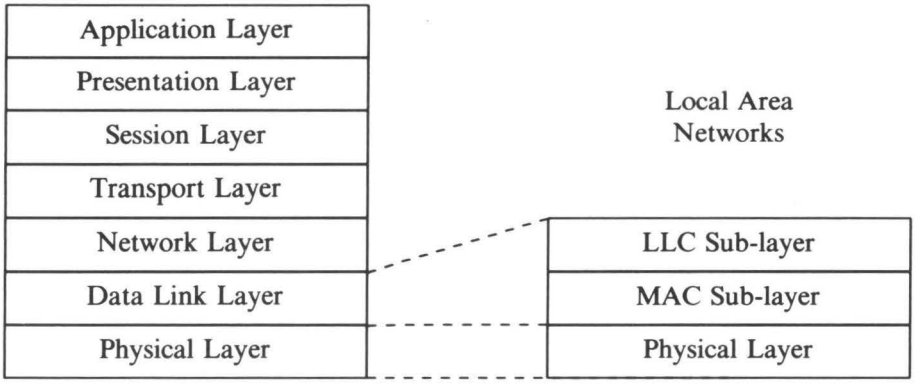

Figure 8.1 Relationship between LAN architecture and OSI Reference Model

The standards for CSMA/CD networks are described in the IEEE Standard 802.3 (IEEE [1983]), for token buses in the IEEE Standard 802.4 (IEEE [1985a]) and for token rings in the IEEE Standard 802.5 (IEEE [1985b]). The most widely used token ring network is the IBM Token-Ring Network. The IBM Token-Ring Network was formally introduced in 1985, first as an interconnection link for personal computers, and later expanded to include the interconnection of terminal controllers, minicomputers, and host systems.

The organization of the remainder of this chapter is as follows. In the next section we provide a technical description of the fundamental operations in a token ring local area network. The token ring that we describe is a 'basic' token ring in that it provides those services that are common in most of the token rings described in the literature, including the token ring as defined by IEEE 802.5 and the IBM Token-Ring Network. The discussion is based on Bux [1981,1985], IEEE [1985b], Reiser [1986] and Strole [1987]. In Section 8.3 it is discussed how the analysis of the cyclic polling model as described in the previous model can aid in the performance analysis of the basic token ring. A performance model for the token ring is formulated and shown to fit the cyclic polling model. Several examples are given, illustrating the method of analysis. Particular attention is paid to the modeling of the 'token-holding' mechanism. Finally, in Section 8.4 the interconnection of token-passing LAN's is studied. In Section 8.4.1 a possible structure of such an interconnected LAN is presented consisting of several local token rings interconnected by a backbone token ring. The backbone ring, which in this framework is supposed to be a token ring employing the FDDI (Fiber Distributed Data Interface) token-

passing protocol is described in detail in Section 8.4.2. In Section 8.4.3 we consider the performance analysis of a local ring in isolation. In Section 8.4.4 the analysis is extended to that of the end-to-end delay in the interconnected network. The chapter is closed with some directions for future research.

8.2 Description of a basic token ring

A token ring consists of a set of stations connected by a transmission medium in a ring topology. Information is transferred sequentially from one station to the next station downstream. Since the transmission is point-to-point between stations, a variety of media can be used on the separate links.

Remark 8.1

Typically, telephone twisted pair (TTP) can be used to carry a baseband signal up to 4 Megabit per second (Mbps). TTP consists of two insulated wires, generally copper or copper coated steel, arranged in a spiral. Spiralling minimizes the electromagnetic interference between wire pairs. 'Data-grade' media like shielded twisted pair or coaxial cable can provide a reliable transmission medium for a local network system operating at baseband rates up to 16 Mbps. For higher speeds optical fiber is the most appropriate transmission medium. The fiber may be composed of glass or plastic. Data rates as high as 1 Gbps (Gigabit/s) are attainable. Fiber optic transmission is not affected by electrical or electromagnetic interference as with twisted pair and coaxial cable.

A special bit sequence, called the *token* is passed from station to station. A station that detects a token passing gains the right to transmit. Any station, upon detection of the token, may seize the token by modifying it to a start-of-frame sequence, and then append appropriate control and address fields, the LLC supplied data, the frame check sequence and the end-of-frame delimiter. The structure of the resulting frame for the IBM Token Ring is depicted in Figure 8.2; the frame format as specified in the IEEE 802.5 standard is slightly different.

All other stations repeat each bit received. The addressed destination station copies the information as it passes. The station that initiates a frame transfer is responsible for removing that frame from the ring. Upon receipt of the physical header and completion of frame transmission the station issues a new token. If a station finishes transmitting the entire frame prior to receiving its own physical header, it continues to transmit idle characters (contiguous 0-bits) until the header is recognized. This ensures that only one token or frame is on the ring at any time. This scheme is called *single token operation* (cf. Bux [1981]).

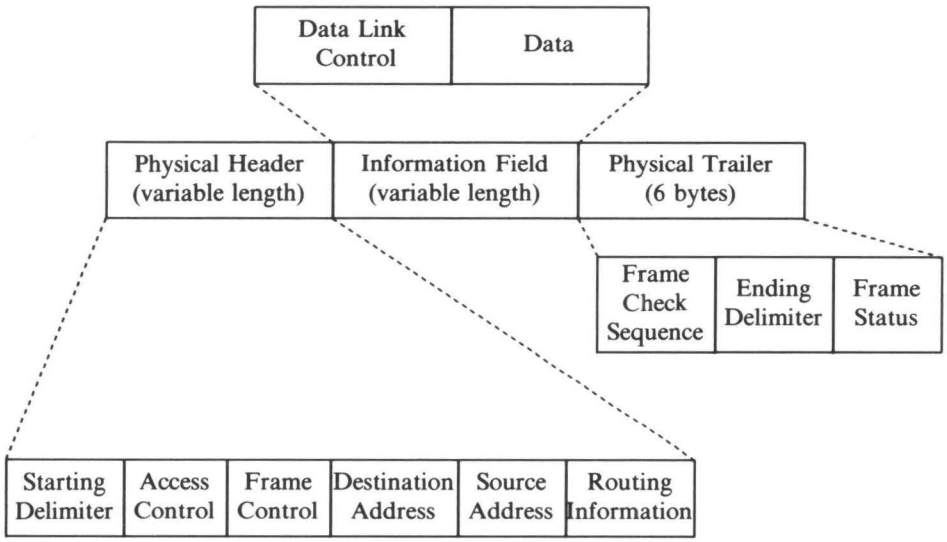

Figure 8.2 Frame format in the IBM Token Ring network

REMARK 8.2
Several other variants have been described in the literature. With *multiple token operation* the token may be sent out immediately after completion of transmission, i.e., at the earliest possible time. With this scheme it is possible to have several frames on the ring at one time. Multiple token operation is interesting because it seems to offer maximal ring utilization and minimal delays for seizing the token. Another alternative, *single frame operation*, requires that the last bit of the frame sent out by the station be received by that station (and removed from the ring) before a new token can be generated. This is the most conservative protocol, and only allows the bits of one message to be on the ring at any one time. For reliability and recovery this is the most attractive choice, as only one message is in jeopardy if the ring should fail. Single token operation provides the intuitive compromise between these two schemes. Both the IEEE 802.5 and the FDDI ring standards provide trailing-indicators within the frame format to incorporate acknowledgements for proper operation. Single token operation has the reliability and recovery advantages of single frame operation, yet approaches the ring utilization of multiple token operation for large frames.

In the basic token ring protocol the length of time a station is allowed to transmit when in possession of the token is controlled by a so-called Token Holding Timer (THT). Long THT timeouts will lead to an exhaustive type of service, in which the station can empty its buffer completely. Short THT timeouts yield an operation similar to limited visit disciplines.

In the IEEE 802.5 Standard (IEEE [1985b]), eight different priority levels are defined, combined with a reservation scheme. It is very hard to incorporate this type of priority scheduling in an analytic model. Only very recently some papers have emerged addressing this issue using an analytical approach (cf. Fischer [1989], Fournier and Rosberg [1989]). For a simulation study see Goyal and Dias [1988]. We shall assume here that all traffic is of the same type and hence that there is only one priority level. Note that an alternative way of providing (global) priority to a station is by increasing the value of its THT timeout, thus controlling the station-specific quality of service.

Token rings as described above usually operate at bandwidths ranging from 1 Mbps to 16 Mbps. Up to 260 stations may be connected to the same ring. The propagation delay for most transmission media is about 5 microseconds (μs) per kilometer of cable. The latency caused by the repeater within each station is usually one bit time. The amount of control information contained in a frame is variable and may consist of about 176 bit (cf. Figure 8.1): Start Delimiter (8 bit), Access Control Field (8 bit), Frame Control Field (8 bit), Destination Address (32 bit), Source Address (32 bit), Routing Information (8 bit), Data Link Control (32 bit), Frame Check Sequence (32 bit), End Delimiter (8 bit) and Frame Status (8 bit). The length of the Information field is also variable and is of course bounded due to the THT. The time to transmit the control information in a token ring with a bandwidth of V Mbps is $176/V$ μs.

8.3 Performance model for the basic token ring

We shall now discuss how the analysis of the cyclic polling model as described in the previous chapters can aid in analyzing the performance of the basic token ring. In the performance model for the token ring the transmit buffers at the stations are represented by the queues and the token is represented by the single server. Since we assume that only the transmitting station can release the token and give the next station downstream opportunity to send, it follows that the queues are served in a cyclic manner. Also, this assumption implies that the performance of the ring is not affected by the location of the frame destinations on the ring relative to the transmitting station.

The arrival rate of messages at the stations corresponds directly to the arrival rate of customers at the queues in the polling system. The transmission time of a frame (message + overhead) corresponds to the service time of a customer in the polling system.

The time needed to pass the token from one station to the next is modeled by the switch-over time between the queues. This delay corresponds to the propagation delay of the signals between two physical connections, the latency at one station and actions such as alteration of the Access Control Field. In the performance model, only the propagation delay and the latency will be incorporated, other factors are assumed to be negligible.

The visit disciplines at the queues are directly related to the ratio of the THT timeouts and the frame transmission times at the various queues. Representing the 'timing mechanisms' of the token ring system by

appropriately chosen visit disciplines at the queues of the polling models is one of the major - and least investigated - problems in modeling token ring networks. We shall consider two different cases: 1) a non-preemptive timer protocol and 2) a preemptive timer protocol. In the non-preemptive protocol transmission of a message is completed even if the THT expires during the transmission. In the preemptive protocol messages are split into fixed-length packets which may be transmitted during different rounds.

Modeling of the non-preemptive timer protocol

We shall consider the modeling of the timer protocol at one particular station; for ease of notation the station index is dropped from the various variables. In the following we aim to determine how many message transmissions can be initiated before the THT expires.

Let x denote the value of the THT timeout at the station. Denote the distribution of the transmission time of a frame (containing exactly one message) by $B(\cdot)$ and let β denote the first moment of this distribution. We define $\{\mathbf{B}_i,\ i=1,2,...\}$ to be an i.i.d. sequence of random variables, each having distribution function $B(\cdot)$ and mean β. We further introduce

$$\mathbf{S}_0 := 0,\quad \mathbf{S}_n := \mathbf{B}_1 + \cdots + \mathbf{B}_n,\quad n=1,2,\ldots . \tag{8.1}$$

We consider the stochastic process $\mathbf{Y}_t := 1+\max\{n: \mathbf{S}_n<t\}$, $t\geqslant 0$. Note that $E\mathbf{Y}_x$ gives the mean number of *renewals* in $[0,x]$ (including a renewal at $t=0$) and thus represents the mean number of message transmissions that can be initiated before expiration of the THT. It is well-known from renewal theory that the renewal function $m(t)$, $t\geqslant 0$ which is defined by $m(t) := E\mathbf{Y}_t - 1$ is given by (cf. Cohen [1982, p.95]),

$$m(t) = \sum_{n=0}^{\infty} B^{(n+1)^*}(t), \tag{8.2}$$

where $B^{n^*}(t)$ denotes the n-fold convolution of $B(t)$ with itself, $n=1,2,\ldots$. Hence the mean number of message transmissions initiated before THT expiration is given by

$$E\mathbf{Y}_x = 1 + \sum_{n=0}^{\infty} B^{(n+1)^*}(x). \tag{8.3}$$

Introducing the Laplace-Stieltjes transforms

$$\beta(s) := \int_0^{\infty} e^{-st}\,dB(t),\quad \mu(s) := \int_0^{\infty} e^{-st}\,dm(t), \tag{8.4}$$

and using (8.2) it can be seen (cf. Cohen [1982, p.96-97]) that

$$\mu(s) = \frac{\beta(s)}{1-\beta(s)}. \tag{8.5}$$

Note that this relation can be very useful for determining $m(t)$.

For a stationary renewal process (cf. Cohen [1982, p.98]) it may be shown that $\mu(s)=1/\beta s$ and hence that

$$m(t) = \frac{t}{\beta}. \tag{8.6}$$

When $B(t)=1-e^{-t/\beta}$ the renewal sequence is stationary; so for that case

$$E\mathbf{Y}_x = 1 + \frac{x}{\beta}. \tag{8.7}$$

Denote by $\{E\mathbf{Y}_x\}$ the value of $E\mathbf{Y}_x$ rounded off to the nearest integer. We propose to model the non-preemptive timed protocol at the station of the token ring by the k-limited visit discipline (as described in Chapter 3) at the corresponding queue in the polling model with $k:=\{E\mathbf{Y}_x\}$. The effect of this modeling assumption will be briefly investigated below. A more thorough investigation is the subject of further study.

When the THT timeout is small compared to the mean frame transmission time ($x<<\beta$), it is obvious that the THT mechanism can be adequately modeled by assuming the 1-limited service discipline at the queue. When the THT timeout is much larger than the average frame transmission time ($x>>\beta$) the THT mechanism is evidently very similar to the exhaustive visit discipline. Hence the most interesting - and difficult - case arises when x is of the same order of magnitude as β. Two such cases are investigated in Tables 8.1 and 8.2 below. In these tables we investigate a polling system under three different visit disciplines: 1) timer protocol, 2) 1-limited and 3) 2-limited. The timer protocol is non-preemptive and at all queues the THT timeout x is equal to β. The cases are considered under a load of $\rho=0.3$, 0.5 and 0.8 respectively. The systems under consideration consist of three queues; the considered performance measures are the mean waiting times $E\mathbf{W}_i$, $i=1,2,3$ of customers. The service time distributions are assumed to be negative exponential and the switch-over time distributions are assumed to be deterministic.

The parameters for the cases are presented immediately above the tables. Note that since $x=\beta$ and the service time distribution is negative exponential we have $\{E\mathbf{Y}_x\}=2$. The considerations above imply that the THT mechanism should be modeled by the 2-limited visit discipline.

The results for the timer protocol have of course been obtained from simulation, since no analytic expressions or even approximations are available for this case. The results for the 1-limited visit discipline are for $\rho=0.3$ obtained from the (basic) approximation described in Chapter 7, in particular from Formulas (7.4) and (7.5) and for $\rho=0.5$ and 0.8 from the refined approximation described in that chapter (Formulas (7.17) and (7.19)-(7.22)). Finally, the results for 2-limited service are obtained from the approximation of Fuhrmann and Wang [1988] (Formulas (8) and (11)).

TABLE 8.1
Simulation comparison of non-preemptive THT protocol and 1-limited service discipline when THT timeout equals mean frame transmission time.
3 queues, $\lambda_1 = 0.6$, $\lambda_2 = \lambda_3 = 0.2$, $\beta_1 = \beta_2 = \beta_3$ (neg. exp.), $x_i = \beta_i$ for $i = 1,2,3$, $s_1 = s_2 = s_3 = 0.1$ (determ.).

	THT timer		1-limited		2-limited	
ρ	EW_1	$EW_{2,3}$	EW_1	$EW_{2,3}$	EW_1	$EW_{2,3}$
0.3	0.421	0.347	0.525	0.370	0.420	0.322
0.5	1.12	0.747	1.51	0.775	1.12	0.670
0.8	10.6	2.33	55.7	2.31	9.67	2.23

TABLE 8.2
Simulation comparison of non-preemptive THT protocol and 1-limited service discipline when THT timeout equals mean frame transmission time.
3 queues, $\lambda_1 = \lambda_2 = \lambda_3 = 1/3$, $\beta_1 = 3\beta_2 = 3\beta_3$ (neg. exp.), $x_i = \beta_i$ for $i = 1,2,3$, $s_1 = s_2 = s_3 = 0.1$ (determ.).

	THT timer		1-limited		2-limited	
ρ	EW_1	$EW_{2,3}$	EW_1	$EW_{2,3}$	EW_1	$EW_{2,3}$
0.3	0.442	0.399	0.506	0.444	0.317	0.558
0.5	1.20	0.989	1.38	1.16	0.861	1.38
0.8	7.52	5.60	10.72	8.30	6.91	6.22

As the tables indicate, modeling the THT mechanism by the 1-limited visit discipline produces results that may be much too large when ρ is high. Note that in particular the station with high arrival rate is affected. The reason for this is that when the queue is non-empty at the polling instant, in the non-preemptive timer protocol always *at least* one message is served, whereas in the 1-limited discipline *precisely* one message is served. Modeling the THT mechanism by the 2-limited visit discipline clearly yields better results, very accurate in the first table and rather less accurate in the second table. We have at present no insight why the results of Table 1 are more accurate than those of Table 2.

REMARK 8.3
Note that the results are strongly dependent on the distributions of the service times at the queues. Although the tables above only present results for negative exponential service times, it may be easily seen that when the service times are, e.g., deterministic it makes a considerable difference to either take the THT timeout slightly smaller or slightly bigger than the service time.

Modeling of the preemptive timer protocol

We shall now consider the following timer protocol: Messages that are to be transmitted from a station are broken into fixed-length packets. Any open space in the last packet of a message is not used for another message. The THT timeout x is assumed to be some multiple of the packet transmission time β, say $x/\beta = k$, where k is a positive integer.

We model this protocol as follows. We assume that a message arrives as a bulk of packets with mean bulk size l. We propose to model the visit discipline at the queue by the k-limited discipline, limited w.r.t. packets, i.e., at most k packets can be transmitted from the station during one round.

There are no readily available approximation procedures for the mean waiting times in the k-limited model with bulk arrivals. For the 1-limited model with bulk arrivals an extension of the basic approximation method of Chapter 7 is indicated in Section 7.2.4. A similar extension to bulk arrivals of the approximation of Fuhrmann and Wang [1988] can be easily derived and will not be further discussed here.

REMARK 8.4

In the literature it is usually assumed that the timer protocol is non-preemptive. When information transfer for a large part consists of file transfers (as when the station is a file server for instance) this approach does not seem very realistic. In such a case the second modeling approach seems more appropriate.

REMARK 8.5

Above we modeled the timer mechanism at the stations of the token ring by an appropriately chosen visit discipline at the queues of the performance model. One could also take the direct approach of analyzing a similar timer mechanism at the queues. However, such a 'timed visit discipline' is very difficult to analyze. The only studies in this direction that we are aware of are Coffman, Fayolle and Mitrani [1988], Ulm [1982] and Yue [1987].

Delay analysis of the token ring

From the discussion above it is now easily seen that the results of the previous chapters, in particular those of Chapters 3 and 7, may be applied to the performance model described above. In the following we shall use the model to investigate the influence of several system parameters on the mean delays at the stations for some specific token rings.

In the next two examples we consider the influence of the THT timeout on the mean waiting times at the queues. Figure 8.3 relates to a system of two queues. The system is completely symmetric, with parameters $\lambda_1 = \lambda_2 = 0.5$, $\beta_1 = \beta_2 = 0.5$ (neg. exp.) and $s_1 = s_2 = 0.1$ (determ.). In the figure the THT timeout is varied from 0 to 20 times the mean service times ($20 \times 0.5 = 10$).

The timer protocol is assumed to be non-preemptive. Note that when the THT timeout goes to zero the mean waiting time $E\mathbf{W}$ $(=E\mathbf{W}_1=E\mathbf{W}_2)$ is in the limit that of the 1-limited visit discipline ($E\mathbf{W}=0.9375$, indicated in the figure by an asterisk). When the THT timeout becomes very large the mean waiting time converges to the result for the exhaustive discipline (in the figure represented by the dotted line, $E\mathbf{W}=0.65$). The mean waiting times for the 1-limited and the exhaustive discipline have been calculated from the pseudoconservation law for these models: Formula (6.39) for the model with 1-limited service at both queues and

$$\rho_1 E\mathbf{W}_1 + \rho_2 E\mathbf{W}_2 = \rho\frac{\lambda_1\beta_1^{(2)}+\lambda_2\beta_2^{(2)}}{2(1-\rho)} + \rho\frac{s^{(2)}}{2s} + \frac{s}{2(1-\rho)}(\rho^2-\rho_1^2-\rho_2^2)$$

for the model with exhaustive service at both queues. The other results have been obtained from simulation.

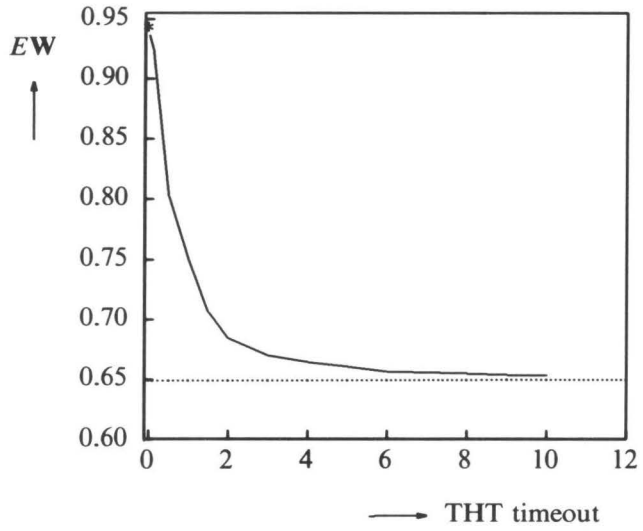

Figure 8.3 Mean waiting time versus THT timeout

As can be observed from Figure 8.3, the mean waiting times at the queues decrease when the THT timeout is increased. We can interpret this as follows. Since the switch-over times in the system are non-zero, increasing the THT timeout at each of the stations improves the efficiency of the system (reduces the workload of the system). The mean workload $E\mathbf{V}_C$ and the mean waiting time satisfy the following linear relation (cf. also (3.15)):

$$E\mathbf{V}_C = \rho E\mathbf{W} + \rho\frac{\beta^{(2)}}{\beta}, \tag{8.8}$$

in which $\beta=\beta_1=\beta_2$ and $\beta^{(2)}=\beta_1^{(2)}=\beta_2^{(2)}$. Therefore, the mean waiting time decreases with the workload when the THT timeout is increased.

Figure 8.4 relates to a non-symmetric system of two queues. The parameters are: $\lambda_1=0.7$, $\lambda_2=0.3$, $\beta_1=\beta_2=0.5$ (neg. exp.) and $s_1=s_2=0.1$ (determ.). In the figure the THT timeout is varied for each of the queues from 0 to 20 times the mean service times. Note that when the THT timeout goes to zero the mean waiting times approach the results for the 1-limited visit discipline ($E\mathbf{W}_1=1.152$, $E\mathbf{W}_2=0.671$, indicated in the figure by asterisks). These results have been calculated from the exact analysis in Chapter 6; they also appear in Table 6.2, Case A. When the THT timeout becomes very large, the mean waiting times converge to those for the exhaustive discipline ($E\mathbf{W}_1=0.585$, $E\mathbf{W}_2=0.792$, indicated in the figure by the dotted lines). These results have been calculated from exact analysis, cf. Ferguson and Aminetzah [1985]. The results for the timer protocol have been obtained from simulation.

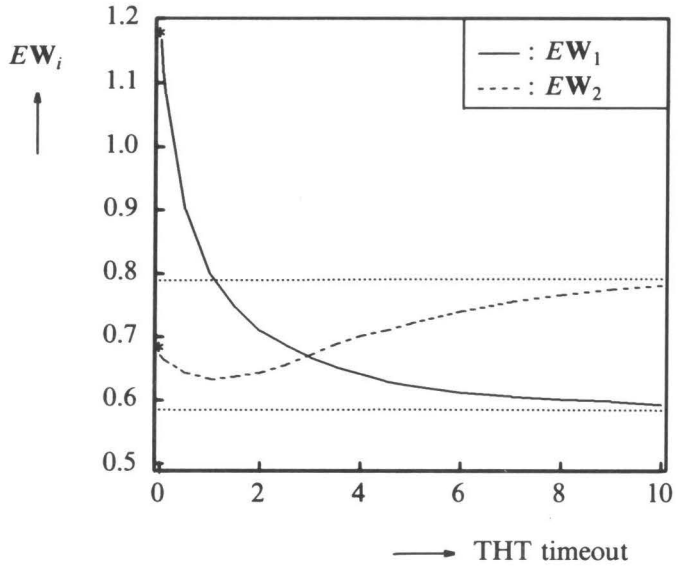

Figure 8.4 Mean waiting times versus THT timeout

As can be observed in Figure 8.4 the mean waiting time for station 2 first decreases and then increases. Apparently the initial decrease is caused by the same phenomenon as in the previous model, namely the increased efficiency of the system at larger values of the THT timeout. When the THT timeout becomes larger the more heavily loaded station 1 starts monopolizing the system at the expense of the waiting times of customers at station 2.

In the next example we consider for a fixed value of the THT timeout the effect of the message length upon the delay at the stations. We first describe the parameters for the token ring network and subsequently illustrate how to

translate the system parameters to those for the polling model.

We consider a token ring with 100 active stations connected via a cable with a length of 5 kilometer. The bandwidth of the network is 10 Mbps. The stations are assumed to have equidistant positions around the ring. The value of the THT timeout at the stations is set to 200 μs. We consider three different cases:

1) the average message length at all stations is 324 bit,

2) the average message length at all stations is 1824 bit,

3) the average message length at all stations is 7296 bit.

The overhead in one frame is 176 bit. The timer protocol is non-preemptive in Cases 1) and 2) and preemptive in Case 3). In this latter case the frame length is fixed and equal to 2000 bit.

Translation of these parameters to those for the polling model is rather straightforward. Since the stations have equidistant positions around the ring the switch-over time from Q_i to Q_{i+1} does not depend on i; it is assumed to be constant. We can calculate its mean as:

$$s_i = 5\times 10^{-6}\times\frac{d}{N} + \frac{1}{V}, \tag{8.9}$$

where d denotes the physical length of the ring in kilometers, N is the number of active stations attached to the ring, and V denotes the bandwidth of the ring (in Mbps). Hence, $s_i = 25\times 10^{-8} + 1\times 10^{-7} = 0.35$ μs; the total switch-over time $s = 35$ μs.

In Case 1) the average frame length is $324+176=500$ bit. Since the bandwidth of the network is 10 Mbps the average frame transmission time (the average service time of a customer in the polling system) is given by $\beta = 500/10\times 10^6 = 50$ μs. For ease of calculations the frame transmission time is assumed to follow an exponential distribution. Based on the discussion above we model the timer protocol of the token ring by the 5-limited visit discipline ($\{1+200/50\}=5$) at all queues in the polling system.

In Case 2) the average frame length is $1824+176=2000$ bit, leading to an average service time in the polling system of $\beta = 2000/10\times 10^6 = 200$ μs. Again the service time distribution is assumed to be exponential. The visit discipline at all queues is 2-limited service ($\{1+200/200\}=2$).

In Case 3) the frame length is fixed and equal to 2000 bit. Hence the transmission time of a frame is 200 μs. During each round, exactly one frame is transmitted from a station. A frame consists of 176 bit overhead and of 1824 bit that can be used for data transmission. This implies that a message of 7296 bit needs on the average 4 frames to complete its transmission ($4\times 1824 = 7296$). These considerations lead us to model this case using a bulk arrival process. Each arrival to the polling system is assumed to consist of a bulk of customers; each customer has a constant service time of $\beta = 2000/10\times 10^6 = 200$ μs. The mean bulk size is 4. For ease of calculations

the bulk size distribution is assumed to be shifted geometric (cf. Remark 7.3).

The systems under consideration are all chosen to be symmetric. For the cases of at most five and at most two message transmissions per station per round we have calculated the results using the approximation method of Fuhrmann and Wang [1988, Formulas (8) and (11)]. In the last case we have calculated the results directly from the pseudoconservation law (7.28) with the terms for exhaustive and gated service omitted. This is possible since the system under consideration is symmetric. Of course we can employ the techniques from Chapter 7 to investigate asymmetric systems as well. In Figure 8.5 the mean message delay (including transmission time) at the queues has been depicted as a function of message throughput for the three message lengths. Note that the delay is depicted on a logarithmic scale. Throughput is defined here as carried load times the bandwidth of the system.

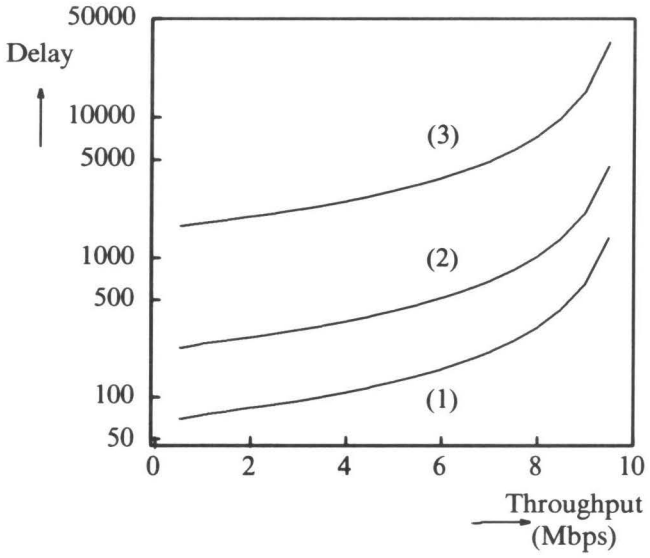

Figure 8.5 Delay-throughput characteristics for three different message lengths

The *power* of the network (cf. Kleinrock [1986]) is defined as the throughput divided by the mean delay. In Figure 8.6 the power of the network, normalized so that the maximum is equal to 1, has been depicted; the figure suggests that the normalized power is rather insensitive to the mean message length for the three (symmetric) systems under consideration. In fact, we can show that for these three cases the normalized power can approximately be written as $4\rho(1-\rho)$. Note that the normalized power is *exactly* equal to $4\rho(1-\rho)$ for a symmetric polling system with zero switch-over times and exponential service-time distribution.

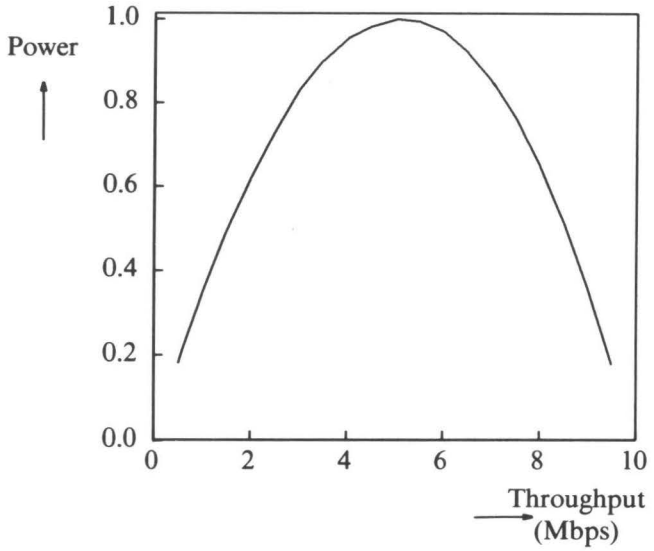

Figure 8.6 Power of the network as function of the throughput

8.4 Performance analysis of interconnected token-passing LAN's

In this section we consider the performance analysis of a network of several token rings interconnected by a backbone ring. In Section 8.4.1 the structure of the interconnected LAN is presented. In Section 8.4.2 the backbone ring, a high-speed fiber ring is described in some detail. In Section 8.4.3 the performance model for the interconnected system is considered. A local ring in which some of the attached devices are bridges is studied in isolation. In Section 8.4.4 we shall consider the analysis of throughput and end-to-end delay for the system of multiple token rings.

8.4.1 The multiple token ring LAN

There may be several reasons for interconnecting LAN's, e.g.:

- The number of attaching stations has become too large to be accommodated by a single LAN: In a local area network which has to support a large number of attaching devices, it is undesirable, if not impossible, to connect all devices to a single subnetwork such as a ring or bus.
- Geographical extension of LAN coverage: LAN's are typically unable to cover areas of more than 10 kilometers in diameter. Interconnection of LAN's provides a means to solve this problem.
- Allowing information transfer between separate LAN's: It may be desirable to interconnect separate networks to allow users on each network to use resources on the other networks as well as allowing data to be transferred between all the networks.

- Reliability: If one subnetwork fails, only a limited group of users is affected.

In existing situations, interconnection of LAN's will often lead to a heterogeneous network: some of the LAN's will be Ethernets, others rings. When a network has to be newly developed, there is considerably more freedom in determining the desired structure of the interconnected LAN. In the remainder of this chapter a communication system will be studied which in our opinion is a promising candidate for future office communication networks. The system consists of several local token rings and one *backbone* token ring. Each local ring is connected to the backbone network via a device called a *bridge*. The backbone network is a special token ring network, operating according to the Fiber Distributed Data Interface (FDDI) protocol. FDDI provides a 100 Mbps communication system using fiber optics as the transmission medium in a ring configuration. For a detailed description of the FDDI protocol we refer to Section 8.4.2. User stations are only connected to the local rings but not to the backbone; the latter serves to interconnect the bridges. In Figure 8.7 the architecture of the multiple token ring network has been depicted.

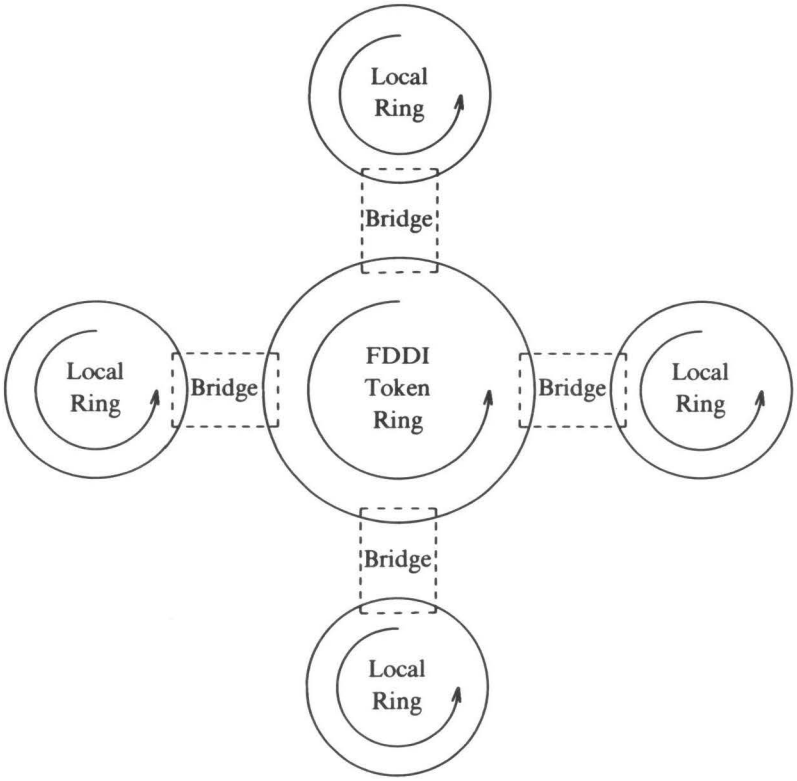

Figure 8.7 FDDI token ring as a backbone to interconnect several 802.5 token rings

In order to achieve high throughput, bridges perform only a basic store-and-forward function and are not involved in error or flow control. In case of congestion, bridges simply discard arriving frames. Lost frames will be recovered through an appropriate end-to-end protocol between the communicating stations. Bridges are assumed to be full duplex and to have separate buffer areas for each direction of data flow. In the queueing model a bridge is modeled as two separate queues, each belonging to a different token ring (cf. also Figure 8.8).

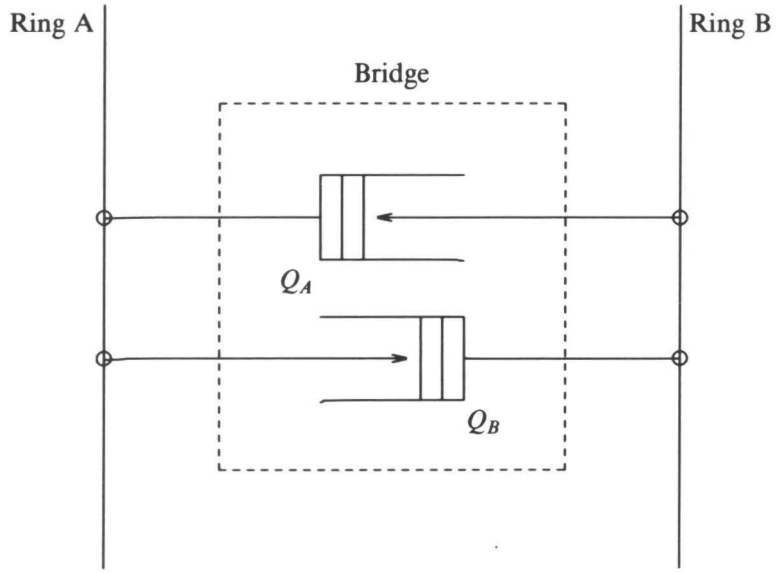

Figure 8.8 Queueing model of a bridge

As a queueing model for the interconnected token ring system we use the model as in Figure 8.9, which represents a system with two local rings interconnected by a backbone token ring. For further details on this model see the discussion in Section 8.4.4.

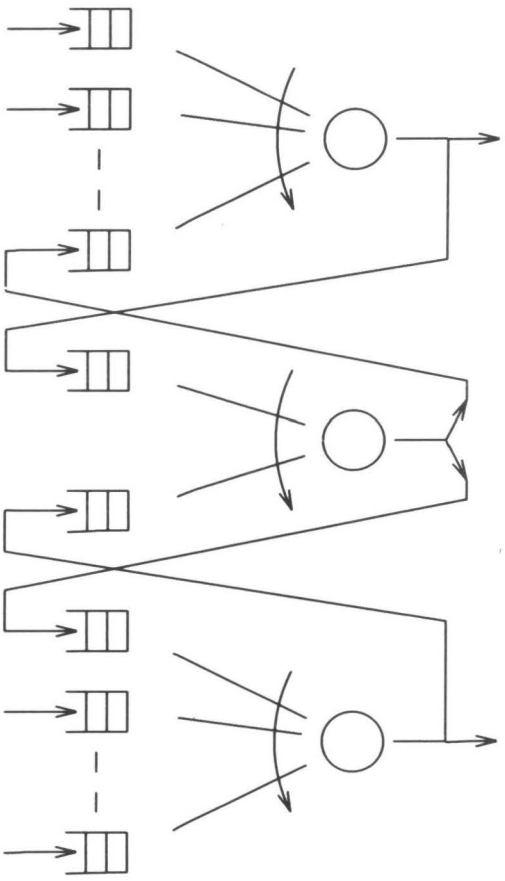

Figure 8.9 Queueing model of two interconnected token rings

Before discussing the performance analysis of multiple token ring networks, we provide a detailed discussion of the FDDI MAC protocol.

8.4.2 The FDDI MAC protocol

FDDI has been proposed by the American National Standards Institute (ANSI) for three areas of applications:

- a back-end network to interconnect processors and associated mass storage devices/peripherals
- a front-end network to interconnect high-performance hosts
- a backbone network for IEEE 802 type LAN's

FDDI is a 100 Mbps token ring which uses optical fiber as its transmission medium. Default timer values allow up to 500 stations with a total of 100 kilometers of cable. The architecture of FDDI consists of two counter-rotating

fiber rings each with an information transmission rate of 100 Mbps. These rings are referred to as *primary* and *secondary* rings and in principle they allow an effective transmission rate of 200 Mbps. However, the secondary ring is only used when a link in the primary ring fails; in that case the stations upstream and downstream from the link will reconfigure themselves to operate in *wrap* mode and use the secondary ring to again establish a logical ring. In normal operation, only the primary ring is used.

The access scheme in FDDI is based on token passing. To take advantage of the high bandwidth, multiple token operation (cf. Remark 8.2) is employed. FDDI has preserved the IEEE 802.5 frame structure and addressing conventions and, in general, deviates from the IEEE 802.5 concepts only where necessitated by its higher data transmission rates.

All traffic in FDDI is classified as either synchronous or asynchronous. The FDDI priority mechanism is designed to provide different classes of service to accommodate these two types of traffic simultaneously. For synchronous traffic, bandwidth and transmission delay are guaranteed; thus the class of synchronous traffic is suitable for applications with real-time delivery requirements, such as voice, real-time animated graphics or process control. For asynchronous traffic FDDI provides eight levels of priority.

To achieve the guaranteed bandwidth and transmission delay for synchronous traffic a so-called 'timed-token' protocol is used. In such a protocol the token-holding time at a station depends in part on the time between successive arrivals of the token at that station (the token rotation time). Below we shall briefly describe a simplified version of the protocol.

At ring initialization time, all stations negotiate a target token rotation time (TTRT), a parameter which specifies the expected token rotation time. Each station requests a value that is small enough to support its synchronous traffic needs. At the end of the negotiation, the shortest TTRT requested becomes the operational TTRT.

A token rotation timer (TRT) is used in each station to measure the time between successive arrivals of the token at that station. At each arrival of the token, the TRT is compared to the operational TTRT. The station is allowed to transmit synchronous frames for its alotted time whenever it receives the token. However, it may initiate transmission of asynchronous frames only when the token is 'ahead of schedule', i.e., when the TRT is smaller than the operational TTRT. A token-holding timer (THT) is used for determining the amount of time the station is allowed to transmit asynchronous frames. As the token is received, the THT is loaded with the amount of time the token is 'ahead of schedule', i.e., with the value of the operational TTRT minus the value of the TRT. Transmissions already in progress when the THT expires are completed. It follows that all bandwidth that is not used for synchronous transmission is available for asynchronous transmission.

A detailed description of the FDDI protocol can be found in ANSI [1986]. For some performance-related aspects of FDDI see Ulm [1982], Johnson [1987], Sevcik and Johnson [1987], Goyal and Dias [1988], Dykeman and Bux [1988] and Bux [1988].

8.4.3 Delay analysis of a local ring in isolation

As an example consider a local token ring with a bandwidth of 10 Mbps and a total length of 2 kilometer. We assume that there are 36 active user stations connected to the ring. In addition, 4 bridge stations attached to the ring provide information transfer between the local ring and other networks. We present a delay analysis of the local ring in isolation, ignoring any dependencies caused by the interaction with other networks. The results will be used for the analysis of the interconnected network in the next section.

Since messages that have already taken one or more hops through the network should be treated preferentially to newly arriving messages, the THT at a bridge is given a large value; we model this by assuming that the queue representing a bridge is served according to the exhaustive visit discipline. Note that only inbound traffic is present at the bridge queues, cf. also the remarks at the end of Section 8.4.1. At the user stations the value of the THT is assumed to be much smaller than the mean transmission time of one frame. Hence the queues representing the user stations are served 1-limited. The transmission time of a frame is assumed to follow a negative exponential distribution with mean 200 μs. From (8.9) we calculate the switch-over time between queue i and queue $i+1$ as $s_i = 5 \times 2/40 + 1/10 = 0.35 \mu s$, $i = 1,...,40$, $s = \sum s_i = 14 \mu s$. The switch-over time is deterministic. Denote the arrival rate at each of the user stations by λ_{user} and the arrival rates at each of the bridge queues by λ_{bridge}. It is assumed that the arrival processes at the user stations as well as the bridge stations is Poisson (cf. also Remark 8.6 below). Further denote by $E\mathbf{T}_{bridge}$ and $E\mathbf{T}_{user}$ the delay averaged over the bridges and the user stations respectively. Note that the delay includes the transmission time. In Figure 8.10 the throughput-delay characteristics as calculated from the approximation procedure described in Chapter 7 have been depicted for the case that $\lambda_{bridge} = 9\lambda_{user}$, and in Figure 8.11 for the case that $\lambda_{bridge} = \lambda_{user}$.

The results presented in the figures are obtained using the (basic) approximation method described in Chapter 7. The parameters used for the polling model are:

40 queues, $Q_1, \ldots, Q_{36}$ served 1-limited, $Q_{37}, \ldots, Q_{40}$ served exhaustively;
$\beta_i = 200$, $i = 1,...,40$ (neg. exp.); $s = 14$ (determ.);
$E\mathbf{T}_{user} = E\mathbf{W}_1 + 200$, $E\mathbf{T}_{bridge} = E\mathbf{W}_{40} + 200$.

REMARK 8.6

The assumption of Poisson arrivals at the bridge queues may be more severe than the assumption of Poisson arrivals at the user stations since arrivals at these bridge queues are strongly correlated.

In Figure 8.12 the power P of the network, *normalized at 1*, has been depicted for both cases; here the power has been defined as the throughput (Λ) of the system divided by the average delay ($E\bar{\mathbf{T}}$): $P = \Lambda / E\bar{\mathbf{T}}$, where $E\bar{\mathbf{T}}$ is given by,

$$E\bar{\mathbf{T}} = \frac{1}{4\lambda_{bridge} + 36\lambda_{user}} \times \left[4\lambda_{bridge} E\mathbf{T}_{bridge} + 36\lambda_{user} E\mathbf{T}_{user} \right]. \tag{8.10}$$

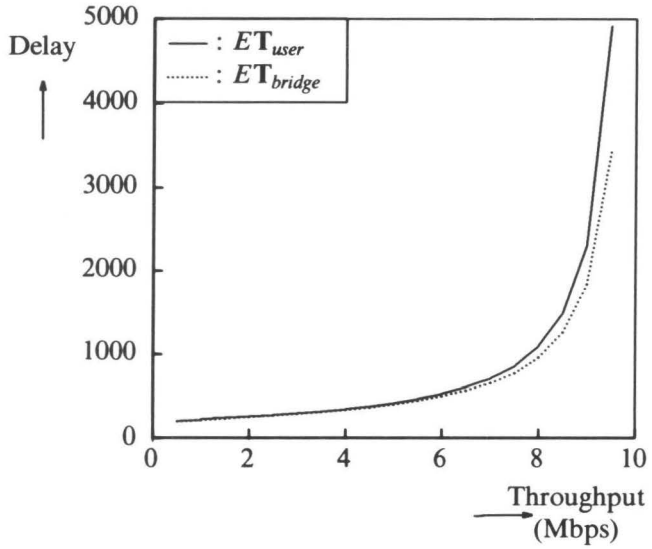

Figure 8.10 Throughput-delay characteristic for the case that $\lambda_{bridge} = 9\lambda_{user}$

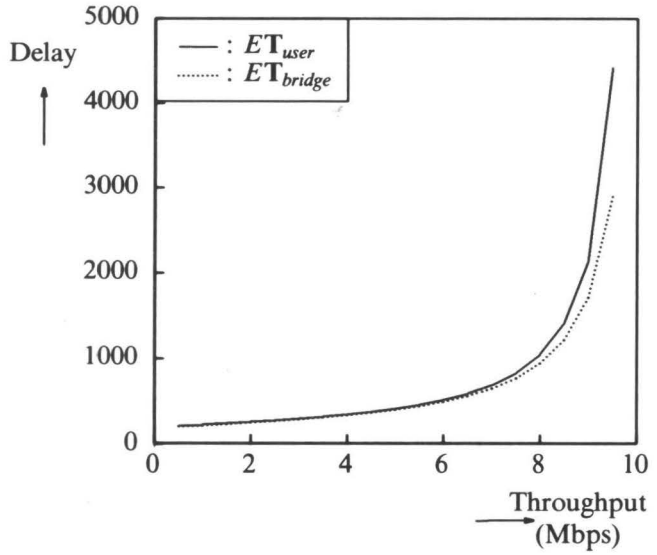

Figure 8.11 Throughput-delay characteristic for the case that $\lambda_{bridge} = \lambda_{user}$

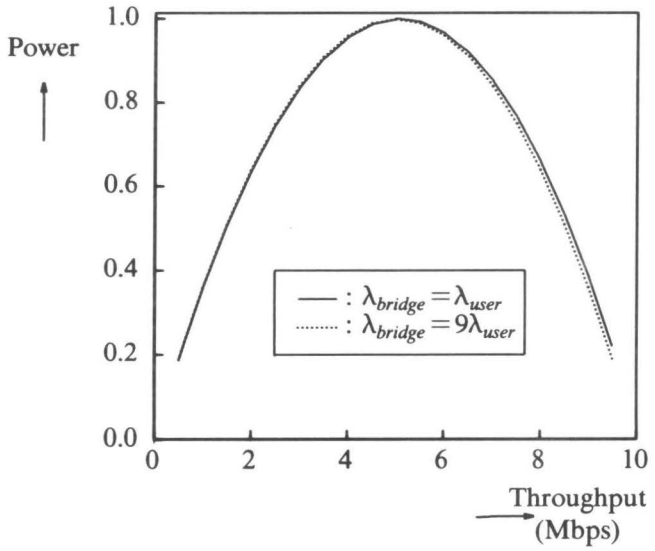

Figure 8.12 Power of the network as function of the throughput for the case that $\lambda_{bridge} = 9\lambda_{user}$ and the case that $\lambda_{bridge} = \lambda_{user}$

Note that the figure suggests that the power of the network is rather insensitive to the ratio of λ_{bridge} and λ_{user}.

8.4.4 Performance analysis of the multiple token ring

Consider a system consisting of several local 802.5 token rings connected via an FDDI token ring backbone as in Figure 8.7. The bridges are full duplex and have at their disposal two processors, each of which controls one direction of data flow. Inter-network messages experience no queueing delay due to processing at the bridge. An important prerequisite for our analysis is that the number of discarded messages at the bridges is sufficiently low so as to avoid the need for excessive retransmissions by the LLC protocol.

REMARK 8.7

In Bux and Grillo [1985] the effect of end-to-end retransmissions by the LLC protocol due to message losses at the bridges is studied via a simulation model. It is shown that such retransmissions should be avoided as much as possible, since they in turn trigger other retransmissions and thus may seriously degrade network performance. In Gerla and Kleinrock [1988] some flow and congestion control mechanisms for implementation at a MAC bridge are considered. The main problem here is to implement such features while retaining the transparency and efficiency of the bridges. In Bux, Meister and Wong [1983] a simulation study is undertaken investigating the effect of various buffering strategies upon the overflow probability at bridges with finite capacity buffers.

The queueing model for the multiple token ring is constructed from the queueing model of the single token ring as follows. Consider a model with M local rings $R_1, \ldots, R_M$ connected by one backbone ring R_{M+1}. Ring R_m, $m = 1,...,M$, consists of N_m stations; ring R_{M+1} consists of M stations (the bridges of the local rings). Each local ring is assumed to have one bridge connecting it to the backbone ring. The stations at each local ring are indexed according to the ring to which they belong and to their position on the ring. For example, $Q_{3,7}$ refers to the seventh station of ring R_3. The station at the last position in each local ring is assumed to be the bridge station; so Q_{m,N_m} is the bridge station of ring R_m. See Figure 8.13 (and also Figure 8.9)) for the queueing model with two local token rings R_1 and R_2 interconnected by a backbone ring R_3.

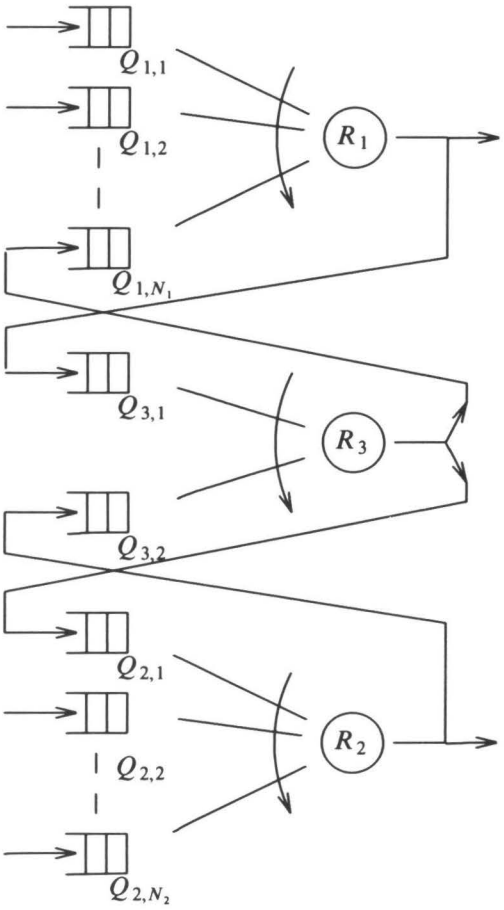

Figure 8.13 Queueing model of two interconnected token rings

The arrival process at $Q_{m,i}$, $m=1,...,M$; $i=1,...,N_m-1$, is assumed to be a Poisson process with parameter $\lambda_{m,i}$. Similarly, all parameters of the service and the switch-over processes are indexed according to the queue index. There are no arrivals from outside the network at the bridges. Denote by λ_{m,N_m} the arrival rate of the bridge in ring R_m (due to inter-network traffic).

Messages arriving at queue $Q_{m,i}$, $m=1,...,M$; $i=1,...,N_m-1$, are assumed to be inter-network messages destined for ring R_j, $j\neq m$, with probability $p_{m,i,j}$.

The user stations in the local rings are served 1-limited; the bridges are served exhaustively (see the the discussion at the beginning of Section 8.4.3). The stations of the backbone ring are all served 1-limited.

It is assumed that all traffic in the network is data traffic and has the same priority. Our performance measure of interest will again be the throughput-delay characteristic.

REMARK 8.8

Note that for large data transfers such as file transfers throughput is usually a more important performance measure than delay. For interactive data traffic both throughput and delay may be important, whereas for synchronous (time-critical) traffic such as voice or animated graphics delay is the most important performance measure.

As an example we shall investigate the simplest possible network: two token rings connected by a backbone ring, cf. Figure 8.13. Let us first consider the arrival rate at the bridges. With a slight simplification of notation, denote by $p_{1,i}$ the probability that a message arriving at $Q_{1,i}$ is a packet destined for ring R_2. Similarly, let $p_{2,i}$ denote the probability that a message arriving at $Q_{2,i}$ is a packet destined for ring R_1. It is easily seen that the arrival rates at the bridge queues Q_{1,N_1} and Q_{2,N_2} can be determined as the solution of the following set of two equations:

$$\begin{aligned}\lambda_{1,N_1} &= p_{2,1}\lambda_{2,1}+p_{2,2}\lambda_{2,2}+\cdots+p_{2,N_2}\lambda_{2,N_2},\\ \lambda_{2,N_2} &= p_{1,1}\lambda_{1,1}+p_{1,2}\lambda_{1,2}+\cdots+p_{1,N_1}\lambda_{1,N_1}.\end{aligned} \tag{8.11}$$

Solving (8.11) for λ_{1,N_1} and λ_{2,N_2} leads to:

$$\begin{aligned}\lambda_{1,N_1} &= \frac{\sum_{j=1}^{N_2-1} p_{2,j}\lambda_{2,j}+p_{2,N_2}\sum_{j=1}^{N_1-1} p_{1,j}\lambda_{1,j}}{1-p_{1,N_1}p_{2,N_2}},\\ \lambda_{2,N_2} &= \frac{\sum_{j=1}^{N_1-1} p_{1,j}\lambda_{1,j}+p_{1,N_1}\sum_{j=1}^{N_2-1} p_{2,j}\lambda_{2,j}}{1-p_{1,N_1}p_{2,N_2}}.\end{aligned} \tag{8.12}$$

Parameters

The data transmission rate in the local rings is 10 Mbps and in the backbone ring 100 Mbps. The mean frame transmission time of a message in one of the local rings is 200 μs and in the backbone ring 20 μs. The frame transmission time in both cases follows a hyperexponential distribution with a squared coefficient of variation equal to 2. Each of the local rings has 20 user stations attached to it and one bridge. So $N_1 = N_2 = 21$. The backbone ring has only two stations attached to it, viz. the bridge stations of the local rings. So $N_3 = 2$. The length of each of the local rings is 1.05 kilometer, leading to a switch-over time of 0.35 μs between stations. The length of the backbone ring is 20 kilometer. The switch-over time between the stations of the backbone ring is 51.5 μs. The switch-over times are assumed to be deterministic. A message arriving at a user station has a probability of 0.5 of being destined for the other local ring. A message arriving from the backbone at a bridge in the local ring has a probability of 0.1 of returning to the other ring. Hence $p_{m,i} = 0.5$ for $m = 1,2$ and $i = 1,...,20$; $p_{m,N_m} = 0.1$ for $m = 1,2$. The arrival rates at the user stations are assumed to be equal to λ for both local rings.

End-to-end delay

For the analysis of the end-to-end delay of a message that arrives at ring R_1 and is destined for ring R_2, we decompose the network into three separate parts, the local ring R_1, the backbone R_3 and the local ring R_2, and analyze the delay in each of them separately using the (basic) approximation procedure of Chapter 7 in the same manner as described in Section 8.4.3. The three parts are then aggregated and the end-to-end delay is approximated by the sum of the delays in the three sub-networks (in the first ring the delay from the user station, and in the last ring the delay from the bridge station). Note that in this approach we assume that messages arrive at the bridges according to a Poisson process whereas, in reality, the arrival process at the bridge is the partial output process of a token ring system (cf. also Remark 8.6). In addition, we are neglecting several dependencies, for instance we may expect that a period of increased traffic in one ring leads to increased traffic in other rings as well, etc.

In Figure 8.14, the end-to-end delay and the individual delays in the sub-networks have been depicted versus the throughput.

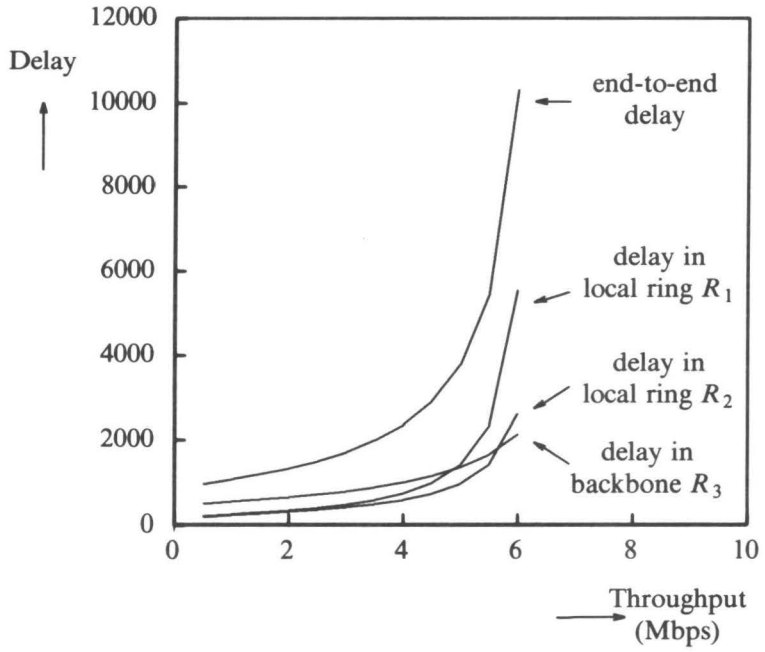

Figure 8.14 Sub-network and end-to-end delay in the multiple token ring

As can be observed from Figure 8.14 a message initiated from ring R_1 suffers the highest delay in this first ring. The reason is that it is transmitted from a user station with low priority (1-limited service). Then this message is transmitted over the backbone ring R_3. On the backbone ring all traffic has the same priority (all traffic on the backbone ring is internetwork traffic). Finally the message is transmitted on its destination ring R_2. The delay it experiences is smaller than that on R_1, since on R_2 it is transmitted from the bridge with high priority (exhaustive service).

Finally we shall briefly mention some of the available literature on performance related aspects of interconnection of token ring LAN's. In Bux [1985] a detailed simulation study of the multiple token ring system is presented. In Welzel [1987] simulation is employed to analyze the behavior of the backbone part of the multiple ring. Several approximations have been proposed to analyze the end-to-end delay in the interconnected network. Most of the studies decouple the network and decompose it into several subnetworks. A few of the recent papers addressing this issue are Chiarawongse, Srinivasan and Teorey [1988], Ibe and Cheng [1988], Kuruppilai and Bengtson [1988] and Yang, Ghosal and Bhuyan [1986]. Menasce and Leite [1984] approximately analyze the end-to-end delays in an interconnected system of token buses. Takine, Takahashi and Hasegawa [1986] consider an interconnected system of token rings with single buffers and exhaustive service at all queues. In Murata

and Takagi [1987] a token ring with a finite capacity bridge is studied and an approximation is derived for the waiting times at the queues.

8.5 Directions for further research

In this chapter we have only made a beginning with a study of the modeling of token-passing networks by polling networks. Many open problems remain which should be investigated; however, a more systematic analysis than the one presented here is not within the scope of the present study.

We mention a few directions for further research. First, the modeling of the timer mechanism in the token ring network by an appropriately chosen visit discipline should be investigated in more detail. It would in this respect be interesting to test the Bernoulli visit discipline (cf. Chapter 3) as a possible candidate, because the choice of the Bernoulli parameter provides some extra flexibility in representing the THT mechanism.

A second issue is the analysis of the bridges in interconnected networks. In the course of our analysis we have made several simplifying assumptions. A detailed simulation study should investigate the effect of these assumptions. It may turn out to be necessary to incorporate also elements from higher layer protocols, such as the Logical Link Control protocol.

Finally the approach of analyzing the delay in an interconnected network by decomposing the network and analyzing the delay in each of the sub-networks has some obvious weaknesses (several dependencies are neglected, unjustified Poisson assumptions), and should leave room for improvement.

References

Abramson, N. (1970). The Aloha system - another alternative for computer communications. *Proceedings of the Fall Joint Computer Conference,* 281-285.

Ali, O.M.E., Neuts, M.F. (1984). A service system with two stages of waiting and feedback of customers. *J. Appl. Prob.,* Vol. 21, 404-413.

ANSI (1986). FDDI token ring media access control (MAC). *Amer. Nat. Stand., draft proposal X3T9/84-100.*

Baker, J.E., Rubin, I. (1987). Polling with a general-service order table. *IEEE Trans. Commun.,* Vol. COM-35, 283-288.

Blanc, J.P.C. (1982). *Applications of the Theory of Boundary Value Problems in the Analysis of a Queueing Model with Paired Services.* Mathematical Centre Tract 153, Amsterdam.

Blanc, J.P.C. (1988). A numerical approach to cyclic-service queueing models. *Research Memorandum FEW 312, Tilburg University.*

Boxma, O.J. (1985). Two symmetric queues with alternating service and switching times. In: *Performance '84,* ed. E. Gelenbe, North-Holland Publ. Cy., Amsterdam, 409-431.

Boxma, O.J. (1986). Models of two queues: a few new views. In: *Teletraffic Analysis and Computer Performance Evaluation,* eds. O.J. Boxma, J.W. Cohen and H.C. Tijms, North-Holland Publ. Cy., Amsterdam, 75-98.

Boxma, O.J. (1989). Workloads and waiting times in single-server systems with multiple customer classes. *To appear in Queueing Systems.*

Boxma, O. J., Groenendijk, W.P. (1987). Pseudo-conservation laws in cyclic-service systems. *J. Appl. Prob.,* Vol. 24, 949-964.

Boxma, O.J., Groenendijk, W.P. (1988a). Waiting times in discrete-time cyclic-service systems. *IEEE Trans. Commun.,* Vol. COM-36, 164-170.

Boxma, O.J., Groenendijk, W.P. (1988b). Two queues with alternating service and switching times. In: *Queueing Theory and its Applications - Liber Amicorum for J.W. Cohen,* eds. O.J. Boxma and R. Syski, North-Holland Publ. Cy., Amsterdam, 261-282.

Boxma, O.J., Groenendijk, W.P., Weststrate, J.A. (1988). A pseudoconservation law for service systems with a polling table. *Report Centre for Mathematics and Computer Science.* To appear in *IEEE Trans. Commun.*

Boxma, O. J., Meister, B. (1986). Waiting-time approximations for cyclic-service systems with switch-over times. *Performance Evaluation Review,* Vol. 14, 254-262.

Boxma, O. J., Meister, B. (1987). Waiting-time approximations in multi-queue systems with cyclic service. *Performance Evaluation,* Vol. 7, 59-70.

Boxma, O.J., Weststrate, J.A. (1989). Waiting times in polling systems with Markovian server routing. *Report Centre for Mathematics and Computer Science.* To appear in *Proc. 1989 Conf. on Measurement, Modeling and Performance Evaluation of Computer Systems, Braunschweig, BRD.*

Browne, S., Yechiali, U. (1989a). Dynamic routing in polling systems. In:

Teletraffic Science for New Cost-Effective Systems, Networks and Services, ITC-12, ed. M. Bonatti, North-Holland Publ. Cy., Amsterdam, 1455-1466.

Browne, S., Yechiali, U. (1989b). Dynamic priority rules for cyclic-type queues. *Adv. Appl. Prob.,* Vol. 21, 432-450.

Bux, W. (1981). Local-area subnetworks: A performance comparison. *IEEE Trans. Commun.,* Vol. COM-29, 1465-1473.

Bux, W. (1985). Performance issues in local-area networks. *IBM Systems Journal,* Vol. 23, 351-374.

Bux, W. (1988). Modeling token ring networks - a survey. In: *Data Communication Systems and Their Performance,* eds. L.F.M. de Moraes, E. de Souza e Silva, L.F.G. Soares. North-Holland Publ. Cy., Amsterdam, 192-221.

Bux, W., Grillo, D. (1985). Flow control in local-area networks of interconnected token rings. *IEEE Trans. Commun.,* Vol. COM-33, 1058-1066.

Bux, W., Meister, B.W., Wong, J.W. (1983). Bridges for interconnection of ring networks - a simulation study. In: *Information Processing '83,* North-Holland Publ. Cy., Amsterdam, 181-185.

Bux, W., Truong, H.L. (1983). Mean-delay approximation for cyclic-service queueing systems. *Performance Evaluation,* 3, 187-196.

Chiarawongse, J., Srinivasan, M.M., Teorey, T.J. (1988). Performance analysis of a large interconnected network by decomposition techniques. *IEEE Network,* Vol. 2, 19-27.

Coffman, E.G., Fayolle, G., Mitrani, I. (1988). Two queues with alternating service periods. In: *Performance '87,* eds. P.-J. Courtois and G. Latouche, North-Holland Publ. Cy., Amsterdam, 227-239.

Coffman, E.G., Hofri, M. (1982). On the expected performance of scanning disks. *SIAM Journal on Computing,* Vol. 11, 60-70.

Cohen, J.W. (1982). *The Single Server Queue.* North-Holland Publ. Cy., Amsterdam; 2nd ed.

Cohen, J.W. (1987). A two-queue, one-server model with priority for the longer queue. *Queueing Systems,* Vol. 2, 261-283.

Cohen, J.W. (1988a). A two-queue model with semi-exhaustive alternating service. In: *Performance '87,* eds. P.-J. Courtois and G. Latouche, North-Holland Publ. Cy., Amsterdam, 19-37.

Cohen, J.W. (1988b). Boundary value problems in queueing theory. *Queueing Systems,* Vol. 3, 97-128.

Cohen, J.W, Boxma, O.J. (1981). The M/G/1 queue with alternating service formulated as a Riemann-Hilbert boundary value problem. In: *Performance '81,* ed. F.J. Kylstra, North-Holland Publ. Cy., Amsterdam, 181-199.

Cohen, J.W, Boxma, O.J. (1983). *Boundary Value Problems in Queueing System Analysis.* North-Holland Publ. Cy., Amsterdam.

Cooper, R.B. (1970). Queues served in cyclic order: Waiting times. *The Bell System Technical Journal,* Vol. 49, 399-413.

Cooper, R.B., Murray, G. (1969). Queues served in cyclic order. *The Bell System Technical Journal,* Vol. 48, 675-689.

Doshi, B.T. (1985). A note on stochastic decomposition in a GI/G/1 queue with vacations or set-up times. *J. Appl. Prob.,* Vol. 22, 419-428.

Doshi, B.T. (1986). Queueing systems with vacations - a survey. *Queueing Systems,* Vol. 1, 29-66.

Doshi, B.T. (1988). Generalizations of the stochastic decomposition results for single server queues with vacations. *Report AT&T Bell Labs, Holmdel, NJ.*

Dykeman, D., Bux, W. (1988). Analysis and tuning of the FDDI media access control protocol. *IEEE J. Sel. Areas Commun.,* Vol. SAC-6, 997-1010.

Eckberg, A.E., Meier-Hellstern, K.S. (1988). An effective method for delay distribution approximation for token-rings with large numbers of nodes. In: *Teletraffic Science for New Cost-Effective Systems, Networks and Services, ITC-12,* ed. M. Bonatti, North-Holland Publ. Cy., Amsterdam, 560-566.

Eisenberg, M. (1971). Two queues with changeover times. *Oper. Res.,* Vol. 19, 386-401.

Eisenberg, M. (1972). Queues with periodic service and changeover time. *Oper. Res.,* Vol. 20, 440-451.

Everitt, D.E. (1986a). Simple approximations for token rings. *IEEE Trans. Commun.,* Vol. COM-34, 719-721.

Everitt, D.E. (1986b). A conservation-type law for the token ring with limited service. *Br. Telecom Technol. J.* 4, 51-61.

Everitt, D.E. (1989). A note on the pseudo-conservation laws for cyclic service systems with limited service disciplines. *IEEE Trans. Commun.,* Vol. COM-37, 781-783.

Feller, W. (1966). *Probability Theory and Its Applications,* Vol. 2. Wiley, New York,

Ferguson, M.J. (1986). Computation of the variance of the waiting time for token rings. *IEEE J. Sel. Areas Commun.,* Vol. SAC-4, 775-782.

Ferguson, M.J., Aminetzah, Y.J. (1985). Exact results for nonsymmetric token ring systems. *IEEE Trans. Commun.,* Vol. COM-33, 223-231.

Fischer, W. (1989). Approximate analysis of a class of priority polling systems with application to the D-channel access protocol. *Report Institute of Communications Switching and Data Techniques, University of Stuttgart.*

Fournier, L., Rosberg, Z. (1989). Expected waiting times in cyclic service systems under priority disciplines. *Report Computer Science Department, Technion, Haifa.*

Franken, P., König, D., Arndt, U., Schmidt, V. (1982). *Queues and Point Processes.* Wiley, New York.

Fuhrmann, S.W. (1984). A note on the M/G/1 queue with server vacations. *Oper. Res.,* Vol. 32, 1368-1373.

Fuhrmann, S.W. (1987). Inequalities for cyclic service systems with limited service. In: *Proc. GLOBECOM '87, Tokyo,* 182-186.

Fuhrmann, S.W., Cooper, R.B. (1985). Stochastic decompositions in the M/G/1 queue with generalized vacations. *Oper. Res.,* Vol. 33, 1117-1129.

Fuhrmann, S.W., Wang, Y.T. (1988). Mean waiting time approximations of cyclic service systems with limited service. In: *Performance '87,* eds. P.-J. Courtois and G. Latouche, North-Holland Publ. Cy., Amsterdam, 253-265.

Gaier, D. (1964). *Konstruktive Methoden der Konformen Abbildung.* Springer Verlag, Berlin.

Gaver, D.P. Jr. (1962). A waiting line with interrupted service, including priorities. *J. Roy. Stat. Soc. B* 24, 73-90.

Gelenbe, E., Iasnogorodski, R. (1980). A queue with server of walking type (autonomous service). *Ann. Inst. Henri Poincaré,* Vol. B16, 63-73.

Gelenbe, E., Mitrani, I. (1980). *Analysis and Synthesis of Computer Systems.* Academic Press, New York.

Gerla, M., Kleinrock, L. (1988). Congestion control in interconnected LAN's. *IEEE Network,* Vol. 2, 72-76.

Gianini, J., Manfield, D.R. (1988). An analysis of symmetric polling systems with two priority classes. *Performance Evaluation,* Vol. 8, 93-115.

Giannakouros, N.P., Laloux, A. (1988). A general conservation law for a priority polling system. *Report Telecommunications Laboratory, Leuven University, Belgium.*

Goyal, A., Dias, D. (1988). Performance of priority protocols on high speed token ring networks. In: *Data Communication Systems and Their Performance,* eds. L.F.M. de Moraes, E. de Souza e Silva, L.F.G. Soares, North-Holland Publ. Cy., Amsterdam, 13-22.

Groenendijk, W.P. (1988a). A conservation-law based approximation algorithm for waiting times in polling systems. *Report Centre for Mathematics and Computer Science, Amsterdam.*

Groenendijk, W.P. (1988b). *Unpublished.*

Groenendijk, W.P. (1989). Waiting-time approximations for cyclic-service systems with mixed service strategies. In: *Teletraffic Science for New Cost-Effective Systems, Networks and Services, ITC-12,* ed. M. Bonatti, North-Holland Publ. Cy., Amsterdam, 1434-1441.

Groenendijk, W.P., Levy, H. (1989). Performance analysis of Transaction Driven Computer Systems via queueing analysis of polling models. *Report Centre for Mathematics and Computer Science, Amsterdam.*

Halfin, S. (1983). Batch delays versus customer delays. *Bell System Tech. J.* 62, 2011-2015.

Harris, C.M., Marchal, W.G. (1988). State dependence in M/G/1 server-vacation models. *Oper. Res.,* Vol. 36, 560-565.

Hashida, O., Ohara, K. (1972). Line accommodation capacity of a communication control unit. *Review of the Electrical Communication Laboratories,* Vol. 20, 231-239.

Heyman, D.P., Sobel, M.J. (1982). *Stochastic Models in Operations Research, Vol. I.* McGraw-Hill Book Company, New York.

Hunter, J.J. (1983). *Mathematical Techniques of Applied Probability, Vol. 2.* Academic Press, New York.

Hofri, M. (1986). Two queues and one server with threshold switching. In: *Teletraffic Analysis and Computer Performance Evaluation,* eds. O.J. Boxma, J.W. Cohen, H.C. Tijms, North-Holland Publ. Cy., Amsterdam, 409-428.

Ibe, O.C., Cheng, X. (1988). Delay analysis of networks of interconnected token rings. *Report Georgia Institute of Technology.*

IEEE (1983). IEEE Standard 802.3, CSMA/CD access method and physical layer specifications.

IEEE (1985a). IEEE Standard 802.4, token-passing bus access method and physical layer specifications.

IEEE (1985b). IEEE Standard 802.5, token ring access method and physical layer specifications.

Jaiswal, N.K. (1968). *Priority Queues.* Academic Press, New York.

Johnson, M.J. (1987). Proof that timing requirements of the FDDI token ring protocol are satisfied, *IEEE Trans. Commun.,* Vol. COM-620-625.

Keilson, J., Servi, L.D. (1986). Oscillating random walk models for GI/G/1 vacation systems with Bernoulli schedules. *J. Appl. Prob.,* Vol. 23, 790-802.

Kella, O., Yechiali, U. (1988). Priorities in the M/G/1 queue with server vacations. *Naval Res. Log. Quart.,* Vol. 35, 23-34.

Kleinrock, L. (1964). *Communication Nets - Stochastic Message Flow and Delay.* Dover, New York.

Kleinrock, L. (1965). A conservation law for a wide class of queueing disciplines. *Naval Res. Log. Quart.,* Vol. 12, 181-192.

Kleinrock, L. (1976). *Queueing Systems, Vol. 2.* Wiley, New York.

Kleinrock, L. (1986). Performance models for distributed systems. In: *Teletraffic Analysis and Computer Performance Evaluation,* eds. O.J. Boxma, J.W. Cohen and H.C. Tijms, North-Holland Publ. Cy., Amsterdam, 1-15.

Kobayashi, H., Konheim, A.G. (1977). Queueing models for computer communications systems analysis. *IEEE Trans. Commun.,* Vol. COM-25, 2-29.

Konheim, A.G., Meister, B. (1974). Waiting lines and times in a system with polling. *J. Ass. Comput. Mach.* 21, 470-490.

Kühn, P.J. (1979). Multiqueue systems with nonexhaustive cyclic service. *The Bell System Techn. J.,* Vol. 58, 671-698.

Kümmerle, K., Reiser, M. (1987). Local-area networks - major technologies and trends. In: *Frontiers in Communications: Advances in Local Area Networks,* IEEE Press, New York, 2-26.

Kuruppilai, R., Bengtson, N. (1988). Performance analysis in local area networks of interconnected token rings. *Computer Communications,* Vol. 11, 59-64.

Levy, H. (1988). Optimization of polling systems via binomial service. *Report Department of Computer Science, Tel-Aviv University, Tel-Aviv.*

Levy, H., Kleinrock, L. (1986). A queue with starter and a queue with vacations: delay analysis by decomposition. *Oper. Res.,* Vol. 34, 426-436.

Levy, H., Kleinrock, L. (1987). Polling systems with zero switch-over times: a general method for analyzing the expected delay. *Technical Report, AT&T Bell Laboratories, Holmdel, NJ.*

Levy, H., Sidi, M. (1988). Correlated arrivals in polling systems. *Report Department of Computer Science, Tel-Aviv University, Tel-Aviv.*

Levy, H., Sidi, M., Boxma, O.J. (1988). Dominance relations in polling systems. *Report Department of Computer Science, Tel-Aviv University, Tel-Aviv.*

Levy, H., Yechiali, U. (1975). Utilization of idle time in an M/G/1 queueing system. *Management Sci.* 22, 202-211.

Lucantoni, D.M., Meier-Hellstern, K.S., Neuts, M.F. (1988). A single server queue with server vacations and a class of non-renewal arrival processes.

Submitted to J. Appl. Prob.

Manfield, D.R. (1985). Analysis of a priority polling system for two-way traffic. *IEEE Trans. Commun.,* Vol. COM-33, 1001-1006.

Mapp, G.E., Manfield, D.R. (1986). Performance analysis of priority polling systems with complex cycles. *International Conference on Computer Communications 1986,* ed. P.J. Kühn, North-Holland Publ. Cy., Amsterdam, 583-588.

Melamed, B., Whitt, W. (1988). On arrivals that see time averages. *Technical Report AT&T Bell Laboratories.*

Menasce, D.A., Leite, L.L.P. (1984). Performance evaluation of isolated and interconnected token bus local area networks. *Performance Evaluation Review,* Vol. 12, 167-175.

Metcalfe, R.M., Boggs, D.R. (1976). ETHERNET: distributed packet switching for local computer networks. *Commun. of the ACM,* Vol. 19, 395-404.

Murata, M., Takagi, H. (1987). Performance of token ring networks with a finite capacity bridge. *TRL Research Report TR87-0027, IBM Japan, Tokyo.*

Nauta, H. (1989). *Ergodicity Conditions for a Class of Two-Dimensional Queueing Problems.* Thesis, University of Utrecht, The Netherlands.

Ott, T.J. (1984). On the M/G/1 queue with additional inputs. *J. Appl. Prob.,* Vol. 21, 129-142.

Ozawa, T. (1987). An analysis for multi-queueing systems with cyclic-service discipline. Models with exhaustive and gated service. *IEICE Technical Report,* Vol. 87, 19-24 (in Japanese).

Pang, J.W.M., Donaldson, R.W. (1986). Approximate delay analysis and results for asymmetric token-passing and polling networks. *IEEE J. Sel. Areas Commun.,* Vol. SAC-4, 783-793.

Reiser, M. (1986). Communication-system models embedded in the OSI-reference model, a survey. *Computer Networking and Performance Evaluation,* eds. T. Hasegawa, H. Takagi and Y. Takahashi, North-Holland Publ. Cy., Amsterdam, 85-111.

Rosberg, Z., Gail, H.R. (1989). ASTA implies an M/G/1-like load decomposition for a server with vacations. *Report Computer Science Department, Technion, Haifa.*

Rubin, I., DeMoraes, L.F. (1983). Message delay analysis for polling and token multiple-access schemes for local communication models. *IEEE J. Sel. Areas Commun.,* Vol. SAC-1, 935-947.

Sarkar, D., Zangwill, W.I. (1987). Expected waiting time for nonsymmetric cyclic queueing systems - Exact results and applications. To appear in *Management Science.*

Scholl, M., Kleinrock, L. (1983). On the M/G/1 queue with rest periods and certain service independent queueing disciplines. *Oper. Res.,* Vol. 31, 705-719.

Schrage, L. (1970). An alternative proof of a conservation law for the queue G/G/1. *Oper. Res.,* Vol. 18, 185-187.

Servi, L.D. (1986). Average delay approximation of M/G/1 cyclic service queues with Bernoulli schedules. *IEEE J. Sel. Areas Commun.,* Vol. SAC-4,

813-822.

Sevcik, K.C., Johnson, M.J. (1987). Cycle time properties of the FDDI token ring protocol. *IEEE Trans. Softw. Eng.*, Vol. SE-13, 376-385.

Shanthikumar, J.G. (1988). On stochastic decomposition in M/G/1 type queues with generalized server vacations. *Oper. Res.*, Vol. 36, 566-569.

Sidi, M., Levy, H. (1988). A queueing network with a single cyclically roving server. *Report Electrical Engineering Department, Technion, Haifa.*

Skinner, C.E. (1967). A priority queueing system with server-walking time. *Oper. Res.*, Vol. 15, 278-285.

Srinivasan, M.M. (1988). An approximation for mean waiting times in cyclic server systems with nonexhaustive service. *Performance Evaluation* 9, 17-33.

Strole, N.C. (1987). The IBM token-ring network - a functional overview. *IEEE Network Magazine,* Vol. 1, 23-30.

Swartz, G.B. (1980). Polling in a loop system. *J. Ass. Comput. Mach.* 27, 42-59.

Sykes, J.S. (1970). Simplified analysis of an alternating-priority queueing model with set-up times. *Oper. Res.*, Vol. 18, 1182-1192.

Szpankowski, W., Rego, V. (1988). Ultimate stability conditions for some multidimensional distributed systems. *Technical Report, Department of Computer Science, Purdue University, West Lafayette, Indiana.*

Takács, L. (1968). Two queues attended by a single server. *Oper. Res.*, Vol. 16, 639-650.

Takagi, H. (1985). Mean message waiting times in a symmetric polling system. In: *Performance '84,* ed. E. Gelenbe, North-Holland Publ. Cy., Amsterdam, 293-302.

Takagi, H. (1986). *Analysis of Polling Systems.* The MIT Press, Cambridge, Massachusetts.

Takagi, H. (1988). Queueing analysis of polling models. *ACM Comp. Surveys* 20, 5-28.

Takagi, H. (1989). Analysis of polling models with a mix of exhaustive and gated service. To appear in: *Journal of Operations Research Society of Japan,* Vol. 32.

Takagi, H., Murata, M. (1986). Queueing analysis of scan-type TDM and polling systems. In *Computer Networking and Performance Evaluation,* eds. T. Hasegawa, H. Takagi, Y. Takahashi, North-Holland Publ. Cy., Amsterdam, 199-211.

Takine, T., Takahashi, Y., Hasegawa, T. (1986). Performance analysis of a polling system with single buffers and its application to interconnected networks. *IEEE J. Sel. Areas Commun.*, Vol. SAC-4, 802-812.

Tedijanto. (1988). Exact results for the cyclic-service queue with a Bernoulli schedule. *Report Electrical Engineering Department and Systems Research Center, University of Maryland.* To appear in *Performance Evaluation.*

Tijms, H.C. (1986). *Stochastic Modelling and Analysis.* Wiley, New York.

Ulm, J.M. (1982). A timed token ring local area network and its performance characteristics. *7th Conference on Local Computer Networks, Minneapolis,* IEEE Press, 50-56.

Watson, K. S. (1985). Performance evaluation of cyclic service strategies - a survey. In: *Performance '84,* ed. E. Gelenbe, North-Holland Publ. Cy., Amsterdam, 521-533.

Welzel, T. (1987). Simulation of a multiple token ring backbone. In: *High Speed Local Area Networks,* North-Holland Publ. Cy., Amsterdam, 99-113.

Whitt, W. (1985). Approximations for the GI/G/m queue. *Technical Report, AT&T Bell Laboratories, Holmdel, NJ.*

Wolff, R.W. (1982). Poisson arrivals see time averages. *Oper. Res.,* Vol. 30, 223-231.

Yang, Q., Ghosal, D., Bhuyan, L.N. (1986). Performance analysis of multiple token ring and multiple slotted ring networks. *IEEE Trans. Comp.,* Vol. 37, 848-853.

Yue, O.C. (1987). Performance analysis of the timed token scheme in MAP. *IEEE/IEICE Global Telecommunications Conference 1987,* 187-192.

SAMENVATTING

De wachtrijtheorie houdt zich bezig met het wiskundig onderzoek naar de prestatie van een systeem dat diensten aanbiedt voor collectief gebruik. In dit proefschrift worden wachtrijsystemen bestudeerd waarin één bediende achtereenvolgens meerdere klassen van klanten bedient. Men spreekt in dit verband van 'polling' modellen. Vele systemen uit de praktijk kunnen beschreven worden als polling modellen, zoals een computer met 'multi-drop' terminals en een token ring local area netwerk.

Bij de modelbeschrijving van polling systemen wordt doorgaans van de veronderstelling uitgegaan dat de klanten van de verschillende klassen arriveren volgens onafhankelijke stochastische processen, met een zekere stochastische bedieningsvraag. Als een klant van een bepaalde klasse niet meteen kan worden bediend neemt deze plaats in de wachtrij voor die klasse. De bediende bezoekt de wachtrijen in een vaste volgorde. Tussen de bezoeken aan twee opeenvolgende wachtrijen wordt dikwijls verondersteld een zgn. *overschakeltijd* benodigd te zijn.

Als de overschakeltijden verwaarloosbaar zijn dan is onder bepaalde voorwaarden de hoeveelheid werk in het systeem onafhankelijk van de precieze bedieningsvolgorde van de klanten. Dit wordt het *principe van behoud van werk* genoemd. Het principe van behoud van werk heeft in het verleden bewezen zeer waardevol te zijn in de analyse van wachtrijsystemen met gecompliceerde bedieningsdisciplines. Wanneer de overschakeltijden *niet* verwaarloosbaar zijn, is dit principe niet meer geldig. In het proefschrift wordt aangetoond dat voor zulke systemen een natuurlijke uitbreiding van dit principe is aan te geven: een *decompositie* van de hoeveelheid aangeboden werk in het systeem. Dit resultaat leidt vervolgens tot de formulering van *pseudobehoudswetten*, exacte uitdrukkingen voor een gewogen som van de gemiddelde wachttijden in het systeem. Pseudobehoudswetten blijken zeer bruikbaar; in het proefschrift wordt in het bijzonder ingegaan op de mogelijkheid ze te gebruiken voor het construeren van benaderingen voor de individuele wachttijden. Zulke benaderingen zijn nodig omdat de wiskundige analyse van polling modellen over het algemeen zeer gecompliceerd is.

Hierna volgt een kort overzicht van de inhoud van de diverse hoofdstukken.

CHAPTER 1

Introduction and overview

In dit inleidende hoofdstuk wordt de probleemstelling beschreven, en een overzicht gegeven van resultaten uit de literatuur. De plaats van het proefschrift binnen de literatuur wordt belicht, en een vooruitblik naar de resultaten in de volgende hoofdstukken wordt gepresenteerd.

CHAPTER 2

Work decomposition: an extension of the principle of work conservation

In hoofdstuk 2 beschouwen we de hoeveelheid werk in het systeem als primaire

grootheid. Het principe van behoud van werk wordt besproken, en er wordt aangetoond dat een natuurlijke uitbreiding van dit principe is te geven voor systemen met overschakeltijden: de hoeveelheid werk in het systeem is verdeeld als de som van twee onafhankelijke stochastische variabelen, te weten 1) de hoeveelheid werk in een corresponderend systeem met dezelfde karakteristieken maar *zonder* overschakeltijden, en 2) de hoeveelheid werk in het oorspronkelijke systeem op een willekeurig moment tijdens het overschakelen. Enkele generalisaties van dit resultaat worden gegeven, en het hoofdstuk wordt afgesloten met een bespreking van decompositie resultaten in "vakantie" modellen.

CHAPTER 3
A pseudoconservation law for cyclic-service systems with switch-over times
In dit hoofdstuk wordt getoond hoe, en onder welke voorwaarden, het principe van behoud van werk leidt tot de formulering van een zgn. behoudswet, een exacte uitdrukking voor een gewogen som van de gemiddelde wachttijden. De werkdecompositie blijkt op een analoge manier aanleiding te geven tot de formulering van een heel algemene "pseudobehoudswet" voor een gewogen som van de gemiddelde wachttijden. Voor verschillende gevallen wordt een expliciete uitdrukking van deze wet gevonden. De pseudobehoudswet bevat in veel gevallen de enige exacte informatie betreffende gemiddelde wachttijden die voorhanden is; derhalve is zij van groot praktisch belang. De pseudobehoudswet kan gebruikt worden om benaderingen voor de individuele gemiddelde wachttijden te ontwikkelen of te testen, of zelfs om simulaties te testen.

CHAPTER 4
Work decomposition and pseudoconservation law for discrete-time cyclic-service systems
Alle resultaten uit de voorgaande hoofdstukken zijn voor continue tijd. Bij de over het algemeen tijd-synchrone configuratie van veel praktische communicatienetwerken is een discrete-tijd formulering meer natuurlijk. Via een limietprocedure kan de continue-tijd pseudobehoudswet afgeleid worden uit die voor de discrete-tijd formulering.

CHAPTER 5
Cyclic-service systems with a polling table
In hoofdstuk 5 wordt de bedieningsvolgorde van de wachtrijen bepaald door een "polling table". Zo'n tabel schrijft een vaste, niet noodzakelijkerwijs cyclische, volgorde voor waarin de bediende de wachtrijen bedient. Iedere klasse van klanten komt tenminste eenmaal in de tabel voor. Het gebruik van polling tabellen opent mogelijkheden voor optimalizering van de bezoekvolgorde van de bediende. Via een uitbreiding van de werkdecompositie eigenschap worden pseudobehoudswetten voor dit systeem afgeleid. Het speciale geval van polling in een stervormig netwerk wordt vergeleken met een puur cyclische strategie.

CHAPTER 6
Exact results for some two-queue models with 1-limited service at one queue
In dit hoofdstuk worden enkele modellen met twee wachtrijen met "1-limited" bediening (de bediende bedient ten hoogste één klant per bezoek) aan één van beide wachtrijen exact geanalyseerd. Het model met 1-limited bediening aan beide wachtrijen geeft aanleiding tot de formulering van een Riemann-Hilbert randwaarde probleem. Het model met "exhaustive" bediening (de bediende bedient de wachtrij tot deze leeg is) aan de andere wachtrij blijkt via een eenvoudige iteratieprocedure exact oplosbaar. De verkregen resultaten geven inzicht in het gedrag van meer algemene systemen en zijn bruikbaar bij het testen van benaderingen.

CHAPTER 7
Approximations for mean waiting times in polling systems
In veel polling modellen vormen de pseudobehoudswetten de enig beschikbare exacte informatie betreffende gemiddelde wachttijden. Het ligt voor de hand deze informatie te gebruiken om benaderingen te vinden voor de individuele wachttijden. Eerst wordt een eenvoudige procedure beschreven, die een uitdrukking voor de gemiddelde wachttijden in gesloten vorm geeft. Hierna wordt kort een iteratieve procedure besproken, die nauwkeuriger resultaten blijkt te geven; de nauwkeurigheid gaat echter ten koste van de eenvoud en helderheid van de voorgaande benadering. Uitgebreide benaderingsresultaten worden gepresenteerd en vergeleken met simulatie.

CHAPTER 8
Performance modeling and analysis of token passing Local Area Networks
In dit hoofdstuk wordt nagegaan hoe de verkregen resultaten, in het bijzonder de op de behoudswet gebaseerde benaderingen, gebruikt kunnen worden in de analyse van lokale netwerken met "multi-access based" protocollen. Dit wordt geïllustreerd aan de hand van het zgn. token-passing protocol. Het polling model zoals beschreven in het proefschrift blijkt een natuurlijk model voor zulke systemen.

CURRICULUM VITAE

De schrijver van dit proefschrift werd op 10 december 1961 geboren te Utrecht. Hij genoot middelbaar onderwijs aan het Sint Bonifacius College te Utrecht, waar hij in 1980 het Gymnasium-β diploma behaalde. Hij begon in 1980 met de studie Wiskunde aan de Rijksuniversiteit Utrecht en studeerde op 20 januari 1986 af in de Toegepaste Wiskunde. Het afstudeerwerk omvatte de analyse van een twee-dimensionaal wachtrijsysteem met behulp van methoden uit de theorie der randwaardeproblemen. Het afstudeerwerk is uitgevoerd onder de leiding van professor dr. ir. J.W. Cohen in het kader van een stage bij het Centrum voor Wiskunde en Informatica (CWI) te Amsterdam. Tijdens zijn studie volgde hij de bijvakken Informatica en Capita Selecta van de Wiskunde.

Hij trad in februari 1986 in dienst van het CWI als wetenschappelijk medewerker. In de periode van 1986 tot einde 1989 heeft hij daar het onderzoek verricht waarop dit proefschrift is gebaseerd. Het onderzoek is verricht onder de supervisie van professor dr. ir. O.J. Boxma en professor dr. ir. J.W. Cohen. In 1988 is de schrijver twee maanden op uitnodiging werkzaam geweest bij AT&T Bell Laboratories in Holmdel, New Jersey, USA.

Op 1 januari 1990 treedt hij in dienst van het Koninklijke/Shell-Laboratorium, Amsterdam, waar hij werkzaam zal als Research Mathematician in de groep Mathematics and Systems Engineering.