Sojourn times in Feedback and Processor Sharing Queues

J.L. van den Berg

Sojourn Times in Feedback and Processor Sharing Queues

Verblijftijden in wachtrijmodellen met terugkoppeling en in 'processor sharing' modellen

(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 maandag 23 april 1990 des morgens te 10.30 uur

door

Johan Leo van den Berg

geboren op 17 november 1961 te Lexmond

Promotoren:	Prof. dr. ir. O.J. Boxma
	Prof. dr. ir. J.W. Cohen

DANKWOORD

Bij het schrijven van dit proefschrift heb ik van velen hulp ondervonden. Een aantal van hen wil ik met name noemen.

Allereerst bedank ik Onno Boxma voor de begeleiding van het onderzoek en voor de aangename persoonlijke contacten; deze zijn voor mij een grote stimulans geweest. Mijn dank gaat ook uit naar prof. J.W. Cohen die het onderzoek mede begeleid heeft; door zijn colleges in Utrecht werd mijn belangstelling voor de wachtrijtheorie gewekt.

Wim Groenendijk, mijn kamergenoot, dank ik voor de vele intensieve discussies die wij gevoerd hebben.

Tenslotte dank ik het Centrum voor Wiskunde en Informatica voor de faciliteiten die mij gedurende het onderzoek beschikbaar zijn gesteld.

TABLE OF CONTENTS

1 INTRODUCTION

2 THE M/M/1 QUEUE WITH GENERAL FEEDBACK

3 THE M/G/1 PROCESSOR SHARING QUEUE AS A LIMITING MODEL OF THE M/M/1 FEEDBACK QUEUE

Chapter 1

INTRODUCTION

1.1 Background

Queueing theory is concerned with the mathematical analysis of the performance of systems that offer services for collective use, like telephone exchanges and computer networks. Due to the finite capacity of such systems queueing arises in many practical situations when too many users require access to the same facility at the same time. The mathematical study of queueing phenomena started with the investigation of telephone call congestion and delay about the beginning of this century. Later, queueing theory was successfully applied in operations research and management science, in particular for production planning; in the past decades it has become an almost indispensable tool for the performance prediction of complex computer communication systems.

A queueing model is usually described in terms of customers requiring service, service facilities providing service, and queues containing customers waiting for service. The present study is devoted to the analysis of queueing models where customers may repeatedly return to some service facility to obtain several phases of service before they finally depart from the system. Such *feedback* phenomena occur in a wide variety of processes arising in computer-communication and in production networks.

The basic queueing model representing the occurrence of feedback consists of a waiting room and a single service facility at which customers arrive according to a stochastic process; after having obtained a random amount of service a customer either returns to the queue of waiting customers to await another service or leaves the system, according to a probabilistic feedback scheme.

An example that illustrates the feedback phenomenon is found in manufacturing processes where quality control inspections are performed after the execution of an operation on a part or product, see Fig. 1.1. A part that does not meet the quality standards is sent back for reworking; this may happen several times until it finally passes the test and proceeds to the next phase of operation.

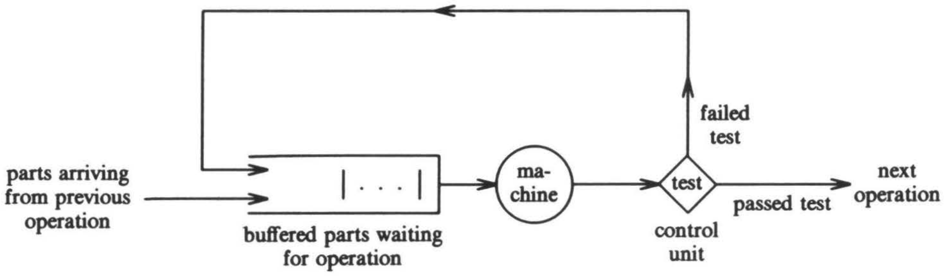
parts arriving from previous operation
buffered parts waiting for operation
ma-chine
test
failed test
passed test
next operation
control unit

Fig. 1.1 Quality control of parts during a production process.

Another important example of the occurrence of feedback is encountered in a computer system which operates in a *time sharing* mode. In Fig. 1.2 a scheme of such a time sharing system is shown. In a time shared computer system each job is allocated a small time interval for uninterrupted processing at the CPU. If the total required processing time of a job exceeds the length of this interval it is fed back to a system of queues containing waiting jobs; here the job waits until it is permitted a second turn in the processing facility, according to a certain scheduling algorithm. This procedure is repeated until the job has obtained its required processing time and leaves the system.

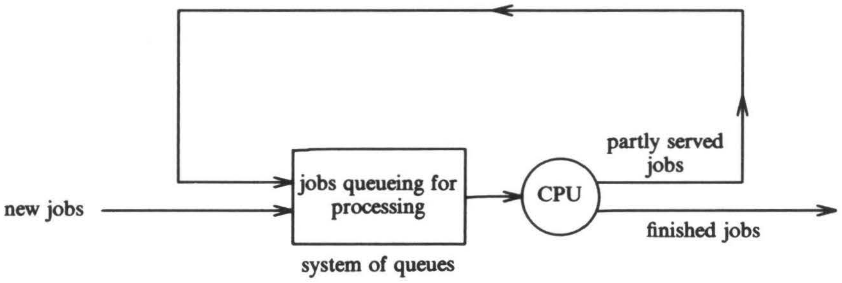

Fig. 1.2 Principle of feedback in a time shared computer system.

The introduction of time sharing systems in the early sixties and the need to determine their performance has led to an extensive study of queueing models with feedback. On the other hand, theoretical investigations concerning networks of queues have also stimulated research on feedback queues. The research of J.R. Jackson (Jackson [1957,1963]) on queueing networks with exponential services and independent external Poisson arrival processes revealed that, under certain assumptions, a queue in such a network in steady state behaves just like an M/M/s queue in isolation. In the case of so called *feedforward* networks where customers never return to a queue they have once left, one can indeed prove that the input process to each queue is a Poisson process. But when feedback is possible the input process is no longer a Poisson process - which makes the M/M/s behaviour all the more surprising. These observations revived the interest in single server queueing models with feedback of customers, like an M/M/1 queue with constant feedback probability. In particular, the stochastic aspects of customer flows in feedback queues were extensively studied (cf. the survey by Disney and König [1985]). Furthermore, the steady state queue length processes in Jackson networks and their generalizations, the so called BCMP networks (see Baskett et al. [1975]), appeared to be amenable to a detailed analysis; however, the analysis of other important characteristics such as the *waiting time* and *sojourn time* processes presented quite some difficulties. Again the case of feedforward Jackson networks was relatively simple - as long as each node contains only a single server; in this case the joint steady state distribution of the successive sojourn times of a particular customer could explicitly be obtained, cf. Lemoine [1979] and Walrand and Varaiya [1980]. But it readily became clear that the possibility of customers *overtaking* one another introduced considerable analytical difficulties. Once more the M/M/1 queue with feedback provided the simplest example to study this 'overtaking' phenomenon in isolation.

The aim of the present study is the analysis of sojourn times in single server queueing models with feedback: we shall derive the joint steady state distribution of the successive sojourn times of a customer in a feedback queue with a quite general feedback mechanism. As an important by-product, our study on feedback queues leads to new insights in the analysis of the well-known and widely used 'processor sharing' model for time sharing systems.

The remainder of this chapter is organized as follows. In Section 1.2 we shall first describe a basic feedback queueing model and discuss its main properties. Next, the central feedback model of this study is introduced. Time sharing systems and models are discussed in Section 1.3. In Section 1.4 we introduce an interesting variant of the 'standard' M/G/1 queueing model, viz. an M/G/1 queue with a fixed number of additional *permanent* customers; that are customers who reside permanently in the system, i.e. they are fed back after each service. Section 1.5 contains an extensive overview of the literature related with the models considered in the present study. Section 1.6 is concerned with assumptions about the notations and terminology used in this thesis. Finally, in Section 1.7 we give an overview of the contents of Chapters 2-5.

1.2 Queueing models with feedback

The basic, in literature most frequently encountered, feedback queueing model is the M/G/1 queue with 'Bernoulli' feedback, see Fig. 1.3. The behaviour of the customers in this model is as follows. New customers, arriving according to a Poisson process, join the end of the queue. Immediately after his service

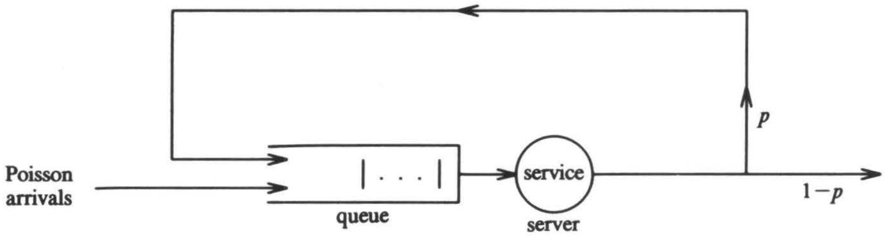

Fig. 1.3 The M/G/1 queue with Bernoulli feedback.

completion a customer returns to the end of the queue with probability p or leaves the system with probability $1-p$, $0 \leqslant p < 1$. It is assumed that all service times are independent, identically distributed, random variables. The customers in the queue are served according to the 'head-of-the-line' discipline.

It is readily seen that the M/G/1 queue with Bernoulli feedback has a stationary queue length process which has the same distribution as an equivalent M/G/1 queue *without* feedback, i.e. an M/G/1 system in which the service time distribution of a customer is equal to that of the total service time a customer obtains in the feedback model. Indeed, because the feedback probabilities are constant the queue length distribution is *independent* of the *order* in which the customers are served; so we may assume that they are served in one stretch with a service time equal to the total service time that they would have if they were served in the original manner.

A much more difficult task is the determination of the distribution of the total *sojourn time* of a customer. The problem is caused by the fact that the total sojourn time of a particular (tagged) customer is the sum of the (partial) sojourn times during his successive passes through the system, which are clearly *not* independent of each other. Moreover, for the analysis of the sojourn time of a tagged customer one has to take into account that new customers may arrive during the presence of the tagged customer and reside in the system during some passes (note that their services contribute to the tagged customer's total sojourn time), but leave the system before the tagged customer. The possibility that customers can *overtake* each other leads to dependencies which almost invariably makes the determination of the sojourn time distribution impossible (see e.g. the survey of Boxma and Daduna [1989] on sojourn times in queueing networks). For the M/G/1 queue with Bernoulli feedback, however, the sojourn time problem could be solved, see Takȧcs [1963]. Takȧcs obtained a recurrence relation for the (Laplace-Stieltjes transform and generating function of the) joint distribution of a tagged customer's total sojourn time and the number of customers present in the system after k services, $k = 1,2,\ldots$. The derivation is based on the observation that for a tagged customer the joint distribution of the duration of his $(i+1)$-

th pass through the system (also called his $(i+1)$-th sojourn time) and the number of customers present at the end of his $(i+1)$-th service is completely determined by his i-th sojourn time and the number of customers present at the end of his i-th service, $i=1,2,...$ (i.e. the joint process of successive service completion epochs and queue length at these epochs is a *Markov renewal* process). In fact, this observation is the basis of the analysis of most of the feedback models discussed in the present study.

The feedback model investigated in this thesis is actually a generalization of the M/M/1 queue with Bernoulli feedback. It is an M/M/1 feedback model in which the probability that a customer is fed back after service completion depends on the number of times he has already been served: when a customer completes his i-th service he departs from the system with probability $1-p(i)$ and he recycles with probability $p(i)$, $i=1,2,...$. Obviously, taking $p(i) \equiv p$ this model reduces to the M/M/1 queue with Bernoulli feedback. In the sequel a customer who is visiting the queue for the i-th time will be called a 'type-i customer', $i=1,2,...$.

It is important to note that the M/M/1 queue with general feedback as described above belongs to the well-known class of so called 'product form' networks (see e.g. Baskett et al. [1975] and Kelly [1979]), i.e. the (stationary) finite dimensional joint queue length distribution of the different types of customers is known and has a product form. It is noted here that due to the general feedback mechanism the distribution of the total number of customers in the system is *not* the same as the queue length distribution in the standard M/G/1 queue with service times equal to the total service time in the feedback queue, as is the case for the M/G/1 queue with Bernoulli feedback.

The main result of the present study of the M/M/1 queue with general feedback is a complete description of the *joint* distribution of the successive sojourn times of a particular customer.

1.3 Time sharing models; round robin and processor sharing queues

In Section 1.1 we already described the principle of time sharing in computer systems. Actually, the motivation for the introduction of time sharing computers has been to provide multiple users simultaneous and (almost) direct access to a single processor unit. In fact this is achieved by giving small jobs (for which the users expect small response times) preferential treatment at the expense of the longer ones (for which the users expect larger response times). It is desirable that this property is reflected by queueing models of time shared systems. Accordingly the performance measure most often used for time sharing models is the response time for a job *conditional* on its required service time. We shall discuss this performance measure for the time sharing model described below.

The M/G/1 queue with the so called 'round robin' (RR) service discipline is the most frequently encountered queueing model for the time shared computer systems as described in Section 1.1. We have pictured this single queue model in Fig. 1.4.

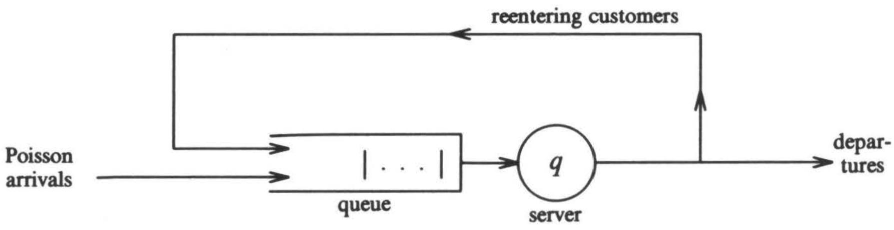

Fig. 1.4 The M/G/1 queue with round robin service.

The customers are served as follows. New customers, arriving according to a Poisson process, join the end of the queue. The customers in the queue are served according to the head-of-the-line discipline receiving a (small) fixed quantum q of service. At the end of his service quantum a customer leaves the system if his total service requirement is met; if not he returns to the end of the queue with his remaining service requirement reduced by an amount q.

To overcome the mathematical problems that arise from the analysis of the RR model with fixed positive quantum size q it is often assumed that $q \to 0$ (an idea originally due to Kleinrock [1967]). RR models under the assumption $q \to 0$ are called *processor sharing* (PS) models and are of great interest; they have pleasing mathematical properties and they also accurately model the performance of many real systems.

In queueing literature the PS service discipline is often described as follows: when there are $n \geqslant 1$ customers present in the system then each customer receives service at a rate which is $1/n$ times the rate of service that a solitary customer in the system would receive. It is easily seen that this alternative description coincides with the original definition of PS as the limiting case of the RR service discipline.

The M/G/1 PS queue has some very interesting mathematical properties. First, the mean conditional sojourn time $ES^{PS}(x)$ of a customer with service demand $x \geqslant 0$ is *linear* in x and depends only on the first moment of the service time distribution: (see Kleinrock [1967], Sakata et al. [1971])

$$ES^{PS}(x) = \frac{x}{1-\rho} ,$$

where ρ denotes the offered load to the system. This formula shows in which way the PS discipline provides preferential treatment to short jobs (customers): a job half as long as an other will spend on the average half as long in the system. The above formula for the mean conditional sojourn time in the M/G/1 PS queue should be compared with the result for the corresponding quantity in the M/G/1 first-come-first-served (FCFS) queue:

$$ES^{FCFS}(x) = \frac{\lambda\beta_2}{2(1-\rho)} + x ,$$

with λ and β_2 denoting the arrival intensity and the second moment of the service time distribution, respectively.

Another important property of the M/G/1 PS queue is that the (stationary) queue length distribution is 'insensitive' to the character of the service time distribution, apart from its first moment:

$$Pr\{k \textit{ customers in the system}\} = (1-\rho)\rho^k , \quad k=0,1,\ldots .$$

This insensitivity property also holds for the joint queue length distribution in networks of PS queues (cf. Baskett et al. [1975], Kelly [1979]).

The usefulness of the PS service discipline for modeling the performance of computer systems, and its mathematical properties, have strongly contributed to the extensive use of PS (network) models in present day performance analysis.

A very difficult mathematical problem for PS models is the determination of the sojourn time *distribution.* The difficulties are caused by the same phenomenon as occurring in the analysis of sojourn times in feedback queues: the PS service discipline allows customers to overtake each other. A solution of the sojourn time problem for the M/G/1 PS queue was first obtained by Yashkov [1983]. However, the derivation of his results is complicated and does not provide much insight into the behaviour of the main sojourn time characteristics.

In this thesis we present a new, more transparent method for the derivation of the (Laplace-Stieltjes transform of the) distribution of the sojourn time in the M/G/1 PS queue. The idea is to consider the M/G/1 PS queue as a limiting case of the M/M/1 queue with general feedback (see Section 1.2). The PS model is obtained by letting the feedback probabilities approach one and the mean service time at each loop approach zero, such that a customer's total required mean service time remains constant. Different choices of the feedback probabilities lead to different service time distributions in the PS queue. Application of this limiting procedure to the sojourn time results obtained for the M/M/1 feedback queue leads to results for the corresponding quantities in the PS queue.

The method described above exploits well-known product form results for the feedback model and gives much insight into the occurrence of many basic sojourn time properties for the limiting PS model.

Many generalizations and variants of the PS (round robin) service discipline have been introduced, see e.g. the surveys by Jaiswal [1982] and Yashkov [1987]. For most of them the sojourn time distribution problem (even for the simplest M/M/1 case) is still unsolved. For one generalization, called *'generalized processor sharing'* (GPS), we shall present in this thesis some new sojourn time results. Therefore, the GPS model will be discussed here in some more detail.

The GPS discipline generalizes the PS discipline as follows: when there are $n \geqslant 1$ customers present in the system then each customer is served with a rate equal to $f(n)$ with $f(\cdot)$ an (almost) arbitrary positive function. Obviously, for $f(n)=1/n$ the GPS discipline reduces to the PS discipline (assuming that the total capacity of the server is normalized to one).

Network models of GPS queues contain many interesting special cases such as the classical Erlang and Engset systems as well as many new processor sharing systems, see Cohen [1979]. The GPS discipline generalizes known results for classical networks: it preserves the product form and insensitivity property of the joint distribution of the queue lengths at the different nodes. At present most sojourn time results for GPS models are limited to the mean conditional sojourn time of a customer with given service demand. In general, sojourn time distributions are still unknown.

In this thesis we propose a new approach to the analysis of GPS queues. The idea is similar to that for the PS case: we consider the M/G/1 GPS queue as a limiting case of the M/M/1 queue with general feedback introduced in Section 1.2 but with *state dependent service rates.* Different choices of the service rates in the feedback model lead to different service rate functions $f(\cdot)$ in the GPS queue. We show that (known) results for the M/G/1 GPS queue can be very easily obtained from the analysis of this (product form) feedback model with state dependent service rates. (In fact, the analysis of the single node GPS model can be easily extended to the analysis of networks of GPS queues). For a special class of GPS disciplines this approach leads to new results for the sojourn time distribution.

1.4 MODELS WITH PERMANENT CUSTOMERS

An extreme case of a feedback queue is a *closed* queueing system, i.e. a queueing system with a fixed number of *'permanent'* customers. Closed queueing systems model the situation where the number of customers in the system is constant (once a customer has obtained his required service he is immediately replaced by another one with the same characteristics). A queue with additional permanent customers is a system where next to the ordinary customer stream(s) also permanent customers are processed. An interesting aspect in such a case is the interference of permanent customers with the other customer streams, in particular the influence of the presence of the permanent customers on the queueing characteristics of the other customers. In the last chapter of this thesis, Chapter 5, we consider some single queue single server models with two types of customers: (i) ordinary customers who arrive according to a Poisson process, and (ii) a fixed number of permanent customers who immediately return to the end of the queue after having received a service. See Fig. 1.5 for the case of an M/G/1 queue with permanent customers. Our main goal is to present a study of the influence of the presence of additional permanent customers on queue lengths and sojourn times of the 'Poisson customers' for the standard M/G/1 queue and for the feedback and PS models discussed in the previous sections.

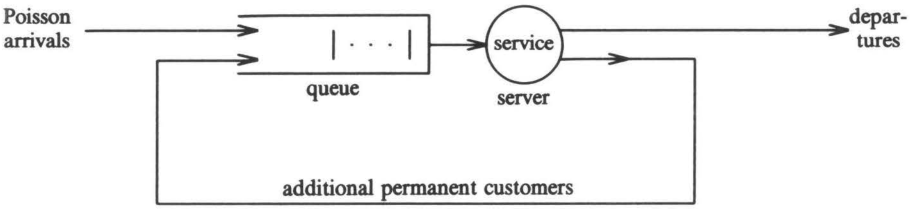

Fig. 1.5 The M/G/1 queue with additional permanent customers.

The main reason for studying these relatively simple models with permanent customers is that these models expose - stripped from all non-essential features - a structure that appears in many representations of computer and communication networks. For example, consider a telephone exchange to which two types of jobs are offered: call requests and operator tasks. To guarantee a certain quality of service to the call requests only a limited number, say K, of operator tasks (which are assumed to be always available) is allowed to be in the system at the same time. Obviously, it is important to know in which way the choice of the control parameter K influences the performance of the system with respect to the waiting times of the call requests and the throughput of the operator tasks.

Another reason for studying models with permanent customers is that there are several interesting relations with other important queueing models. For example, the M/G/1 queue with one additional permanent customer behaves exactly like a *vacation* queue, a queueing model where the server interrupts the service to a customer stream at certain epochs to take a vacation. Other related models are discussed in Chapter 5.

1.5 Review of related literature

In this section we discuss some literature related to this study and indicate the place of the study within the literature. We restrict ourself mainly to literature concerning single server models where the (external) arrival process is a Poisson process. Work on multi resource systems (networks) and finite source models is not discussed.

1.5.1 Feedback queues

A pioneering study on feedback systems is Takàcs [1963]. He considered the M/G/1 FCFS queue with Bernoulli feedback. His main result is a recurrence relation for the Laplace-Stieltjes transform (LST) and generating function of the joint distribution of a customer's total sojourn time and the number of customers present in the system after k services, $k=1,2,\ldots$. The key observation leading to this result is that for a tagged customer the joint process of successive service completion epochs and queue length at these epochs is a Markov renewal process, see Section 1.2. In fact, this observation is the basis of the analysis in most of the feedback studies discussed below.

Disney [1981], Disney et al. [1984] and Doshi and Kaufman [1988] have also studied queue length and sojourn time distributions in the M/G/1 Bernoulli feedback queue in some detail. In particular Doshi and Kaufman derive the LST of the joint distribution of the sojourn times of a customer on his successive passes through the system. Disney et al. [1980] is mainly concerned with a fundamental study of several traffic flow processes in the system and queue length distributions at different embedded stochastic epochs, see also Disney and König [1985]. They show for the M/M/1 queue with Bernoulli feedback that the input process (the successive epochs at which a customer (re)enters the queue) and the output process (the successive service completion epochs) are Markov renewal processes; for general service times the output process is also Markov renewal. It is shown that for positive feedback probabilities these processes are never renewal. Disney and König [1985] also present an overview of literature concerned with the analysis of Bernoulli feedback models.

Fontana and Diaz Berzosa [1984,1985] extend some results obtained for the M/G/1 model with Bernoulli feedback to a more general feedback model with non preemptive priorities. However, one should take care in applying their results because some of them do not agree with those in Disney et al. [1980] (e.g. Fontana and Diaz Berzosa [1984,1985] erroneously conclude that for the M/G/1 Bernoulli feedback queue the queue length distribution at output and arbitrary epochs are equal; Disney et al. [1980] prove that this does *not* hold).

Simon [1984] analyzes an M/G/1 feedback queue with multiple customer types and preemptive and non preemptive priority levels that may change after a service completion; the customers are fed back a fixed number of times. The main result of his paper is the derivation of a set of linear equations for the mean sojourn time of each visit.

The feedback model studied by Lam and Shankar [1981] is basically the same as the M/M/1 queue with general feedback analyzed in the present study. They consider the model as a time sharing model with exponentially distributed service quanta. Lam and Shankar derive the total sojourn time distribution. This distribution is a special case of our result for the joint distribution of the sojourn times of a customer on his successive cycles. Nelson [1987] also considers queues with general feedback, but with varying service times, to study the effect that assigning increasing service times to customers has on mean sojourn times.

Ali and Neuts [1984] consider an M/G/1 Bernoulli feedback queue with a waiting room and a service room. Newly arriving customers and customers who have been fed back join the waiting room. Whenever the service room becomes empty all customers from the waiting room, together with a random number of overhead customers, are transferred to it. Ali and Neuts determine the stationary distribution of queue lengths at various embedded random epochs and the distribution of a customer's waiting time until his first service.

Hunter [1989] considers single server queues with state dependent feedback and finite waiting room. In particular, he studies an appropriately constructed Markov renewal process which describes the behaviour of the system starting at the arrival of a tagged customer; the sojourn time of the tagged customer relates to a first passage time in this process. For some special cases (e.g. the M/M/1/2 queue with Bernoulli feedback) this approach leads to the derivation of explicit expressions for the LST of the distribution of the total sojourn time. Mean sojourn times are obtained for the M/M/1/N ($N \geqslant 1$) queue with Bernoulli feedback. Hunter also gives a brief survey of the literature on sojourn times in feedback models.

1.5.2 Time sharing and processor sharing queues

The first queueing model for a time shared computer system was presented by Kleinrock [1964]. He studied a single queue single server system under the round robin discipline as a discrete time Markovian model. Since then until the early seventies many papers on (variants of) this model have been published, see e.g. Schrage [1967], Kleinrock [1967], Coffman and Kleinrock [1968], Adiri and Avi-Itzhak [1969A,1969B], Sakata et al. [1971], Adiri [1972] and Muntz [1972]. All these papers are concerned with the derivation of *mean* queue lengths and sojourn times apart from Muntz [1972]. Muntz derives the queue length and sojourn time distribution for an M/M/1 queue under the RR discipline with fixed quantum size and overhead due to switching between customers. An extensive discussion on time sharing systems and models, with many references, is given by Kleinrock [1976, Chapter 4] and Jaiswal [1982].

To simplify the mathematical analysis of RR models Kleinrock [1967] proposed to study the case that the size of the service quanta shrinks to zero, thus obtaining the *processor sharing* discipline. In particular the derivation of queue length distributions and mean (conditional) sojourn times appeared to be much easier for the PS model than for the corresponding RR system with positive service quanta. Kleinrock [1967] showed that for the M/M/1 PS queue the mean conditional sojourn time is linear in the required service time (cf. the formula in Section 1.3). Sakata et al. [1971] obtained the same result for the case of general service times by letting in their RR model the quantum size tend to zero. O'Donovan [1974] and Asare and Foster [1983] derived the mean conditional sojourn time directly from the behaviour of the M/G/1 PS queue. A very recent paper on this subject is Foley and Klutke [1989]. Foley and Klutke show that the (non stationary) queue length process starting at the arrival of a tagged customer is stochastically strictly increasing during the presence of the tagged customer, which makes it quite surprising that the mean conditional sojourn time is a linear function of the required service time. Their approach, based on introducing different time scales for different processes, provides insight into this somewhat paradoxical property. In this thesis we shall give an explanation of it which is based on our 'feedback approach' to PS queues.

The derivation of the insensitivity property of the distribution of the queue length in the M/G/1 PS queue (given in Section 1.3) is originally due to Sakata et al. [1969]. In fact, they obtain their result as a special case of the queue length distribution in a multi server processor sharing model.

In contrast to the derivation of queue length distributions it appeared to be much more difficult to derive sojourn time distributions for PS queues. Coffman et al. [1970] obtained the distribution of the conditional sojourn time for the M/M/1 PS queue; for general service times it remained an unsolved problem until the early eighties. The LST of the distribution of the conditional sojourn time for the M/G/1 PS queue was obtained by Yashkov [1983], Ott [1984] and Schassberger [1984]. The approaches used by Yashkov and Ott are quite similar. The essence is a decomposition of the sojourn time of a (tagged) customer as the sum of 'time delays' which are induced by the customers present in the system at the arrival of the tagged customer (and by the tagged customer himself); these time delays include the influence of customers who arrive *during* the sojourn time of the tagged customer. It is shown that the time delays can be interpreted as lifetimes of some terminating branching process. The dynamics of the time delays is described by some integro-differential equations derived by using ideas from branching theory. Schassberger [1984] obtained his result by means of the analysis of a discrete time queue under a slight variation of the standard RR discipline in which a newly arriving customer immediately receives a quantum of service and only then joins the tail of the queue. Using known sojourn time results for this RR model (obtained in Schassberger [1981]) and letting the quantum size shrink to zero he finds results for the corresponding sojourn times in the M/G/1 PS queue. Schassberger [1984] also gives the theoretical background of the weak convergence of the sojourn time distribution for the discrete time RR model to the distribution of the sojourn time in the PS queue.

The analysis of the sojourn time distribution in the M/G/1 PS queue proposed in this thesis is in some sense similar to the method used by Schassberger [1984]. We first analyze a kind of RR (feedback) queue with exponentially distributed service quanta and then take appropriate limits such that the behaviour of the system approaches that of the M/G/1 PS queue. The advantage of this method is that it exploits well-known product form results for the feedback model which give much insight into the behaviour of the sojourn time in the limiting PS model. The proof that the distribution of the sojourn time in the feedback model converges weakly to that in the PS model has been given recently by Resing et al. [1989].

From the results obtained by Yashkov [1983], Ott [1984] and Schassberger [1984] expressions for the second and higher moments of the sojourn time distribution can be derived. The resulting formulas are very complex and their (numerical) evaluation requires quite some effort. As far as we know no attention has been paid in the processor sharing literature to the derivation of approximations or asymptotic formulas which are useful for practical evaluation, apart from a paper by Yashkov [1986]. He derived some asymptotic estimates for the conditional sojourn time variance for customers with very small or very large service demands. In this thesis simple approximations for the second moment of the sojourn time in the M/G/1 PS queue are presented. The derivations of these approximations are mainly based on new asymptotic results (e.g. heavy traffic and an extension of Yashkov's results) and on exact expressions for specific service time distributions.

A fundamental study of the generalized processor sharing service discipline is given by Cohen [1979]. Cohen studies the class of GPS disciplines in a very general model of closed and open networks with multiple customer types. This model contains as special cases the classical Erlang and Engset systems, the multi server M/G/*s* PS queue as well as many new PS systems. He obtains generalizations of known results for classical networks such as the product form and insensitivity property of the joint distribution of the queue lengths at the nodes and the mean conditional sojourn time of a customer with given service demand; sojourn time distributions are not studied.

In the present study it is shown that most of Cohen's GPS results can be obtained by using an approach similar to the one used for the analysis of the PS case, see the discussion at the end of Section 1.3. In addition, we shall derive the LST of the distribution of the sojourn time in the M/G/1 GPS queue for a special class of GPS disciplines.

Next to the GPS discipline there exist many other generalizations and variants of the PS discipline. For an overview of the various models and a discussion of the results we refer to the surveys by Jaiswal [1982] and Yashkov [1987]. Recently some studies on two interesting variants of the M/G/1 PS queue appeared, which are not covered by Jaiswal [1982] and Yashkov [1987]. Avi-Itzhak and Halfin [1988] consider an M/M/1 PS queue with a limited number (*r*) of service positions and preemption when there are at least *r* customers in the system upon the arrival of a new customer. (The case without preemption is treated in Avi-Itzhak and Halfin [1989A]). They present methods for calculating the LST and the moments of the (conditional) sojourn time distribution; it is shown that for $r \to \infty$ the results coincide with the results obtained by Coffman et al. [1970]. In Rege and Sengupta [1989] a 'gated' M/M/1 PS queue is studied. In this model a gate controlling the access to the service facility opens when the server becomes idle, admitting at most $m \geqslant 1$ waiting customers, and then closes again. Here, the case $m = 1$ corresponds to the M/M/1 FCFS queue. (Avi-Itzhak and Halfin [1989B] consider the case $m = \infty$ for general service times). The authors derive the LST of the waiting time distribution, the LST of the distribution of the time in service conditional on the required amount of service and the mean conditional sojourn time. Both PS variants can be used for modeling the performance of certain multiprogrammed (time shared) computer systems in which the degree of multiprogramming is limited to some maximum due to the constraints of finite memory.

1.6 Assumptions and notations

Throughout this study we assume that all the systems considered are stable (in statistical equilibrium), i.e. we assume that all involved stochastic processes (e.g. the queue length process and the sojourn time process) are stationary. For the systems studied in this thesis necessary and sufficient conditions for stability are well-known or else can be easily obtained from existing results for related models. So, when we refer to the 'sojourn time of an arbitrary customer' then we mean an independent copy of the sojourn time of the nth (newly) arriving customer, for some $n \geqslant 1$. Similarly, the 'queue length at an arbitrary epoch' refers to an independent copy of the queue length at some time t.

Throughout, random variables representing service times are indicated by bold printed Greek letters tau ($\boldsymbol{\tau}$); all other random variables are indicated by capitals also printed in bold type. Sections, formulas, theorems, figures, etc. are referred to by a numeral indicating the chapter in which they originally occur followed by their number within that chapter.

1.7 Overview of the contents of the next chapters

Chapter 2 is concerned with a fundamental analysis of sojourn times in the M/M/1 queue with general feedback. We first derive, in the form of Laplace-Stieltjes transforms and generating functions, a recursive expression for the joint steady state distribution of the successive sojourn times and the number of customers present in the system at each service completion of a tagged customer who has been fed back k times, $k = 0,1,\ldots$. Using this result it is shown that the successive sojourn times have the same marginal distribution, which is negative exponential. We also derive some other sojourn time characteristics, such as the distribution and the variance of the total sojourn time after k services and the correlation coefficient of the i-th and the j-th sojourn time of a tagged customer, $i < j$. In particular, we prove the intuitively appealing properties that the latter quantity is positive and that it decreases if $j-i$ grows. It is shown that for some interesting special choices of the feedback probabilities (e.g. Bernoulli feedback) the general expressions reduce to simple, explicit formulas. Chapter 2 concludes with the analysis of an M/G/1 queue with general feedback where the service time of a customer at each service depends on the number of times that he has been fed back. The results for this model are restricted to *mean* queue lengths and sojourn times. For the special case of deterministic feedback (i.e. each customer is fed back a fixed number of times) and all mean service times equal it is shown that from the second visit on the successive mean sojourn times of a tagged customer are all equal.

In Chapter 3 the sojourn times in the M/G/1 PS queue are analyzed by taking appropriate limits in the M/M/1 queue with general feedback. We first formulate the limiting procedure and then show how this procedure can be applied to the sojourn time formulas for the M/M/1 feedback model obtained in Chapter 2. It appears that some well-known results for the M/G/1 PS queue (e.g. the mean conditional sojourn time) follow immediately from the product form properties of the joint queue length distribution in the feedback model. Next to the mean sojourn time we also derive the variance of the sojourn time and the (LST of the) sojourn time distribution. In particular, a new asymptotic result for the variance of the conditional sojourn time for customers with a very small service demand is obtained. Subsequently it is pointed out how the analysis of the M/G/1 PS queue can be extended to the analysis of the M/G/1 queue with generalized processor sharing by applying a similar limiting procedure to the M/M/1 queue with general feedback and state dependent service rates. Using known product form results for the latter model we present a new, simple, derivation of the queue length distribution and mean conditional sojourn time in the M/G/1 GPS queue. The last section of Chapter 3 is devoted to the analysis of sojourn times in the M/G/1 PS queue with Bernoulli feedback. From the results for the M/M/1 queue with general feedback obtained in Chapter 2 and application of the limiting procedure we derive new results for the correlation coefficients of the successive sojourn times of a tagged customer in the PS feedback model.

In Chapter 4 we develop some simple approximation formulas for the second moment of the conditional and unconditional sojourn time in the M/G/1 PS queue. The main reason for the development of these approximations is that the exact expressions can in general only be evaluated numerically and require perfect information about the service time distribution (which is almost never available in practical situations). The approximations depend on the service time distribution only through its first and second moment. They are mainly based on new asymptotic results (e.g. heavy traffic) and on simple exact expressions for some specific service time distributions. The many numerical examples show that these simple two-moment approximations are sufficiently accurate for many practical purposes. A refinement of the approximations is obtained by taking the third moment of the service time distribution into account.

As mentioned before, Chapter 5 is concerned with the study of some queueing models with additional permanent customers. First a detailed study of the basic model, the M/G/1 queue with permanent customers, is presented. The analysis is largely based on a decomposition of the queue length as a sum of independent random variables; the distributions of these random variables are obtained from known results for a related M/G/1 model with vacations. We derive queue length and sojourn time distributions for the Poisson customers and the permanent customers and we obtain simple explicit expressions for their first moments. Next, the M/M/1 queue with general feedback and additional permanent customers is studied. We obtain the rather remarkable result that for the case with $K \geqslant 1$ additional permanent customers the sojourn time distribution is the $(K+1)$-fold convolution of the sojourn time distribution in the original system (i.e. without permanent customers). Application of the limiting procedure as described in Chapter 3 leads to a similar result for the M/G/1 PS queue with additional permanent customers.

Most results presented in this thesis are new except from that concerning the distribution of the sojourn time in the M/G/1 PS queue; actually the analysis of PS queues as presented in this study is new and our approach provides much more insight into the main sojourn time and queue length characteristics than previous methods.

The results of Chapter 2 are based on Van den Berg et al. [1989] and Van den Berg and Boxma [1989A]. The first three sections of Chapter 3 are mainly based on Van den Berg and Boxma [1989B]. Chapter 4 is based on Van den Berg [1989]. Chapter 5 is based on Van den Berg [1990].

Chapter 2

THE M/M/1 QUEUE WITH GENERAL FEEDBACK

2.1 Introduction

In this chapter we consider an M/M/1 queue with a very general feedback mechanism. When a newly arriving customer, to be called a type-1 customer, has received his service, he departs from the system with probability $1-p(1)$ and is fed back to the end of the queue with probability $p(1)$; in the latter case he becomes a type-2 customer. When he has received his i-th service, he leaves with probability $1-p(i)$ and he recycles with probability $p(i)$, in the latter case becoming a type-$(i+1)$ customer. The service times of each customer at all visits are independent, identically, negative exponentially distributed random variables. The resulting queueing model has the property that the joint queue-length distribution of type-i customers, $i=1,2,...$, is of product-form type. This property will be exploited to analyze the sojourn time process. In particular, we present a complete description of the joint distribution of the sojourn times of a customer on his successive cycles.

In the queueing literature, research on feedback queues has been mainly restricted to single server queues with Bernoulli feedback (see Takács [1963], Disney [1981], Disney et al. [1984] and Doshi and Kaufman [1988]). The Bernoulli feedback mechanism is a special case of the one in the present study: take $p(i) \equiv p$ in the general model. Lam and Shankar [1981] have studied a feedback model with basically the same feedback procedure as described above; they derive the total sojourn time distribution. This distribution comes out as a special case of our result for the joint distribution of a customer's successive sojourn times.

The organization of this chapter is as follows. In Section 2.2 the model is described in detail and some preliminary results are given. Section 2.3 contains our main result. We derive a formula for (the transform of) the joint distribution of the successive sojourn times of a tagged customer in the system and the numbers of customers of the various types present at his successive departure epochs. In Section 2.4 it is shown that the sojourn times in all individual cycles are identically, negative exponentially, distributed. Also, the correlation between the sojourn times of the j-th and k-th cycle of the tagged customer is calculated; furthermore, the distribution of the total sojourn time is derived. In Section 2.5 two special feedback mechanisms are studied:

Bernoulli feedback (Subsection 2.5.1) and deterministic feedback (Subsection 2.5.2). Finally, in Section 2.6 a similar model with *generally* distributed service times at each visit is considered. We derive a set of linear equations from which the mean sojourn time per visit can be calculated. Sections 2.2 through 2.5 are based on Van den Berg and Boxma [1989A]; Section 2.6 is based on Van den Berg et al. [1989].

2.2 Model description

We consider a single server queueing system with infinite waiting room, see Fig. 2.1. Customers arrive at the system according to a Poisson process with intensity $\lambda > 0$. After having received a service, a customer may either leave the system or be fed back. When a customer has completed his i-th service, he departs from the system with probability $1-p(i)$ and is fed back with probability $p(i)$. Fed back customers return instantaneously, joining the end of the queue. A customer who is visiting the queue for the i-th time will be called a type-i customer. To avoid the problems that occur in dealing with an infinite number of different customer types, it is assumed that after a certain number of services the feedback probabilities of a customer remain constant. Thus $p(i) = p(N) := p$, $i = N, N+1, \ldots$ for some $N \geqslant 1$. The service discipline is first-come-first-served (FCFS).

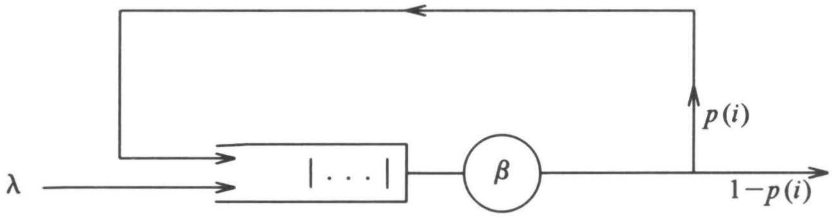

Fig. 2.1 The M/M/1 queue with general feedback.

It is assumed that the successive service times of a customer are independent, negative exponentially distributed, random variables, with mean β. These service times are also independent of the service times of other customers.
Introduce

$$q(0) := 1, \tag{2.1}$$

$$q(i) := \prod_{j=0}^{i-1} p(j), \quad i = 1, \ldots, N-1,$$

$$q(N) := \sum_{m=N}^{\infty} \prod_{j=0}^{m-1} p(j) = q(N-1)p(N-1)/(1-p),$$

with

$$p(0) := 1.$$

Note that $\lambda q(i)$ is the arrival rate of type-i customers, $i=1,...,N$. The total offered load to the queue per unit of time, denoted by ρ, is given by

$$\rho = \lambda\beta\sum_{i=1}^{N} q(i). \tag{2.2}$$

For stability it is required that $\rho<1$.

We are interested in the following steady-state quantities:

- $\mathbf{X}_i$: number of type-i customers in the system at an arbitrary epoch, $i=1,...,N$;
- $\mathbf{X}_i^{(j)}$: number of type-i customers in the system at the j-th service completion of a customer, $i=1,...,N$, $j=1,2,...$;
- $\mathbf{X}_i^{(0)}$: number of type-i customers in the system at the arrival of a new customer, $i=1,...,N$;
- $\mathbf{S}_j$: time required for the j-th pass through the system (j-th sojourn time), $j=1,2,...$;
- $\mathbf{S}^{(k)}$: total sojourn time after k services: $\mathbf{S}^{(k)} = \sum_{j=1}^{k}\mathbf{S}_j$, $k=1,2...$.

It is important to note that the system described above can be considered as a queueing network consisting of one queue with N types of customers. Type-i customers are fed back with probability $p(i)$ after service, and then change into type-$(i+1)$ customers, $i=1,...,N-1$. Type-N customers are fed back with probability p after service, and do not change their type. Because the service times are assumed to be independent exponentially distributed, the joint distribution of the number of type-i customers in the system at an arbitrary epoch, $i=1,2,...,N$, is of product-form type, see Baskett et al. [1975]. It is found that, for $x_1,...,x_N = 0,1,...$,

$$P(x_1,\ldots,x_N) := Pr\{\mathbf{X}_1=x_1,\ldots,\mathbf{X}_N=x_N\} = \tag{2.3}$$

$$(1-\rho)\,\frac{(x_1+\cdots+x_N)!}{x_1!\cdots x_N!}\prod_{i=1}^{N}(\lambda\beta q(i))^{x_i}\,.$$

It is convenient to have at our disposal the generating function of the joint queue length distribution. We have, for $|z_i|\leqslant 1$, $i=1,...,N$,

$$E\{z_1^{\mathbf{X}_1} \cdots z_N^{\mathbf{X}_N}\} = \sum_{x_1=0}^{\infty} \cdots \sum_{x_N=0}^{\infty} z_1^{\mathbf{X}_1} \cdots z_N^{\mathbf{X}_N} P(x_1, \cdots, x_N) = \tag{2.4}$$

$$(1-\rho) \sum_{m=0}^{\infty} \sum_{\substack{x_1 \\ x_1+...+x_N=m}} \cdots \sum_{x_N} \frac{m!}{x_1! \cdots x_N!} \prod_{i=1}^{N} (\lambda\beta q(i) z_i)^{x_i} =$$

$$(1-\rho) \sum_{m=0}^{\infty} \left[\sum_{i=1}^{N} \lambda\beta q(i) z_i \right]^m = \frac{1-\rho}{1-\sum_{i=1}^{N} \lambda\beta q(i) z_i}.$$

The distribution of the total number of customers in the system coincides with the queue length distribution in an ordinary M/M/1 model:

$$E\{z^{\mathbf{X}_1+...+\mathbf{X}_N}\} = \frac{1-\rho}{1-\rho z}, \quad |z| \leqslant 1,$$

i.e.

$$Pr\{\mathbf{X}_1 + \cdots + \mathbf{X}_N = j\} = (1-\rho)\rho^j, \quad j = 0,1,.... \tag{2.5}$$

We shall use these results in the next section.

2.3 Main results

In this section we present, in the form of Laplace-Stieltjes transforms and generating functions, an expression for the joint steady-state distribution of the successive sojourn times $\mathbf{S}_j$, $j = 1,...,k$, and the number of type-i customers, $\mathbf{X}_i^{(j)}$, $i = 1,...,N$, present at the j-th service completion of a customer who is fed back at least $k-1$ times, $k = 1,2,...$.

Let us follow a tagged customer from the moment he arrives as a type-1 customer until he completes his k-th service. Obviously, the k successive sojourn times of the tagged customer depend on the number of customers of each type present in the system upon his arrival, the behaviour of these customers and the behaviour of subsequent arrivals. The PASTA property (Wolff [1982]) implies the equality of the joint queue length distribution at the epoch of a new arrival and at an arbitrary epoch:

$$Pr\{\mathbf{X}_1^{(0)} = x_1, \ldots, \mathbf{X}_N^{(0)} = x_N\} = P(x_1, \ldots, x_N), \quad x_1, \ldots, x_N = 0,1,.... \tag{2.6}$$

Hence, for Re $\omega_i \geqslant 0$, $|z_{i,j}| \leqslant 1$, $i = 1,...,N$, $j = 0,...,k$,

$$E\{e^{-(\omega_1 \mathbf{S}_1 + ... + \omega_k \mathbf{S}_k)} (z_{1,0}^{\mathbf{X}_1^{(0)}} \cdots z_{N,0}^{\mathbf{X}_N^{(0)}}) \cdots (z_{1,k}^{\mathbf{X}_1^{(k)}} \cdots z_{N,k}^{\mathbf{X}_N^{(k)}})\} = \tag{2.7}$$

$$\sum_{x_1=0}^{\infty} \cdots \sum_{x_N=0}^{\infty} P(x_1, \ldots, x_N) \times$$

$$E\{e^{-(\omega_1 \mathbf{S}_1 + \ldots + \omega_k \mathbf{S}_k)} (z_{1,0}^{\mathbf{X}_1^{(0)}} \cdots z_{N,0}^{\mathbf{X}_N^{(0)}}) \cdots (z_{1,k}^{\mathbf{X}_1^{(k)}} \cdots z_{N,k}^{\mathbf{X}_N^{(k)}}) \mid \mathbf{X}_1^{(0)} = x_1, \ldots, \mathbf{X}_N^{(0)} = x_N\}.$$

The conditional expectation in the RHS of (2.7) can be evaluated by using the following property, which is easily seen to hold: $(\mathbf{X}_1^{(i+1)}, \cdots, \mathbf{X}_N^{(i+1)})$, which determines the distribution of $\mathbf{S}_{i+2}$, is conditionally independent of $\{(\mathbf{X}_1^{(j)}, \cdots, \mathbf{X}_N^{(j)}),\ j=0,...,i-1;\ \ \mathbf{S}_1, \cdots, \mathbf{S}_i\}$ given $\{(\mathbf{X}_1^{(i)}, \cdots, \mathbf{X}_N^{(i)});\ \mathbf{S}_{i+1}\}$, $i=1,...,k-1$, i.e. the joint process of successive service completion epochs and queue length vector at these service completion epochs is a *Markov renewal* process (cf. Çinlar [1975, Ch. 10]). The calculations, which are very lengthy, are omitted here; they can be found in Appendix 2.1 at the end of this chapter. There it is shown that

$$E\{e^{-(\omega_1 \mathbf{S}_1 + \ldots + \omega_k \mathbf{S}_k)} (z_{1,0}^{\mathbf{X}_1^{(0)}} \cdots z_{N,0}^{\mathbf{X}_N^{(0)}}) \cdots (z_{1,k}^{\mathbf{X}_1^{(k)}} \cdots z_{N,k}^{\mathbf{X}_N^{(k)}}) \mid \mathbf{X}_1^{(0)} = x_1, \ldots, \mathbf{X}_N^{(0)} = x_N\}$$

$$= \prod_{j=1}^{k} A_k^N(j, \omega, z) \prod_{i=1}^{N} (z_{i,0} f_k^N(i, \omega, z))^{x_i}, \tag{2.8}$$

with $\omega := (\omega_1, \ldots, \omega_k)$, $z := ((z_{1,0}, \ldots, z_{N,0}), \ldots, (z_{1,k}, \ldots, z_{N,k}))$, and

$$A_k^N(1,\omega,z) := [1+\beta\{\omega_k + \lambda(1-z_{1,k})\}]^{-1}, \tag{2.9}$$

$$A_k^N(2,\omega,z) := [1+\beta\{\omega_{k-1} + \lambda - \lambda z_{1,k-1} A_k^N(1,\omega,z)[p(1) z_{2,k} + 1 - p(1)]\}]^{-1},$$

$$\begin{aligned} A_k^N(i,\omega,z) := {} & [1+\beta\{\omega_{k-i+1} + \lambda - \lambda z_{1,k-i+1} A_k^N(i-1,\omega,z)[A_k^N(i-2,\omega,z)[\cdots \\ & [A_k^N(2,\omega,z)[A_k^N(1,\omega,z)[p(i-1) z_{i,k} + 1 - p(i-1)] \\ & p(i-2) z_{i-1,k-1} + 1 - p(i-2)] p(i-3) z_{i-2,k-2} + 1 - p(i-3)] \cdots] \\ & p(1) z_{2,k-i+2} + 1 - p(1)]\}]^{-1}, \quad i=3,...,k, \end{aligned}$$

$$\begin{aligned} f_k^N(i,\omega,z) := {} & A_k^N(k,\omega,z)[A_k^N(k-1,\omega,z)[\cdots [A_k^N(2,\omega,z)[A_k^N(1,\omega,z) \\ & [p(k+i-1) z_{k+i,k} + 1 - p(k+i-1)] p(k+i-2) z_{k+i-1,k-1} + \\ & 1 - p(k+i-2)] p(k+i-3) z_{k+i-2,k-2} + 1 - p(k+i-3)] \cdots] \\ & p(i) z_{i+1,1} + 1 - p(i)], \quad i=1,...,N. \end{aligned} \tag{2.10}$$

Here we have defined

$$z_{i,j} := z_{N,j}, \quad i = N+1,...,N+k.$$

REMARK 2.1

From the calculations in Appendix 2.1 it is seen that the factor $(z_{i,0}f_k^N(i,\omega,z))^{x_i}$ in the RHS of (2.8) is due to the contribution to $\{(\mathbf{X}_1^{(j)},\ldots,\mathbf{X}_N^{(j)}),\ j=0,\ldots,k;\ \mathbf{S}_1,\ldots,\mathbf{S}_k\}$ induced by the x_i type-i customers present in the system upon the first arrival of the tagged customer, $i=1,\ldots,N$; the factor $\prod_{j=1}^{k} A_k^N(j,\omega,z)$ is due to the contribution induced by the tagged customer himself. These contributions are all independent, cf. (2.8).

Substituting (2.8) and (2.3) into (2.7) and evaluating the summations (use (2.4)) we obtain our main result:

THEOREM 2.1

$$E\{e^{-(\omega_1\mathbf{S}_1+\ldots+\omega_k\mathbf{S}_k)}(z_{1,0}^{\mathbf{X}_1^{(0)}}\cdots z_{N,0}^{\mathbf{X}_N^{(0)}})\cdots(z_{1,k}^{\mathbf{X}_1^{(k)}}\cdots z_{N,k}^{\mathbf{X}_N^{(k)}})\} = \tag{2.11}$$

$$\frac{(1-\rho)\prod_{j=1}^{k}A_k^N(j,\omega,z)}{1-\lambda\beta\sum_{i=1}^{N}q(i)z_{i,0}f_k^N(i,\omega,z)}, \qquad \operatorname{Re}\omega_j\geqslant 0,\ \ |z_{i,j}|\leqslant 1,\ \ i=1,\ldots,N,\ j=0,\ldots,k.$$

COROLLARY 2.1

The Laplace-Stieltjes transform of the joint distribution of the first k successive sojourn times of a customer, who is fed back at least $k-1$ times, is given by

$$E\{e^{-(\omega_1\mathbf{S}_1+\ldots+\omega_k\mathbf{S}_k)}\} = \frac{(1-\rho)\prod_{j=1}^{k}A_k^N(j,\omega)}{1-\lambda\beta\sum_{i=1}^{N}q(i)f_k^N(i,\omega)}, \tag{2.12}$$

with,

$$A_k^N(1,\omega) := [1+\beta\omega_k]^{-1}, \tag{2.13}$$

$$A_k^N(2,\omega) := [1+\beta\{\omega_{k-1}+\lambda-\lambda A_k^N(1,\omega)\}]^{-1},$$

$$\begin{aligned} A_k^N(i,\omega) := {} & [1+\beta\{\omega_{k-i+1}+\lambda-\lambda A_k^N(i-1,\omega)[A_k^N(i-2,\omega)[\cdots[A_k^N(2,\omega)[A_k^N(1,\omega) \\ & p(i-2)+1-p(i-2)]p(i-3)+1-p(i-3)]\cdots]p(1)+1-p(1)]\}]^{-1}, \\ & \qquad\qquad i=3,\ldots,k, \end{aligned}$$

$$f_k^N(i,\omega) := A_k^N(k,\omega)[A_k^N(k-1,\omega)[\cdots \\ [A_k^N(2,\omega)[A_k^N(1,\omega)p(k+i-2)+1-p(k+i-2)] \\ p(k+i-3)+1-p(k+i-3)]\cdots]p(i)+1-p(i)], \qquad i=1,...,N. \tag{2.14}$$

PROOF
Substitute $z_{i,j}=1$ into (2.9)-(2.11), $i=1,...,N$, $j=0,...,k$.

COROLLARY 2.2
The joint distribution of the number of type-i customers, $i=1,...,N$, present in the queue at the end of the j-th service of a tagged customer is independent of j and given by

$$E\{z_1^{\mathbf{X}_1^{(j)}}\cdots z_N^{\mathbf{X}_N^{(j)}}\} = E\{z_1^{\mathbf{X}_1}\cdots z_N^{\mathbf{X}_N}\} = \frac{1-\rho}{1-\lambda\beta\sum_{i=1}^{N}q(i)z_i}, \tag{2.15}$$

$$|z_i|\leqslant 1,\ i=1,...,N,\ \ j=0,1,...,k.$$

PROOF
It follows from (2.3) and (2.6) that (2.15) holds for $j=0$. If (2.15) also holds for $j=1$, then it clearly holds for all $j=0,1,...$. The validity of (2.15) for $j=1$ follows by a simple calculation.

REMARK 2.2
Corollary 2.2 is, in a more general context, known as the *'arrival theorem'* for product form networks, see e.g. Walrand [1988, Section 4.4]. This theorem implies that an arriving type-*i* customer (who has just completed his $(i-1)$-th service) 'sees' the system as at an arbitrary epoch.

The Laplace-Stieltjes transform of the joint sojourn time distribution ((2.12)-(2.14)) can be presented in a form which is more suitable for obtaining sojourn time moments. For this purpose we first rewrite (2.13) and (2.14):

$$A_k^N(i,\omega) := [1+\beta\omega_{k-i+1}+\lambda\beta\{1-\tilde{q}(i-1)\prod_{j=1}^{i-1}A_k^N(j,\omega)- \\ \sum_{l=2}^{i-1}\tilde{q}(i-l)(1-p(i-l))\prod_{j=l}^{i-1}A_k^N(j,\omega)\}]^{-1}, \qquad i=1,...,k, \tag{2.16}$$

$$f_k^N(i,\omega) := \frac{1}{\tilde{q}(i)}[\tilde{q}(k+i-1)\prod_{j=1}^{k}A_k^N(j,\omega)+ \tag{2.17}$$

$$\sum_{l=2}^{k}\tilde{q}(k+i-l)(1-p(k+i-l))\prod_{j=l}^{k}A_k^N(j,\omega)], \qquad i=1,...,N,$$

with

$$\tilde{q}(0) := 1,$$

$$\tilde{q}(i) := \prod_{j=0}^{i-1}p(j), \quad i=1,2,...,$$

and an empty product being one by definition. From (2.17) it is found that the summation in the denominator of (2.12) can be written as

$$\lambda\beta\sum_{i=1}^{N}q(i)f_k^N(i,\omega) = (\rho-\lambda\beta\sum_{i=1}^{k-1}\tilde{q}(i))\prod_{j=1}^{k}A_k^N(j,\omega) + \lambda\beta\sum_{l=2}^{k}\tilde{q}(k-l+1)\prod_{j=l}^{k}A_k^N(j,\omega)\,.$$

Substituting this into (2.12) and introducing

$$M_k(i,\omega) := \prod_{j=1}^{i}\frac{1}{A_k^N(j,\omega)}, \quad i=1,...,k,$$

$$M_k(0,\omega) := 1,$$

we obtain

$$E\{e^{-(\omega_1\mathbf{S}_1+...+\omega_k\mathbf{S}_k)}\} = \tag{2.18}$$

$$\frac{1-\rho}{M_k(k,\omega) - \lambda\beta\sum_{l=2}^{k}\tilde{q}(k-l+1)M_k(l-1,\omega) - (\rho-\lambda\beta\sum_{i=1}^{k-1}\tilde{q}(i))},$$

with, from (2.16),

$$M_k(i,\omega) = (1+\beta\omega_{k-i+1})M_k(i-1,\omega)+\lambda\beta\Big[M_k(i-1,\omega)-\tilde{q}(i-1)- \tag{2.19}$$

$$\sum_{j=2}^{i-1}\tilde{q}(i-j)(1-p(i-j))\,M_k(j-1,\omega)\Big], \quad i=1,...,k.$$

Note that (2.18) and (2.19) are *independent* of N - the number of different customer types in the system. Hence, this result for the joint sojourn time distribution is also valid without the assumption made in Section 2.2 that the feedback probabilities remain constant after a finite number N of services. Moreover, it follows from (2.18) and (2.19) that the joint distribution of the first k successive sojourn times of a particular customer depends on the feedback probabilities *only* through $p(1),\ldots,p(k-3)$ and ρ (which reflects the influence of $p(k-2)$, $p(k-1),\ldots$, cf. (2.1), (2.2)). In the next section we shall use (2.18) and (2.19) to derive some important sojourn time characteristics.

2.4 Sojourn time characteristics

In this section we derive expressions for some important sojourn time characteristics such as the marginal distribution of the successive sojourn times, the correlation coefficient of the i-th and the j-th sojourn time of a particular customer, and the mean and variance of the total sojourn time after k services. As an example we study the case that a customer receives exactly 2 services; for this case simple explicit results for the above mentioned sojourn time characteristics are obtained.

The fact that the joint queue length distribution at the arrival of a customer and after each of his passes is the same (cf. Corollary 2.2), implies that the sojourn times $\mathbf{S}_j$, $j=1,\ldots,k$ have the same marginal distribution. $\mathbf{S}_1$ can easily be obtained from (2.18) and (2.19) by taking $k=1$. It is found that the sojourn times are negative exponentially distributed with mean $\beta/(1-\rho)$:

$$E\{e^{-\omega_j \mathbf{S}_j}\} = \frac{1-\rho}{1-\rho+\beta\omega_j}, \quad j=1,\ldots,k. \tag{2.20}$$

Note that this coincides with the sojourn time transform in an ordinary M/M/1 queue with mean service time β and arrival rate $\lambda\sum_{i=1}^{N} q(i)$, cf. (2.5).

In order to investigate the dependence between the i-th and j-th sojourn times we have computed the Laplace-Stieltjes transform of the joint distribution of $\mathbf{S}_i$ and $\mathbf{S}_j$, $1 \leqslant i < j \leqslant k$. It is found from (2.18) and (2.19) that

$$E\{e^{-(\omega_i \mathbf{S}_i+\omega_j \mathbf{S}_j)}\} = \frac{1-\rho}{1-\rho+\beta\omega_i+\beta\omega_j+\beta^2\omega_i\omega_j C_{j-i}}, \quad 1 \leqslant i < j \leqslant k, \tag{2.21}$$

where C_{j-i} is determined by

$$\begin{aligned} C_1 &= 1, \\ C_n &= (1+\lambda\beta)C_{n-1} - \lambda\beta \sum_{l=2}^{n-1} \tilde{q}(n-l)(1-p(n-l))C_{l-1}, \quad n=2,\ldots,k-1. \end{aligned} \tag{2.22}$$

Note that $E\{e^{-(\omega_i \mathbf{S}_i+\omega_j \mathbf{S}_j)}\}$ only depends on i and j through the difference $j-i$. This property might also have been derived from Corollary 2.2.

REMARK 2.3
It was pointed out by Prof. J.W. Cohen that the two-dimensional Laplace-Stieltjes transform given by (2.21) is of a type for which the corresponding joint probability density function, $f_{i,j}(\cdot\,,\cdot)$, is known. From the formula given in entry 8 of Table B in Voelker and Doetsch [1950, p. 208] it is found that, for $1 \leqslant i < j \leqslant k$,

$$f_{i,j}(x,y) = \tag{2.23}$$

$$\frac{1-\rho}{\beta^2 C_{j-i}} e^{-(x+y)/(\beta C_{j-i})} \sum_{m=0}^{\infty} \left[\frac{-xy}{\beta^2 C_{j-i}} \right]^m (1-\rho+1/C_{j-i})^m (1/m!)^2, \quad x,y \geqslant 0.$$

From (2.21) the correlation coefficient, $corr(\mathbf{S}_i,\mathbf{S}_j)$, can easily be obtained:

$$corr(\mathbf{S}_i,\mathbf{S}_j) = 1 - C_{j-i}(1-\rho), \quad 1 \leqslant i < j \leqslant k. \tag{2.24}$$

It follows from (2.22) and (2.24) that $corr(\mathbf{S}_i,\mathbf{S}_j)$ as a function of i and j only depends on $j-i$. Noting that in (2.22) $\sum_{l=2}^{n-1}\tilde{q}(n-l)(1-p(n-l)) \leqslant 1$ (remember that $\tilde{q}(n-l)(1-p(n-l))$ is the probability that a customer receives exactly $n-l$ services) it follows by induction that the row $\{C_n,\ n=1,2,...\}$ is monotonically increasing. Hence, from (2.24), $corr(\mathbf{S}_i,\mathbf{S}_j)$ decreases if $j-i$ grows. In particular it can be proven that $\lim_{n\to\infty} C_n = 1/(1-\rho)$, see Chapter 3, yielding $\lim_{j-i\to\infty} corr(\mathbf{S}_i,\mathbf{S}_j) = 0$. For $j-i=1$, $corr(\mathbf{S}_i,\mathbf{S}_j) = \rho$. So, the successive sojourn times of a tagged customer are always correlated positively.

The Laplace-Stieltjes transform of the distribution of a customer's total time spent in the system until the end of his k-th pass $\mathbf{S}^{(k)} := \mathbf{S}_1 + ... + \mathbf{S}_k$, can be obtained from (2.18) by substituting $\omega_j = \omega_0$, $j=1,...,k$. From (2.20) it follows immediately that $E\{\mathbf{S}^{(k)}\}$ is linear in k:

$$E\{\mathbf{S}^{(k)}\} = \sum_{i=1}^{k} E\{\mathbf{S}_i\} = k\frac{\beta}{1-\rho}. \tag{2.25}$$

To derive an expression for the variance of this sojourn time, $var(\mathbf{S}^{(k)})$, it is convenient to use the formula

$$var(\mathbf{S}^{(k)}) = \sum_{i=1}^{k} var(\mathbf{S}_i) + 2\sum_{i=1}^{k}\sum_{j=i+1}^{k} cov(\mathbf{S}_i,\mathbf{S}_j).$$

Hence, from (2.20),

$$var(\mathbf{S}^{(k)}) = k\ var(\mathbf{S}_1)+2\sum_{i=1}^{k}\sum_{j=i+1}^{k} cov(\mathbf{S}_i,\mathbf{S}_j).$$

The covariance of $\mathbf{S}_i$ and $\mathbf{S}_j$, $cov(\mathbf{S}_i,\mathbf{S}_j)$, and $var(\mathbf{S}_1)$ can easily be obtained from the results (2.20) and (2.24). It is found that

$$var(\mathbf{S}^{(k)}) = \left[\frac{\beta}{1-\rho}\right]^2 \left[k^2-2(1-\rho)\sum_{j=1}^{k-1} jC_{k-j}\right], \tag{2.26}$$

with $C_1, \ldots, C_{k-1}$ given by (2.22).

The Laplace-Stieltjes transform of the distribution of the total sojourn time $\mathbf{S}$ of an arbitrary customer is now given by

$$E\{e^{-\omega_0\mathbf{S}}\} = \sum_{k=1}^{\infty}\tilde{q}(k)(1-p(k))E\{e^{-\omega_0\mathbf{S}^{(k)}}\}. \tag{2.27}$$

In an example, we shall examine the case $k=2$ for which explicit closed form results can easily be obtained.

EXAMPLE 2.1 (*The case $k=2$*)
From (2.21) and (2.22) it follows that for the case $k=2$,

$$E\{e^{-(\omega_1\mathbf{S}_1+\omega_2\mathbf{S}_2)}\} = \frac{1-\rho}{1-\rho+\beta\omega_1+\beta\omega_2+\beta^2\omega_1\omega_2}. \tag{2.28}$$

Note that the feedback probabilities $p(i)$, $i=1,...,N$, enter into the joint distribution of $\mathbf{S}_1$ and $\mathbf{S}_2$ only via the offered load ρ. *Thus, as long as ρ remains constant, the joint distribution of $\mathbf{S}_1$ and $\mathbf{S}_2$ is independent of the individual values of $p(i)$, $i=1,...,N$.* (Consequently, this also holds for $\mathbf{S}_i$ and $\mathbf{S}_{i+1}$, $i=1,2,...$, cf. (2.21)). Doshi and Kaufman [1988] derived (2.28) for the (special) case of Bernoulli feedback ($p(i)\equiv p$).

From (2.24) it follows that

$$corr(\mathbf{S}_1,\mathbf{S}_2) = \rho. \tag{2.29}$$

Let $F_2(t)$ denote the distribution function of the sojourn time until the end of the second pass:

$$F_2(t) := Pr\{\mathbf{S}_1+\mathbf{S}_2<t\}, \quad t\geqslant 0.$$

From (2.28) we find

$$E\{e^{-\omega_0(\mathbf{S}_1+\mathbf{S}_2)}\} = \frac{1+\sqrt{\rho}}{2\sqrt{\rho}}\frac{1-\sqrt{\rho}}{1-\sqrt{\rho}+\beta\omega_0} - \frac{1-\sqrt{\rho}}{2\sqrt{\rho}}\frac{1+\sqrt{\rho}}{1+\sqrt{\rho}+\beta\omega_0}.$$

Hence

$$F_2(t) = \frac{1+\sqrt{\rho}}{2\sqrt{\rho}}(1-e^{-t(1-\sqrt{\rho})/\beta}) - \frac{1-\sqrt{\rho}}{2\sqrt{\rho}}(1-e^{-t(1+\sqrt{\rho})/\beta}), \quad t\geqslant 0. \quad (2.30)$$

In Doshi and Kaufman [1988], $F_2(\cdot)$ is compared with the distribution of $\mathbf{S}_1+\mathbf{S}_2$ that results when one assumes that $\mathbf{S}_1$ and $\mathbf{S}_2$ are independent. Due to the positive correlation between $\mathbf{S}_1$ and $\mathbf{S}_2$ (cf. (2.29)), it is found that $F_2(\cdot)$ has a longer "tail" than this approximate distribution.

Finally, the variance of $\mathbf{S}_1+\mathbf{S}_2$ is obtained from (2.26):

$$var(\mathbf{S}_1+\mathbf{S}_2) = 2(1+\rho)\left[\frac{\beta}{1-\rho}\right]^2. \quad (2.31)$$

2.5 Special cases: Bernoulli feedback and deterministic feedback

In this section we study two feedback systems which are special cases of the general model described in Section 2.2, viz., Bernoulli feedback (Subsection 2.5.1) and deterministic feedback (Subsection 2.5.2). For these models we obtain simple, explicit expressions for most of the quantities analyzed in Section 2.3. The results for the deterministic feedback model have been published before in Van den Berg et al. [1989]. The Laplace-Stieltjes transform of the joint sojourn time distribution in the Bernoulli feedback model has also been derived by Doshi and Kaufman [1988].

2.5.1 Bernoulli feedback

The Bernoulli feedback model is obtained from the general model by taking $p(i)\equiv p$: when a customer completes his service he departs from the system with probability $1-p$ and is fed back with probability p.
Obviously

$$\rho = \frac{\lambda\beta}{1-p}.$$

The Laplace-Stieltjes transform of the joint distribution of the successive sojourn times $\mathbf{S}_1,\ldots,\mathbf{S}_k$ can be obtained from (2.12)-(2.14) (or from (2.18) and (2.19)) by substituting $p(i)\equiv p$. The expression that results from (2.12)-

(2.14) has also been derived by Doshi and Kaufman [1988].

To obtain explicit expressions for $E\{e^{-(\omega_i\mathbf{S}_i+\omega_j\mathbf{S}_j)}\}$, $corr(\mathbf{S}_i,\mathbf{S}_j)$, and $var(\mathbf{S}^{(k)})$, (see (2.21), (2.24) and (2.26)) we have to derive C_n, $n=1,...,k-1$, from the set of difference equations (2.22). After the substitution $\tilde{q}(j)=p^{j-1}$, $j=1,2,...$, (2.22) reduces to

$$\begin{aligned} C_1 &= 1, \qquad (2.32)\\ C_n &= (1+\lambda\beta)C_{n-1}-\lambda\beta\sum_{l=2}^{n-1}(p^{n-l-1}-p^{n-l})C_{l-1}, \quad n=2,...,k-1. \end{aligned}$$

From (2.32) it follows that

$$\begin{aligned} C_1 &= 1,\\ C_2 &= 1+\lambda\beta,\\ C_n-pC_{n-1} &= (1+\lambda\beta)C_{n-1}-p(1+\lambda\beta)C_{n-2}-\lambda\beta(1-p)C_{n-2},\\ &\qquad n=3,...,k-1. \end{aligned}$$

Hence

$$\begin{aligned} C_1 &= 1, \qquad (2.33)\\ C_2 &= 1+\lambda\beta,\\ C_n &= (1+\lambda\beta+p)C_{n-1}-(\lambda\beta+p)C_{n-2}, \quad n=3,...,k-1. \end{aligned}$$

The general solution of (2.33) is given by

$$C_n = U_1y_1^n+U_2y_2^n,$$

where $y_1=1$ and $y_2=\lambda\beta+p$ are the roots of

$$y^2-(1+\lambda\beta+p)y+(\lambda\beta+p)=0,$$

and U_1 and U_2 are determined by

$$\begin{aligned} &U_1y_1+U_2y_2=1,\\ &U_1y_1^2+U_2y_2^2=1+\lambda\beta. \end{aligned}$$

After some calculations it is found that

$$C_n = \frac{1-\rho(\lambda\beta+p)^{n-1}}{1-\rho}, \quad n=1,...,k-1. \qquad (2.34)$$

Substitution of (2.34) in (2.21), (2.24) and (2.26) yields, respectively

$$E\{e^{-(\omega_i\mathbf{S}_i+\omega_j\mathbf{S}_j)}\} = \frac{1-\rho}{1-\rho+\beta\omega_i+\beta\omega_j+\beta^2\omega_i\omega_j\dfrac{1-\rho(\lambda\beta+p)^{j-i-1}}{1-\rho}}, \tag{2.35}$$

$$1\leqslant i<j\leqslant k, \quad \text{Re }\omega_i,\omega_j\geqslant 0,$$

$$corr(\mathbf{S}_i,\mathbf{S}_j) = \rho(\lambda\beta+p)^{j-i-1}, \quad 1\leqslant i<j\leqslant k, \tag{2.36}$$

$$var(\mathbf{S}^{(k)}) = \tag{2.37}$$

$$\left[\frac{\beta}{1-\rho}\right]^2\left[k+2\frac{\rho}{1-p}\left\{\frac{k}{1-\rho}-\frac{1}{1-p}\frac{1}{(1-\rho)^2}(1-(p+\lambda\beta)^k)\right\}\right],$$

$$k=1,2,\dots .$$

It follows from (2.36) that $\lim\limits_{j-i\uparrow\infty} corr(\mathbf{S}_i,\mathbf{S}_j)=0$ (cf. Section 2.3). It is also seen that $corr(\mathbf{S}_i,\mathbf{S}_j)$ is an increasing function of $\lambda\beta$ for fixed i and j. These intuitively appealing properties are illustrated in Fig. 2.2.

The Laplace-Stieltjes transform of the distribution of $\mathbf{S}^{(k)}$ can be obtained from (2.18) and (2.19) by substituting $\omega_j=\omega_0,\ j=1,\dots,k$. The resulting set of difference equations (2.19) can be solved in the same way as (2.33). After extensive but straightforward calculations it is found that

$$E\{e^{-\omega_0\mathbf{S}^{(k)}}\} = \frac{1-p-\lambda\beta}{Q_1(1-x_2)x_1^k+Q_2(1-x_1)x_2^k}, \quad \text{Re }\omega_0\geqslant 0, \quad k=1,2,\dots, \tag{2.38}$$

where,

$$x_1 = \frac{1+\beta\omega_0+\lambda\beta+p+\sqrt{(1+\beta\omega_0+\lambda\beta+p)^2-4(p+p\beta\omega_0+\lambda\beta)}}{2},$$

$$x_2 = \frac{1+\beta\omega_0+\lambda\beta+p-\sqrt{(1+\beta\omega_0+\lambda\beta+p)^2-4(p+p\beta\omega_0+\lambda\beta)}}{2},$$

$$Q_1 = \frac{x_2-(1+\beta\omega_0)}{x_2-x_1},$$

$$Q_2 = \frac{x_1-(1+\beta\omega_0)}{x_1-x_2}.$$

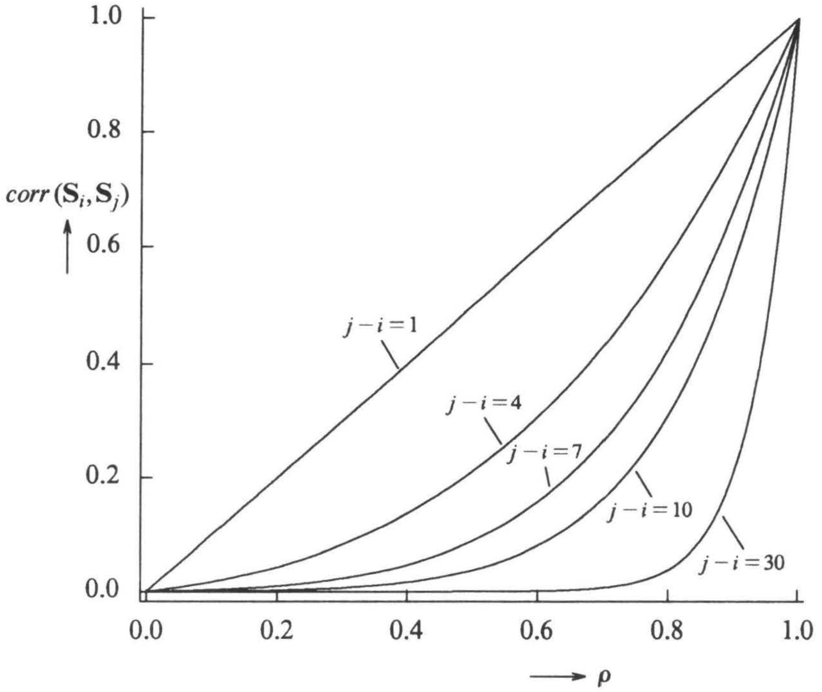

Fig. 2.2 $corr(\mathbf{S}_i,\mathbf{S}_j)$ as a function of offered load $\rho=\dfrac{\lambda\beta}{1-p}$, with $p=0.5$.

2.5.2 Deterministic feedback

Taking $p(i)=1$, $i=1,...,N-1$, $p(N)=p=0$, we obtain the deterministic feedback model in which each customer is fed back exactly $N-1$ times and leaves the system after the N-th service.

Obviously

$$\rho = N\lambda\beta.$$

Noting that

$$\begin{aligned} q(j) = \tilde{q}(j) &= 1, \quad 0 \leqslant j \leqslant N, \\ &= 0, \quad j > N, \end{aligned}$$

it is easily seen from (2.18) and (2.19) that

$$E\{e^{-(\omega_1\mathbf{S}_1+...+\omega_k\mathbf{S}_k)}\} = \frac{1-\rho}{M_k(k,\omega) - \lambda\beta \sum_{i=0}^{k-1} M_k(i,\omega) - (N-k)\lambda\beta}, \tag{2.39}$$

with

$$M_k(0,\omega) = 1, \tag{2.40}$$

$$M_k(i,\omega) = (1+\lambda\beta+\beta\omega_{k-i+1})M_k(i-1,\omega)-\lambda\beta, \quad i=1,...,k, \quad k \leqslant N.$$

At the end of this subsection we shall use (2.39) and (2.40) to obtain an explicit expression for the Laplace-Stieltjes transform of the total sojourn time distribution.

As in Subsection 2.5.1 we solve the set of difference equations (2.22) to obtain explicit expressions for $E\{e^{-(\omega_i \mathbf{S}_i + \omega_j \mathbf{S}_j)}\}$, $corr(\mathbf{S}_i,\mathbf{S}_j)$ and $var(\mathbf{S}^{(k)})$ from the general formulas (2.21), (2.24) and (2.26). Substituting in (2.22) $p(i)=\tilde{q}(i)=1$, $i=1,...,N-1$, we get

$$C_1 = 1,$$

$$C_n = (1+\lambda\beta)C_{n-1}, \quad n=2,...,N-1.$$

Hence

$$C_n = (1+\lambda\beta)^{n-1}, \quad n=1,...,N-1.$$

Now it follows from (2.21), (2.24) and (2.26) that

$$E\{e^{-(\omega_i \mathbf{S}_i + \omega_j \mathbf{S}_j)}\} = \frac{1-\rho}{1-\rho+\beta\omega_i+\beta\omega_j+\beta^2\omega_i\omega_j(1+\lambda\beta)^{j-i-1}}, \tag{2.41}$$

$$1 \leqslant i < j \leqslant N,$$

$$corr(\mathbf{S}_i,\mathbf{S}_j) = 1-(1-\rho)(1+\lambda\beta)^{j-i-1}, \quad 1 \leqslant i < j \leqslant N, \tag{2.42}$$

$$var(\mathbf{S}^{(k)}) = \left[\frac{\beta}{1-\rho}\right]^2 \left[k^2-2(1-\rho)\frac{(1+\lambda\beta)^k-k\lambda\beta-1}{(\lambda\beta)^2}\right], \quad k \leqslant N. \tag{2.43}$$

The Laplace-Stieltjes transform of the distribution of the total sojourn time after k services is obtained from (2.39) and (2.40) by substituting $\omega_j=\omega_0$, $j=1,...,k$. The resulting set of relations yields

$$M_k(i,\omega) = (1+\lambda\beta+\beta\omega_0)^i - \lambda\beta\sum_{j=0}^{i-1}(1+\lambda\beta+\beta\omega_0)^j = \tag{2.44}$$

$$\frac{\omega_0}{\lambda+\omega_0}(1+\lambda\beta+\beta\omega_0)^i + \frac{\lambda}{\lambda+\omega_0}, \quad i=0,...,k, \quad k \leqslant N.$$

Using (2.44) it follows from (2.39) that

$$E\{e^{-\omega_0 \mathbf{S}^{(k)}}\} = \tag{2.45}$$

$$\frac{(1-\rho)(\lambda+\omega_0)^2}{\omega_0^2(1+\lambda\beta+\beta\omega_0)^k+\lambda(\lambda+\omega_0)(1-\rho-(N-k)\beta\omega_0)+\lambda\omega_0}, \quad \mathrm{Re}\,\omega_0 \geqslant 0, \quad k \leqslant N.$$

Remark 2.4

The simple explicit formula (2.45) for the LST of the total sojourn time distribution has been obtained under the assumption that $k \leqslant N$. Unfortunately, this result can not be extended to the sojourn time of a (special) customer who is fed back $k > N$ times. The problem is that, for $k > N$, the set of difference equations (2.19) for the $M_k(i,\omega)$'s can not be explicitly solved (cf. (2.44) for the case $k \leqslant N$). The same holds for the solution of the C_n's from (2.22) for the analysis of the sojourn time variance.

2.6 Further extensions

In this section we consider the feedback model described in Section 2.2 with the following extension: the successive service times of a customer are generally distributed and may depend on the number of times he has already been fed back. We also omit the assumption made in Section 2.2 that the feedback probabilities remain constant after a finite number (N) of services (cf. the discussion below (2.19)). For this extended model the joint stationary distribution of the number of type-i customers in the system is no longer of product form type. In fact no results concerning the distribution of the queue length are available. Consequently, it can not be expected that we are able to obtain sojourn time distributions. Therefore, in this section, we restrict ourself to the derivation of *mean* queue lengths and *mean* sojourn times. First, as in Simon [1984] (cf. Subsection 1.5.1), we derive a set of linear equations from which the mean sojourn time per visit can be calculated. Next, we show that for the special case of deterministic feedback with all mean service times equal (but not necessarily negative exponentially distributed), this set of linear equations can be solved explicitly. It appears that from the second visit on, all mean sojourn times are equal. Finally, explicit results are obtained for the case of Bernoulli feedback.

2.6.1 Derivation of a set of linear equations

We consider the case that the service time distribution of a customer who has been fed back $i-1$ times is given by $B_i(\cdot)$, with mean β_i and second moment $\beta_i^{(2)}$, $i=1,2,\ldots$. The definitions of type-i customers and their characteristic quantities, as given in Section 2.2, are extended in an obvious way. Denote by

$\rho_i := \lambda q(i)\beta_i$ the offered traffic due to type-i customers. Obviously the stability condition for this system is that $\rho = \sum_{i=1}^{\infty}\rho_i < 1$. We start by obtaining a relation for $E\mathbf{S}_1$. Note that a newly arriving customer is a Poisson arrival and hence PASTA (see Wolff [1982]) applies. Consider the mean amount of work that has to be handled before this newly arriving customer (in the following: the tagged customer) receives his first service. This quantity consists of two components:

1. the mean amount of waiting work found upon his arrival that is handled before his first service, given by: $\sum_{i=1}^{\infty}\beta_i E\mathbf{X}_i^w$;
2. the mean amount of work currently in service: $\sum_{i=1}^{\infty}\rho_i \frac{\beta_i^{(2)}}{2\beta_i}$;

where $\mathbf{X}_i^w$ denotes the number of *waiting* type-i customers. The expression for the second component follows by noting that, at an arbitrary epoch, a type-i customer is being served with probability ρ_i, while his residual service time has mean $\beta_i^{(2)}/2\beta_i$. It may now be seen that,

$$E\mathbf{S}_1 = \sum_{i=1}^{\infty}\beta_i E\mathbf{X}_i^w + \sum_{i=1}^{\infty}\rho_i \frac{\beta_i^{(2)}}{2\beta_i} + \beta_1 .$$

With $E\mathbf{X}_i^w = E\mathbf{X}_i - \rho_i$ we obtain:

$$E\mathbf{S}_1 = \sum_{j=1}^{\infty}\beta_j E\mathbf{X}_j + \sum_{j=1}^{\infty}(\frac{\beta_j^{(2)}}{2\beta_j} - \beta_j)\rho_j + \beta_1 . \qquad (2.46)$$

$E\mathbf{S}_{i+1}$ is composed of mean service times of "old" customers (customers who were already present at the first arrival of the tagged customer) and of customers who have arrived during the first i sojourn times. It is easily seen that the mean number of old type-j customers still present in the queue (as type-$j+i$ customers) immediately after the i-th service of the tagged customer is given by $\frac{q(j+i)}{q(j)}E\mathbf{X}_j$. The mean number of customers that arrived during the tagged customer's j-th sojourn time and that are still present (as type-$(i-j+1)$ customers) at the end of his i-th service is $\lambda q(i-j+1)E\mathbf{S}_j$. Hence

$$E\mathbf{S}_{i+1} = \sum_{j=1}^{\infty}\frac{q(j+i)}{q(j)}\beta_{j+i}E\mathbf{X}_j + \lambda\sum_{j=1}^{i}q(i-j+1)\beta_{i-j+1}E\mathbf{S}_j + \beta_{i+1} , \qquad (2.47)$$

$$i = 1,2,\ldots .$$

The mean number of type-j customers in the system and the j-th sojourn time can be related to each other by Little's formula (see e.g. Kleinrock [1975]):

$$E\mathbf{X}_j = \lambda q(j)E\mathbf{S}_j .$$

Substituting this into (2.46) and (2.47) leads to

$$E\mathbf{S}_1 = \sum_{j=1}^{\infty}\rho_j E\mathbf{S}_j + \sum_{j=1}^{\infty}(\frac{\beta_j^{(2)}}{2\beta_j} - \beta_j)\rho_j + \beta_1 . \tag{2.48}$$

$$E\mathbf{S}_{i+1} = \sum_{j=1}^{\infty}\rho_{j+i}E\mathbf{S}_j + \sum_{j=1}^{i}\rho_{i-j+1}E\mathbf{S}_j + \beta_{i+1} , \quad i=1,2,\ldots . \tag{2.49}$$

Formulas (2.48) and (2.49) represent an infinite set of linear equations in $E\mathbf{S}_1$, $E\mathbf{S}_2, \ldots$. For some special cases this set of equations can be easily solved explicitly. In the next subsections we shall consider two cases which yield interesting results for the successive mean sojourn times of a customer.

2.6.2 Special case: M/G/1 queue with deterministic feedback

In this subsection we assume that the customers require exactly N services and that all service time distributions are the same, i.e. $B_i(.) \equiv B(.)$ and $p(i)=1$, $i=1,\ldots,N-1$, $p(N)=0$, in the general model. Let $\mathbf{S}$ denote the total sojourn time after N services. The equations (2.48) and (2.49) now become:

$$E\mathbf{S}_1 = \lambda\beta E\mathbf{S} + \frac{\lambda}{2}N(\beta^{(2)} - 2\beta^2) + \beta, \tag{2.50}$$

$$E\mathbf{S}_{i+1} = \lambda\beta\sum_{j=1}^{N-i}E\mathbf{S}_j + \lambda\beta\sum_{j=1}^{i}E\mathbf{S}_j + \beta, \quad i=1,\ldots,N-1. \tag{2.51}$$

Due to the symmetry in (2.51) we have that

$$E\mathbf{S}_{i+1} = E\mathbf{S}_{N-i+1}, \quad i=1,\ldots,N-1.$$

Subtracting $E\mathbf{S}_i$ from $E\mathbf{S}_{i+1}$, we obtain

$$E\mathbf{S}_{i+1} - E\mathbf{S}_i = -\lambda\beta E\mathbf{S}_{N-i+1} + \lambda\beta E\mathbf{S}_i = -\lambda\beta(E\mathbf{S}_{i+1} - E\mathbf{S}_i), \quad i=2,\ldots,N-1.$$

Hence, $E\mathbf{S}_i = E\mathbf{S}_{i+1}$, $i=2,\ldots,N-1$, and, interestingly, we have

$$E\mathbf{S}_2 = E\mathbf{S}_3 = \cdots = E\mathbf{S}_N. \tag{2.52}$$

Now from (2.50) and (2.52):

$$\boldsymbol{ES}_1 = \lambda\beta \boldsymbol{ES}_1 + (N-1)\lambda\beta \boldsymbol{ES}_2 + \frac{\lambda}{2}N(\beta^{(2)}-2\beta^2) + \beta. \tag{2.53}$$

And from (2.51) and (2.52):

$$\boldsymbol{ES}_2 = 2\lambda\beta \boldsymbol{ES}_1 + (N-2)\lambda\beta \boldsymbol{ES}_2 + \beta. \tag{2.54}$$

Solving equations (2.53) and (2.54) yields:

$$\boldsymbol{ES}_1 = \frac{\beta}{1-N\lambda\beta} + \frac{(1-(N-2)\lambda\beta)\frac{\lambda}{2}N(\beta^{(2)}-2\beta^2)}{(1+\lambda\beta)(1-N\lambda\beta)}, \tag{2.55}$$

$$\boldsymbol{ES}_2 = \boldsymbol{ES}_3 = \cdots = \boldsymbol{ES}_N = \frac{\beta}{1-N\lambda\beta} + \frac{\lambda^2\beta N(\beta^{(2)}-2\beta^2)}{(1+\lambda\beta)(1-N\lambda\beta)}. \tag{2.56}$$

Hence

$$\boldsymbol{ES} = \frac{N\beta}{1-N\lambda\beta} + \frac{1+N\lambda\beta}{1-N\lambda\beta}\frac{\lambda}{2}\frac{N(\beta^{(2)}-2\beta^2)}{1+\lambda\beta}. \tag{2.57}$$

For $N=1$ (2.57) gives the well-known formula for the mean sojourn time in the standard M/G/1 queue, (see e.g. Cohen [1982])

$$\boldsymbol{ES} = \frac{\lambda\beta^{(2)}}{2(1-\lambda\beta)} + \beta, \tag{2.58}$$

as could be expected.
Observe from (2.55) and (2.56) that $\boldsymbol{ES}_1 = \boldsymbol{ES}_2$ if the service times are negative exponentially distributed (i.e. $\beta^{(2)}=2\beta^2$), cf. Section 2.3. Noting that $(1-(N-2)\lambda\beta)\lambda N/2 > \lambda^2\beta N$ (use $N\lambda\beta=\rho<1$) it follows from (2.55) and (2.56) that $\boldsymbol{ES}_1<\boldsymbol{ES}_2$ if $\beta^{(2)}<2\beta^2$ and $\boldsymbol{ES}_1>\boldsymbol{ES}_2$ if $\beta^{(2)}>2\beta^2$.

REMARK 2.5
Note that, in fact, for the derivation of (2.52) it suffices to assume that all mean sojourn times are equal, i.e. $\beta_i=\beta$, $i=1,...,N$.

2.6.3 Special case: M/G/1 queue with Bernoulli feedback

In this subsection we consider the M/G/1 queue with Bernoulli feedback, i.e. $B_i(\cdot)\equiv B(\cdot)$ and $p(i)\equiv p$ in the general model. For this case the set of equations (2.48) and (2.49) reads

$$E\mathbf{S}_1 = \lambda\beta\sum_{j=1}^{\infty} p^{j-1} E\mathbf{S}_j + \frac{\lambda}{2}(\beta^{(2)} - 2\beta^2)\sum_{j=1}^{\infty} p^{j-1} + \beta . \tag{2.59}$$

$$E\mathbf{S}_{i+1} = \lambda\beta\sum_{j=1}^{\infty} p^{j+i-1} E\mathbf{S}_j + \lambda\beta\sum_{j=1}^{i} p^{i-j} E\mathbf{S}_j + \beta , \quad i=1,2,\dots . \tag{2.60}$$

Introducing

$$M := \frac{\lambda}{2}(\beta^{(2)} - 2\beta^2)\frac{1}{1-p},$$

$$M_j := \frac{1}{M}\left[E\mathbf{S}_j - \frac{\beta}{1-\lambda\beta/(1-p)}\right], \quad j=1,2,\dots,$$

we can rewrite (2.59) and (2.60) into

$$M_1 = \lambda\beta\sum_{j=1}^{\infty} p^{j-1} M_j + 1, \tag{2.61}$$

$$M_{i+1} = \lambda\beta\sum_{j=0}^{i-1} p^j M_{i-j} + \lambda\beta\sum_{j=1}^{\infty} p^{i+j-1} M_j, \quad i=1,2,\dots . \tag{2.62}$$

From (2.62),

$$M_{i+2} = (\lambda\beta+p)M_{i+1} = \cdots = (\lambda\beta+p)^i M_2, \quad i=0,1,\dots . \tag{2.63}$$

Substitution of (2.63) into (2.61) and (2.62) leads to a set of two linear equations with two unknowns M_1 and M_2; these equations yield

$$M_1 = \frac{1-p-\lambda\beta p}{1-p-\lambda\beta},$$

$$M_2 = \lambda\beta\frac{1-p(\lambda\beta+p)}{1-p-\lambda\beta},$$

so

$$E\mathbf{S}_1 = \frac{\beta}{1-\lambda\beta/(1-p)} + \frac{\lambda}{2}(\beta^{(2)} - 2\beta^2)\frac{1}{1-p}\,\frac{1-p-\lambda\beta p}{1-p-\lambda\beta}, \tag{2.64}$$

$$ES_k = \frac{\beta}{1-\lambda\beta/(1-p)} + \tag{2.65}$$

$$\frac{\lambda}{2}(\beta^{(2)}-2\beta^2)\frac{1}{1-p}\lambda\beta\frac{1-p(\lambda\beta+p)}{1-p-\lambda\beta}(\lambda\beta+p)^{k-2}, \quad k=2,3,\ldots.$$

From (2.64) and (2.65) it follows that the successive mean sojourn times are all equal if the service times are exponentially distributed ($\beta^{(2)}=2\beta^2$), cf. Section 2.3. Using $\lambda\beta/(1-p)=\rho<1$ it can be shown that $ES_1<ES_2$ if $\beta^{(2)}<2\beta^2$ and $ES_1>ES_2$ if $\beta^{(2)}>2\beta^2$. In the previous subsection we have observed similar properties for the first two sojourn times in the M/G/1 queue with deterministic feedback. Apparently, the difference between ES_1 and ES_2 is due to the fact that a customer's first sojourn time may contain a residual service (of mean length $\beta^{(2)}/(2\beta)$) while the second sojourn time only consists of complete service times, cf. (2.50), (2.51) and (2.59), (2.60). From (2.65) it is seen that, for $k\rightarrow\infty$,

$$ES_k\rightarrow\frac{\beta}{1-\lambda\beta/(1-p)}, \tag{2.66}$$

which is the mean sojourn time per visit in the case of a negative exponential service time distribution, cf. (2.20).

Finally, the mean total sojourn time, ES, is given by

$$ES = \sum_{i=1}^{\infty}p^{i-1}ES_i = \frac{\beta}{1-p-\lambda\beta} + \frac{\lambda}{2}(\beta^{(2)}-2\beta^2)\frac{1}{1-p-\lambda\beta}. \tag{2.67}$$

This result has been obtained before by Takács [1963].

Appendix 2.1

In this appendix we derive Formula (2.8). For ease of notation we introduce the service time distribution function $B(t) := 1 - e^{-t/\beta}$; the n-fold convolution of $B(\cdot)$ is denoted by $B(\cdot)^{n^*}$, $n = 1,2,\ldots$.

The derivation of (2.8) is based on the fact that $(\mathbf{X}_1^{(i+1)}, \ldots, \mathbf{X}_N^{(i+1)})$, which determines the distribution of $\mathbf{S}_{i+2}$, is conditionally independent of $\{(\mathbf{X}_1^{(0)}, \ldots, \mathbf{X}_N^{(0)}), \ldots, (\mathbf{X}_1^{(i-1)}, \ldots, \mathbf{X}_N^{(i-1)}); \mathbf{S}_1, \ldots, \mathbf{S}_i\}$ given $\{(\mathbf{X}_1^{(i)}, \ldots, \mathbf{X}_N^{(i)}); \mathbf{S}_{i+1}\}$, $i = 1,\ldots,k-1$. Using this property it is easily seen that, conditioning on the number of type-j arrivals, $n_j^{(m)}$, during the m-th sojourn time,

$$E\{e^{-(\omega_1 \mathbf{S}_1 + \ldots + \omega_k \mathbf{S}_k)} (z_{1,0}^{\mathbf{X}_1^{(0)}} \cdots z_{N,0}^{\mathbf{X}_N^{(0)}}) \cdots (z_{1,k}^{\mathbf{X}_1^{(k)}} \cdots z_{N,k}^{\mathbf{X}_N^{(k)}}) \mid \mathbf{X}_1^{(0)} = x_1, \ldots, \mathbf{X}_N^{(0)} = x_N\} =$$

$$z_{1,0}^{x_1} \cdots z_{N,0}^{x_N} \int_{t_1=0}^{\infty} e^{-\omega_1 t_1} \int_{t_2=0}^{\infty} e^{-\omega_2 t_2} \cdots \int_{t_{k-1}=0}^{\infty} e^{-\omega_{k-1} t_{k-1}} \int_{t_k=0}^{\infty} e^{-\omega_k t_k}$$

$$\sum_{n_1^{(1)}=0}^{\infty} e^{-\lambda t_1} \frac{(\lambda t_1)^{n_1^{(1)}}}{n_1^{(1)}!} z_{1,1}^{n_1^{(1)}} \left[\prod_{j=1}^{N} \sum_{n_{j+1}^{(1)}=0}^{x_j} \begin{bmatrix} x_j \\ n_{j+1}^{(1)} \end{bmatrix} p(j)^{n_{j+1}^{(1)}} (1-p(j))^{x_j - n_{j+1}^{(1)}} z_{j+1,1}^{n_{j+1}^{(1)}} \right]$$

$$\prod_{m=2}^{k} \left\{ \sum_{n_1^{(m)}=0}^{\infty} e^{-\lambda t_m} \frac{(\lambda t_m)^{n_1^{(m)}}}{n_1^{(m)}!} z_{1,m}^{n_1^{(m)}} \left[\prod_{j=1}^{N+m-1} \sum_{n_{j+1}^{(m)}=0}^{n_j^{(m-1)}} \begin{bmatrix} n_j^{(m-1)} \\ n_{j+1}^{(m)} \end{bmatrix} p(j)^{n_{j+1}^{(m)}} (1-p(j))^{n_j^{(m-1)} - n_{j+1}^{(m)}} z_{j+1,m}^{n_{j+1}^{(m)}} \right] \right\}$$

$$dB(t_k)^{(1+n_1^{(k-1)}+\ldots+n_{N+k-1}^{(k-1)})^*} dB(t_{k-1})^{(1+n_1^{(k-2)}+\ldots+n_{N+k-2}^{(k-2)})^*} \cdots$$

$$dB(t_2)^{(1+n_1^{(1)}+\ldots+n_{N+1}^{(1)})^*} dB(t_1)^{(1+x_1+\ldots+x_N)^*}.$$

Note that by definition $z_{i,j} := z_{N,j}$, $i = N+1,\ldots,N+k$, $j = 1,\ldots,k$.

We first evaluate the integral with respect to t_k, obtaining

$$E\{e^{-(\omega_1 \mathbf{S}_1 + \ldots + \omega_k \mathbf{S}_k)} (z_{1,0}^{\mathbf{X}_1^{(0)}} \cdots z_{N,0}^{\mathbf{X}_N^{(0)}}) \cdots (z_{1,k}^{\mathbf{X}_1^{(k)}} \cdots z_{N,k}^{\mathbf{X}_N^{(k)}}) \mid \mathbf{X}_1^{(0)} = x_1, \ldots, \mathbf{X}_N^{(0)} = x_N\} =$$

$$z_{1,0}^{x_1} \cdots z_{N,0}^{x_N} \int_{t_1=0}^{\infty} e^{-\omega_1 t_1} \int_{t_2=0}^{\infty} e^{-\omega_2 t_2} \cdots \int_{t_{k-1}=0}^{\infty} e^{-\omega_{k-1} t_{k-1}}$$

$$\sum_{n_1^{(1)}=0}^{\infty} e^{-\lambda t_1} \frac{(\lambda t_1)^{n_1^{(1)}}}{n_1^{(1)}!} z_{1,1}^{n_1^{(1)}} \left[\prod_{j=1}^{N} \sum_{n_{j+1}^{(1)}=0}^{x_j} \begin{bmatrix} x_j \\ n_{j+1}^{(1)} \end{bmatrix} p(j)^{n_{j+1}^{(1)}} (1-p(j))^{x_j - n_{j+1}^{(1)}} z_{j+1,1}^{n_{j+1}^{(1)}} \right]$$

$$\prod_{m=2}^{k-1} \left\{ \sum_{n_1^{(m)}=0}^{\infty} e^{-\lambda t_m} \frac{(\lambda t_m)^{n_1^{(m)}}}{n_1^{(m)}!} z_{1,m}^{n_1^{(m)}} \left[\prod_{j=1}^{N+m-1} \sum_{n_{j+1}^{(m)}=0}^{n_j^{(m-1)}} \begin{bmatrix} n_j^{(m-1)} \\ n_{j+1}^{(m)} \end{bmatrix} p(j)^{n_{j+1}^{(m)}} (1-p(j))^{n_j^{(m-1)} - n_{j+1}^{(m)}} z_{j+1,m}^{n_{j+1}^{(m)}} \right] \right\}$$

$$[1+\beta\{\omega_k+\lambda(1-z_{1,k})\}]^{-(1+n_1^{(k-1)}+\ldots+n_{N+k-1}^{(k-1)})}\prod_{j=1}^{N+k-1}(p(j)z_{j+1,k}+1-p(j))^{n_j^{(k-1)}}$$

$$dB(t_{k-1})^{(1+n_1^{(k-2)}+\ldots+n_{N+k-2}^{(k-2)})^*}\cdots dB(t_2)^{(1+n_1^{(1)}+\ldots+n_{N+1}^{(1)})^*}dB(t_1)^{(1+x_1+\ldots+x_N)^*} =$$

$$z_{1,0}^{x_1}\cdots z_{N,0}^{x_N}A_k^N(1,\omega,z)\int_{t_1=0}^{\infty}e^{-\omega_1 t_1}\int_{t_2=0}^{\infty}e^{-\omega_2 t_2}\cdots\int_{t_{k-1}=0}^{\infty}e^{-\omega_{k-1}t_{k-1}}$$

$$\sum_{n_1^{(1)}=0}^{\infty}e^{-\lambda t_1}\frac{(\lambda t_1)^{n_1^{(1)}}}{n_1^{(1)}!}z_{1,1}^{n_1^{(1)}}\left[\prod_{j=1}^{N}\sum_{n_{j+1}^{(1)}=0}^{x_j}\left[\begin{matrix}x_j\\ n_{j+1}^{(1)}\end{matrix}\right]p(j)^{n_{j+1}^{(1)}}(1-p(j))^{x_j-n_{j+1}^{(1)}}z_{j+1,1}^{n_{j+1}^{(1)}}\right]$$

$$\prod_{m=2}^{k-2}\left\{\sum_{n_1^{(m)}=0}^{\infty}e^{-\lambda t_m}\frac{(\lambda t_m)^{n_1^{(m)}}}{n_1^{(m)}!}z_{1,m}^{n_1^{(m)}}\left[\prod_{j=1}^{N+m-1}\sum_{n_{j+1}^{(m)}=0}^{n_j^{(m-1)}}\left[\begin{matrix}n_j^{(m-1)}\\ n_{j+1}^{(m)}\end{matrix}\right]p(j)^{n_{j+1}^{(m)}}(1-p(j))^{n_j^{(m-1)}-n_{j+1}^{(m)}}z_{j+1,m}^{n_{j+1}^{(m)}}\right]\right\}$$

$$e^{-\lambda t_{k-1}(1-z_{1,k-1}A_k^N(1,\omega,z)[p(1)z_{2,k}+1-p(1)])}$$

$$\prod_{j=1}^{N+k-2}(A_k^N(1,\omega,z)[p(j+1)z_{j+2,k}+1-p(j+1)]p(j)z_{j+1,k-1}+1-p(j))^{n_j^{(k-2)}}$$

$$dB(t_{k-1})^{(1+n_1^{(k-2)}+\ldots+n_{N+k-2}^{(k-2)})^*}\cdots dB(t_2)^{(1+n_1^{(1)}+\ldots+n_{N+1}^{(1)})^*}dB(t_1)^{(1+x_1+\ldots+x_N)^*}.$$

Next the integral with respect to t_{k-1} is evaluated, yielding

$$E\{e^{-(\omega_1\mathbf{S}_1+\ldots+\omega_k\mathbf{S}_k)}(z_{1,0}^{\mathbf{X}_1^{(0)}}\cdots z_{N,0}^{\mathbf{X}_N^{(0)}})\cdots(z_{1,k}^{\mathbf{X}_1^{(k)}}\cdots z_{N,k}^{\mathbf{X}_N^{(k)}})\,|\mathbf{X}_1^{(0)}=x_1,\ldots,\mathbf{X}_N^{(0)}=x_N\} =$$

$$z_{1,0}^{x_1}\cdots z_{N,0}^{x_N}A_k^N(1,\omega,z)\int_{t_1=0}^{\infty}e^{-\omega_1 t_1}\int_{t_2=0}^{\infty}e^{-\omega_2 t_2}\cdots\int_{t_{k-2}=0}^{\infty}e^{-\omega_{k-2}t_{k-2}}$$

$$\sum_{n_1^{(1)}=0}^{\infty}e^{-\lambda t_1}\frac{(\lambda t_1)^{n_1^{(1)}}}{n_1^{(1)}!}z_{1,1}^{n_1^{(1)}}\left[\prod_{j=1}^{N}\sum_{n_{j+1}^{(1)}=0}^{x_j}\left[\begin{matrix}x_j\\ n_{j+1}^{(1)}\end{matrix}\right]p(j)^{n_{j+1}^{(1)}}(1-p(j))^{x_j-n_{j+1}^{(1)}}z_{j+1,1}^{n_{j+1}^{(1)}}\right]$$

$$\prod_{m=2}^{k-2}\left\{\sum_{n_1^{(m)}=0}^{\infty}e^{-\lambda t_m}\frac{(\lambda t_m)^{n_1^{(m)}}}{n_1^{(m)}!}z_{1,m}^{n_1^{(m)}}\left[\prod_{j=1}^{N+m-1}\sum_{n_{j+1}^{(m)}=0}^{n_j^{(m-1)}}\left[\begin{matrix}n_j^{(m-1)}\\ n_{j+1}^{(m)}\end{matrix}\right]p(j)^{n_{j+1}^{(m)}}(1-p(j))^{n_j^{(m-1)}-n_{j+1}^{(m)}}z_{j+1,m}^{n_{j+1}^{(m)}}\right]\right\}$$

$$[1+\beta\{\omega_{k-1}+\lambda-\lambda z_{1,k-1}A_k^N(1,\omega,z)[p(1)z_{2,k}+1-p(1)]\}]^{-(1+n_1^{(k-2)}+\ldots+n_{N+k-2}^{(k-2)})}$$

$$\prod_{j=1}^{N+k-2}(A_k^N(1,\omega,z)[p(j+1)z_{j+2,k}+1-p(j+1)]p(j)z_{j+1,k-1}+1-p(j))^{n_j^{(k-2)}}$$

$$dB(t_{k-2})^{(1+n_1^{(k-3)}+\ldots+n_{N+k-3}^{(k-3)})^*} \cdots dB(t_2)^{(1+n_1^{(1)}+\ldots+n_{N+1}^{(1)})^*} dB(t_1)^{(1+x_1+\ldots+x_N)^*} =$$

$$z_{1,0}^{x_1} \cdots z_{N,0}^{x_N} A_k^N(1,\omega,z) A_k^N(2,\omega,z) \int_{t_1=0}^{\infty} e^{-\omega_1 t_1} \int_{t_2=0}^{\infty} e^{-\omega_2 t_2} \cdots \int_{t_{k-2}=0}^{\infty} e^{-\omega_{k-2} t_{k-2}}$$

$$\sum_{n_1^{(1)}=0}^{\infty} e^{-\lambda t_1} \frac{(\lambda t_1)^{n_1^{(1)}}}{n_1^{(1)}!} z_{1,1}^{n_1^{(1)}} \left[\prod_{j=1}^{N} \sum_{n_{j+1}^{(1)}=0}^{x_j} \begin{pmatrix} x_j \\ n_{j+1}^{(1)} \end{pmatrix} p(j)^{n_{j+1}^{(1)}} (1-p(j))^{x_j - n_{j+1}^{(1)}} z_{j+1,1}^{n_{j+1}^{(1)}} \right]$$

$$\prod_{m=2}^{k-2} \left\{ \sum_{n_1^{(m)}=0}^{\infty} e^{-\lambda t_m} \frac{(\lambda t_m)^{n_1^{(m)}}}{n_1^{(m)}!} z_{1,m}^{n_1^{(m)}} \left[\prod_{j=1}^{N+m-1} \sum_{n_{j+1}^{(m)}=0}^{n_j^{(m-1)}} \begin{pmatrix} n_j^{(m-1)} \\ n_{j+1}^{(m)} \end{pmatrix} p(j)^{n_{j+1}^{(m)}} (1-p(j))^{n_j^{(m-1)} - n_{j+1}^{(m)}} z_{j+1,m}^{n_{j+1}^{(m)}} \right] \right\}$$

$$\prod_{j=1}^{N+k-2} (A_k^N(2,\omega,z)[A_k^N(1,\omega,z)[p(j+1)z_{j+2,k}+1-p(j+1)]p(j)z_{j+1,k-1}+1-p(j)])^{n_j^{(k-2)}}$$

$$dB(t_{k-2})^{(1+n_1^{(k-3)}+\ldots+n_{N+k-3}^{(k-3)})^*} \cdots dB(t_2)^{(1+n_1^{(1)}+\ldots+n_{N+1}^{(1)})^*} dB(t_1)^{(1+x_1+\ldots+x_N)^*} .$$

We now sum over $n_1^{(k-2)}, \ldots, n_{N+k-2}^{(k-2)}$ and subsequently integrate with respect to t_{k-2}, thus obtaining $A_k^N(3,\omega,z)$ terms; etc.; finally the summations over $n_1^{(1)}, \ldots, n_N^{(1)}$ and the integration over t_1 are performed, which gives rise to the $(f_k^N(i,\omega,z))^{x_i}$ contribution in Formula (2.8).

Chapter 3

THE M/G/1 PROCESSOR SHARING QUEUE AS A LIMITING MODEL OF THE M/M/1 FEEDBACK QUEUE

3.1 Introduction

In the previous chapter we have regarded the feedback model as an M/M/1 queue in which after each service it is decided whether or not the customer is fed back. In this chapter we consider the same model from another point of view, viz. as a round robin (time sharing) model in which a customer's service demand requires a stochastic number of exponentially distributed service quanta with mean length β. Obviously, the service requirements are completely determined by the feedback probabilities $p(1)$, $p(2)$, $\cdots$, as defined in Chapter 2. From this point of view it is intuitively clear that if the mean service time β shrinks to zero while the feedback probabilities go to one such that a customer's total required service time remains unchanged, the behaviour of the feedback queue approaches that of the M/G/1 processor sharing (PS) queue. Different choices of the feedback probabilities lead to different service time distributions in the PS queue.

The queue length process in a round robin type of queue is usually less amenable to mathematical analysis than the queue length process in its limiting case, a PS queue. This has been the main reason for the queueing analysis of processor sharing, see Kleinrock [1976]. Sakata et al. [1969] showed that the distribution of the queue length, $\mathbf{X}^{PS}$, in the M/G/1 PS queue is independent of the distribution of the required service time apart from its first moment:

$$Pr\{\mathbf{X}^{PS}=j\} = (1-\rho)\rho^j, \quad j=0,1,2,\ldots, \tag{3.1}$$

with ρ the offered load per unit of time. The determination of the *sojourn time* distribution in a PS queue has turned out to be a much harder problem. Only recently the sojourn time distribution in the M/G/1 PS queue has been derived, cf. Yashkov [1983], Ott [1984], Schassberger [1984], and the survey of Yashkov [1987]. We refer to Section 1.5 for a brief description of the approaches used by these authors. The approach presented in this chapter is new: via a limiting procedure we obtain sojourn time results for the M/G/1 PS queue from known sojourn time results (obtained in Chapter 2) for the M/M/1 queue with general feedback.

The limiting procedure described above was first proposed by Van den Berg et al. [1989A]. In that paper it is shown how the distribution of the sojourn time in the M/D/1 PS queue follows immediately (by taking appropriate limits) from the sojourn time distribution in the M/M/1 queue with deterministic feedback. In Van den Berg and Boxma [1989B] this method has been extended to the analysis of the processor sharing queue with general service times. In these papers the authors concluded on *intuitive* grounds that the performance measures such as the sojourn time in the feedback model converge to the corresponding performance measures in the processor sharing queue. Only very recently a formal proof of this convergence has been given by Resing et al. [1989]. They present a probabilistic coupling between the M/G/1 PS queue and the approximating sequence of M/M/1 feedback queues, which shows that the sojourn time of the n-th customer in the feedback model converges almost surely to the corresponding quantity in the PS model. From this result they conclude the distributional convergence of the steady state sojourn times. The proof partially follows the same line of thought as Schassberger [1984].

The organization of the rest of this chapter is as follows. Section 3.2 contains some definitions and restates those results of Chapter 2 that are essential for the analysis in Section 3.3. In the latter section we study sojourn times in the M/G/1 PS queue, by taking appropriate limits in the M/M/1 queue with feedback. We first derive the mean sojourn time (Subsection 3.3.2) and the sojourn time variance (Subsection 3.3.3). Next, in Subsection 3.3.4, it is shown how the LST of the distribution of the sojourn time in the M/G/1 PS queue can be obtained. Section 3.2, Subsection 3.3.2 and Subsection 3.3.3 are mainly based on Van den Berg and Boxma [1989B]. In Section 3.4 we consider the same feedback model as in Chapter 2 but with state dependent service rates. It is shown how a similar limiting procedure leads to the analysis of the M/G/1 queue with the so called 'generalized processor sharing' service discipline. This section is restricted to the derivation of mean sojourn times. In the last section of this chapter, Section 3.5, we analyze sojourn times in the M/G/1 PS queue with feedback. Using the (M/M/1 FCFS) feedback results obtained in Chapter 2 and applying the limiting procedure we derive new results for the correlation coefficients of the successive sojourn times of a tagged customer.

3.2 Preliminary results

We consider the M/M/1 feedback queue introduced and analyzed in Chapter 2. In Section 2.2 we made the assumption that the feedback probabilities of a customer remain constant after a finite number of services. However, from the ultimate result for the joint sojourn time distribution ((2.18)) it appeared that this assumption is not needed, see the discussion below (2.19). In the following no restrictions will be put on the structure of the feedback probabilities. So, we rewrite Definition (2.1):

$$q(1) := 1, \qquad (3.2)$$
$$q(i) := \prod_{j=1}^{i-1} p(j), \ \ i=2,3,\dots .$$

Obviously, for stability it is required that the $q(i)$'s satisfy:

$$\lambda\beta\sum_{i=1}^{\infty} q(i) =: \rho < 1 .$$

For future reference we introduce the generating function of the probabilities of visiting the queue exactly i times, $i=1,2,\dots$:

$$Q(z) := \sum_{i=1}^{\infty} q(i)(1-p(i))z^{i}, \quad |z| \leqslant 1. \qquad (3.3)$$

We now recall those feedback results obtained in Chapter 2 which are essential for the analysis in the next sections. Some of the expressions will be rewritten such that they are more suitable for analyzing sojourn times in the M/G/1 PS queue.

First, we slightly rewrite formula (2.18) and use it to obtain a convenient expression for the LST of the total sojourn time after k services. Replacing the term $M_k(k,\omega)$ in the denominator of (2.18) by the RHS of (2.19) (with $i=k$) and substituting $\omega_j = \omega_0$, $j=1,\dots,k$, it is easily seen that (using the notation introduced in (3.2)), for Re $\omega_0 \geqslant 0$, $k=1,2,\dots$,

$$E\{e^{-\omega_0 \mathbf{S}^{(k)}}\} = \qquad (3.4)$$

$$\frac{1-\rho}{(1+\beta\omega_0)M_{k-1} - \lambda\beta\sum_{j=1}^{k-2} q(k-j-1)M_j - (\rho-\lambda\beta\sum_{i=1}^{k-2} q(i))},$$

where,

$$M_0 := 1, \qquad (3.5)$$
$$M_n := (1+\beta\omega_0+\lambda\beta)M_{n-1} - \lambda\beta\left[q(n-1)+\sum_{l=2}^{n-1} q(n-l)(1-p(n-l))M_{l-1}\right],$$
$$n=1,2,\dots .$$

$(q(0) := 1)$

In Chapter 2 it has been found that the mean total sojourn time after k services is linear in k, cf. (2.25):

$$E\{\mathbf{S}^{(k)}\} = k\frac{\beta}{1-\rho}, \quad k=1,2,\ldots . \tag{3.6}$$

Formula (2.26) gives the variance of the total sojourn time after k services:

$$var(\mathbf{S}^{(k)}) = (\frac{\beta}{1-\rho})^2[k^2-2(1-\rho)\sum_{j=1}^{k-1} jC_{k-j}], \quad k=1,2,..., \tag{3.7}$$

where $C_1,\ldots,C_{k-1}$ can be successively obtained from (2.22). This recurrence relation for the C_n can be simplified in the following way. Noting that $q(n-l)(1-p(n-l))=q(n-l)-q(n-l+1)$, and splitting the sum in (2.22) we obtain

$$C_n-\lambda\beta\sum_{l=2}^{n}q(n-l+1)C_{l-1} = C_{n-1}-\lambda\beta\sum_{l=2}^{n-1}q(n-l)C_{l-1}.$$

Now, using $C_1=1$, it is easily seen that

$$C_1 = 1, \tag{3.8}$$

$$C_n = 1 + \lambda\beta\sum_{l=1}^{n-1}q(n-l)C_l, \quad n=2,3\ldots .$$

Taking generating functions and using (3.3) leads to

$$C(z) := \sum_{n=1}^{\infty}C_nz^n = \frac{z}{(1-z)(1-\lambda\beta\frac{z}{1-z}(1-Q(z)))}, \quad |z|<1. \tag{3.9}$$

The sequence $C_1,C_2,\ldots$ is non-decreasing and, cf. (2.24), limited from above. Hence $\lim_{n\to\infty} C_n$ exists; an Abelian theorem now implies that (cf. Titchmarsh [1952])

$$\lim_{n\to\infty} C_n = \lim_{z\to 1}(1-z)C(z) = \frac{1}{1-\rho}. \tag{3.10}$$

For future use we also introduce the generating function of the M_n's. From (3.5) it follows that

$$M(z) := \sum_{n=1}^{\infty}M_nz^n = z\frac{1+\beta\omega_0-\lambda\beta\frac{z}{1-z}(1-Q(z))}{1-z(1+\beta\omega_0+\lambda\beta)+\lambda\beta zQ(z)}, \quad |z|<1. \tag{3.11}$$

3.3 The M/G/1 processor sharing queue

3.3.1 The limiting procedure

In this section we show how the feedback results collected in Section 3.2 can be used to analyze the sojourn time in the M/G/1 PS queue. We apply a limiting procedure, in which $\beta \to 0$ while the feedback probabilities approach one in such a way that the mean total required service time, $\hat{\beta}$, remains a positive constant. We restrict ourself to those service times, $\boldsymbol{\tau}^{PS}$, in the PS queue which are composed of negative exponentially distributed stages:

$$E\{\exp(-\omega_0 \boldsymbol{\tau}^{PS})\} = \sum_{j=1}^{m} \alpha_j \prod_{i=1}^{r_j} \frac{1}{1+\hat{\beta}_{ij}\omega_0}, \tag{3.12}$$

with $\alpha_1, \ldots, \alpha_m > 0$, $\sum_{j=1}^{m} \alpha_j = 1$, $r_1, \ldots, r_m$ positive integers (cf. Kleinrock [1975], p. 145); note that this class of distributions contains the Erlang, hyperexponential and Coxian distributions, and that arbitrary probability distributions of nonnegative random variables can be arbitrarily closely approximated by distributions from this class (cf. Tijms [1986], p. 398). This choice of service time distribution for the PS queue enables us to choose the feedback probabilities (hence $Q(z)$) such that $\boldsymbol{\tau}^{PS}$ and the total required service time $\boldsymbol{\tau}^{FB}$ in the feedback queue have exactly the same distribution - not just in the limit $\beta \to 0$, but for a wide range of values of β. Observe that, cf. (3.3),

$$E\{\exp(-\omega_0 \boldsymbol{\tau}^{FB})\} = \sum_{i=1}^{\infty} q(i)(1-p(i))(\frac{1}{1+\beta\omega_0})^i = Q(\frac{1}{1+\beta\omega_0}), \tag{3.13}$$

$$\text{Re } \omega_0 \geqslant 0.$$

Now choose

$$Q(z) = \sum_{j=1}^{m} \alpha_j \prod_{i=1}^{r_j} \frac{(1-p_{ij})z}{1-p_{ij}z}, \tag{3.14}$$

with

$$p_{ij} = 1 - \beta / \hat{\beta}_{ij} > 0, \quad i=1,...,r_j,\ j=1,...,m. \tag{3.15}$$

Then

$$E\{\exp(-\omega_0 \boldsymbol{\tau}^{FB})\} = \sum_{j=1}^{m} \alpha_j \prod_{i=1}^{r_j} \frac{\beta / \hat{\beta}_{ij}}{1+\beta\omega_0 - (1-\beta/\hat{\beta}_{ij})} = \tag{3.16}$$

$$\sum_{j=1}^{m} \alpha_j \prod_{i=1}^{r_j} \frac{1}{1+\hat{\beta}_{ij}\omega_0} = E\{\exp(-\omega_0 \boldsymbol{\tau}^{PS})\}.$$

As an example, consider the case of Bernoulli feedback:

$Q(z) = (1-p)z/(1-pz)$. In this case,

$$E\{\exp(-\omega_0 \boldsymbol{\tau}^{PS})\} = E\{\exp(-\omega_0 \boldsymbol{\tau}^{FB})\} = \frac{1}{1+(\beta/(1-p))\omega_0} = \frac{1}{1+\hat{\beta}\omega_0}. \quad (3.17)$$

Hence the total required service times in both the feedback queue and the PS queue are negative exponentially distributed with mean $\hat{\beta}=\beta/(1-p)$.

When $\beta \to 0$, performance measures in the feedback queue clearly approach corresponding performance measures in the PS queue. Resing et al. [1989] give a formal proof of the convergence of the sojourn time. Note that the queue length distribution in both models is the same for the *whole* range of possible β values, cf. (3.1) and (2.5). Below we shall focus mainly on sojourn times. In particular we are interested in the sojourn time of a customer *conditioned* on his required service time. This is an important performance measure for time sharing systems like PS queues, cf. Kleinrock [1976]. We define for the PS queue

- $\mathbf{S}^{PS}(x)$: conditional sojourn time of a customer with service demand x;
- $\mathbf{S}^{PS}$: sojourn time of an arbitrary customer.

Obviously,

$$Pr\{\mathbf{S}^{PS}<s\} = \int_{x=0}^{\infty} Pr\{\mathbf{S}^{PS}(x)<s\} dPr\{\boldsymbol{\tau}^{PS}<x\}\ , \quad s \geq 0. \quad (3.18)$$

The conditional sojourn time $\mathbf{S}^{PS}(x)$ can be derived from the total sojourn time after k services, $\mathbf{S}^{(k)}$, in the feedback queue as follows. Choose $Q(z)$ for the feedback queue as in (3.14), (3.15), and consider a newly arriving customer, say C, who requires exactly k services. Then take $\beta=x/k$ and let $k \to \infty$. It is easily seen that the total required service time of C approaches the constant x. Indeed, the LST of C's total required service time equals $(1+\beta\omega_0)^{-k} = (1+x\omega_0/k)^{-k} \to e^{-x\omega_0}$. *Hence, for $k \to \infty$, C can be viewed as a customer with service request x in the $M/G/1$ PS queue with service time distribution characterized by (3.12).*

The limiting procedure described above will be applied below. We shall obtain results for the mean, the variance and the LST of the sojourn time in the PS queue from $E\{\mathbf{S}^{PS}(x)\} = \lim_{k\to\infty} E\{\mathbf{S}^{(k)}\}$, $var(\mathbf{S}^{PS}(x)) = \lim_{k\to\infty} var(\mathbf{S}^{(k)})$ and $E\{e^{-\omega_0 \mathbf{S}^{PS}(x)}\} = \lim_{k\to\infty} E\{e^{-\omega_0 \mathbf{S}^{(k)}}\}$ respectively. The results to be presented for the mean and the variance of the sojourn time are more general and more detailed than the results for the LST.

3.3.2 The mean sojourn time

In the M/G/1 PS queue, the mean sojourn time of a customer with service demand x is linear in x (cf. Kleinrock [1976]):

$$E\{\mathbf{S}^{PS}(x)\} = \frac{x}{1-\rho}. \tag{3.19}$$

We now show how this well known result can be easily obtained from the feedback results collected in Section 3.2. The mean total sojourn time $E\{\mathbf{S}^{(k)}\}$ of a customer who requires k services is linear in k, see (3.6). Apply the limiting procedure described in Subsection 3.3.1, taking $\beta = x/k$ and letting $k \rightarrow \infty$. Formula (3.19) now immediately follows from (3.6).

3.3.3 The variance of the sojourn time

The sojourn time variance for a customer with service request x in the M/G/1 PS queue, $var(\mathbf{S}^{PS}(x))$, can be obtained by applying the limiting procedure to (3.7). First, as an example, we derive $var(\mathbf{S}^{PS}(x))$ for the M/M/1 PS queue. Next the analysis is extended to the PS queue with general service times. This leads to a simple explicit expression for the asymptotic behaviour of $var(\mathbf{S}^{PS}(x))$ for very large $(x \rightarrow \infty)$ and very small $(x \rightarrow 0)$ service requests.

The M/M/1 PS queue

As observed in (3.17), the choice $Q(z) = (1-p)z/(1-pz)$ leads, in the feedback queue as well as the PS queue, to a negative exponentially distributed total service time with mean $\beta/(1-p) = \hat{\beta}$. To obtain an explicit expression for $var(\mathbf{S}^{(k)})$, see (3.7), we derive C_n, $n = 1,2,...,$ from (3.9). Substituting $Q(z) = (1-p)z/(1-pz)$ into (3.9) yields

$$C(z) = z\frac{1-pz}{(1-z)(1-(\lambda\beta+p)z)}. \tag{3.20}$$

Rewriting the right-hand side of (3.20) as

$$z(U_1\frac{1}{1-z} + U_2\frac{1}{1-(\lambda\beta+p)z}),$$

it follows that

$$C_n = U_1 + U_2 x_2^{n-1}, \quad n = 1,2,..., \tag{3.21}$$

with $U_1 = 1/(1-\rho)$, $U_2 = -\rho/(1-\rho)$, $x_2 = \lambda\beta+p$. Substituting (3.21) into (3.7) yields

$$var(\mathbf{S}^{(k)}) = (\frac{\beta}{1-\rho})^2[k - 2(1-\rho)U_2\frac{x_2^k + k(1-x_2) - 1}{(1-x_2)^2}] = \tag{3.22}$$

$$(\frac{\beta}{1-\rho})^2[k + \frac{2\rho}{1-p}(\frac{k}{1-\rho} - \frac{1-(\lambda\beta+p)^k}{(1-p)(1-\rho)^2})].$$

Let $\hat{\beta}$ be the mean service time for the M/M/1 PS queue and let x be the service time of a tagged customer (cf. Subsection 3.3.2). Substitute $\beta=x/k$ and $p=1-x/k\hat{\beta}$ into (3.22). Letting $k \to \infty$ leads to $var(\mathbf{S}^{PS}(x))$:

$$var(\mathbf{S}^{PS}(x)) = \lim_{k\to\infty} var(\mathbf{S}^{(k)}) = \frac{2\rho\hat{\beta}x}{(1-\rho)^3} - \frac{2\rho\hat{\beta}^2}{(1-\rho)^4}[1-e^{-x(1-\rho)/\hat{\beta}}], \tag{3.23}$$

a result previously obtained by Ott [1984]. Note that the sojourn time variance depends linearly on x for $x \to \infty$:

$$var(\mathbf{S}^{PS}(x)) \sim \frac{2\rho\hat{\beta}}{(1-\rho)^3}x - \frac{2\rho\hat{\beta}^2}{(1-\rho)^4}, \quad x\to\infty, \tag{3.24}$$

(see also Kleinrock [1976], p. 170), whereas it depends quadratically on x for $x \to 0$:

$$var(\mathbf{S}^{PS}(x)) \sim \frac{\rho}{(1-\rho)^2}x^2 - \frac{\lambda}{3(1-\rho)}x^3, \quad x\to 0. \tag{3.25}$$

The M/G/1 PS queue

We now derive an expression for $var(\mathbf{S}^{PS}(x))$ for the M/G/1 PS queue, in particular showing that the above asymptotic properties hold for general service time distributions. We consider service time distributions with LST as in (3.12), by choosing $Q(z)$ as in (3.14), (3.15):

$$Q(z) = \sum_{j=1}^{m} \alpha_j \prod_{i=1}^{r_j} \frac{(1-p_{ij})z}{1-p_{ij}z} = \sum_{j=1}^{m} \alpha_j \prod_{i=1}^{r_j} \frac{\beta z/\hat{\beta}_{ij}}{1-(1-\beta/\hat{\beta}_{ij})z}. \tag{3.26}$$

Analogously to the M/M/1 case analyzed above, (3.9) and (3.26) lead to:

$$C_n = U_1 + U_2x_2^{n-1} + \cdots + U_Lx_L^{n-1}, \quad n=1,2,..., \tag{3.27}$$

where $1/x_2, \ldots, 1/x_L$ are the roots of

$$1-\lambda\beta\frac{z}{1-z}(1-Q(z)) = 0. \tag{3.28}$$

$U_1, \ldots, U_L$ are determined by

$$\frac{U_1 z}{1-x_1 z} + \ldots + \frac{U_L z}{1-x_L z} = C(z). \tag{3.29}$$

Note that in (3.27)-(3.29) we have used the following assumption:

ASSUMPTION 3.1
It is assumed that the roots $1/x_2, \ldots, 1/x_L$ of (3.28) are all distinct.

REMARK 3.1
Assumption 3.1 can be easily proved to hold for the case of Erlang and hyperexponential service times; we have found no example for which the roots are *not* distinct.

REMARK 3.2
$1/x_2, \ldots, 1/x_L$ are the roots of a polynomial of degree $L-1 \leqslant \sum_{j=1}^{m} r_j$, see (3.26), (3.28); for example, for m-stage hyperexponential and m-stage Erlang service time distributions $L = m+1$. Note that (3.29) leads to a set of L linear equations from which $U_1, \ldots, U_L$ can be obtained.

We now prove some properties of x_i and U_i that will be used in the sequel.

LEMMA 3.1
(i) $|x_i| < 1, \quad i = 2,...,L;$
(ii) *x_i can be written as*

$$x_i = 1 - \beta a_i, \tag{3.30}$$

with a_i independent of β, and Re $a_i > 0, \quad i = 2,...,L;$
(iii) *U_i is independent of β, $i = 1,...,L$, and $U_1 = 1/(1-\rho)$.*

PROOF
Noting that (see (3.9)),

$$1 - \lambda\beta\frac{z}{1-z}(1-Q(z)) = 1 - \lambda\beta\sum_{i=1}^{\infty} q(i) z^i,$$

and $\lambda\beta\sum_{i=1}^{\infty} q(i) = \rho < 1$, it follows immediately that $|x_i| < 1$, $i = 2,...,L$. To prove (ii), substitute (3.26) into (3.28) and replace z by $1/(1-\beta\tilde{z})$. Then (3.28) reduces to

$$1 + \frac{\lambda}{\tilde{z}} - \frac{\lambda}{\tilde{z}}\sum_{j=1}^{m} \alpha_j \prod_{i=1}^{r_j} \frac{1}{1-\hat{\beta}_{ij}\tilde{z}} = 0. \tag{3.31}$$

Since $1/x_i$ is a root of (3.28), $(1-x_i)/\beta = a_i$ is a root of (3.31). The fact that

β does not occur in the left-hand side of (3.31) implies that $1-x_i$ depends linearly on β. The statement concerning Re $a_i > 0$ now follows from (i).
It follows from (3.29) that

$$U_i \frac{1}{x_i} = \lim_{z \to 1/x_i} (1-zx_i)C(z) = \lim_{\tilde{z} \to a_i} (1-\frac{x_i}{1-\beta\tilde{z}})C(\frac{1}{1-\beta\tilde{z}}). \tag{3.32}$$

Observing that $\beta C(\frac{1}{1-\beta\tilde{z}})$ is independent of β, it is found that

$$U_i = \lim_{\tilde{z} \to a_i} x_i(1-\frac{x_i}{1-\beta\tilde{z}})C(\frac{1}{1-\beta\tilde{z}})$$

is independent of β.
Finally, $U_1 = 1/(1-\rho)$ follows from (3.10), (3.27) and (i).

Substituting (3.27) into (3.7) yields (cf. (3.22))

$$var(\mathbf{S}^{(k)}) = (\frac{\beta}{1-\rho})^2[k \ - \ 2(1-\rho)\sum_{j=2}^{L} U_j \frac{x_j^k + k(1-x_j)-1}{(1-x_j)^2}]. \tag{3.33}$$

Now, let x be the service time of a tagged customer, and take $\beta = x/k$. For $k \to \infty$, $var(\mathbf{S}^{PS}(x))$ follows from (3.33) and (i) of Lemma 3.1; integrating $E\{(\mathbf{S}^{PS}(x))^2\} = var(\mathbf{S}^{PS}(x)) + x^2/(1-\rho)^2$ over x and subtracting $(E\{\mathbf{S}^{PS}\})^2 = \hat{\beta}^2/(1-\rho)^2$ yields the unconditional sojourn time variance. We collect these results in

THEOREM 3.1
In the M/G/1 PS queue with service time LST given by (3.12),

$$var(\mathbf{S}^{PS}(x)) = \frac{2}{1-\rho}\sum_{j=2}^{L}(1/a_j)^2 U_j[1-xa_j-e^{-xa_j}], \tag{3.34}$$

$$var(\mathbf{S}^{PS}) = \frac{2}{1-\rho}\sum_{j=2}^{L}(1/a_j)^2 U_j[1-\hat{\beta}a_j - E\{e^{-a_j\boldsymbol{\tau}^{PS}}\}] + \frac{E\{(\boldsymbol{\tau}^{PS})^2\}-\hat{\beta}^2}{(1-\rho)^2}, \tag{3.35}$$

with $a_2, \ldots, a_L$ the roots of (3.31) and $U_2, \ldots, U_L$ determined by (3.29), cf. Remark 3.2; a_j and U_j are independent of x, $j=2,\ldots,L$.

Formula (3.34) shows that $var(\mathbf{S}^{PS}(x))$ depends on the required service time x in a very simple way. It is convenient to use this formula for the analysis of the behaviour of the sojourn time variance when x varies. Below we shall derive asymptotic results for $x\rightarrow\infty$ and $x\rightarrow 0$.

From (ii) of Lemma 3.1 it follows that

$$\lim_{x\rightarrow\infty} \frac{2}{1-\rho}\sum_{j=2}^{L}(1/a_j)^2 U_j e^{-xa_j} = 0 .$$

Hence, the sojourn time variance is asymptotically linear in x:

$$var(\mathbf{S}^{PS}(x)) \sim \frac{2}{1-\rho}\sum_{j=2}^{L}(1/a_j)^2 U_j(1-xa_j), \qquad x\rightarrow\infty . \tag{3.36}$$

From (3.27) and (3.30),

$$\sum_{j=2}^{L}(1/a_j)U_j = \beta\sum_{j=2}^{L}U_j\frac{1}{1-x_j} = \beta\sum_{n=1}^{\infty}(C_n-\frac{1}{1-\rho}) ,$$

and

$$\sum_{j=2}^{L}(1/a_j)^2 U_j = \beta^2\sum_{j=2}^{L}U_j\frac{1}{(1-x_j)^2} = \beta^2\sum_{n=1}^{\infty}n(C_n-\frac{1}{1-\rho}) .$$

It can be derived from (3.9) that

$$\beta\sum_{n=1}^{\infty}(C_n-\frac{1}{1-\rho}) = -\frac{\lambda\hat{\beta}_2}{2(1-\rho)^2} ,$$

and

$$\beta^2\sum_{n=1}^{\infty}n(C_n-\frac{1}{1-\rho}) = -\frac{1}{2(1-\rho)^3}[\frac{\lambda}{3}\hat{\beta}_3+\lambda^2(\frac{1}{2}\hat{\beta}_2^2-\frac{1}{3}\hat{\beta}\hat{\beta}_3)] ,$$

with $\hat{\beta}_i := E\{(\tau^{PS})^i\}$, $i=2,3$.
Hence, from (3.36),

$$var(\mathbf{S}^{PS}(x)) \sim \frac{\lambda\hat{\beta}_2}{(1-\rho)^3}x - \frac{1}{(1-\rho)^4}[\frac{\lambda}{3}\hat{\beta}_3+\lambda^2(\frac{1}{2}\hat{\beta}_2^2-\frac{1}{3}\hat{\beta}\hat{\beta}_3)] , \qquad x\rightarrow\infty. \tag{3.37}$$

(Note that, formally, this asymptotic result should have been written as $var(\mathbf{S}^{PS}(x))-(\lambda\hat{\beta}_2/(1-\rho)^3)x \sim -[\frac{\lambda}{3}\hat{\beta}_3+\lambda^2(\frac{1}{2}\hat{\beta}_2^2-\frac{1}{3}\hat{\beta}\hat{\beta}_3)]/(1-\rho)^4$, $x\rightarrow\infty$).

Noting that, in (3.34),

$$1-xa_j-e^{-xa_j} = -\sum_{i=2}^{\infty}\frac{(-xa_j)^i}{i!}, \quad j=2,...,L,$$

and using

$$\sum_{j=2}^{L}U_j = C_1-\frac{1}{1-\rho} = -\frac{\rho}{1-\rho}, \qquad \sum_{j=2}^{L}U_jx_j = C_2-\frac{1}{1-\rho} = \lambda\beta-\frac{\rho}{1-\rho},$$

it is found that (cf. the remark below (3.37))

$$var(\mathbf{S}^{PS}(x)) \sim \frac{\rho}{(1-\rho)^2}x^2-\frac{\lambda}{3(1-\rho)}x^3, \quad x\to 0. \tag{3.38}$$

This expression appears to be independent of the service time distribution, apart from its first moment (cf. also (3.25)). The quadratic behaviour of $var(\mathbf{S}^{PS}(x))$ for small service requests x should be contrasted with the linear behaviour for large x.

REMARK 3.3
Formula (3.38) slightly generalizes Theorem 1 of Yashkov [1986]. Formula (3.37) is contained in Theorem 2 of the same paper; but for the service time distributions defined by (3.12), Yashkov's theorem follows immediately from (3.34).

3.3.4 The distribution of the sojourn time

Application of the limiting procedure to (3.4) yields the LST ρf the distribution of the sojourn time in the M/G/1 PS queue. The analysis can be performed along the same lines as the analysis of the sojourn time variance. It appears that the M_n's in (3.4) have similar properties as the C_n's in the previous subsection. However, there are some difficulties which did not arise in the analysis of the variance. These problems are due to the presence of the individual feedback probabilities contained in the $q(n)$'s in the denominator of (3.4). In general the $q(n)$'s are given by very complicated expressions and can not be explicitly determined for the whole class of service time distributions given by (3.12) (cf. (3.14), (3.15) and (3.16)). Therefore, we shall restrict ourself below to a subclass of these service times, viz. mixtures of Erlang distributions: (cf. (3.12))

$$E\{e^{-\omega_0\tau^{PS}}\} = \sum_{j=1}^{m}\alpha_j\left[\frac{1}{1+\hat{\beta}_j\omega_0}\right]^{r_j}, \tag{3.39}$$

with $\alpha_1, \ldots, \alpha_m \geqslant 0$, $\sum_{j=1}^{m} \alpha_j = 1$, $r_1, \ldots, r_m$ positive integers. The corresponding feedback probabilities are determined by (cf. (3.14), (3.15))

$$Q(z) = \sum_{j=1}^{m} \alpha_j \left[\frac{(1-p_j)z}{1-p_j z} \right]^{r_j}, \tag{3.40}$$

with $p_j = 1 - \beta / \hat{\beta}_j > 0$.

From (3.40) we find

$$q(l)(1-p(l)) = \sum_{j=1}^{m} \alpha_j (1-p_j)^{r_j} \begin{bmatrix} l-1 \\ r_j - 1 \end{bmatrix} p_j^{l-r_j}, \tag{3.41}$$

from which the $q(n)$'s can be obtained via:

$$q(n) = \sum_{l=n}^{\infty} q(l)(1-p(l)), \quad n = 1,2,\ldots . \tag{3.42}$$

Note that the (sub)class of distribution functions determined by (3.39) is still large enough to approximate the distribution of any nonnegative random variable arbitrarily closely (cf. Tijms [1986], p. 398).

We shall start the analysis with a lemma that states some properties of the M_n's given by (3.5) (see also (3.11)). Then, as an example, we consider the M/M/1 PS queue and show how these properties can be exploited to derive from (3.4) the LST of the sojourn time distribution. Next, the general case is treated. Finally, we consider the M/D/1 PS queue. Although the deterministic distribution is not contained in the class of service time distributions determined by (3.39) it appears that this case can be (partially) analyzed and yields simple expressions.

Let $1/y_1, \ldots, 1/y_L$ be the zeros of the denominator,

$$1 - z(1 + \beta\omega_0 + \lambda\beta) + \lambda\beta z Q(z), \tag{3.43}$$

of the generating function $M(z)$ of the M_n's, cf. (3.11). To obtain closed expressions for the M_n's we introduce the following assumption: (cf. Assumption 3.1)

ASSUMPTION 3.2
We assume that the zeros $1/y_1, \ldots, 1/y_L$ of (3.43) are all distinct.

Under this assumption it is easily seen that we can write, cf. (3.27),

$$M_n = A_1 y_1^n + \cdots + A_L y_L^n, \quad n = 1,2,\ldots, \tag{3.44}$$

with $A_1, \ldots, A_L$ determined by

$$\frac{A_1 y_1}{1-y_1 z} + \cdots + \frac{A_L y_L}{1-y_L z} = M(z). \tag{3.45}$$

Remark 3.4
$1/y_1, \ldots, 1/y_L$ are the roots of a polynomial of degree $L \leqslant 1+\sum_{j=1}^{m} r_j$, see (3.11), (3.40); for example, for m-stage hyperexponential and m-stage Erlang service time distributions $L=m+1$. Note that (3.45) leads to a set of L linear equations from which $A_1, \ldots, A_L$ can be obtained (cf. Remark 3.2).

Analogously to the proof of (ii) and (iii) of Lemma 3.1 it can be shown that

Lemma 3.2
(i) *y_i can be written as*

$$y_i = 1-\beta d_i, \tag{3.46}$$

with d_i independent of β, $i=1,\ldots,L$;

(ii) *A_i is independent of β, $i=1,\ldots,L$.*

Note that, in fact, $d_i=(1-y_i)/\beta$, $i=1,\ldots,L$ are the roots of: (cf. (3.43) and the derivation of (3.31))

$$\tilde{z} + \omega_0 + \lambda - \lambda\sum_{j=1}^{m}\alpha_j\left[\frac{1}{1-\hat{\beta}_j\tilde{z}}\right]^{r_j} = 0. \tag{3.47}$$

The properties stated in Lemma 3.2 will be used below. Before treating the general case we first give an example.

The M/M/1 PS queue
For exponential service times ($Q(z)=(1-p)z/(1-pz)$, with $p=1-\beta/\hat{\beta}$)

$$M(z) = z\frac{1+\beta\omega_0-\lambda\beta\frac{z}{1-z}(1-(1-p)z/(1-pz))}{1-z(1+\beta\omega_0+\lambda\beta)+\lambda\beta z(1-p)z/(1-pz)}. \tag{3.48}$$

It is easily seen that the zeros $1/y_1$ and $1/y_2$ of the denominator of (3.48) are given by

$$\begin{aligned} y_1 &= \frac{1}{2}\left[1+\beta\omega_0+\lambda\beta+p+\sqrt{(1+\beta\omega_0+\lambda\beta+p)^2-4(p+p\beta\omega_0+\lambda\beta)}\right] \\ &= 1+\frac{1}{2}\beta\left[\omega_0+\lambda-1/\hat{\beta}+\sqrt{(\omega_0+\lambda-1/\hat{\beta})^2+4\omega_0/\hat{\beta}}\right], \end{aligned}$$

$$y_2 = \frac{1}{2}\left[1+\beta\omega_0+\lambda\beta+p-\sqrt{(1+\beta\omega_0+\lambda\beta+p)^2-4(p+p\beta\omega_0+\lambda\beta)}\right]$$

$$= 1+\frac{1}{2}\beta\left[\omega_0+\lambda-1/\hat{\beta}-\sqrt{(\omega_0+\lambda-1/\hat{\beta})^2+4\omega_0/\hat{\beta}}\right].$$

We can write (cf. (3.44))

$$M_n = A_1 y_1^n + A_2 y_2^n\ , \quad n=1,2,\ldots\ . \tag{3.49}$$

For the determination of A_1 and A_2 it is more convenient to use (3.5) instead of (3.45):

$$A_1+A_2 = M_0 = 1\ ,$$
$$A_1y_1+A_2y_2 = M_1 = 1+\beta\omega_0\ .$$

Hence

$$A_1 = \frac{y_2-(1+\beta\omega_0)}{y_2-y_1}\ , \qquad A_2 = \frac{y_1-(1+\beta\omega_0)}{y_1-y_2}\ .$$

Now substitute (3.49) into (3.4) and evaluate the summations in the denominator (take $\underset{\wedge}{q}(i)=p^{i-1}$). Taking in the resulting expressions $y_i=1-\beta d_i$, $i=1,2$, $p=1-\beta/\beta$, $\beta=x/k$ and using that d_i is independent of β it is easily seen that

$$\lim_{k\to\infty}(1+\beta\omega_0)M_{k-1} = \sum_{h=1}^{2}A_h e^{-xd_h}\ ,$$

$$\lim_{k\to\infty}\lambda\beta\sum_{j=1}^{k-2}q(k-j-1)M_j = \lambda\sum_{h=1}^{2}A_h\hat{\beta}\frac{1}{1-\hat{\beta}d_h}[e^{-xd_h}-e^{-x/\hat{\beta}}]\ ,$$

$$\lim_{k\to\infty}\lambda\beta\sum_{i=1}^{k-2}q(i) = \lambda\hat{\beta}(1-e^{-x/\hat{\beta}})\ .$$

Hence, cf. (3.4),

$$E\{e^{-\omega_0\mathbf{S}^{PS}(x)}\} = \lim_{k\to\infty}E\{e^{-\omega_0\mathbf{S}^{(k)}}\} = \tag{3.50}$$

$$\frac{1-\rho}{\sum_{h=1}^{2}A_h e^{-xd_h}-\rho e^{-x/\hat{\beta}}-\lambda\sum_{h=1}^{2}A_h\hat{\beta}\frac{1}{1-\hat{\beta}d_h}[e^{-xd_h}-e^{-x/\hat{\beta}}]}\ .$$

It is easily shown that this result coincides with the result obtained in Coffman et al. [1970] (formula (30) on page 128). Note that formula (30) of that paper represents the LST of the distribution of the total *delay* of a customer with a specific service demand. To match it with our result it has to be multiplied by the LST of the required service time (given by $e^{-\omega_0 x}$).

The M/G/1 PS queue

Now we shall treat the general case, i.e. the case that the service times are determined by (3.39). Consider in the corresponding feedback queue the total sojourn time after k services given by (3.4). As in the M/M/1 case, we evaluate the terms $(1+\beta\omega_0)M_{k-1}$, $\lambda\beta\sum_{j=1}^{k-2} q(k-j-1)M_j$ and $\lambda\beta\sum_{i=1}^{k-2} q(i)$ in the denominator and take the limit $k \to \infty$ independently for each term. The first term is simple: from (3.44) and (3.46) it is easily seen that

$$\lim_{k\to\infty}(1+\beta\omega_0)M_{k-1} = \lim_{k\to\infty}(1+\frac{x}{k}\omega_0)\sum_{h=1}^{L} A_h(1-\frac{x}{k}d_h)^{k-1} = \sum_{h=1}^{L} A_h e^{-xd_h} . \qquad (3.51)$$

The second one needs more effort. Using (3.41)-(3.44) and (3.46) it is found after extensive calculations that

$$\lambda\beta\sum_{j=1}^{k-2} q(k-j-1)M_j =$$

$$\lambda\beta\sum_{j=1}^{k-2} M_j \sum_{n=1}^{m}\alpha_n \sum_{i=0}^{r_n-1}\binom{k-j-2}{r_n-1-i}(1-p_n)^{r_n-1-i} p_n^{k-j-2-(r_n-1-i)} =$$

$$\lambda\beta\sum_{h=1}^{L} A_h \sum_{n=1}^{m}\alpha_n \sum_{i=0}^{r_n-1}\sum_{j=1}^{k-2}\binom{k-j-2}{r_n-1-i}(1-p_n)^{r_n-1-i} p_n^{k-j-2-(r_n-1-i)} y_h^j =$$

$$\lambda\sum_{h=1}^{L} A_h \sum_{n=1}^{m}\alpha_n \sum_{i=0}^{r_n-1}\frac{(1-x/(k\hat{\beta}_n))^{k-2-(r_n-1-i)}}{(r_n-1-i)!}\times$$

$$\sum_{j=1}^{k-r_n-1+i}\frac{x}{k}(k-j-2)\cdots(k-j-2-(r_n-2-i))(x/(k\hat{\beta}_n))^{r_n-1-i}\left[\frac{1-xd_h/k}{1-x/(k\hat{\beta}_n)}\right]^j .$$

The last equality is obtained by substituting $p_n = 1-\beta/\hat{\beta}_n$, $\beta = x/k$ and noting that $\binom{k-j-2}{r_n-1-i} = 0$ if $k-j-2 < r_n-1-i$.

Using that, actually by definition,

$$\lim_{k\to\infty}\sum_{j=1}^{k-r_n-1+i}\frac{x}{k}(k-j-2)(k-j-3)\cdots(k-j-2-(r_n-2-i))\times \qquad (3.52)$$

$$(x/(k\hat{\beta}_n))^{r_n-1-i}\left[\frac{1-xd_h/k}{1-x/(k\hat{\beta}_n)}\right]^j =$$

$$\int_{s=0}^{x}(\frac{x}{\hat{\beta}_n}-\frac{s}{\hat{\beta}_n})^{r_n-1-i}e^{-s(d_h-1/\hat{\beta}_n)}ds ,$$

we obtain

$$\lim_{k\to\infty}\lambda\beta\sum_{j=1}^{k-2}q(k-j-1)M_j = \tag{3.53}$$

$$\lambda\sum_{h=1}^{L}A_h\sum_{n=1}^{m}\alpha_n\sum_{i=0}^{r_n-1}\frac{e^{-x/\hat{\beta}_n}}{(r_n-1-i)!}\int_{s=0}^{x}(\frac{x}{\hat{\beta}_n}-\frac{s}{\hat{\beta}_n})^{r_n-1-i}e^{-s(d_h-1/\hat{\beta}_n)}ds .$$

The evaluation of the third term is analogous to that of the second term:

$$\lambda\beta\sum_{i=1}^{k-2}q(i) = \lambda\beta\sum_{i=1}^{k-2}\sum_{n=1}^{m}\alpha_n\sum_{j=0}^{r_n-1}\left[\begin{matrix}i-1\\r_n-1-j\end{matrix}\right](1-p_n)^{r_n-1-j}p_n^{i-1-(r_n-1-j)} =$$

$$\lambda\sum_{n=1}^{m}\alpha_n\sum_{j=0}^{r_n-1}\frac{(1-x/(k\hat{\beta}_n))^{-(r_n-j)}}{(r_n-1-j)!}\times$$

$$\sum_{i=r_n-j}^{k-2}\frac{x}{k}(i-1)\cdots(i-1-(r_n-2-j))(x/(k\hat{\beta}_n))^{r_n-1-j}(1-x/(k\hat{\beta}_n))^i .$$

Hence, cf. (3.52),

$$\lim_{k\to\infty}\lambda\beta\sum_{i=1}^{k-2}q(i) = \lambda\sum_{n=1}^{m}\alpha_n\sum_{j=0}^{r_n-1}\frac{1}{(r_n-1-j)!}\int_{s=0}^{x}(\frac{s}{\hat{\beta}_n})^{r_n-1-j}e^{-s/\hat{\beta}_n}ds . \tag{3.54}$$

(In the derivation of (3.54) one recognizes the convergence of the binomial probability $\left[\begin{matrix}i-1\\r_n-1-j\end{matrix}\right](1-p_n)^{r_n-1-j}p_n^{i-1-(r_n-1-j)}$ to the Poisson probability $\frac{1}{(r_n-1-j)!}(\frac{s}{\hat{\beta}_n})^{r_n-1-j}e^{-s/\hat{\beta}_n}$, cf. Feller [1950, Ch. 6]; a similar phenomenon occurs in the derivation of (3.53)).

The integrals in (3.53) and (3.54) can be evaluated by noting that

$$\int_{s=0}^{x} s^n e^{-s/c} ds = n!c^{n+1}(1-e^{-x/c})-e^{-x/c}\sum_{j=0}^{n-1}\frac{n!}{(n-j)!}x^{n-j}c^{j+1} .$$

Using the resulting expressions and (3.51) we obtain from (3.4):

THEOREM 3.2
In the M/G/1 PS queue with service time LST given by (3.39), for Re $\omega_0 \geqslant 0$,

$$E\{e^{-\omega_0 \mathbf{S}^{PS}(x)}\} = \lim_{k\to\infty} E\{e^{-\omega_0 \mathbf{S}^{(k)}}\} = \tag{3.55}$$

$$(1-\rho)\left[\sum_{h=1}^{L} A_h e^{-xd_h} - \lambda\sum_{n=1}^{m}\alpha_n\hat{\beta}_n e^{-x/\hat{\beta}_n}\sum_{j=0}^{r_n-1}\sum_{i=0}^{j}\frac{(x/\hat{\beta}_n)^i}{i!} - \right.$$

$$\left.\lambda\sum_{h=1}^{L} A_h \sum_{n=1}^{m}\alpha_n\hat{\beta}_n(\frac{1}{1-\hat{\beta}_n d_h})^{r_n}\sum_{j=0}^{r_n-1}(1-\hat{\beta}_n d_h)^j(e^{-xd_h}-e^{-x/\hat{\beta}_n}\sum_{i=0}^{r_n-1-j}\frac{(x(1-\hat{\beta}_n d_h)/\hat{\beta}_n)^i}{i!})\right]^{-1} ,$$

with $d_1, . . . , d_L$ *the roots of (3.47) and* $A_1, . . . , A_L$ *determined by (3.45), cf. Remark 3.4;* d_h *and* A_h *are independent of* x, $h=1,...,L$.

For hyperexponentially (H_m) distributed service times ($r_j=1$, $j=1,...,m$, cf. (3.39)) (3.55) reduces to

$$E\{e^{-\omega_0 \mathbf{S}^{PS}(x)}\} = (1-\rho)\left[\sum_{h=1}^{L} A_h e^{-xd_h} - \lambda\sum_{n=1}^{m}\alpha_n\hat{\beta}_n e^{-x/\hat{\beta}_n} - \right. \tag{3.56}$$

$$\left.\lambda\sum_{h=1}^{L} A_h e^{-xd_h}\sum_{n=1}^{m}\alpha_n\hat{\beta}_n\frac{1}{1-\hat{\beta}_n d_h}(1-e^{-x(1-\hat{\beta}_n d_h)/\hat{\beta}_n})\right]^{-1} .$$

It is easily verified that for $m=1$ (the M/M/1 case) (3.56) coincides with (3.50).

REMARK 3.5
Our formulas for the variance ((3.34), (3.35)) and the LST ((3.55)) of the sojourn time are given in terms of the roots of a polynomial and the solution of a set of linear equations. The corresponding formulas presented in Yashkov [1987] are given in terms of multiple integrals. In general both types of formulas can only be evaluated numerically. For obtaining numerical results it seems in our case to be more convenient to use the feedback results (2.22), (2.26) and (3.4), (3.5) and to evaluate a finite number of steps of the limiting

procedure described in Subsection 3.3.1.

As we remarked at the beginning of this subsection the analysis above does not apply to the M/D/1 PS queue. In fact, the deterministic service times can be approximated by an Erlang-n distribution (for large n) but this leads to the problem of finding the roots of an $(n+1)$-th degree polynomial and the solution of a set of $n+1$ linear equations (cf. (3.44), (3.45)). Below we shall show how explicit formulas for the sojourn time in the M/D/1 PS queue can be easily obtained from the sojourn time in the M/M/1 queue with deterministic feedback analyzed in Chapter 2.

The M/D/1 PS queue
Consider the M/M/1 queue with deterministic feedback in which each customer receives exactly N services, see Subsection 2.5.2. Taking $N = \lceil \hat{\beta}/\beta \rceil$ and $\beta = x/k$ it is clear that the total sojourn time after k services in the feedback queue approaches, for $k \to \infty$, the sojourn time of a (special) customer with service demand x in the M/D/1 PS queue with service time $\hat{\beta}$. Application of this limiting procedure to (2.45) yields immediately the LST of the distribution of $\mathbf{S}^{PS}(x)$:

$$E\{e^{-\omega_0 \mathbf{S}^{PS}(x)}\} = \frac{(1-\lambda\hat{\beta})(\lambda+\omega_0)^2}{\omega_0^2 e^{x(\lambda+\omega_0)} + \lambda(\lambda+\omega_0)[1-\lambda\hat{\beta}+\omega_0(x-\hat{\beta})]+\lambda\omega_0}, \tag{3.57}$$

$$\text{Re } \omega_0 \geqslant 0, \quad x \leqslant \hat{\beta}.$$

This result has been obtained before by Ott [1984].

Remark 3.6
(3.57) holds only for $0 \leqslant x \leqslant \hat{\beta}$. To obtain a similar formula for $x > \hat{\beta}$ we need an explicit expression for $E\{e^{-\omega_0 \mathbf{S}^{(k)}}\}$ for $k > N$, cf. (2.45); but as noted in Remark 2.4 this seems to be impossible for the case of deterministic feedback. Ott's method also precludes the derivation of an explicit formula for $x > \hat{\beta}$ (cf. Remark 5.2 in Ott [1984]).

We conclude this section with a remark on the state of the PS system just after the departure of a tagged customer.

Remark 3.7
From Corollary 2.2 and application of the limiting procedure, see Subsection 3.3.1, it follows that for the M/G/1 PS queue the state of the system (the number of customers present *and* their residual service requests) just after the departure of a tagged customer who has received an amount $x \geqslant 0$ of service is described by the stationary distribution of the state of the system at an arbitrary epoch, *independent* of x. This result slightly extends Theorem 2.3 of Ott [1984]. Ott's theorem concerns only the distribution of the number of

customers at a departure epoch of a tagged customer with initial service demand x.

3.4 The M/G/1 queue with generalized processor sharing

3.4.1 Introduction

In the previous sections it has been shown that the M/G/1 PS queue can be considered as a limiting case of an M/M/1 queue with feedback. We exploited the well known product form property of the joint queue length distribution to derive some sojourn time characteristics in the feedback queue and we used these results to obtain corresponding performance measures in the PS queue. In this section we shall show how a completely similar method can be applied to analyze an interesting generalization of the PS service discipline denoted as 'generalized processor sharing' (GPS). The GPS service discipline, investigated by Cohen [1979], generalizes the PS discipline as follows: when there are j customers present in the system then the service rate for each of these customers is $f(j)>0$, i.e. during a small time Δt the attained service of each customer increases with $f(j)\Delta t$, $j=1,2,...$. Note that, when j customers are present, the capacity of the server (the total service rate) is equal to $jf(j)$, $j=1,2,...$. If $f(j)=1/j$ then the GPS model clearly reduces to the PS model. It will be shown below that results for the M/G/1 GPS queue can be easily obtained from the analysis of the M/M/1 feedback queue *with state dependent service rates.* Indeed, it is intuitively clear that when we choose the service rate in the feedback queue equal to $jf(j)$ when there are j customers present and let the mean service times $\beta \rightarrow 0$ $(p(i)\rightarrow 1)$ as described in Subsection 3.3.1 then the behaviour of the feedback queue approaches that of the GPS queue. The formal proof of this convergence is completely analogous to that of the PS case, see Section 3.1. To illustrate this new approach to the GPS service discipline we shall show how the following two basic GPS results can be easily obtained from corresponding results for the feedback queue with state dependent service rates. Let

$$\phi(n) := \left[\prod_{j=1}^{n} f(j)\right]^{-1}, \quad n=1,2,..., \tag{3.58}$$

$$:= 1, \quad n=0.$$

For the M/G/1 GPS queue, cf. Cohen [1979],

(i) the distribution of the number of customers $\mathbf{X}^{GPS}$ present in the system is given by

$$Pr\{\mathbf{X}^{GPS}=n\} = \frac{\frac{\rho^n}{n!}\phi(n)}{\sum_{j=0}^{\infty}\frac{\rho^j}{j!}\phi(j)}, \tag{3.59}$$

(ii) the mean conditional sojourn time, $E\{\mathbf{S}^{GPS}(x)\}$ of a customer with service demand x is given by

$$E\{\mathbf{S}^{GPS}(x)\} = x \frac{\sum_{n=0}^{\infty} \frac{\rho^n}{n!}\phi(n+1)}{\sum_{j=0}^{\infty} \frac{\rho^j}{j!}\phi(j)}, \tag{3.60}$$

where ρ denotes, as usual, the offered load to the system per unit of time.

The formulas (3.59) and (3.60) have been derived under the stability condition $\sum_{j=0}^{\infty} \frac{\rho^j}{j!}\phi(j) < \infty$. Note that both the queue length distribution and the mean conditional sojourn time are independent of the service time distribution apart from its first moment; $E\{\mathbf{S}^{GPS}(x)\}$ is linear in x. Apparently, the GPS service discipline generalizes similar properties of the M/G/1 PS queue, cf. (3.19).

In fact, Cohen [1979] studies the GPS discipline in a very general model of open and closed networks with different job classes. This general model contains e.g. the classical Erlang and Engset systems. For the analysis Cohen uses the technique of the supplementary variable. He obtains generalizations of known results for closed and open networks such as the product form and the insensitivity property of the probabilities of the network states, cf. Baskett et al. [1975]; sojourn time results are restricted to means. We have found that most of these results can also be obtained from the M/M/1 feedback queue with state dependent service rates or from networks of these feedback queues. In fact the GPS network can be considered as a limiting model of a network consisting of M/M/1 feedback queues with state dependent service rates and suitably chosen routing probabilities. Each feedback queue corresponds to a node in the GPS network. This network of feedback queues is contained in the well known class of product form networks analyzed by Baskett et al. [1975]. Application of the limiting procedure to their results yields Cohen's GPS results. Here we shall restrict ourself to the derivation of (3.59) and (3.60).

3.4.2 Analysis

Consider the M/M/1 feedback model described in Section 2.2 but with one difference: when there are j customers present in the system then the server works with a rate $\mu(j)$, $j = 1,2,\ldots$. Note that the amount of service that a customer receives during each pass is still exponentially distributed with mean β; only the speed with which he is served may change during time. The 'old' case is obtained by taking $\mu(j) \equiv 1$. As in Section 2.2 we start the analysis assuming that the feedback probabilities remain constant after a finite number of services, i.e. $p(i) \equiv p$, $i \geqslant N$ for some $N \geqslant 1$. (The notation introduced before is extended in an obvious way). Thus, the total number of different customer types is limited to N. It will appear later on that the results are independent of N; so, this assumption is no restriction. We define

$$\rho_i := \lambda\beta\prod_{j=1}^{i-1} p(j), \quad i=1,...,N-1,$$

$$\rho_N := \frac{\lambda\beta}{1-p}\prod_{j=1}^{N-1} p(j).$$

Thus, ρ_i is the offered load to the system per unit of time due to type-i customers. Obviously, the *total* offered load to the system per unit of time, ρ, is equal to $\sum_{i=1}^{N}\rho_i$.

The (stationary) joint queue length distribution is found from the general network results obtained by Baskett et al. [1975]:

$$Pr\{\mathbf{X}_1=x_1, \ldots, \mathbf{X}_N=x_N\} = \tag{3.61}$$

$$C\,\frac{(x_1+\cdots+x_N)!}{x_1!\cdots x_N!}\prod_{i=1}^{N}\rho_i^{x_i}\prod_{j=1}^{x_1+...+x_N}(\mu(j))^{-1},$$

with

$$C = \frac{1}{\sum_{m=0}^{\infty}\rho^m\prod_{j=1}^{m}(\mu(j))^{-1}}, \tag{3.62}$$

an empty product being one by definition. (Note, for the derivation of (3.61) and (3.62), that in Baskett et al. [1975] the service rate is defined as the mean number of customers that can be served per unit of time; multiplication by β yields our definition of the service rate.)

It follows from (3.61) and (3.62) that the distribution of the *total* number of customers in the system is given by

$$Pr\{\mathbf{X}=n\} = Pr\{\mathbf{X}_1+\cdots+\mathbf{X}_N=n\} = \frac{\rho^n\prod_{j=1}^{n}(\mu(j))^{-1}}{\sum_{m=0}^{\infty}\rho^m\prod_{j=1}^{m}(\mu(j))^{-1}}, \tag{3.63}$$

$$n=0,1,\ldots .$$

The mean number of type-h customers in the system can also be obtained from (3.61) and (3.62):

$$E\{\mathbf{X}_h\} = \tag{3.64}$$

$$C\sum_{x_1=0}^{\infty}\cdots\sum_{x_N=0}^{\infty}x_h\frac{(x_1+\cdots+x_N)!}{x_1!\cdots x_N!}\prod_{i=1}^{N}\rho_i^{x_i}\prod_{j=1}^{x_1+\ldots+x_N}(\mu(j))^{-1} =$$

$$\rho_h C\sum_{x_1=0}^{\infty}\cdots\sum_{x_N=0}^{\infty}\frac{(x_1+\cdots+x_N+1)!}{x_1!\cdots x_N!}\prod_{i=1}^{N}\rho_i^{x_i}\prod_{j=1}^{x_1+\ldots+x_N+1}(\mu(j))^{-1} =$$

$$\rho_h C\sum_{m=0}^{\infty}(m+1)\rho^m\prod_{j=1}^{m+1}(\mu(j))^{-1}\ , \quad h=1,\ldots,N.$$

Using Little's formula it follows immediately that the mean duration of the h-th sojourn time, $E\{\mathbf{S}_h\}$, of a customer is given by

$$E\{\mathbf{S}_h\} = \beta C\sum_{m=0}^{\infty}(m+1)\rho^m\prod_{j=1}^{m+1}(\mu(j))^{-1}\ , \quad h=1,\ldots,N. \tag{3.65}$$

Note that both (3.63) and (3.65) are independent of N. In addition $E\{\mathbf{S}_h\}$ is independent of h. Hence, the mean total sojourn time after k services is linear in k:

$$E\{\mathbf{S}^{(k)}\} = k\frac{\beta\sum\limits_{m=0}^{\infty}(m+1)\rho^m\prod\limits_{j=1}^{m+1}(\mu(j))^{-1}}{\sum\limits_{m=0}^{\infty}\rho^m\prod\limits_{j=1}^{m}(\mu(j))^{-1}}\ , \quad k=1,2,\ldots . \tag{3.66}$$

Now, choose the feedback probabilities such that the total required service time has the same distribution as the service time in the GPS queue, cf. (3.14)-(3.16). Substitution of $\mu(j)=jf(j)$ in (3.63) and (3.66) and application of the limiting procedure described in Subsection 3.3.1 yield immediately the GPS results (3.59) and (3.60).

Although our approach to the GPS queue yields simple derivations of the results previously obtained by Cohen [1979] it seems to be very hard to derive new results such as the sojourn time distribution. The problem is the untractability of the distribution of the successive sojourn times of a customer in the feedback queue with state dependent service rate. For the derivation of the distribution of the sojourn time in the feedback queue with constant service rate, which led to the sojourn time distribution in the PS queue, we used the property that the joint process of successive departure epochs and queue length vectors at these departure epochs is a Markov renewal process, cf. the derivation of (2.8). However, this property does not hold in the feedback queue with state dependent service rate. Indeed, in the latter model a customer's sojourn time does not only depend on the number of customers of each type present at the beginning of the previous sojourn time and the *number* of new arrivals during it but it depends also on the *epochs* at which these arrivals occur. Moreover, the sojourn time is also dependent on the order of the different types of customers in the queue.

3.5 The M/G/1 processor sharing queue with feedback

3.5.1 Introduction

In this section we consider an M/G/1 PS queue with feedback. The feedback mechanism has the same structure as described in Chapter 2 for the M/M/1 FCFS queue, i.e. the probability that a customer is fed back after completing his service may depend on the number of times he has already been served. We shall study the successive sojourn times of a tagged customer. In particular we are interested in dependencies between these sojourn times.

The PS queue with feedback has been studied before by Klutke et al. [1988]. They consider the special case of Bernoulli feedback and analyze the behaviour of the internal input and output processes. In particular they study the influence of the shape of the service time distribution on the interoutput time distribution. Their main result is that when service time distributions with the same mean are convexly ordered (see Stoyan [1983]), so are interoutput time distributions. The purpose of their study is to gain insight into the properties of traffic processes in general queueing networks with processor sharing nodes.

In Klutke et al. [1988] it is remarked that the study of flow processes is crucial to understanding the behaviour of more complicated processes in the system. As an example the authors mention the sojourn time process and say that *"this is still an open problem"*. In this section we shall show that sojourn time results for the M/G/1 PS queue with feedback can be obtained from the sojourn time results for the M/M/1 FCFS feedback queue derived in Chapter 2.

3.5.2 Model description and notations

We consider an M/G/1 PS queue with feedback (PSFB), see Fig. 3.1. When a customer in the system has completed his i-th service, he departs from the system with probability $1-\tilde{p}(i)$ and is fed back with probability $\tilde{p}(i)$, $i=1,2,...$. Fed back customers return instantaneously, i.e. due to the PS service discipline a returning customer is immediately taken into service again. The successive service requests $\tilde{\boldsymbol{\tau}}_1, \tilde{\boldsymbol{\tau}}_2, \ldots$ of a customer are independent random variables with distribution functions $\tilde{B}_1(\cdot), \tilde{B}_2(\cdot), \ldots$ and means $\tilde{\beta}_1, \tilde{\beta}_2, \ldots$ respectively. New customers arrive according to a Poisson process with intensity λ. Obviously, for stability it is required that the offered load $\rho = \lambda\sum_{j=1}^{\infty}((1-\tilde{p}(j))\prod_{i=1}^{j-1}\tilde{p}(i))(\tilde{\beta}_1 + \cdots + \tilde{\beta}_j) < 1.$

We are interested in the successive sojourn times $\tilde{\mathbf{S}}_1(T_1), \ldots, \tilde{\mathbf{S}}_N(T_N)$ of a (tagged) customer in the PSFB queue who requires at least $N \geqslant 1$ services of length $T_1, \ldots, T_N \geqslant 0$ respectively. In particular we shall derive an expression

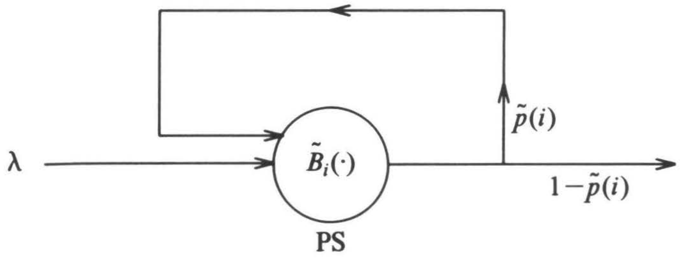

Fig. 3.1 The M/G/1 PS queue with feedback.

for the correlation coefficient, $corr(\tilde{\mathbf{S}}_i(T_i),\tilde{\mathbf{S}}_j(T_j))$, of the i-th and the j-th sojourn time of a tagged customer, $i,j=1,...,N$.

For the analysis of the successive sojourn times in the PSFB queue we shall consider corresponding sojourn times in an associated processor sharing queue *without* feedback. Let $B(\cdot)$ denote the distribution function of the total required service time, i.e.

$$\hat{B}(t) := \sum_{j=1}^{\infty}((1-\tilde{p}(j))\prod_{i=1}^{j-1}\tilde{p}(i))(\tilde{B}_1(t)* \cdots *\tilde{B}_j(t)), \quad t\geqslant 0. \tag{3.67}$$

It is easily seen that the behaviour of the M/G/1 PS queue with service time distribution $\hat{B}(\cdot)$ is exactly the same as the behaviour of the PSFB queue described above. In the sequel the PS queue with service time distribution $\hat{B}(\cdot)$ will be called "the associated PS queue" (or simply "the PS queue"). For a tagged customer with initial service demand $\boldsymbol{\tau}^{PS}\geqslant T_1+\cdots+T_N$, $T_1,\ldots,T_N\geqslant 0$, in the associated PS queue we define:

- $\mathbf{S}_i^{PS}(T_i)$: time during which the remaining service demand of the tagged customer is in the range $(\boldsymbol{\tau}^{PS}-\sum_{j=1}^{i}T_j,\boldsymbol{\tau}^{PS}-\sum_{j=1}^{i-1}T_j]$, $i=1,...,N$.

Obviously, the joint distribution of $\mathbf{S}_1^{PS}(T_1),\ldots,\mathbf{S}_i^{PS}(T_i)$ does not depend on $T_{i+1},\ldots,T_N$, $i=1,...,N-1$; $\mathbf{S}_1^{PS}(T_1)$ is distributed as the conditional sojourn time of a tagged customer with service demand T_1:

$$E\{e^{-\omega_0\mathbf{S}_1^{PS}(T_1)}\} = E\{e^{-\omega_0\mathbf{S}^{PS}(T_1)}\}, \quad T_1\geqslant 0, \quad \mathrm{Re}\,\omega_0\geqslant 0. \tag{3.68}$$

It is clear that the quantities $\mathbf{S}_i^{PS}(T_i)$, $i=1,...,N$ in the associated PS queue correspond to the successive sojourn times $\tilde{\mathbf{S}}_1(T_1),\ldots,\tilde{\mathbf{S}}_N(T_N)$ in the PSFB queue, i.e., for $t_1,\ldots,t_N\geqslant 0$,

$$Pr\{\tilde{\mathbf{S}}_1(T_1)<t_1,\ldots,\tilde{\mathbf{S}}_N(T_N)<t_N\} = \tag{3.69}$$

$$Pr\{\mathbf{S}_1^{PS}(T_1)<t_1,\ldots,\mathbf{S}_N^{PS}(T_N)<t_N\},\qquad T_1,\ldots,T_N\geqslant 0.$$

Specifically,

$$corr(\tilde{\mathbf{S}}_i(T_i),\tilde{\mathbf{S}}_j(T_j)) = corr(\mathbf{S}_i^{PS}(T_i),\mathbf{S}_j^{PS}(T_j)),\quad i,j=1,\ldots,N. \tag{3.70}$$

So below we shall focus on the sojourn times $\mathbf{S}_i^{PS}(T_i)$, $T_i\geqslant 0$, $i=1,\ldots,N$, in the PS queue.

3.5.3 Analysis

Consider the M/G/1 PS queue with service time distribution $\hat{B}(\cdot)$. We assume that $\hat{B}(\cdot)$ belongs to the class of phase-type distributions given by (3.12). The first moment of $\hat{B}(\cdot)$ is denoted by $\hat{\beta}$. From Remark 3.7 it follows immediately that for $2\leqslant i\leqslant N$ the joint distribution of $\mathbf{S}_i^{PS}(T_i),\ldots,\mathbf{S}_N^{PS}(T_N)$ does not depend on T_j, $j=1,\ldots,i-1$; in fact Remark 3.7 implies that, cf. (3.68),

$$E\{e^{-\omega_0\mathbf{S}_i^{PS}(T_i)}\} = E\{e^{-\omega_0\mathbf{S}^{PS}(T_i)}\},\quad T_i\geqslant 0,\ i=1,\ldots,N,\quad \mathrm{Re}\,\omega_0\geqslant 0. \tag{3.71}$$

Hence means are simply given by, see (3.19),

$$E\{\mathbf{S}_i^{PS}(T_i)\} = \frac{T_i}{1-\rho},\quad i=1,\ldots,N, \tag{3.72}$$

with offered load $\rho=\lambda\hat{\beta}$. It also follows that $corr(\mathbf{S}_i^{PS}(T_i),\mathbf{S}_j^{PS}(T_j))$, $T_1,\ldots,T_N\geqslant 0$ depends only on T_i, T_j and $\sum_{n=i+1}^{j-1}T_n$, $1\leqslant i<j\leqslant N$. Hence, for the analysis of $corr(\mathbf{S}_i^{PS}(T_i),\mathbf{S}_j^{PS}(T_j))$, $T_1,\ldots,T_N\geqslant 0$, $i,j=1,\ldots,N$ we can restrict ourself to the determination of $corr(\mathbf{S}_1^{PS}(T_1),\mathbf{S}_3^{PS}(T_3))$, $T_1,T_2,T_3\geqslant 0$ without loss of generality. Below we shall derive an expression for $corr(\mathbf{S}_1^{PS}(T_1),\mathbf{S}_3^{PS}(T_3))$, $T_1,T_2,T_3\geqslant 0$. We shall consider corresponding sojourn times in the M/M/1 FCFS feedback queue and apply the limiting procedure described in Subsection 3.3.1. The analysis is largely analogous to the derivation of the sojourn time variance in the M/G/1 PS queue, see Subsection 3.3.3.

Consider the M/M/1 FCFS feedback queue with mean service time β and feedback probabilities $p(i)$, $i=1,2,\ldots$ related with β such that the total required service time has distribution function $\hat{B}(\cdot)$, see (3.14)-(3.16). We follow a tagged customer during his first $k=k_1+k_2+k_3$ successive sojourn times $\mathbf{S}_1,\ldots,\mathbf{S}_k$. Define

$$\mathbf{S}_1(k_1):=\mathbf{S}_1+\cdots+\mathbf{S}_{k_1},$$

$$\mathbf{S}_2(k_2):=\mathbf{S}_{k_1+1}+\cdots+\mathbf{S}_{k_1+k_2},$$

$$\mathbf{S}_3(k_3):=\mathbf{S}_{k_1+k_2+1}+\cdots+\mathbf{S}_{k_1+k_2+k_3}.$$

Clearly, when we take $k_2 = \lceil T_2/\beta \rceil$, $k_3 = \lceil T_3/\beta \rceil$, $\beta = T_1/k_1$ and let $k_1 \to \infty$ then $\mathbf{S}_1(k_1)$, $\mathbf{S}_2(k_2)$ and $\mathbf{S}_3(k_3)$ correspond to the PS quantities $\mathbf{S}_1^{PS}(T_1)$, $\mathbf{S}_2^{PS}(T_2)$ and $\mathbf{S}_3^{PS}(T_3)$ respectively (cf. Subsection 3.3.1; note that, for $k_1 \to \infty$, $k_i\beta \to T_i$, $i = 1,2,3$). We shall first derive $corr(\mathbf{S}_1(k_1),\mathbf{S}_3(k_3))$ for general $k_1,k_2,k_3 \geqslant 0$. Next, taking k_2, k_3 and β as indicated above we use

$$corr(\mathbf{S}_1^{PS}(T_1),\mathbf{S}_3^{PS}(T_3)) = \lim_{k_1 \to \infty} corr(\mathbf{S}_1(k_1),\mathbf{S}_3(k_3)). \tag{3.73}$$

From the definition of $\mathbf{S}_i(k_i)$, $i = 1,2,3$, it follows that the covariance of $\mathbf{S}_1(k_1)$ and $\mathbf{S}_3(k_3)$ can be written as

$$cov(\mathbf{S}_1(k_1),\mathbf{S}_3(k_3)) = \sum_{i=1}^{k_1}\sum_{j=1}^{k_3} cov(\mathbf{S}_i,\mathbf{S}_{k_1+k_2+j}), \quad k_1,k_2,k_3 \geqslant 0. \tag{3.74}$$

The covariance of $\mathbf{S}_i$ and $\mathbf{S}_j$ is found from (2.20) and (2.24):

$$cov(\mathbf{S}_i,\mathbf{S}_j) = \left[\frac{\beta}{1-\rho}\right]^2 (1-(1-\rho)C_{j-i}), \quad 1 \leqslant i < j \leqslant k, \tag{3.75}$$

with C_n, $n = 1,\dots,k$ as in Subsection 3.3.3, see (3.27)-(3.30).
Substituting (3.75) into (3.74) and writing C_n as in (3.27) it follows that, for $k_1,k_2,k_3 \geqslant 0$,

$$\begin{aligned} cov(\mathbf{S}_1(k_1),\mathbf{S}_3(k_3)) &= \left[\frac{\beta}{1-\rho}\right]^2 \sum_{i=1}^{k_1}\sum_{j=1}^{k_3}(1-(1-\rho)C_{k_1+k_2+j-i}) \qquad (3.76) \\ &= \left[\frac{\beta}{1-\rho}\right]^2 \left[k_1k_3 - (1-\rho)\sum_{i=1}^{k_1}\sum_{j=1}^{k_3}\sum_{l=1}^{L} U_l x_l^{k_1+k_2+j-i-1}\right] \\ &= -\frac{\beta^2}{1-\rho}\sum_{l=2}^{L} U_l x_l^{k_1+k_2-1} \sum_{i=1}^{k_1} x_l^{-i} \sum_{j=1}^{k_3} x_l^j \\ &= -\frac{\beta^2}{1-\rho}\sum_{l=2}^{L} U_l x_l^{k_2}\left(\frac{x_l^{k_1}-x_l^{-1}}{x_l-1} - x_l^{k_1-1}\right)\left(\frac{1-x_l^{k_3+1}}{1-x_l} - 1\right). \end{aligned}$$

The third equality of (3.76) follows from (iii) of Lemma 3.1. Replacing in (3.76) x_l by $1-\beta a_l$, $l = 2,\dots,L$, see (3.30), we obtain

$$cov(\mathbf{S}_1(k_1),\mathbf{S}_3(k_3)) = -\frac{\beta^2}{1-\rho} \times \tag{3.77}$$

$$\sum_{l=2}^{L} U_l (1-\beta a_l)^{k_2} \left(\frac{(1-\beta a_l)^{k_1} - (1-\beta a_l)^{-1}}{-\beta a_l} - (1-\beta a_l)^{k_1 -1} \right) \left(\frac{1-(1-\beta a_l)^{k_3+1}}{\beta a_l} - 1 \right),$$

$$k_1, k_2, k_3 \geqslant 0.$$

Now, taking in (3.77) appropriate limits, i.e. $k_2 = \lceil T_2 / \beta \rceil$, $k_3 = \lceil T_3 / \beta \rceil$, $\beta = T_1 / k_1$ and $k_1 \to \infty$, we find, cf. (3.73):

$$cov(\mathbf{S}_1^{PS}(T_1), \mathbf{S}_3^{PS}(T_3)) = \tag{3.78}$$

$$-\frac{1}{1-\rho} \sum_{l=2}^{L} U_l (1/a_l)^2 e^{-T_2 a_l} (1-e^{-T_1 a_l})(1-e^{-T_3 a_l}), \quad T_1, T_2, T_3 \geqslant 0.$$

Hence, from (3.71) and (3.34),

$$corr(\mathbf{S}_1^{PS}(T_1), \mathbf{S}_3^{PS}(T_3)) = \tag{3.79}$$

$$\frac{-\sum_{l=2}^{L} U_l (1/a_l)^2 e^{-T_2 a_l} (1-e^{-T_1 a_l})(1-e^{-T_3 a_l})}{2 \left[\sum_{l=2}^{L} (1/a_l)^2 U_l (1 - T_1 a_l - e^{-T_1 a_l}) \right]^{\frac{1}{2}} \left[\sum_{l=2}^{L} (1/a_l)^2 U_l (1 - T_3 a_l - e^{-T_3 a_l}) \right]^{\frac{1}{2}}},$$

$$T_1, T_2, T_3 \geqslant 0.$$

Returning to the PS queue with feedback we have from (3.79) (cf. (3.70), the discussion below (3.72) and Lemma 3.1):

THEOREM 3.3
For the successive sojourn times $\tilde{\mathbf{S}}_1(T_1), \ldots, \tilde{\mathbf{S}}_N(T_N)$, $T_1, \ldots, T_N \geqslant 0$, *of a tagged customer in the M/G/1 PSFB queue with total service request LST given by (3.12),*

$$corr(\tilde{\mathbf{S}}_i(T_i), \tilde{\mathbf{S}}_j(T_j)) = \tag{3.80}$$

$$\frac{-\sum_{l=2}^{L} U_l (1/a_l)^2 e^{-T_{i,j} a_l} (1-e^{-T_i a_l})(1-e^{-T_j a_l})}{2 \left[\sum_{l=2}^{L} (1/a_l)^2 U_l (1 - T_i a_l - e^{-T_i a_l}) \right]^{\frac{1}{2}} \left[\sum_{l=2}^{L} (1/a_l)^2 U_l (1 - T_j a_l - e^{-T_j a_l}) \right]^{\frac{1}{2}}},$$

$$1 \leqslant i < j \leqslant N,$$

with $T_{i,j} := \sum_{n=i+1}^{j-1} T_n$. $a_2, \ldots, a_L$ *are the roots of (3.31) and* $U_2, \ldots, U_L$ *are determined by (3.29), cf. Remark 3.2;* a_l *and* U_l *are independent of* T_n, $n = 1,...,N$, $l = 2,...,L$.

It is interesting to consider some asymptotic properties of $corr(\tilde{\mathbf{S}}_i(T_i), \tilde{\mathbf{S}}_j(T_j))$. First, noting that in (3.80) $\mathrm{Re}\, a_l > 0$, $l = 2,...,L$, see Lemma 3.1, we obtain

$$corr(\tilde{\mathbf{S}}_i(T_i), \tilde{\mathbf{S}}_j(T_j)) \to 0, \quad T_i, T_j \geqslant 0,\ T_{i,j} \to \infty,\ 1 \leqslant i < j \leqslant N, \tag{3.81}$$

which is intuitively clear. Another asymptotic result applies to the case that T_i, T_j and $T_{i,j}$ become very small. Using $\sum_{l=2}^{L} U_l = 1 - 1/(1-\rho)$ and (3.34), (3.38) it follows from (3.80) that

$$corr(\tilde{\mathbf{S}}_i(T_i), \tilde{\mathbf{S}}_j(T_j)) \to 1, \quad T_i, T_j, T_{i,j} \to 0,\ 1 \leqslant i < j \leqslant N. \tag{3.82}$$

This result can be explained as follows. Suppose a tagged customer starts his i-th service at time t. For T_i, T_j and $T_{i,j}$ close to zero it may be expected that the successive sojourn times $\tilde{\mathbf{S}}_i(T_i), \ldots, \tilde{\mathbf{S}}_j(T_j)$ of the tagged customer are small (cf. (3.72)) and no new arrivals or departures occur during the time interval $[t, t + \tilde{\mathbf{S}}_i(T_i) + ... + \tilde{\mathbf{S}}_j(T_j)]$. Hence, due to the PS service discipline $\tilde{\mathbf{S}}_j(T_j) = T_j \tilde{\mathbf{S}}_i(T_i) / T_i$, i.e. $\tilde{\mathbf{S}}_j(T_j)$ is completely determined by $\tilde{\mathbf{S}}_i(T_i)$.

We conclude this section with an example.

The M/M/1 PS queue with Bernoulli feedback

Consider the M/M/1 PS queue with Bernoulli feedback, i.e. $\tilde{B}(t) = 1 - e^{-t/\tilde{\beta}}$, $\tilde{p}(i) \equiv p$, $0 \leqslant p < 1$, see Subsection 3.5.2. For this case the total required service time is exponentially distributed with mean $\hat{\beta} = \tilde{\beta}/(1-p)$, cf. (3.67). From the calculations for the determination of the sojourn time variance in the M/M/1 PS queue, see Subsection 3.3.3, we have in (3.80) $L = 2$, $U_2 = -1/(1-\rho)$, $a_2 = (1-x_2)/\beta = (1-\rho)/\hat{\beta}$. Hence, for the M/M/1 queue with Bernoulli feedback (3.80) reduces to

$$corr(\tilde{\mathbf{S}}_i(T_i), \tilde{\mathbf{S}}_j(T_j)) = \tag{3.83}$$

$$\frac{e^{-T_{i,j}(1-\rho)/\hat{\beta}}(1 - e^{-T_i(1-\rho)/\hat{\beta}})(1 - e^{-T_j(1-\rho)/\hat{\beta}})}{2\left[e^{-T_i(1-\rho)/\hat{\beta}} - 1 + T_i(1-\rho)/\hat{\beta}\right]^{\frac{1}{2}} \left[e^{-T_j(1-\rho)/\hat{\beta}} - 1 + T_j(1-\rho)/\hat{\beta}\right]^{\frac{1}{2}}}, \quad 1 \leqslant i < j \leqslant N.$$

Chapter 4

SIMPLE APPROXIMATIONS FOR SECOND MOMENT CHARACTERISTICS OF THE SOJOURN TIME IN THE M/G/1 PROCESSOR SHARING QUEUE

4.1 Introduction

Although in literature considerable attention has been paid to the exact analysis of the sojourn time in the M/G/1 PS queue, little work has been done on the investigation of the practical implications of the results. The expressions for the second moment and the LST of the sojourn time distribution obtained by Yashkov [1983], Ott [1984] and Schassberger [1984], see also (3.34), (3.35) and (3.55), are complex and not very attractive for practical applications. Only for the mean sojourn time a simple explicit expression exists; this expression is insensitive to the service time distribution apart from its first moment, see (3.19). The formulas for the second moment of the sojourn time require perfect information about the service time distribution, which is almost never available in practice. Moreover, in general these formulas can only be evaluated numerically. As far as we know no attention has been paid to the derivation of approximations or asymptotic formulas which are useful for practical evaluation, apart from a paper by Yashkov [1986]. Yashkov derives some asymptotic estimates for the conditional sojourn time variance for customers with small or large service times. We have obtained similar results in Section 3.3, see (3.37) and (3.38). In particular the asymptotic formula (3.38), for customers with small service times, slightly generalizes Yashkov's result, cf. Remark 3.3.

Actually, our interest in approximations for second moment characteristics of the sojourn time in the M/G/1 PS queue started with the derivation of (3.37) and (3.38). We found that these asymptotic formulas yield reasonable estimates for a wide range of the parameter values. The discovery of simple bounds for the second moment of the sojourn time also stimulated the investigation of approximations. Indeed, noting that in (3.7) $C_n \geqslant 1$, $n = 1,2,...$, it is easily seen that, for the total sojourn time after k services in the feedback queue,

$$var(\mathbf{S}^{(k)}) \leqslant \frac{\rho}{(1-\rho)^2}(k\beta)^2 \ , \quad k=1,2,\ldots .$$

Hence, applying the limiting procedure described in Section 3.3,

$$var(\mathbf{S}^{PS}(x)) \leqslant \frac{\rho}{(1-\rho)^2}x^2 \ . \tag{4.1}$$

Note that this bound depends on the service time distribution only through its first moment. Obviously, (4.1) implies an upper bound for the unconditional sojourn time variance which depends on the first two moments of the service time distribution. After having noticed the existence of these bounds we found out that they can also be easily obtained from the results in Yashkov [1983] and Ott [1984]. However, neither Yashkov nor Ott points at this interesting property of the sojourn time variance in the PS queue.

The aim of the present study is to derive approximations for the second moment of the sojourn time distribution, which are quite simple and yet accurate enough for most practical purposes. Some very simple approximation formulas based on the first and second moment of the service time are presented. The accuracy of the approximations is tested for a large number of different service time distributions and a wide range of traffic intensities. A refinement of the approximation is obtained by taking the third moment of the service time into account. This refinement yields remarkably accurate results with relative errors less than 1.5 percent in most cases.

The organization of the rest of this chapter is as follows. In Section 4.2 we introduce the notations and give a summary of those known sojourn time results which are relevant for our study. We also present some extensions and new results. In particular the heavy traffic behaviour of the second moment of the sojourn time is derived. Section 4.3 is concerned with the second moment of the *conditional* sojourn time of a customer with service demand x. We show that the asymptotic result (3.38) yields reasonable approximations for a wide range of x values. In Section 4.4 approximations are developed for the second moment of the *unconditional* sojourn time. We first propose an approximation which uses only information about the first and second moment of the service time distribution (Subsection 4.4.1). In Subsection 4.4.2 we construct a more detailed (and more accurate) approximation formula, which is based on the first three moments of the service time distribution.

4.2 NOTATIONS AND PRELIMINARY RESULTS

We consider an M/G/1 PS queue with arrival rate λ and service time distribution $\hat{B}(\cdot)$ with first and second moment $\hat{\beta}$ and $\hat{\beta}_2$. It is assumed that $\rho := \lambda\hat{\beta} < 1$ and that the system is in steady state. Since confusion with the sojourn time in the feedback queue is not possible we use in the rest of this chapter $\mathbf{S}(x)$ and $\mathbf{S}$ instead of $\mathbf{S}^{PS}(x)$ and $\mathbf{S}^{PS}$ to denote the conditional and unconditional sojourn time in the PS queue.

The sojourn time formulas obtained by Yashkov [1983] and Ott [1984] are more suitable for heavy traffic analysis than our expressions derived in Chapter 3. They have for the second moment of the conditional sojourn time,

$$E\{\mathbf{S}^2(x)\} = \frac{x^2}{(1-\rho)^2} + \frac{2}{(1-\rho)^2}\int_{t=0}^{x}(x-t)(1-R(t))\,dt\ , \tag{4.2}$$

where $R(t)$ represents the waiting time distribution for the M/G/1 first come first served (FCFS) queue with service time distribution $\hat{B}(\cdot)$,

$$R(t) = (1-\rho)\sum_{n=0}^{\infty}\rho^n F^{n^*}(t)\ , \tag{4.3}$$

$$F(t) = \frac{1}{\hat{\beta}}\int_{u=0}^{t}(1-\hat{B}(u))\,du\ .$$

Note, for the waiting time distribution $R(t)$ in (4.2), that $1-R(t) \leqslant 1-R(0)=\rho$, $t \geqslant 0$. Hence, cf. (4.1),

$$E\{\mathbf{S}^2(x)\} \leqslant \frac{1+\rho}{(1-\rho)^2}x^2\ . \tag{4.4}$$

A lower bound for $E\{\mathbf{S}^2(x)\}$ follows immediately from the mean sojourn time, given by (3.19), and Schwartz' inequality (or alternatively from (4.2)),

$$E\{\mathbf{S}^2(x)\} \geqslant \frac{x^2}{(1-\rho)^2}\ . \tag{4.5}$$

So,

$$\frac{1}{(1-\rho)^2}x^2 \leqslant E\{\mathbf{S}^2(x)\} \leqslant \frac{1+\rho}{(1-\rho)^2}x^2\ . \tag{4.6}$$

Note that the upper bound is $100\rho\%$ higher than the lower bound and that these bounds depend only on the mean service time; this supports a certain robustness of $E\{\mathbf{S}^2(x)\}$ for the service time distribution.

The heavy traffic behaviour of $E\{\mathbf{S}^2(x)\}$ can be derived from (4.2) by noting that the heavy traffic behaviour of the waiting time distribution for the M/G/1 FCFS queue is, for $\rho \to 1$, negative exponential, i.e. (see Cohen [1982, p. 596])

$$R(\frac{t}{1-\rho}) \sim 1-e^{-t/d}\ , \qquad \text{for } \rho\to 1\ , \tag{4.7}$$

where $d=\frac{1}{2}\rho\hat{\beta}_2/\hat{\beta}$.

Substituting (4.7) into (4.2) yields

$$E\{\mathbf{S}^2(x)\} \sim \frac{1+\rho}{(1-\rho)^2}x^2, \qquad \text{for } \rho\to 1. \tag{4.8}$$

The asymptotic behaviour of $E\{\mathbf{S}^2(x)\}$ for $x\to 0$ is given by (3.38) and (3.19):

$$E\{\mathbf{S}^2(x)\} \sim \frac{1+\rho}{(1-\rho)^2}x^2 - \frac{\rho}{3\hat{\beta}(1-\rho)}x^3, \qquad \text{for } x\to 0. \tag{4.9}$$

For exponential and deterministic service times, simple explicit expressions for $E\ S^2(x)$ exist. For future use we state these expressions. From (3.23) and (3.57) it follows that

$$E\{\mathbf{S}^2(x)\}_{EXP} = \frac{2\rho\hat{\beta}}{(1-\rho)^3}x - \frac{2\rho\hat{\beta}^2}{(1-\rho)^4}(1-e^{-x(1-\rho)/\hat{\beta}}), \qquad x\geqslant 0, \tag{4.10}$$

$$E\{\mathbf{S}^2(x)\}_{DET} = \frac{2}{(1-\rho)^2}x^2 - \frac{2\hat{\beta}^2}{\rho^2(1-\rho)}(e^{\rho x/\hat{\beta}}-1-\rho x/\hat{\beta}), \quad 0\leqslant x\leqslant\hat{\beta}. \tag{4.11}$$

From (3.19) and the above results for $\mathbf{S}(x)$ it follows immediately that for the unconditional sojourn time $\mathbf{S}$, cf. (3.18),

$$E\{\mathbf{S}\} = \frac{\hat{\beta}}{1-\rho}, \tag{4.12}$$

$$\frac{\hat{\beta}_2}{(1-\rho)^2} \leqslant E\{\mathbf{S}^2\} \leqslant \frac{1+\rho}{(1-\rho)^2}\hat{\beta}_2, \tag{4.13}$$

$$E\{\mathbf{S}^2\} \sim \frac{1+\rho}{(1-\rho)^2}\hat{\beta}_2, \qquad \text{for } \rho\to 1. \tag{4.14}$$

For exponential and deterministic service times,

$$E\{\mathbf{S}^2\}_{EXP} = (1+\frac{2+\rho}{2-\rho})\frac{\hat{\beta}^2}{(1-\rho)^2}, \tag{4.15}$$

$$E\{\mathbf{S}^2\}_{DET} = \frac{2\hat{\beta}^2}{(1-\rho)^2} - \frac{2\hat{\beta}^2}{\rho^2(1-\rho)}(e^\rho - 1 - \rho)\,. \tag{4.16}$$

REMARK 4.1
(4.13) implies that, for the M/G/1 PS queue, the dependence of $E\{\mathbf{S}^2\}$ on the third moment of the service time distribution is limited. This should be contrasted with the behaviour of the second moment of the sojourn time distribution for the M/G/1 FCFS queue. For the FCFS discipline it depends linearly on the third moment of the service time distribution, see e.g. Cohen [1982].

In the following sections the above results are exploited to develop simple approximations for $E\{\mathbf{S}^2(x)\}$ and $E\{\mathbf{S}^2\}$. We present extensive tables comparing the approximations with exact values. The service time distributions which we have chosen to test the approximations are:

- exponential distribution
- deterministic distribution
- k-stage Erlang distribution (E_k)
- two-stage hyperexponential distribution (H_2), in particular
  H_2 with balanced means (H_2^{BM}), and
  H_2 with gamma normalization (H_2^{GN})
- two-stage Coxian distribution (C_2)
- three-stage hyperexponential distribution (H_3)

These types of service time distributions are often used for practical applications in queueing theory, see Tijms [1986] and Whitt [1982, 1984].

In practice service times are often characterized by the mean, $\hat{\beta}$, and the squared coefficient of variation, c^2, defined by

$$c^2 = \frac{\sigma^2}{\hat{\beta}^2}\,,$$

where σ^2 denotes the service time variance, see Tijms [1986]. Here we shall use c^2 rather than σ^2 to characterize the variability of the service times.

The H_2^{BM} and H_2^{GN} distributions have been introduced to reduce the number of parameters of the H_2 distribution, see Tijms [1986]; they are uniquely determined by their first two moments. In particular, the H_2^{GN} distribution with mean $\hat{\beta}$ and $c^2 \geqslant 1$ has the same third moment as the gamma distribution with mean $\hat{\beta}$ and squared coefficient of variation c^2. In Section 4.4 the class of H_2 distributions will be considered in more detail.

The tables presented at the end of this chapter contain relative errors of the approximations for various service time distributions. The relative error is defined as

$$100\%\ \frac{\textit{approximation result} - \textit{exact result}}{\textit{exact result}}.$$

The exact values of $E\{\mathbf{S}^2(x)\}$ and $E\{\mathbf{S}^2\}$ have been obtained from (3.34) and

(3.35). For H_k and C_k (and E_k) service time distributions these formulas require the roots of a polynomial of degree k and the solution of a set of k linear equations. Even for the case $k=2$, the resulting expressions are very large and complicated and do not give much insight into the influence of the parameters.

4.3 Approximation of $E\{\mathbf{S}^2(x)\}$

In this section we show that the asymptotic result (4.9) yields a good approximation for $E\{\mathbf{S}^2(x)\}$ for an important range of x-values.

We define (cf. (4.9)),

$$E\{\mathbf{S}^2(x)\}_{APPX} := \frac{1+\rho}{(1-\rho)^2}x^2 - \frac{\rho}{3\hat{\beta}(1-\rho)}x^3 . \tag{4.17}$$

Note that $E\{\mathbf{S}^2(x)\}_{APPX}$ satisfies the heavy traffic behaviour of $E\{\mathbf{S}^2(x)\}$ (see (4.8)) and that $E\{\mathbf{S}^2(x)\}_{APPX}$ is smaller than the upper bound of $E\{\mathbf{S}^2(x)\}$ given by (4.4). Approximation $E\{\mathbf{S}^2(x)\}_{APPX}$ is independent of the service time distribution apart from its first moment. Obviously it can not be applied for too large values of x because it becomes negative for $x>3\hat{\beta}(1+\rho)/(\rho(1-\rho))$. Moreover, assuming that the variance of $\mathbf{S}(x)$ is a convex function of x (cf. (4.9) and (3.37)), we may not expect that $E\{\mathbf{S}^2(x)\}_{APPX}$ is a good approximation for $x>x_1$, where $x_1=\hat{\beta}/(1-\rho)=E\{\mathbf{S}\}$ is the point of inflection of (cf. (3.38))

$$f(x) = \frac{\rho}{(1-\rho)^2}x^2 - \frac{\rho}{3\hat{\beta}(1-\rho)}x^3 . \tag{4.18}$$

For $x<x_1$, $E\{\mathbf{S}^2(x)\}_{APPX}$ is within the bounds of $E\{\mathbf{S}^2(x)\}$ given by (4.6).

In Table 4.1 approximation results are compared with exact results for a number of different values of x ($x=\frac{1}{2}\hat{\beta}, \hat{\beta}, \frac{3}{2}\hat{\beta}, 2\hat{\beta}, \frac{\hat{\beta}/2}{1-\rho}, \frac{\hat{\beta}}{1-\rho}$) and different service time distributions. For each of these cases ρ varies from 0.1 to 0.9. For the sake of clarity only the relative approximation errors are given. It appears that for most cases the relative approximation errors are negative. As expected, the approximation becomes less accurate when x grows. For $0\leqslant x\leqslant\hat{\beta}$ the relative errors are less than 2.34% in absolute value. For $0\leqslant x\leqslant 2\hat{\beta}$ the maximum relative error is 6.56%. When x remains constant the maximum errors occur for $\rho\approx 0.3$. For $x=\hat{\beta}/(2(1-\rho))$ and $x=\hat{\beta}/(1-\rho)$ the relative errors tend to increase when ρ grows. For $x=\hat{\beta}/(1-\rho)$ the maximum error is 11.29%.

It is seen from the results for different service time distributions that the accuracy of the approximation tends to decrease when c^2 becomes larger.

REMARK 4.2
In Van den Berg [1988] we have derived an approximation for $E\{\mathbf{S}^2(x)\}$ for the *whole* range of possible x-values ($x \geqslant 0$) by appropriately combining the two asymptotic formulas (4.9) and (3.37). The idea is as follows. Two values $\tilde{x}_1$ and $\tilde{x}_2$ are determined, such that for $x \leqslant \tilde{x}_1$ (4.9) yields a good approximation and for $x \geqslant \tilde{x}_2$ (3.37) yields a good approximation. For $\tilde{x}_1 \leqslant x \leqslant \tilde{x}_2$, $E\{\mathbf{S}^2(x)\}$ is approximated by the term $x^2/(1-\rho)^2$ plus a linear function of x, cf. (3.37). We took $\tilde{x}_1 = \hat{\beta}/(1-\rho)$. Details about the determination of $\tilde{x}_2$ are given in Van den Berg [1988]. The approximation yields reasonably good results for service time distributions with c^2 not too large ($0 \leqslant c^2 \leqslant 2$). For these cases we found relative errors which are typically less than 10%. A minor drawback of this approximation is that it needs the first *three* moments of the service time distribution (cf. (3.37)).

4.4 APPROXIMATION OF $E\{\mathbf{S}^2\}$

In this section we propose two different approximations for the second moment of the unconditional sojourn time $\mathbf{S}$. First we derive a simple approximation which is based on the exact formula of $E\{\mathbf{S}^2\}$ for the case of deterministic service times and for exponentially distributed service times. This approximation uses only the first two moments of the service time distribution. Next it is shown how this simple approximation can be improved. We derive a (second) approximation based on exact expressions of $E\{\mathbf{S}^2\}$ for two classes of H_2 distributions. This latter approximation also takes the *third* moment of the service time distribution into account.

4.4.1 Simple approximation

It follows from (4.13) that an approximation $E\{\mathbf{S}^2\}_{APP}$ of $E\{\mathbf{S}^2\}$, which satisfies

$$\frac{\hat{\beta}_2}{(1-\rho)^2} \leqslant E\{\mathbf{S}^2\}_{APP} \leqslant \frac{1+\rho}{(1-\rho)^2}\hat{\beta}_2\,, \tag{4.19}$$

yields relative errors which are bounded by $100\rho\%$ in absolute value. This observation and the relations for $E\{\mathbf{S}^2\}$ given in Section 4.2 support the idea to derive an approximation for $E\{\mathbf{S}^2\}$ which is based only on the first two moments of the service time distribution. We propose an approximation which is a linear interpolation on the service time squared coefficient of variation such that it yields exact results for the case of exponential and deterministic service times. This type of two-moment approximations is often used to estimate performance measures (e.g. mean sojourn times) in (complex) queueing systems, see Tijms [1986, Ch. 4]. The rationale behind it is that the Pollaczek-Khinchine formula for the mean sojourn time $E\{\mathbf{S}^{FCFS}\}$ in the M/G/1 FCFS queue allows the representation

$$E\{\mathbf{S}^{FCFS}\} = c^2 E\{\mathbf{S}^{FCFS}\}_{EXP} + (1-c^2)E\{\mathbf{S}^{FCFS}\}_{DET}\ , \tag{4.20}$$

where $E\{\mathbf{S}^{FCFS}\}_{EXP}$ and $E\{\mathbf{S}^{FCFS}\}_{DET}$ denote the mean sojourn time for the special cases of exponential and deterministic service times (with the same means). Note that this kind of representation is also allowed for the mean sojourn time in the M/G/1 PS queue, cf. (4.12). For the present case of the *second* moment of the sojourn time distribution in the M/G/1 PS queue this idea leads to an approximation $E\{\mathbf{S}^2\}_{APP1}$ for $E\{\mathbf{S}^2\}$ which reads as follows:

$$E\{\mathbf{S}^2\}_{APP1} = c^2 E\{S^2\}_{EXP} + (1-c^2)E\{\mathbf{S}^2\}_{DET} = \tag{4.21}$$

$$c^2(1+\frac{2+\rho}{2-\rho})\frac{\hat{\beta}^2}{(1-\rho)^2} + (1-c^2)(\frac{2\hat{\beta}^2}{(1-\rho)^2} - \frac{2\hat{\beta}^2}{\rho^2(1-\rho)}(e^{\rho}-1-\rho)).$$

Note that this approximation has the following appealing properties:

APPROXIMATION PROPERTIES

(1) The approximation is exact for deterministic service times.
(2) The approximation is exact for exponentially distributed service times.
(3) The approximation yields values between the lower and upper bound of $E\{\mathbf{S}^2\}$ given by (4.13).
(4) The approximation satisfies the heavy traffic behaviour of $E\{\mathbf{S}^2\}$ (see (4.14)).
(5) The approximation yields the exact value of $E\{\mathbf{S}^2\}$ for $\rho=0$:
$E\{\mathbf{S}^2\}_{APP} = E\{\mathbf{S}^2\} = \hat{\beta}_2\ ,\quad$ for $\rho=0$.

The approximation results for the test set of service time distributions and traffic intensities are presented in Table 4.2. It appears that the approximation yields reasonably good results. In all tested cases the relative approximation error is smaller than 5%. In particular for service time distributions with c^2 close to one ($0 \leqslant c^2 \leqslant 2$) the relative errors are less than 1.89%. Obviously this small error is due to the fact that the approximation is exact for exponential and deterministic service times. For larger values of c^2 ($c^2 > 4$) the approximation becomes worse. It is noticeable that the approximation is significantly better for the H_2 distribution with gamma normalization (H_2^{GN}) than for the H_2 distribution with balanced means (H_2^{BM}). In the next subsection we shall show that this is due to the influence of the third moment of the service time distribution on $E\{\mathbf{S}^2\}$.

4.4.2 Detailed approximation

The simple approximation (4.21) tends to be less accurate if the squared coefficient of variation of the service time distribution becomes larger.

Therefore, for service time distributions with $c^2 \geqslant 1$ we shall develop a new approximation, $APP2$, for $E\{\mathbf{S}^2\}$. This approximation is based on simple exact formulas for two classes of extreme H_2 distributions. It contains the first three moments of the service time distribution.

We start with recalling some characteristics of the class of H_2 distributions. The H_2 distribution function is given by

$$B_{H_2}(t) = \alpha(1-e^{-t/\hat\beta^{(1)}}) + (1-\alpha)(1-e^{-t/\hat\beta^{(2)}}), \tag{4.22}$$

where $0 \leqslant \alpha \leqslant 1$, $0 \leqslant \hat\beta^{(1)} \leqslant \hat\beta^{(2)}$.
So, there are three parameters. Given the mean $\hat\beta = \alpha\hat\beta^{(1)} + (1-\alpha)\hat\beta^{(2)}$ and $c^2 \geqslant 1$ there is thus one remaining degree of freedom, r, defined by

$$r = \frac{\alpha\hat\beta^{(1)}}{\alpha\hat\beta^{(1)} + (1-\alpha)\hat\beta^{(2)}} .$$

$r = 1/2$ yields the class of H_2 distributions with balanced means (H_2^{BM}).

Obviously, if $\hat\beta$ and c^2 are given, r determines the third moment, $\hat\beta_3$, of the H_2 distribution. For fixed $\hat\beta$ and $\hat\beta_2$ (c^2), the smallest possible value of $\hat\beta_3$ is obtained for $r=0$. In that case $\hat\beta_3 = \frac{3}{2}\,\hat\beta_2^2/\hat\beta$. For $r \to 1$, $\hat\beta_3 \to \infty$ (see Whitt [1982,1984]).

Our numerical experience with respect to H_2 distributions indicates that $E\{\mathbf{S}^2\}$ becomes smaller when $\hat\beta_3$ grows ($\hat\beta$ and $\hat\beta_2$ constant). So (cf. (4.13)), we expect that $E\{\mathbf{S}^2\}_{H_2}$ has a limit for $\hat\beta_3 \to \infty$, $\hat\beta$ and $\hat\beta_2$ fixed. From (3.35) it is found that, for $\hat\beta$ and $\hat\beta_2$ fixed,

$$E\{\mathbf{S}^2\}_{H_2^{r=1}} = \lim_{\hat\beta_3 \to \infty (r \to 1)} E\{\mathbf{S}^2\}_{H_2} = \frac{\hat\beta_2}{(1-\rho)^2} + \frac{2\rho}{2-\rho}\frac{\hat\beta^2}{(1-\rho)^2} , \tag{4.23}$$

and, for $\hat\beta_3 = \frac{3}{2}\hat\beta_2^2/\hat\beta$ $(r=0)$,

$$E\{\mathbf{S}^2\}_{H_2^{r=0}} = (1+\frac{\rho}{2-\rho})\frac{\hat\beta_2}{(1-\rho)^2} . \tag{4.24}$$

It is easily seen that, for $c^2 \geqslant 1$,

$$E\{\mathbf{S}^2\}_{H_2^{r=1}} \leqslant E\{\mathbf{S}^2\}_{H_2^{r=0}} , \qquad 0 \leqslant \rho \leqslant 1 . \tag{4.25}$$

In (4.25), equality holds if $c^2 = 1$ ($\hat\beta_2 = 2\hat\beta^2$), hence if the service times are exponentially distributed.

Now we introduce two approximation assumptions to extend the above results with respect to H_2 distributions to general service time distributions.

Assumption 1: $E\{\mathbf{S}^2\}$ depends only on the first three moments ($\hat{\beta}$, $\hat{\beta}_2$, $\hat{\beta}_3$) of the service time distribution.
Assumption 2: $E\{\mathbf{S}^2\}$ decreases if $\hat{\beta}_3$ grows ($\hat{\beta}$ and $\hat{\beta}_2$ fixed).

Under these assumptions it follows from (4.23) and (4.24) that, for $c^2 \geqslant 1$, $\hat{\beta}_3 \geqslant \frac{3}{2}\hat{\beta}_2^2/\hat{\beta}$,

$$\frac{\hat{\beta}_2}{(1-\rho)^2} + \frac{2\rho}{2-\rho}\frac{\hat{\beta}^2}{(1-\rho)^2} \leqslant E\{\mathbf{S}^2\} \leqslant (1+\frac{\rho}{2-\rho})\frac{\hat{\beta}_2}{(1-\rho)^2} . \tag{4.26}$$

(4.23), (4.24) and (4.26) suggest an approximation, $APP2$, for $E\{\mathbf{S}^2\}$ which reads as follows:

$$E\{\mathbf{S}^2\}_{APP2} = \gamma(1+\frac{\rho}{2-\rho})\frac{\hat{\beta}_2}{(1-\rho)^2} + (1-\gamma)(\frac{\hat{\beta}_2}{(1-\rho)^2} + \frac{2\rho}{2-\rho}\frac{\hat{\beta}^2}{(1-\rho)^2}), \tag{4.27}$$

where $\gamma := \gamma(\rho,\hat{\beta},\hat{\beta}_2,\hat{\beta}_3)$, $0 \leqslant \gamma \leqslant 1$.

The choice of the weight factor γ will be partially determined by the approximation properties listed below (4.21). Besides the properties (2)-(5) we require that

(6) for $\hat{\beta}$, $\hat{\beta}_2$ fixed,

$$\lim_{\hat{\beta}_3 \to \infty} E\ S^2{}_{APP2} = \frac{\hat{\beta}_2}{(1-\rho)^2} + \frac{2\rho}{2-\rho}\frac{\hat{\beta}^2}{(1-\rho)^2} ,$$

(7) for $\hat{\beta}_3 = \frac{3}{2}\hat{\beta}_2^2/\hat{\beta}$,

$$E\{\mathbf{S}^2\}_{APP2} = (1+\frac{\rho}{2-\rho})\frac{\hat{\beta}_2}{(1-\rho)^2} .$$

Note, that, without any further specification of γ, $APP2$ satisfies the approximation properties (2) and (5). Considering the other required properties ((3), (4), (6) and (7)) it is natural to choose γ as follows,

$$\gamma = \frac{1}{1+\gamma_1(\hat{\beta}_3 - \frac{3}{2}\hat{\beta}_2^2/\hat{\beta})(1-\rho)} , \tag{4.28}$$

where γ_1 represents the relative influence of $\hat{\beta}_3$ on $E\{\mathbf{S}^2\}$. γ_1 remains to be specified. We assume that γ_1 depends only on $\hat{\beta}$ and $\hat{\beta}_2$. Note that γ_1 has to be chosen such that γ is dimensionless. The most obvious choices are $\gamma_1 = 1/\hat{\beta}^3$ or $\gamma_1 = 1/(\hat{\beta}\hat{\beta}_2)$. For both cases we compared approximation results with exact results. Our test set consisted of H_2 service time distributions with c^2 ranging from 1 to 20. For each value of c^2 a large number of $\hat{\beta}_3$ values was considered. It appeared that the choice $\gamma_1 = 1/(\hat{\beta}\hat{\beta}_2)$ yields much better results than $\gamma_1 = 1/\hat{\beta}^3$. However, in most cases the choice $\gamma_1 = 1/(\hat{\beta}\hat{\beta}_2)$ underestimated $E\{\mathbf{S}^2\}$. In particular for larger values of c^2 the approximation results became worse. Extensive tests of the approximation for some variants of $\gamma_1 = 1/(\hat{\beta}\hat{\beta}_2)$ led to a modification which yields remarkably accurate results:

$$\gamma_1 = \frac{1}{(c^2-1)} \frac{1}{\hat{\beta}\hat{\beta}_2} .$$

So, the ultimate approximation formula is given by (4.27), with

$$\gamma = \frac{1}{1+(1-\rho)(\frac{\hat{\beta}_3}{\hat{\beta}\hat{\beta}_2} - \frac{3}{2}\frac{\hat{\beta}_2}{\hat{\beta}^2})/(\frac{\hat{\beta}_2}{\hat{\beta}^2} - 2)} . \tag{4.29}$$

It is seen from Table 4.3 that for H_2^{BM} and H_2^{GN} service time distributions (with $c^2 = 2$, 4, 6) $APP2$ yields very accurate results with relative errors less than 1%.

Table 4.4 illustrates the influence of $\hat{\beta}_3$ on $E\{\mathbf{S}^2\}$. This table shows exact values of $E\{\mathbf{S}^2\}$ for a number of H_2 distributions with the same first and second moment but with a different third moment. The traffic intensity varies from 0.1 to 0.95. The relative approximation errors of $APP2$ are indicated below the exact values of $E\{\mathbf{S}^2\}$. As we stated before $E\{\mathbf{S}^2\}$ decreases when $\hat{\beta}_3$ grows. Note that even for large c^2 ($c^2 = 10$) the relative approximation errors are less than 1.5%. It may be concluded from Table 4.4 that the influence of the third moment of the service time distribution on $E\{\mathbf{S}^2\}$ increases when c^2 grows, cf. (4.13).

In Table 4.5 $APP2$ is tested for some arbitrarily chosen H_3 and C_2 service time distributions. The relative errors are in all cases less than 1.5%.

Originally, $APP2$ has been developed for service time distributions with $c^2 \geqslant 1$, $\hat{\beta}_3 \geqslant \frac{3}{2} \hat{\beta}_2^2/\hat{\beta}$. For these cases $E\{\mathbf{S}^2\}_{APP2}$ can be interpreted as an interpolation formula, see (4.27). Nevertheless, approximation formula (4.27) (together with (4.29)) can be applied to service time distributions with $c^2 < 1$ or $\hat{\beta}_3 < \frac{3}{2} \hat{\beta}_2^2/\hat{\beta}$ as well. In Table 4.6 some results are shown for deterministic, C_2 and E_k service time distributions with $c^2 < 1$. It appears that the accuracy of $APP2$ for these cases is about the same as the accuracy of $APP1$.

4.5 Conclusions

In this section we briefly review the results and sum up the main characteristics of the different approximation formulas proposed in Section 4.3 and Section 4.4.

We have studied the second moment of the conditional and unconditional sojourn time, $E\{\mathbf{S}^2(x)\}$ and $E\{\mathbf{S}^2\}$, for the M/G/1 processor sharing queue. An upper bound and some asymptotic properties (like the heavy traffic behaviour) have been derived. Based on these properties and on exact expressions for specific service time distributions we developed some simple approximations. The approximations have been compared with exact results for a large number of different service time distributions and a wide range of traffic intensities. We conclude as follows.

- The influence of the third and higher moments of the service time distribution on $E\{\mathbf{S}^2(x)\}$ and $E\{\mathbf{S}^2\}$ is limited. An upper and a lower bound for $E\{\mathbf{S}^2(x)\}$ can be expressed in terms of x (the service demand of a tagged customer) and the traffic intensity ρ, see (4.6). The corresponding upper and lower bound for $E\{\mathbf{S}^2\}$ contain only the second moment of the sojourn time distribution and ρ, see (4.13).
- Approximation $APPX$ for $E\{\mathbf{S}^2(x)\}$, given by (4.17), is based on the asymptotic result (3.38) for $x \to 0$. It depends on the service time distribution only through its first moment. $APPX$ yields reasonably good results for not too large values of x, see Table 4.1. For $0 \leqslant x \leqslant \hat{\beta}$ the relative error of the approximation is a few percent. The approximation becomes less accurate when x increases. For $x = 2\hat{\beta}$ the relative errors are typically less than 7%. $APPX$ satisfies the heavy traffic behaviour of $E\ S^2(x)$, see (4.8).
- The approximations for $E\{\mathbf{S}^2\}$, $APP\,1$ and $APP\,2$, given by (4.21) and (4.27), have been constructed in such a way that they have the following appealing properties:
  * they are exact for exponential service times
  * they yield values between the lower and upper bound of $E\{\mathbf{S}^2\}$
  * they satisfy the heavy traffic behaviour of $E\{\mathbf{S}^2\}$
  * they yield the exact value of $E\{\mathbf{S}^2\}$ for $\rho=0$.

In addition, $APP\,1$ yields exact results for deterministic service times; $APP\,2$ is exact for two classes of extreme H_2 distributions.

- Approximation $APP\,1$ is the most simple approximation. It depends on the first two moments of the service time distribution. For not too large values of c^2 ($c^2 \leqslant 6$) it yields fairly accurate results, see Table 4.2. In practical situations $APP\,1$ may be applied as a first order approximation for $E\{\mathbf{S}^2\}$.
- $APP\,2$ depends on the first three moments of the service time distribution. It is based on exact formulas of $E\{\mathbf{S}^2\}$ for two classes of extreme H_2 distributions. The details of the construction of $APP\,2$ are rather heuristic. Nevertheless, it yields remarkably accurate results. $APP\,2$ has been tested for a large number of different service time distributions with c^2 ranging from 0 to 10, see Tables 4.3 through 4.6. In all of these cases the relative error is less than 1.5%.

TABLE 4.1. Approximation of $E\{\mathbf{S}^2(x)\}$. The table contains relative errors (%) of approximation APPX (given by (4.17)) for various service time distributions.

Service time distribution: H_2^{BM}, $cv = 2$.

ρ	$x=\frac{1}{2}\hat{\beta}$	$x=\hat{\beta}$	$x=\frac{3}{2}\hat{\beta}$	$x=2\hat{\beta}$	$x=\frac{1}{2}\frac{\hat{\beta}}{1-\rho}$	$x=\frac{\hat{\beta}}{1-\rho}$
0.1	-0.19	-0.66	-1.35	-2.18	-0.23	-0.80
0.3	-0.32	-1.16	-2.40	-3.94	-0.62	-2.20
0.5	-0.27	-0.99	-2.08	-3.46	-0.99	-3.46
0.7	-0.15	-0.56	-1.19	-1.99	-1.44	-4.80
0.9	-0.04	-0.15	-0.30	-0.51	-2.31	-6.61

Service time distribution: H_2^{BM}, $cv = 4$.

ρ	$x=\frac{1}{2}\hat{\beta}$	$x=\hat{\beta}$	$x=\frac{3}{2}\hat{\beta}$	$x=2\hat{\beta}$	$x=\frac{1}{2}\frac{\hat{\beta}}{1-\rho}$	$x=\frac{\hat{\beta}}{1-\rho}$
0.1	-0.22	-0.77	-1.55	-2.46	-0.27	-0.93
0.3	-0.39	-1.39	-2.83	-4.58	-0.76	-2.61
0.5	-0.35	-1.26	-2.58	-4.23	-1.26	-4.23
0.7	-0.21	-0.77	-1.60	-2.64	-1.93	-6.11
0.9	-0.06	-0.23	-0.48	-0.79	-3.39	-8.98

Service time distribution: H_2^{BM}, $cv = 6$.

ρ	$x=\frac{1}{2}\hat{\beta}$	$x=\hat{\beta}$	$x=\frac{3}{2}\hat{\beta}$	$x=2\hat{\beta}$	$x=\frac{1}{2}\frac{\hat{\beta}}{1-\rho}$	$x=\frac{\hat{\beta}}{1-\rho}$
0.1	-0.24	-0.82	-1.63	-2.59	-0.29	-0.99
0.3	-0.42	-1.50	-3.02	-4.86	-0.81	-2.78
0.5	-0.38	-1.37	-2.80	-4.56	-1.37	-4.56
0.7	-0.24	-0.86	-1.78	-2.92	-2.14	-6.67
0.9	-0.07	-0.27	-0.55	-0.91	-3.85	-10.01

Service time distribution: exponential.

ρ	$x=\frac{1}{2}\hat{\beta}$	$x=\hat{\beta}$	$x=\frac{3}{2}\hat{\beta}$	$x=2\hat{\beta}$	$x=\frac{1}{2}\frac{\hat{\beta}}{1-\rho}$	$x=\frac{\hat{\beta}}{1-\rho}$
0.1	-0.14	-0.53	-1.11	-1.85	-0.17	-0.64
0.3	-0.23	-0.86	-1.86	-3.17	-0.45	-1.70
0.5	-0.17	-0.66	-1.46	-2.53	-0.66	-2.53
0.7	-0.08	-0.30	-0.67	-1.19	-0.83	-3.19
0.9	-0.01	-0.04	-0.09	-0.16	-0.96	-3.74

TABLE 4.1 (Cont'd)

Service time distribution: H_2^{GN}, $cv = 2$.

ρ	$x=\frac{1}{2}\hat{\beta}$	$x=\hat{\beta}$	$x=\frac{3}{2}\hat{\beta}$	$x=2\hat{\beta}$	$x=\frac{1}{2}\frac{\hat{\beta}}{1-\rho}$	$x=\frac{\hat{\beta}}{1-\rho}$
0.1	-0.25	-0.83	-1.58	-2.45	-0.30	-0.98
0.3	-0.45	-1.51	-2.91	-4.56	-0.85	-2.69
0.5	-0.41	-1.38	-2.68	-4.21	-1.38	-4.21
0.7	-0.26	-0.87	-1.69	-2.64	-1.99	-5.66
0.9	-0.08	-0.27	-0.52	-0.79	-2.82	-7.15

Service time distribution: H_2^{GN}, $cv = 4$.

ρ	$x=\frac{1}{2}\hat{\beta}$	$x=\hat{\beta}$	$x=\frac{3}{2}\hat{\beta}$	$x=2\hat{\beta}$	$x=\frac{1}{2}\frac{\hat{\beta}}{1-\rho}$	$x=\frac{\hat{\beta}}{1-\rho}$
0.1	-0.34	-1.07	-1.98	-2.98	-0.41	-1.26
0.3	-0.64	-2.03	-3.79	-5.75	-1.17	-3.52
0.5	-0.61	-1.96	-3.68	-5.61	-1.96	-5.61
0.7	-0.41	-1.33	-2.50	-3.81	-2.93	-7.72
0.9	-0.14	-0.46	-0.86	-1.30	-4.33	-10.02

Service time distribution: H_2^{GN}, $cv = 6$.

ρ	$x=\frac{1}{2}\hat{\beta}$	$x=\hat{\beta}$	$x=\frac{3}{2}\hat{\beta}$	$x=2\hat{\beta}$	$x=\frac{1}{2}\frac{\hat{\beta}}{1-\rho}$	$x=\frac{\hat{\beta}}{1-\rho}$
0.1	-0.38	-1.18	-2.16	-3.23	-0.45	-1.39
0.3	-0.72	-2.26	-4.17	-6.29	-1.31	-3.88
0.5	-0.69	-2.21	-4.11	-6.22	-2.21	-6.21
0.7	-0.48	-1.53	-2.85	-4.32	-3.33	-8.63
0.9	-0.17	-0.54	-1.00	-1.52	-4.98	-11.29

Service time distribution: E_2.

ρ	$x=\frac{1}{2}\hat{\beta}$	$x=\hat{\beta}$	$x=\frac{3}{2}\hat{\beta}$	$x=2\hat{\beta}$	$x=\frac{1}{2}\frac{\hat{\beta}}{1-\rho}$	$x=\frac{\hat{\beta}}{1-\rho}$
0.1	-0.03	-0.24	-0.68	-1.32	-0.05	-0.32
0.3	-0.00	-0.24	-0.86	-1.88	-0.06	-0.75
0.5	0.07	0.05	-0.26	-0.91	0.05	-0.91
0.7	0.11	0.27	0.33	0.24	0.32	-0.84
0.9	0.06	0.19	0.34	0.48	0.85	-0.53

TABLE 4.2. Approximation of $E\{\mathbf{S}^2\}$. The table contains relative errors (%) of approximation APP1 (given by (4.21)) for various service time distributions.

Service time distribution: H_2^{BM}.

ρ	$cv=2$	$cv=4$	$cv=6$
0.10	0.29	0.52	0.62
0.30	0.91	1.67	2.01
0.50	1.52	2.88	3.49
0.70	1.89	3.65	4.48
0.90	1.26	2.50	3.11
0.95	0.75	1.50	1.87

Service time distribution:H_2^{GN}.

ρ	$cv=2$	$cv=4$	$cv=6$
0.10	-0.13	-0.22	-0.28
0.30	-0.29	-0.52	-0.61
0.50	-0.30	-0.54	-0.63
0.70	-0.17	-0.27	-0.31
0.90	0.00	0.05	0.07
0.95	0.01	0.05	0.08

Service time distributions with $c^2<1$.

ρ	E_4	E_3	E_2	$C_2^{(1)}$	$C_2^{(2)}$
0.10	0.18	0.14	0.13	0.00	0.00
0.30	0.35	0.35	0.28	0.00	-0.02
0.50	0.38	0.38	0.32	-0.07	-0.09
0.70	0.25	0.24	0.21	-0.13	-0.11
0.90	0.05	-0.01	0.04	-0.11	-0.07
0.95	0.02	-0.01	0.02	-0.08	-0.04

$C_2^{(1)}$: $cv=0.75$, $C_2^{(2)}$: $cv=0.92$.

TABLE 4.3. Approximation of $E\{\mathbf{S}^2\}$. The table contains relative errors (%) of approximation APP2 (given by (4.27) and (4.29)) for various service time distributions with $c^2 \geqslant 1$, $\hat{\beta}_3 \geqslant \frac{3}{2}\hat{\beta}_2^2 / \hat{\beta}$.

Service time distribution: H_2^{BM}.

ρ	$cv=2$	$cv=4$	$cv=6$
0.10	-0.15	-0.27	-0.32
0.30	-0.32	-0.58	-0.69
0.50	-0.31	-0.53	-0.61
0.70	-0.12	-0.09	-0.04
0.90	0.08	0.34	0.51
0.95	0.07	0.26	0.39

Service time distribution: H_2^{GN}.

ρ	$cv=2$	$cv=4$	$cv=6$
0.10	-0.12	-0.21	-0.25
0.30	-0.23	-0.41	-0.48
0.50	-0.17	-0.30	-0.35
0.70	-0.01	0.01	0.03
0.90	0.09	0.20	0.26
0.95	0.07	0.14	0.18

TABLE 4.4. The influence of the third moment of the service time distribution ($\hat{\beta}_3$) on $E\{\mathbf{S}^2\}$. In the table the exact values of $E\{\mathbf{S}^2\}$ are given. The relative approximation errors (%) of $APP2$ are indicated in parentheses below the exact values of $E\{\mathbf{S}^2\}$.

H_2 service time distributions with $\hat{\beta}=1$, $cv=4$.

$\hat{\beta}_3$	$\rho=0.10$	$\rho=0.30$	$\rho=0.50$	$\rho=0.70$	$\rho=0.90$	$\rho=0.95$
37.500*	6.498	12.00	26.67	85.47	909.1	3810
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
49.084	6.434	11.68	25.65	82.10	889.2	3762
	(-0.25)	(-0.50)	(-0.40)	(-0.01)	(0.26)	(0.19)
68.329	6.388	11.43	24.78	78.74	864.3	3698
	(-0.26)	(-0.58)	(-0.56)	(-0.15)	(0.34)	(0.29)
105.42	6.353	11.24	24.02	75.35	830.6	3609
	(-0.19)	(-0.47)	(-0.54)	(-0.28)	(0.24)	(0.26)
190.02	6.329	11.09	23.41	72.17	785.3	3441
	(-0.11)	(-0.29)	(-0.38)	(-0.31)	(0.01)	(0.08)
310.86	6.318	11.02	23.12	70.47	751.6	3295
	(-0.07)	(-0.19)	(-0.26)	(-0.25)	(-0.13)	(-0.09)
716.53	6.309	10.97	22.86	68.85	709.3	3063
	(-0.03)	(-0.08)	(-0.12)	(-0.14)	(-0.17)	(-0.22)
1391.8	6.306	10.95	22.77	68.21	689.0	2927
	(-0.02)	(-0.04)	(-0.06)	(-0.08)	(-0.13)	(-0.20)
2291.9	6.305	10.94	22.73	67.94	679.6	2856
	(-0.01)	(-0.03)	(-0.04)	(-0.05)	(-0.09)	(-0.16)
4541.9	6.304	10.93	22.70	67.74	671.9	2794
	(-0.00)	(-0.01)	(-0.02)	(-0.03)	(-0.05)	(-0.10)
∞	6.303	10.92	22.67	67.52	663.6	2724
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

* $\hat{\beta}_3 = \frac{3}{2}\hat{\beta}_2^2/\hat{\beta}$

TABLE 4.4 (Cont'd)

H_2 service time distributions with $\hat{\beta}=1$, $cv=10$.

$\hat{\beta}_3$	$\rho=0.10$	$\rho=0.30$	$\rho=0.50$	$\rho=0.70$	$\rho=0.90$	$\rho=0.95$
181.50*	14.29 (0.00)	26.41 (0.00)	58.67 (0.00)	188.0 (0.00)	2000 (0.00)	8381 (0.00)
223.21	14.17 (-0.29)	25.80 (-0.50)	56.78 (-0.35)	181.9 (0.05)	1965 (0.28)	8298 (0.19)
330.00	14.01 (-0.40)	24.95 (-0.80)	53.89 (-0.68)	171.3 (0.03)	1891 (0.70)	8110 (0.53)
502.34	13.90 (-0.31)	24.34 (-0.73)	51.62 (-0.72)	161.6 (-0.04)	1802 (1.00)	7859 (0.86)
710.16	13.84 (-0.24)	24.00 (-0.60)	50.26 (-0.65)	155.0 (-0.09)	1724 (1.13)	7612 (1.09)
1323.2	13.78 (-0.14)	23.61 (-0.37)	48.65 (-0.44)	146.4 (-0.11)	1588 (1.13)	7093 (1.37)
1845.6	13.76 (-0.10)	23.49 (-0.27)	48.11 (-0.34)	143.2 (-0.10)	1523 (1.04)	6797 (1.42)
4162.2	13.73 (-0.05)	23.31 (-0.13)	47.31 (-0.17)	138.4 (-0.06)	1401 (0.69)	6129 (1.26)
12263	13.72 (-0.02)	23.22 (-0.04)	46.89 (-0.06)	135.7 (-0.03)	1315 (0.30)	5543 (0.71)
30488	13.71 (-0.01)	23.19 (-0.02)	46.76 (-0.02)	134.8 (-0.01)	1285 (0.13)	5305 (0.34)
∞	13.71 (0.00)	23.17 (0.00)	46.67 (0.00)	134.2 (0.00)	1264 (0.00)	5124 (0.00)

* $\hat{\beta}_3 = \frac{3}{2}\,\hat{\beta}_2^2/\hat{\beta}$

TABLE 4.5. Relative approximation errors (%) of $APP2$ for H_3 and C_2 service time distributions.

ρ	$H_3^{(1)}$	$H_3^{(2)}$	$C_2^{(1)}$	$C_2^{(2)}$	$C_2^{(3)}$
0.10	-0.20	-0.29	-0.15	-0.30	-0.32
0.30	-0.54	-0.71	-0.31	-0.63	-0.63
0.50	-0.65	-0.82	-0.25	-0.54	-0.48
0.70	-0.44	-0.48	-0.04	-0.04	0.05
0.90	-0.00	0.20	0.12	0.40	0.39
0.95	0.00	0.10	0.09	0.31	0.28

$H_3^{(1)}$: $cv = 2.778$, $\hat{\beta}_3 = 40.963$
$H_3^{(2)}$: $cv = 4.130$, $\hat{\beta}_3 = 85.622$
$C_2^{(1)}$: $cv = 2.200$, $\hat{\beta}_3 = 18.240$
$C_2^{(2)}$: $cv = 5.000$, $\hat{\beta}_3 = 84.000$
$C_2^{(3)}$: $cv = 8.556$, $\hat{\beta}_3 = 187.33$
In all cases: $\hat{\beta} = 1$

TABLE 4.6. Relative approximation errors (%) of $APP2$ for various service time distributions with $c^2 \leqslant 1$ and $\hat{\beta}_3 \leqslant \frac{3}{2}\,\hat{\beta}_2^2 / \hat{\beta}$.

ρ	DET	E_4	E_2	C_2^*
0.10	-0.02	0.16	0.12	0.06
0.30	-0.18	0.23	0.22	0.11
0.50	-0.36	0.16	0.18	0.10
0.70	-0.44	-0.02	0.05	0.04
0.90	-0.25	-0.10	-0.04	-0.02
0.95	-0.13	-0.08	-0.03	-0.02

* $cv = 0.75$

Chapter 5

QUEUEING MODELS WITH ADDITIONAL PERMANENT CUSTOMERS

5.1 Introduction

In the previous chapters we considered models with a single Poissonian external arrival stream. However, in many practical situations a service facility is shared by two or more classes of customers originating from different sources. An interesting aspect is the influence of the interference of the different customer streams on their queueing behaviour. A simple case occurs when it is assumed that the arrival processes are independent Poisson processes and the service discipline does not depend on the origin of the customers. Indeed, in that case the resulting 'overall' arrival process is also Poisson and hence we can use known results for the corresponding single arrival stream model.

In this chapter we consider single server queueing models with two classes of customers, viz. (i) ordinary customers who arrive according to a Poisson process, and (ii) permanent customers who immediately return to the end of the queue after having received a service. In Fig. 5.1 we have depicted the basic model: an M/G/1 FCFS queue with K permanent customers all having their own service time distribution $B_i(\cdot)$, $i=1,...,K$; the ordinary customers, in the sequel called 'Poisson customers', have service time distribution $B_0(\cdot)$.

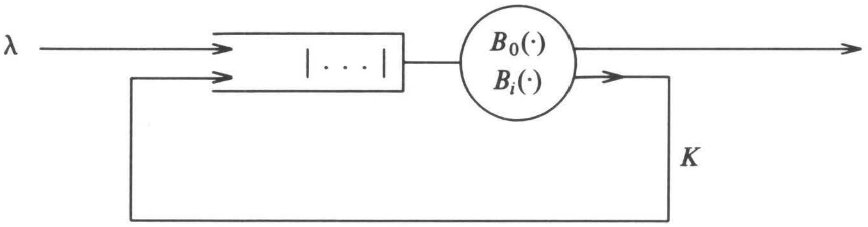

Fig. 5.1 The M/G/1 queue with K additional permanent customers.

Note that the presence of permanent customers may represent the existence of an 'infinite customer pool' (or several infinite customer pools) of which only a fixed number (K) is allowed to be in the system at the same time; when one of these customers departs upon service completion then a new customer from the pool immediately enters the queue. The main goal of this chapter is to study the influence of the K permanent customers on queue length and sojourn time of the Poisson customers.

Besides the basic model described above we shall also analyze some variants related with the models considered in the previous chapters, viz. the M/M/1 queue with general feedback and its limiting case, the M/G/1 processor sharing queue. The analysis of these models with additional permanent customers yields very interesting new results for the sojourn time distribution of the Poisson customers.

The M/G/1 queue with additional permanent customers is related with a class of models referred to as *vacation queues.* These are queueing models where a server now and then interrupts the service to a customer stream to take a vacation, see e.g. Fuhrmann and Cooper [1985] and the survey of Doshi [1986]. For the special case $K=1$ our model behaves exactly like an M/G/1 queue with vacations and so called 'gated service'. This is an M/G/1 vacation queue in which, after a server vacation, the server handles exactly those customers who are present at the end of the vacation, etc.. Clearly, a service of the permanent customer in our model (with $K=1$) corresponds to a server vacation in the gated vacation model.

The M/G/1 queue with vacations is a special case of a *cyclic service model,* a single server multi queue model in which the server attends to the queues in cyclic order, see e.g. Takagi [1986]. In Section 5.5 it will be pointed out that for general $K \geqslant 1$ the M/G/1 queue with K permanent customers can be viewed as a cyclic service model of a type not yet studied before.

Other related queueing models are the finite and infinite source interaction model studied by Kaufman [1985] (see also Boxma [1986A, 1986B] and Doshi and Wong [1987]), and a model with two stages of waiting introduced and analyzed in Ali and Neuts [1984].

The M/G/1 queue with K permanent customers is also studied by Boxma and Cohen [1989]. In particular, they present a fundamental analysis of the Markov chain of queue lengths at service completion epochs. In this chapter we present a different approach which is based on the observation that the queue length at an arbitrary epoch is the sum of K independent random variables which are related with queue lengths in the model with *one* permanent customer; the distributions of these random variables are obtained from known results for the M/G/1 queue with vacations.

As mentioned above we shall also consider the M/M/1 queue with general feedback and additional permanent customers. This is a very interesting model. Under the assumption that the service times of the permanent customers and the service times of the Poisson customers have the same exponential distribution, the addition of permanent customers to the M/M/1 feedback queue preserves the product form property of the joint queue length distribution, cf. (2.3). Using this property we can derive the joint distribution of the successive sojourn times of a tagged Poisson customer. The analysis is largely analogous to the case without permanent customers, see Chapter 2. It appears that the queue length and sojourn time distributions become the $(K+1)$-fold convolution of the queue length and sojourn time distribution in the original system. Application of the limiting procedure described in Section 3.3 leads to similar results for the M/G/1 PS queue with K additional permanent customers. Note that, actually, the service discipline in the latter model is a special case of generalized processor sharing, i.e. an M/G/1 GPS queue with service rate function $f(j) = 1/(j+K)$ (cf. Section 3.4).

The organization of this chapter is as follows. First, in Section 5.2, we give a detailed description of the basic model (the M/G/1 queue with K additional permanent customers) and introduce some notations. Section 5.3 is concerned with the derivation of *mean* queue lengths and sojourn times. We show that these quantities can be obtained from simple balance arguments. In Section 5.4 *distributions* of queue lengths and sojourn times are obtained. We start with the case of only one permanent customer (Subsection 5.4.1). The results for this case are obtained from known results for the M/G/1 queue with vacations. Next, in Subsection 5.4.2, the general model with $K \geqslant 1$ permanent customers is studied. In the remaining part of Section 5.4 we consider the results for some special choices of the service time distributions (Subsection 5.4.3), and we analyze a generalization, viz. the M/G/1 queue with Bernoulli feedback and additional permanent customers (Subsection 5.4.4). Section 5.5 is concerned with the relation with cyclic service models. Finally, in Section 5.6 we study the M/M/1 queue with general feedback and additional permanent customers together with the corresponding M/G/1 (G)PS queue.

5.2 Model description and notations

We consider the single server queueing system with infinite waiting room pictured in Fig. 5.1. There are two classes of customers, viz. (i) ordinary customers who arrive according to a Poisson process with intensity λ, and (ii) a class of K permanent customers who immediately return to the end of the queue after having received a service. The service discipline is first come first served (FCFS). The order of the permanent customers in the system is fixed; they are numbered from 1 to K. The service times of the Poisson customers and of the permanent customers are assumed to be independent random variables; those of the Poisson customers all have the same distribution $B_0(\cdot)$ and the i-th permanent customer has service time distribution $B_i(\cdot)$, $i=1,...,K$. The first two moments of $B_i(\cdot)$ are denoted by β_i and $\beta_i^{(2)}$ respectively, $i=0,...,K$. $\beta_i(\cdot)$ denotes the Laplace-Stieltjes transform of $B_i(\cdot)$, $i=0,...,K$. Obviously, the total offered load to the queue per unit of time due to the Poisson customers, ρ_0, is given by

$$\rho_0 = \lambda\beta_0. \tag{5.1}$$

For stability it is required that $\rho_0 < 1$. (For a formal derivation of this stability condition see Boxma and Cohen [1989]).

Observe that ρ_0 can also be viewed as the long range fraction of time spent on serving Poisson customers (as can be formally proved using the theory of regenerative processes, see Cohen [1976]). Similarly, we define for the permanent customers:

- ρ_i: fraction of time that (permanent) customer i is in service, $i = 1,...,K$.

Noting that the total fraction of time spent on serving permanent customers is equal to $1-\rho_0$ it is easily seen that

$$\rho_i = \frac{\beta_i}{\bar{\beta}}(1-\rho_0), \quad i = 1,...,K, \tag{5.2}$$

with

$$\bar{\beta} := \sum_{j=1}^{K} \beta_j.$$

We shall use (5.2) in the next sections.

We are interested in the following steady-state quantities:

- $\mathbf{X}_0$: number of Poisson customers in the system at an arbitrary epoch;
- $\mathbf{S}_0$: sojourn time of a Poisson customer;
- $\mathbf{C}_i$: sojourn (cycle) time of permanent customer i, $i = 1,...,K$ (i.e. time between two successive service completions of customer i, $i = 1,...,K$).

The generating function of the distribution of $\mathbf{X}_0$ is denoted by $X_0(\cdot)$. In the next section we shall derive the mean values of these performance measures; in Section 5.4 distributions are obtained.

5.3 Mean queue lengths and sojourn times

In this section we derive expressions for the *mean* sojourn times and queue lengths. It is shown how simple (balance) arguments can be used to derive these quantities. We start with the analysis of the mean sojourn time of the Poisson customers. Next, the mean cycle times of the permanent customers are derived.

Poisson customers

The PASTA property (see Wolff [1982]) implies that a newly arriving (tagged)

Poisson customer 'sees' the system as at an arbitrary epoch. Hence, the probability that the tagged customer arrives during the service of a Poisson customer is equal to ρ_0; the probability that permanent customer i is in service upon arrival of the tagged customer is equal to ρ_i, $i=1,...,K$. Now, considering the mean amount of work in the system that has to be handled before the tagged customer receives his service it follows by a similar argument as used for the derivation of (2.46), that

$$E\mathbf{S}_0 = (E\mathbf{X}_0-\rho_0)\beta_0 + \sum_{i=1}^{K}(1-\rho_i)\beta_i + \sum_{i=0}^{K}\rho_i\frac{\beta_i^{(2)}}{2\beta_i} + \beta_0. \tag{5.3}$$

Using Little's formula,

$$E\mathbf{X}_0 = \lambda E\mathbf{S}_0,$$

it follows from (5.3) that

$$E\mathbf{S}_0 = \frac{1}{1-\rho_0}\left[\sum_{i=0}^{K}(1-\rho_i)\beta_i + \sum_{i=0}^{K}\rho_i\frac{\beta_i^{(2)}}{2\beta_i}\right]. \tag{5.4}$$

Substituting (5.2) into (5.4) we find

$$E\mathbf{S}_0 = \frac{\rho_0}{1-\rho_0}\frac{\beta_0^{(2)}}{2\beta_0} + \beta_0 + \frac{\overline{\beta}}{1-\rho_0} + \sum_{i=1}^{K}\frac{\beta_i}{\overline{\beta}}(\frac{\beta_i^{(2)}}{2\beta_i}-\beta_i). \tag{5.5}$$

Again applying Little's formula yields the mean queue length $E\mathbf{X}_0$.

Note that the first two terms in the right-hand side of (5.5) represent the mean sojourn time in a standard M/G/1 queue (i.e. without permanent customers). The expression becomes very simple when all service times are exponentially distributed (i.e. $\beta_i^{(2)}=2\beta_i^2$, $i=0,...,K$). In that case,

$$E\mathbf{S}_0 = \frac{\beta_0+\overline{\beta}}{1-\rho_0}. \tag{5.6}$$

REMARK 5.1

It follows from (5.5) that the mean sojourn time $E\mathbf{S}_0$ of the Poisson customers in the M/G/1 queue with K permanent customers is larger than the mean sojourn time $E\mathbf{S}_0^{super}$ of the Poisson customers in the same M/G/1 queue but with only *one* ('super') permanent customer who has a service time which is equal to the sum of the service times of the K permanent customers in the original model. Noting that the second moment of the service time of the super permanent customer is given by $\sum_{i=1}^{K}(\beta_i^{(2)}+2\beta_i\sum_{j=i+1}^{K}\beta_j)$ it is easily found that

$$ES_0 - ES_0^{super} = \sum_{i=1}^{K} \frac{\beta_i}{\bar{\beta}} \sum_{j=1}^{i-1} \beta_j .$$

We shall now consider the mean cycle times of the permanent customers.

Permanent customers
From the fact that the order of the permanent customers in the system is fixed it follows immediately that their mean cycle times are all equal:

$$EC_i = EC_j, \quad 1 \leqslant i,j \leqslant K. \tag{5.7}$$

It is easily seen that, in steady-state, the mean amount of work that *arrives* during a cycle of customer i is equal to the mean amount of work that is *served* during a cycle. This balance argument leads to the following equation for EC_i:

$$EC_i = \bar{\beta} + (\lambda EC_i)\beta_0, \quad i=1,...,K, \tag{5.8}$$

yielding

$$EC_i = \frac{\bar{\beta}}{1-\rho_0}, \quad i=1,...,K. \tag{5.9}$$

Note that in the right-hand side only the first moments of the various service time distributions do occur. Apparently, this is due to the fact that the successive cycle times of a permanent customer consist of a random sum of *complete* service times.

There are some other simple and interesting methods to derive formula (5.9). Here, we mention two of these methods.

(i) From the definition of ρ_i it follows that ρ_i / β_i equals the throughput of customer i. Hence, from Little's formula: $EC_i = 1/(\rho_i/\beta_i)$, $i=1,...,K$. Substituting (5.2) into this expression yields (5.9).

(ii) The service of a permanent customer induces an amount of work to be handled by the server which consists of the permanent customer's own service time plus the sum of the lengths of a (random) number of standard M/G/1 (λ, $B_0(\cdot)$) busy periods. This number is equal to the number of Poisson customers who arrive during the service time of the permanent customer. So, the mean amount of work induced by a service of customer j is given by $\beta_j + \lambda\beta_j(\beta_0/(1-\rho_0))$, $j=1,...,K$. Noting that during a cycle of customer i all permanent customers receive exactly one service it follows from a balance argument that, cf. (5.9),

$$EC_i = \sum_{j=1}^{K}[\beta_j + \lambda\beta_j(\beta_0/(1-\rho_0))] = \frac{\bar{\beta}}{1-\rho_0}, \quad i=1,...,K.$$

The methods used in this section for the derivation of mean sojourn (cycle) times are based on mean value analysis. The derivation of queue length and sojourn time distributions requires a more detailed study of the model. This will be presented in the next section. It will also give us more insight into the results obtained above.

5.4 Distributions of queue lengths and sojourn times

In this section we derive expressions for the generating functions and Laplace-Stieltjes transforms of the distributions of the queue lengths and the sojourn (cycle) times of the different customers in the model. We shall start in Subsection 5.4.1 with the case of only one permanent customer ($K=1$); in Subsection 5.4.2 the general case ($K \geqslant 1$) is considered. A close study of the behaviour of the system with K permanent customers will show that its queue length at an arbitrary epoch can be written as the sum of K independent random variables, which are related with queue lengths in the model with *one* permanent customer. The distributions of these random variables can be obtained from the results in Subsection 5.4.1. In Subsection 5.4.3 we shall consider the results for some special choices of the service time distributions. Finally, Subsection 5.4.4 is concerned with the analysis of the M/G/1 queue with Bernoulli feedback and additional permanent customers; it appears that this generalization can be analyzed completely analogously to the basic model.

5.4.1 The case $K=1$

In this subsection we consider the M/G/1 queue with one permanent customer. As pointed out in Section 5.1 this special case behaves exactly like an M/G/1 queue with vacations and gated service. A service of the permanent customer corresponds to a server vacation in the vacation model. For the analysis we shall use the following *decomposition* result for the distribution of the queue length in a vacation queue. Define for the M/G/1 queue with vacations,

- $X(\cdot)$: generating function of the distribution of the queue length at an arbitrary epoch,
- $X_V(\cdot)$: generating function of the distribution of the queue length at an arbitrary epoch given that the server is on vacation,

and let

- $\pi(\cdot)$: generating function of the distribution of the queue length at an arbitrary epoch in the corresponding standard M/G/1 queue (i.e. the same model without vacations).

Then,

THEOREM 5.1 (Fuhrmann and Cooper [1985])

$$X(z) = X_V(z)\pi(z), \quad |z| \leqslant 1. \tag{5.10}$$

Actually, this result does not only hold for the M/G/1 vacation queue with *gated* service; it is valid for a very general class of vacation models, see Fuhrmann and Cooper [1985].

Now, we return to the M/G/1 queue with one permanent customer. Define

- $\mathbf{X}_P$: number of Poisson customers in the system at an arbitrary epoch given that the permanent customer is in service,

and let $X_P(\cdot)$ denote the generating function of the distribution of $\mathbf{X}_P$. From (5.10) it follows immediately that the generating function $X_0(\cdot)$ of the distribution of the queue length at an arbitrary epoch can be written as

$$X_0(z) = X_P(z)\pi(z), \quad |z| \leqslant 1. \tag{5.11}$$

The generating function of the queue length distribution in the standard M/G/1 queue, $\pi(\cdot)$, is given by the well-known formula (see e.g. Cohen [1982], p. 238):

$$\pi(z) = (1-\rho_0)\frac{(1-z)\beta_0\{\lambda(1-z)\}}{\beta_0\{\lambda(1-z)\}-z}, \quad |z| \leqslant 1. \tag{5.12}$$

For the derivation of $X_P(z)$ we need the following definitions:

- $\mathbf{X}_{PB}$: number of Poisson customers in the system just after the start of a service of the permanent customer;
- $\mathbf{X}_{PE}$: number of Poisson customers in the system at a service completion epoch of the permanent customer.

$X_{PB}(\cdot)$ and $X_{PE}(\cdot)$ will denote the generating functions of the distributions of $\mathbf{X}_{PB}$ and $\mathbf{X}_{PE}$, respectively.

The number of Poisson customers present at an arbitrary epoch during the service of the permanent customer is equal to the number of Poisson customers present at the start of that service *plus* the number of Poisson customers that has arrived during the past service time. It is easily seen that these quantities are independent. Hence,

$$\chi_P(z) = \chi_{PB}(z)\frac{1-\beta_1\{\lambda(1-z)\}}{\beta_1\lambda(1-z)}, \quad |z|\leqslant 1. \tag{5.13}$$

(Remember that, for Re $\eta \geqslant 0$, $\dfrac{1-\beta_1(\eta)}{\beta_1\eta}$ is the LST of the distribution of the past part of the service time.)

Analogously, we have

$$\chi_{PE}(z) = \chi_{PB}(z)\beta_1\{\lambda(1-z)\}, \quad |z|\leqslant 1. \tag{5.14}$$

Now, from (5.11)-(5.14) it follows that

$$\chi_0(z) = \frac{\chi_{PE}(z)}{\beta_1\{\lambda(1-z)\}}\frac{1-\beta_1\{\lambda(1-z)\}}{\beta_1\lambda(1-z)}\pi(z), \quad |z|\leqslant 1. \tag{5.15}$$

It remains to determine $\chi_{PE}(\cdot)$.

The discrete time stochastic process constituted by the number of Poisson customers present in the system at the successive service completion epochs of the permanent customer is easily seen to be a Markov chain with state space $\{0, 1, \ldots\}$ and stationary transition probabilities p_{ij} given by

$$p_{ij} = \int_{t=0}^{\infty} e^{-\lambda t}\frac{(\lambda t)^j}{j!}\, d[B_0^{i*}(t)*B_1(t)]\,, \quad i,j\geqslant 0, \tag{5.16}$$

with * the convolution operator. Application of standard Markov chain theory leads to the following functional equation for $\chi_{PE}(\cdot)$:

$$\chi_{PE}(z) = \beta_1\{\lambda(1-z)\}\chi_{PE}(\beta_0\{\lambda(1-z)\}), \quad |z|\leqslant 1. \tag{5.17}$$

From this equation the moments of $\mathbf{X}_{PE}$ can be obtained. For example, differentiating both sides once and taking $z=1$ yields,

$$E\mathbf{X}_{PE} = \frac{\lambda\beta_1}{1-\rho_0}. \tag{5.18}$$

Analogously, the moments of the distribution of $\mathbf{X}_0$ can be obtained from (5.15) and (5.17). It is easily found that, for the case $K=1$,

$$E\mathbf{X}_0 = \frac{\lambda\beta_1}{1-\rho_0} - \lambda\beta_1 + \lambda\frac{\beta_1^{(2)}}{2\beta_1} + \frac{\lambda\rho_0}{1-\rho_0}\frac{\beta_0^{(2)}}{2\beta_0} + \rho_0. \tag{5.19}$$

This is in agreement with (5.5) for $K=1$ (noting that $E\mathbf{X}_0 = \lambda E\mathbf{S}_0$).

REMARK 5.2
The generating function of the distribution of the queue length at the *departure* epochs of the Poisson customers is also given by (5.15). Indeed, an up-and-down-crossing argument shows that this distribution equals the queue length distribution seen by an *arriving* Poisson customer; the PASTA property implies that this is also the distribution of the queue length at an *arbitrary* epoch.

REMARK 5.3
In Boxma and Cohen [1989] it is shown that the unique (non trivial) solution of (5.17) is given by:

$$\chi_{PE}(z) = \prod_{h=0}^{\infty}\beta_1\{\lambda(1-\delta_h(z))\}, \quad |z|\leq 1, \tag{5.20}$$

with

$$\begin{aligned} \delta_0(z) &:= z, \\ \delta_{h+1}(z) &:= \beta_0\{\lambda(1-\delta_h(z))\}, \quad h=0,1,\dots . \end{aligned} \tag{5.21}$$

The derivation is based on iteration of (5.17) and on some well-known results from branching theory; the condition $\rho_0<1$ guarantees the convergence of the infinite product in (5.20).

The LST's of the sojourn time of the Poisson customers and the cycle time of the permanent customer can be derived from the above queue length results. Noting that the cycle time of the permanent customer consists of his own service time plus the service times of the Poisson customers present in the system just after the end of his previous service, we have,

$$E\{e^{-\eta \mathbf{C}_1}\} = \chi_{PE}(\beta_0(\eta))\beta_1(\eta), \quad \text{Re } \eta \geq 0, \tag{5.22}$$

with $\chi_{PE}(\cdot)$ determined by (5.17). The LST of the sojourn time distribution of the Poisson customers can be derived from (5.15). The classical observation for FCFS queues that the number of Poisson customers left behind by a departing customer equals the number of (Poisson) arrivals during the sojourn time of that departing customer, leads to (cf. Remark 5.2)

$$E\{e^{-\lambda(1-z)\mathbf{S}_0}\} = \chi_0(z), \quad |z|\leq 1.$$

Hence,

$$E\{e^{-\eta \mathbf{S}_0}\} = \chi_0(1-\eta/\lambda), \quad \text{Re } \eta \geq 0. \tag{5.23}$$

From the above results the moments of the distributions of $\mathbf{S}_0$ and $\mathbf{C}_1$ can be obtained. Differentiating (5.22) and (5.23) once with respect to η and taking $\eta=0$ yields the mean cycle time and sojourn time (use (5.18) and (5.19)); it is easily verified that these results coincide with the results obtained in Section 5.3.

We conclude this subsection with a remark on notation. For the analysis in the next subsection it is convenient to have at our disposal a specific notation for some of the quantities in the model with one permanent customer. Therefore, for the M/G/1 queue with *one* permanent customer and service time distributions $B_0(\cdot)$ and $B_i(\cdot)$ for the Poisson customers and the permanent customer respectively we define, for $i=1,...,K$,

- $X_0^{(1,i)}(\cdot)$: generating function of the distribution of the number of Poisson customers in the system at an arbitrary epoch;
- $X_{PE}^{(1,i)}(\cdot)$: generating function of the distribution of the number of Poisson customers in the system at a service completion epoch of the permanent customer.

Obviously, these quantities can be obtained from (5.15) and (5.17) by taking $B_1(\cdot)\equiv B_i(\cdot)$, $i=1,...,K$.

5.4.2 The case $K \geqslant 1$

We now turn to the case of an arbitrary number $K \geqslant 1$ of permanent customers. In Fig. 5.2(a) we have pictured the composition of the queue just after a service completion epoch of permanent customer $i-1$. It can be described as follows: counted from the head of the queue there is a group of Poisson customers followed by permanent customer i; subsequently there is a second group of Poisson customers followed by permanent customer $i+1$, etc., until finally the queue is ended by a K-th group of Poisson customers followed by permanent customer $i-1$ (who has just returned from the head of the queue). The group of customers consisting of the Poisson customers at the head of the queue and permanent customer i will be called '*group i*'; the group of customers consisting of permanent customer j and the Poisson customers ahead of j and behind $j-1$ will be called '*group j*', $j=1,...,K$, $j \neq i$. (Note that *during* a service time the last permanent customer in the queue may be followed by one or more Poisson customers; in that case these Poisson customers are assumed to be members of the group at the head of the queue, cf. Fig. 5.2(b)). The (random) number of Poisson customers in a particular group will be referred to as the '*size*' of that group. So, the total number of Poisson customers in the system is equal to the sum of the sizes of the K groups. To determine the distribution of the size of each group we shall first investigate how these groups develop. Let us start with the present situation, i.e. just after a service completion of permanent customer $i-1$. Now, the first customer of group i is taken into service; next the second one, etc, until finally permanent customer i

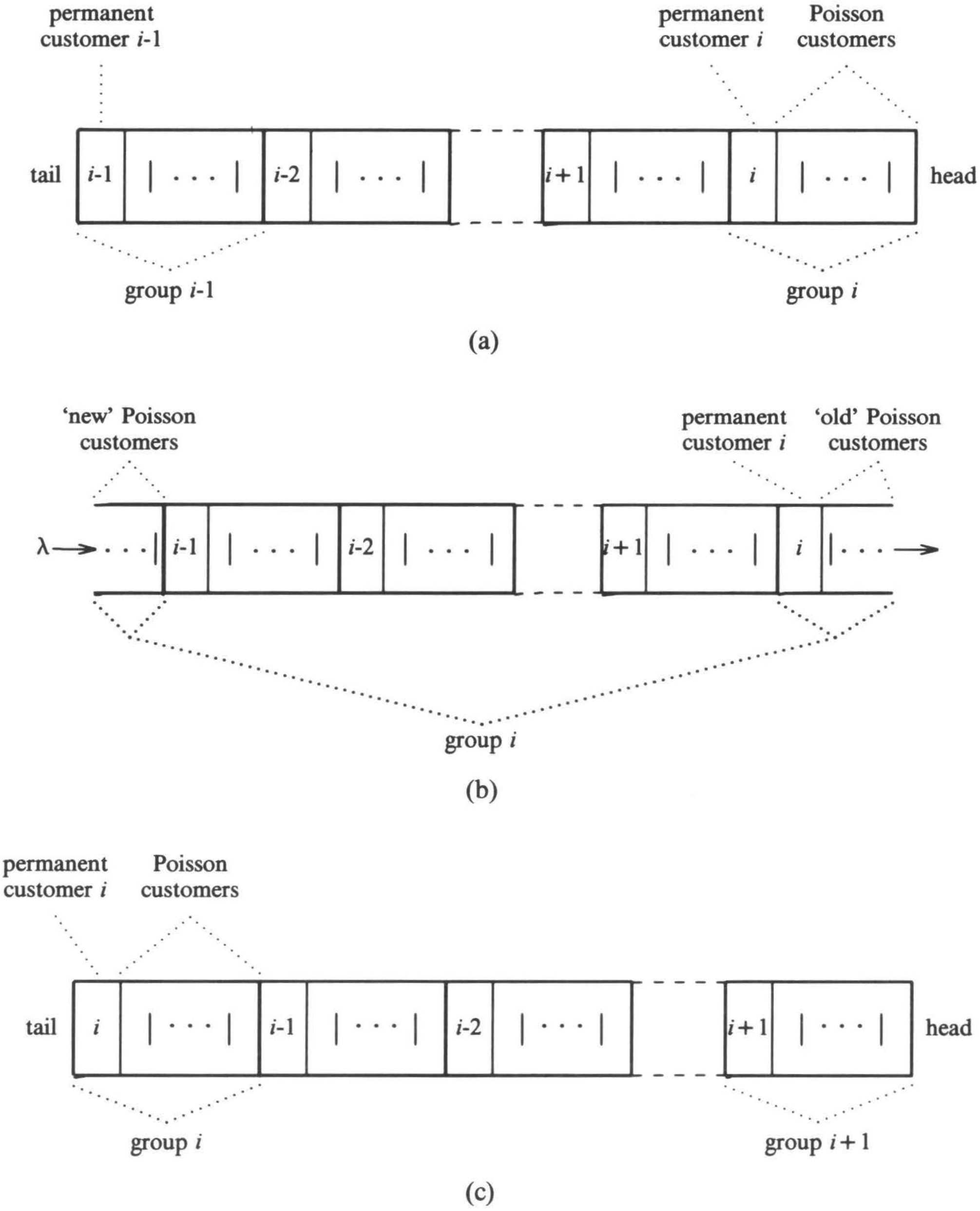

Fig. 5.2 Composition of the queue
(a) just after the service completion of permanent customer $i-1$,
(b) during the service of group i,
(c) just after the service completion of permanent customer i.

completes his service. During these services group i consists of two parts, see Fig. 5.2(b): (i) permanent customer i together with the ('old') Poisson customers in front of him who have not yet been served to completion, and (ii) Poisson customers at the end of the queue (behind permanent customer $i-1$) who have newly arrived during the past service time of group i. Just after the service completion of permanent customer i this second part has developed into a complete, 'new', group i at the end of the queue, see Fig. 5.2(c); its size is equal to the number of Poisson customers who have arrived during the service of the 'old' group i. Note that the size of the other groups has *not* changed. After the service completion of group i the service of group $i+1$ is started, next the service of group $i+2$, etc. During the service of groups $i+1, \ldots, i-1$ group i moves to the head of the queue, *while its size remains unchanged.* Just after the service of permanent customer $i-1$ the order of the different groups within the queue is again as in Fig. 5.2(a). Now, a second service of group i is started, and the whole procedure as described above is repeated, etc.

From the above discussion and from the memoryless property of the Poisson arrival process it is clear that *during the service of group i this group behaves exactly like an $M/G/1$ queue with one permanent customer and service time distributions $B_0(\cdot)$ and $B_i(\cdot)$ for the Poisson customers and the permanent customer respectively,* $i=1,...,K$. It is also seen that *at a service completion epoch of one of the permanent customers the sizes of the groups are independent; this independence property also holds at an arbitrary epoch given that group i is in service,* $i=1,...,K$. Now define, for $i=1,...,K$,

- $\chi_0^{(i)}(\cdot)$: generating function of the distribution of the number of Poisson customers present in the system at an arbitrary epoch given that group i is in service;
- $\chi_{PE}^{(i)}(\cdot)$: generating function of the distribution of the number of Poisson customers present in the system at a service completion epoch of permanent customer i.

It is easily seen that

$$\chi_{PE}^{(i)}(z) = \prod_{j=1}^{K} \chi_{PE}^{(1,j)}(z), \quad i=1,...,K, \quad |z| \leqslant 1, \tag{5.24}$$

and

$$\chi_0^{(i)}(z) = \chi_0^{(1,i)}(z) \prod_{\substack{j=1 \\ j\neq i}}^{K} \chi_{PE}^{(1,j)}(z), \quad i=1,...,K, \quad |z| \leqslant 1, \tag{5.25}$$

where the generating functions $\chi_{PE}^{(1,j)}(\cdot)$ and $\chi_0^{(1,j)}(\cdot)$ of the queue length distributions in the model with one permanent customer are given by (5.20) and (5.15) respectively (see also the definitions at the end of Subsection 5.4.1). Substituting (5.15) into (5.25) yields

$$\chi_0^{(i)}(z) = \frac{1}{\beta_i\{\lambda(1-z)\}}\frac{1-\beta_i\{\lambda(1-z)\}}{\beta_i\lambda(1-z)}\pi(z)\prod_{j=1}^{K}\chi_{PE}^{(1,j)}(z), \tag{5.26}$$

$$i=1,...,K, \quad |z|\leqslant 1.$$

Obviously, the probability p_i that group i is in service at an arbitrary epoch is equal to the long range fraction of time that group i is in service. Noting that the mean time spent on serving group i during a cycle is equal to $\beta_i/(1-\rho_0)$ (cf. the derivation of (5.24) and (5.18)) we have from (5.9),

$$p_i = \frac{\beta_i/(1-\rho_0)}{EC_i} = \frac{\beta_i/(1-\rho_0)}{\bar{\beta}/(1-\rho_0)} = \frac{\beta_i}{\bar{\beta}}, \quad i=1,...,K. \tag{5.27}$$

Hence, for $K\geqslant 1$,

$$\chi_0(z) = \sum_{i=1}^{K}p_i\chi_0^{(i)}(z) = \tag{5.28}$$

$$\pi(z)\{\prod_{j=1}^{K}\chi_{PE}^{(1,j)}(z)\}\sum_{i=1}^{K}\frac{\beta_i}{\bar{\beta}}\frac{1}{\beta_i\{\lambda(1-z)\}}\frac{1-\beta_i\{\lambda(1-z)\}}{\beta_i\lambda(1-z)}, \quad i=1,...,K, \quad |z|\leqslant 1.$$

Remember that $\pi(z)$ represents the generating function of the queue length distribution in the M/G/1 queue without permanent customers, cf. (5.12). Note that it is not allowed to take $K=0$ in (5.28); in its derivation K is explicitly assumed to be positive.

The LST's of the sojourn time and cycle time distributions can be easily obtained from the above queue length results. Analogously to the derivations of the sojourn time and cycle time in the model with one permanent customer we have, from (5.24) (cf. (5.22)),

$$E\{e^{-\eta C_i}\} = \chi_{PE}^{(i)}(\beta_0(\eta))\prod_{j=1}^{K}\beta_j(\eta) = \prod_{j=1}^{K}\{\beta_j(\eta)\chi_{PE}^{(1,j)}(\beta_0(\eta))\}, \tag{5.29}$$

$$i=1,...,K, \quad \text{Re}\,\eta\geqslant 0,$$

and, from (5.28) (cf. (5.23) and Remark 5.2),

$$E\{e^{-\eta \mathbf{S}_0}\} = \chi_0(1-\eta/\lambda) = \tag{5.30}$$

$$\pi(1-\eta/\lambda)(\prod_{j=1}^{K}\chi_{PE}^{(1,j)}(1-\eta/\lambda))\sum_{i=1}^{K}\frac{\beta_i}{\bar{\beta}}\frac{1}{\beta_i(\eta)}\frac{1-\beta_i(\eta)}{\beta_i\eta}, \quad \text{Re}\ \eta \geqslant 0.$$

Differentiating these expressions once with respect to η and taking $\eta=0$ yields the mean cycle time of the permanent customers and the mean sojourn time of the Poisson customers; it is easily verified that these results coincide with (5.9) and (5.5).

5.4.3 Results for some special choices of the service time distributions

For some special choices of the different service time distributions in the model the queue length and sojourn time formulas (5.24) and (5.28)-(5.30) reduce to much simpler expressions. The following two cases are worth mentioning.

(i) Equal service time distributions for the permanent customers

If $B_1(\cdot) \equiv \cdots \equiv B_K(\cdot)$, then, from (5.24),

$$\chi_{PE}^{(i)}(z) = (\chi_{PE}(z))^K, \quad i=1,...,K, \quad |z| \leqslant 1, \tag{5.31}$$

and, from (5.28),

$$\chi_0(z) = \pi(z)\frac{(\chi_{PE}(z))^K}{\beta_1\{\lambda(1-z)\}}\frac{1-\beta_1\{\lambda(1-z)\}}{\beta_1\lambda(1-z)}, \quad |z| \leqslant 1, \tag{5.32}$$

with $\chi_{PE}(\cdot)$ determined by (5.20).

(ii) All service time distributions equal and negative exponential

If $B_i(t) = 1-e^{-t/\beta_0}$, $i=0,1,...,K$, then the solution of the functional equation (5.17) has an explicit form (cf. Remark 5.3, for the solution of the general case). Defining, for $|z| \leqslant 1$, $n=1,2,...$,

$$f^{(1)}(z) := f(z) := \beta_0\{\lambda(1-z)\}, \tag{5.33}$$

$$f^{(n+1)}(z) := f(f^{(n)}(z)),$$

one can iterate (5.17) in the following way:

$$X_{PE}(z) = f^{(1)}(z)X_{PE}(f^{(1)}(z)) = f^{(1)}(z)f^{(2)}(z)X_{PE}(f^{(2)}(z)) = \cdots = \tag{5.34}$$

$$(\prod_{j=1}^{m} f^{(j)}(z))X_{PE}(f^{(m)}(z)).$$

For our case of exponentially distributed service times, i.e. (cf. (5.33))

$$f(z) = \frac{1}{1+\beta_0\lambda(1-z)} = \frac{1}{1+\rho_0(1-z)},$$

it is easily found that

$$\prod_{j=1}^{m} f^{(j)}(z) = \frac{1}{1+(1-z)\sum_{h=1}^{m}\rho_0^h} = \frac{1}{1+(1-z)(\frac{1-\rho_0^{m+1}}{1-\rho_0}-1)}, \tag{5.35}$$

and hence,

$$f^{(m)}(z) \rightarrow 1, \quad \text{for } m\rightarrow\infty. \tag{5.36}$$

Now, from (5.34)-(5.36) we obtain (cf. (5.20))

$$X_{PE}(z) = \lim_{m\rightarrow\infty}(\prod_{j=1}^{m} f^{(j)}(z))X_{PE}(f^{(m)}(z)) = \frac{1-\rho_0}{1-\rho_0 z}, \tag{5.37}$$

which equals the generating function of the queue length distribution in an ordinary M/M/1 queue. Substituting (5.37) into (5.31) and (5.32) we find, for $|z|\leqslant 1$,

$$X_{PE}^{(i)}(z) = \left[\frac{1-\rho_0}{1-\rho_0 z}\right]^K, \quad i=1,...,K, \tag{5.38}$$

$$X_0(z) = \left[\frac{1-\rho_0}{1-\rho_0 z}\right]^{K+1}. \tag{5.39}$$

Formula (5.39) exposes a remarkable effect of the presence of permanent customers on the queue length of the Poisson customers in an M/M/1 queue: *their queue length distribution becomes the $(K+1)$-fold convolution of the queue length distribution in the system without those permanent customers.*

From (5.29), (5.30), (5.38) and (5.39) the LST's of the cycle time and sojourn time distribution are obtained: for Re $\eta \geqslant 0$,

$$E\{e^{-\eta \mathbf{C}_i}\} = \left[\frac{1-\rho_0}{1-\rho_0+\beta_0\eta}\right]^K, \quad i=1,...,K, \tag{5.40}$$

$$E\{e^{-\eta \mathbf{S}_0}\} = \left[\frac{1-\rho_0}{1-\rho_0+\beta_0\eta}\right]^{K+1}. \tag{5.41}$$

So, the presence of K permanent customers also leads to a $(K+1)$-fold increase of the *sojourn times* of the Poisson customers.

REMARK 5.4
The above mentioned phenomenon (expressed by (5.41)) for the case of identical, exponential service times can be explained as follows. Consider the model with one permanent customer, i.e. the case $K=1$ (the general case follows immediately from this special one). It is seen from Formula (5.40) that the successive cycle times of the permanent customer are exponentially distributed with mean $\beta_0/(1-\rho_0)$. A newly arriving (tagged) Poisson customer enters the queue during one of these cycles; it is clear that his sojourn time, $\mathbf{S}_0$, is equal to the residual cycle time $(\mathbf{C}_1^{(R)})$ *plus* the sum of the service times of the customers who have arrived during the past cycle time $(\mathbf{V})$ *plus* his own service time $(\boldsymbol{\tau}_0)$: $\mathbf{S}_0=\mathbf{C}_1^{(R)}+\mathbf{V}+\boldsymbol{\tau}_0$. From standard probabilistic arguments it follows that the residual and past cycle time are exponentially distributed with mean $\beta_0/(1-\rho_0)$, and it is found that the number of customers who have arrived during the past cycle time has a geometric distribution with mean $\rho_0/(1-\rho_0)$. Moreover, $\mathbf{C}_1^{(R)}$, $\mathbf{V}$ and $\boldsymbol{\tau}_0$ are mutually independent. Now it is easily seen that $\mathbf{V}+\boldsymbol{\tau}_0$ is exponentially distributed with mean $\beta_0/(1-\rho_0)$ and, hence, $\mathbf{S}_0$ has a 2-stage Erlang distribution (E_2) with mean $2\beta_0/(1-\rho_0)$, cf. (5.41).

REMARK 5.5
Note that in the present exponential (product form) case a departing (and hence again arriving) permanent customer sees the system in equilibrium with one less customer of his own type (see e.g. Walrand [1988, Section 3.4]), which confirms the relation between (5.38) and (5.39), and between (5.40) and (5.41).

In Section 5.6 it will be shown that generalizations of (5.38)-(5.41) hold for the M/M/1 queue with general feedback and additional permanent customers.

5.4.4 Generalizations; the M/G/1 queue with Bernoulli feedback and additional permanent customers

The results obtained in the previous subsections can be generalized in several ways. Firstly one may allow different arrival rates during the service of the permanent customers. Another interesting possibility is the inclusion of a (Bernoulli) feedback mechanism for the Poisson customers. For both generalizations the basic decomposition formula (5.11), which is implied by Theorem 5.1, remains valid (see Fuhrmann and Cooper [1985] and Shanthikumar [1988]), and the analysis can be carried out in almost exactly the same way as in Subsection 5.4.1 and Subsection 5.4.2. We shall consider the feedback case in some more detail.

The M/G/1 queue with Bernoulli feedback and additional permanent customers
Consider the M/G/1 queue with K permanent customers described in Section 5.2 and assume in addition that the Poisson customers, after having received a service, are fed back to the end of the queue with probability p and leave the system with probability $1-p$.

As for the model without feedback we first analyze the case $K=1$. Clearly, this case behaves exactly like an M/G/1 Bernoulli feedback queue with vacations and gated service. In Fuhrmann and Cooper [1985] it is pointed out that for this generalization of the 'standard' M/G/1 vacation queue (without feedback) Theorem 5.1 remains valid with, in (5.10), $\pi(\cdot)$ denoting the generating function of the queue length distribution in the corresponding M/G/1 queue with Bernoulli feedback but without vacations. It is well-known that $\pi(\cdot)$ is given by: (see e.g. Takács [1963])

$$\pi(z) = (1-\hat{\rho}_0)\frac{(1-z)(1-p)\beta_0\{\lambda(1-z)\}}{(1-p+pz)\beta_0\{\lambda(1-z)\}-z}\,, \tag{5.42}$$

provided that

$$\hat{\rho}_0 := \frac{\rho_0}{1-p} < 1. \tag{5.43}$$

(Note that $\pi(\cdot)$ equals the generating function of the queue length distribution in the standard M/G/1 queue with service time distribution $\hat{B}_0(t) = \sum_{j=1}^{\infty}(1-p)p^{j-1}B_0^{j*}(t)$, cf. (5.12)). Now, it is easily seen that for the present model with Bernoulli feedback the rest of the analysis in Subsection 5.4.1 does not change with the exception that the functional equation (5.17) for the generating function $X_{PE}(\cdot)$ of the queue length distribution just after a service completion of the permanent customer becomes

$$X_{PE}(z) = \beta_1\{\lambda(1-z)\}\,X_{PE}((1-p+pz)\beta_0\{\lambda(1-z)\}), \quad |z|\leqslant 1. \tag{5.44}$$

With this adaptation Formula (5.15) for the generating function $X_0(\cdot)$ of the queue length distribution at an arbitrary epoch remains valid. Now, from equation (5.44), (cf. (5.18))

$$E\mathbf{X}_{PE} \;=\; \frac{\lambda\beta_1/(1-p)}{1-\hat{\rho}_0}\,, \tag{5.45}$$

and hence, from (5.15), (5.42) and (5.45) we have, for the case $K=1$, (cf. (5.19))

$$E\mathbf{X}_0 \;=\; \frac{\lambda\beta_1/(1-p)}{1-\hat{\rho}_0} \;-\; \lambda\beta_1 \;+\; \lambda\frac{\beta_1^{(2)}}{2\beta_1} \;+\; \frac{\lambda\hat{\rho}_0}{1-\hat{\rho}_0}[p\frac{\beta_0}{1-p}+\frac{\beta_0^{(2)}}{2\beta_0}] \;+\; \hat{\rho}_0\,. \tag{5.46}$$

Remark 5.6
It follows from (5.45) that the mean cycle time, $E\mathbf{C}_1$, of the permanent customer is given by $E\mathbf{C}_1 = \beta_0 E\mathbf{X}_{PE} + \beta_1 = \beta_1/(1-\hat{\rho}_0)$. As might be expected, for $\beta_0=\beta_1=\beta$ this result coincides with (2.66) which gives the mean k-th sojourn time of a tagged customer in the M/G/1 queue with Bernoulli feedback for $k\rightarrow\infty$.

It is easily seen that, using the above formulas for the case $K=1$, the analysis for the case $K\geqslant 1$ can be carried out in exactly the same way as for the model without feedback and that all queue length formulas in Subsection 5.4.2 remain unchanged. For example, from the generating function of the queue length distribution given by (5.28) and from (5.45), (5.46) we obtain for general $K\geqslant 1$,

$$\begin{aligned} E\mathbf{X}_0 \;=\; & \frac{\lambda\bar{\beta}/(1-p)}{1-\hat{\rho}_0} \;+\; \lambda\sum_{i=1}^{K}\frac{\beta_i}{\bar{\beta}}(\frac{\beta_i^{(2)}}{2\beta_i}-\beta_i) \;+ \\ & \frac{\lambda\hat{\rho}_0}{1-\hat{\rho}_0}[p\frac{\beta_0}{1-p}+\frac{\beta_0^{(2)}}{2\beta_0}] \;+\; \hat{\rho}_0\,. \end{aligned} \tag{5.47}$$

Hence, using Little's formula, the mean *total* sojourn time of the Poisson customers is given by: (cf. (5.5))

$$\begin{aligned} E\mathbf{S}_0 \;=\; & \frac{\bar{\beta}/(1-p)}{1-\hat{\rho}_0} \;+\; \sum_{i=1}^{K}\frac{\beta_i}{\bar{\beta}}(\frac{\beta_i^{(2)}}{2\beta_i}-\beta_i) \;+ \\ & \frac{\hat{\rho}_0}{1-\hat{\rho}_0}[p\frac{\beta_0}{1-p}+\frac{\beta_0^{(2)}}{2\beta_0}] \;+\; \frac{\beta_0}{1-p}\,. \end{aligned} \tag{5.48}$$

REMARK 5.7
Having obtained the generating function of the distribution of the queue length at an arbitrary epoch we can also derive the LST of the joint distribution of the successive sojourn times of a tagged Poisson customer. The analysis can be done almost completely analogously to the case without permanent customers as treated by Doshi and Kaufmann [1988]. We shall only consider the case of negative exponential service times in some detail; in Section 5.6 it is shown how the LST of the joint sojourn time distribution in the M/M/1 queue with general feedback and additional permanent customers can be obtained from the corresponding result derived in Chapter 2 for the same model without permanent customers.

5.5 RELATION WITH CYCLIC SERVICE MODELS

In Section 5.1 it has been pointed out that the M/G/1 queue with *one* permanent customer behaves exactly like an M/G/1 queue with vacations and gated service. In this section we shall show that the M/G/1 queue with $K \geqslant 1$ permanent customers can be viewed as a variant of a *cyclic service model.*

A cyclic service model (also called a polling model) is a single server multi queue model in which the server attends to the queues in cyclic order, see e.g. Takagi [1986] and Groenendijk [1990]. From the analysis in Subsection 5.4.2 it follows that the M/G/1 queue with K permanent customers can also be viewed as a cyclic service model with K queues - albeit a rather special one, see Fig. 5.3.

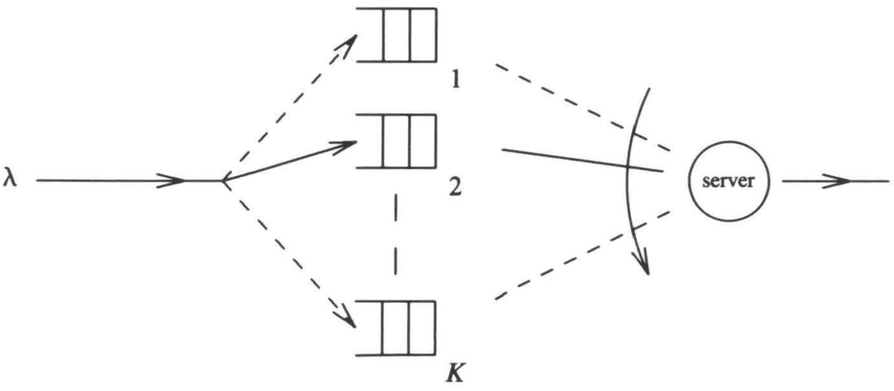

Fig. 5.3 The M/G/1 queue with permanent customers viewed as a cyclic service model.

The service times of the permanent customers correspond to the switch-over times of the server between successive queues; the customers in these queues represent the (Poisson) customers in the different groups in the single queue M/G/1 model (cf. Fig. 5.2). To take into account that during the service of a particular group all new arrivals are attached to that group (cf. Fig. 5.2(b)), we have to assume that arrivals at a queue can only take place during its service and during the subsequent switch-over time.

The resulting polling model is non-standard, but it is well-known that under very general conditions the mean cycle time in a polling model is given by the sum of the mean switch-over times, divided by one minus the load of the system (see e.g. Takagi [1986]). Indeed this result holds here; it coincides with Formula (5.9) for the mean cycle time of a permanent customer.

5.6 The M/M/1 feedback queue with additional permanent customers

5.6.1 Introduction

In this section we consider the same M/M/1 queue with general feedback as in Chapter 2 but with $K \geqslant 1$ additional permanent customers. This model is pictured in Fig. 5.4.

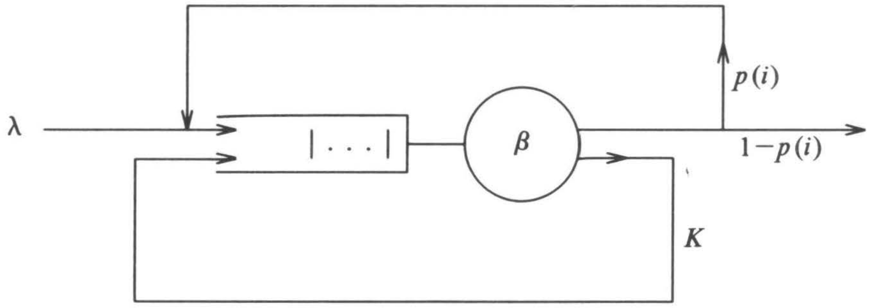

Fig. 5.4 The M/M/1 feedback queue with K additional permanent customers.

It is assumed that the service times of the Poisson customers and the permanent customers are independent, negative exponentially distributed random variables, all with mean β. For the Poisson customers the assumptions about the feedback mechanism, notations, terminology, etc. are the same as for the model without permanent customers, see Section 2.2. Our main goal is to study the influence of the presence of the permanent customers on the joint distribution of the successive sojourn times of a (tagged) Poisson customer and to use the results for the analysis of the sojourn time in the M/G/1 PS queue with additional permanent customers. The results for this latter model are obtained by applying the same limiting procedure as used for the case without permanent customers, see Chapter 3.

Because the Poisson customers and the permanent customers have the *same* exponential service time distribution, the joint stationary distribution of the number of type-i (Poisson) customers, $\mathbf{X}_i$, $i=1,...,N$, in the system at an arbitrary epoch is of product form type. From the queue length results for general product form networks (see Baskett et al. [1975]) it is found that for our model, cf. (2.3),

$$P(x_1,\ldots,x_N) := Pr\{\mathbf{X}_1=x_1,\ldots,\mathbf{X}_N=x_N\} = \tag{5.49}$$

$$(1-\rho)^{K+1}\frac{(x_1+\cdots+x_N+K)!}{x_1!\cdots x_N!K!}\prod_{i=1}^{N}(\lambda\beta q(i))^{x_i}, \qquad x_1,...,x_N=0,1,\ldots.$$

(Remember that $q(i)$ represents the relative arrival rate of type-i (Poisson) customers, $i=1,...,N$ (cf. (2.1)), and that ρ denotes the total offered load to the system per unit of time due to the Poisson customers: $\rho=\sum_{i=1}^{N}\lambda\beta q(i)$).

The generating function of the joint queue length distribution is given by: (cf. the derivation of (2.4))

$$E\{z_1^{\mathbf{X}_1}\cdots z_N^{\mathbf{X}_N}\} = \sum_{x_1=0}^{\infty}\cdots\sum_{x_N=0}^{\infty} z_1^{x_1}\cdots z_N^{x_N}P(x_1,\cdots,x_N) = \tag{5.50}$$

$$(1-\rho)^{K+1}\frac{1}{K!}\sum_{m=0}^{\infty}\underset{x_1+\ldots+x_N=m}{\sum_{x_1}\cdots\sum_{x_N}}\frac{(m+K)!}{x_1!\cdots x_N!}\prod_{i=1}^{N}(\lambda\beta q(i)z_i)^{x_i} =$$

$$(1-\rho)^{K+1}\sum_{m=0}^{\infty}\binom{m+K}{K}\left[\sum_{i=1}^{N}\lambda\beta q(i)z_i\right]^m = \left[\frac{1-\rho}{1-\sum_{i=1}^{N}z_i\lambda\beta q(i)}\right]^{K+1},$$

$$|z_i|\leqslant 1,\ i=1,...,N.$$

Comparing this result with (2.4) we observe a similar phenomenon as for the standard M/M/1 queue (cf. (5.39)): *the presence of the K permanent customers in the M/M/1 feedback queue leads to a joint queue length distribution which is the $(K+1)$-fold convolution of the joint queue length distribution in the same model without permanent customers.* In the next subsection we shall use (5.50) for the analysis of the sojourn time distribution.

5.6.2 Sojourn time distribution

In this section we present, in the form of Laplace-Stieltjes transforms and generating functions, an expression for the joint steady state distribution of the successive sojourn times $\mathbf{S}_j$, $j=1,...,k$, and the number of type-i customers,

$\mathbf{X}_i^{(j)}$, $i=1,...,N$, present at the j-th service completion of a customer who is fed back at least $k-1$ times, $k=1,2,...$. It will appear that for the derivation of this quantity we can largely rely on the analysis of the sojourn time in the model without permanent customers given in Section 2.3.

Consider a newly arriving (tagged) customer, say C, and suppose that he finds $\mathbf{X}_i^{(0)}=x_i$ type-i (Poisson) customers in the system, $i=1,...,N$, together with the K permanent customers. It is easily seen that the determination of the (conditional) joint sojourn time distribution of C can be performed in almost exactly the same way as for the original M/M/1 feedback queue without permanent customers, see Appendix 2.1, the only difference being that for the present model one has to take into account that after *each* of his services C finds K additional permanent customers in the queue (besides the different types of Poisson customers). Realizing this it follows immediately from the analysis in Appendix 2.1 that, for $\mathrm{Re}\,\omega_j \geqslant 0$, $|z_{i,j}| \leqslant 1$, $i=1,...,N$, $j=0,...,k$, (cf. (2.8))

$$E\{e^{-(\omega_1\mathbf{S}_1+...+\omega_k\mathbf{S}_k)}(z_{1,0}^{\mathbf{X}_1^{(0)}} \cdots z_{N,0}^{\mathbf{X}_N^{(0)}}) \cdots (z_{1,k}^{\mathbf{X}_1^{(k)}} \cdots z_{N,k}^{\mathbf{X}_N^{(k)}}) \mid \mathbf{X}_1^{(0)}=x_1, \ldots, \mathbf{X}_N^{(0)}=x_N\}$$

$$= \left[\prod_{j=1}^{k} A_k^N(j,\omega,z)\right]^{K+1} \prod_{i=1}^{N} (z_{i,0} f_k^N(i,\omega,z))^{x_i}, \tag{5.51}$$

with $\omega:=(\omega_1, \ldots, \omega_k)$, $z:=((z_{1,0}, \ldots, z_{N,0}), \ldots, (z_{1,k}, \ldots, z_{N,k}))$, and with $A_k^N(\cdot,\cdot,\cdot)$ and $f_k^N(\cdot,\cdot,\cdot)$ defined by (2.9) and (2.10). Comparing (5.51) with the result ((2.8)) for the corresponding quantity in the model without permanent customers it appears that these results differ only by a factor $(\prod_{j=1}^{k} A_k^N(j,\omega,z))^K$; note that this could have been obtained directly from the discussion in Remark 2.1.

Using the PASTA property and deconditioning we obtain from (5.50) and (5.51) our main result:

THEOREM 5.2
The joint distribution of the successive sojourn times and the number of Poisson customers of each type present in the system at the service completion epochs of a tagged Poisson customer is the $(K+1)$-fold convolution of the corresponding (joint) distribution in the model without permanent customers, cf. (2.11):

$$E\{e^{-(\omega_1\mathbf{S}_1+...+\omega_k\mathbf{S}_k)}(z_{1,0}^{\mathbf{X}_1^{(0)}} \cdots z_{N,0}^{\mathbf{X}_N^{(0)}}) \cdots (z_{1,k}^{\mathbf{X}_1^{(k)}} \cdots z_{N,k}^{\mathbf{X}_N^{(k)}})\} = \tag{5.52}$$

$$\left[\frac{(1-\rho)\prod_{j=1}^{k} A_k^N(j,\omega,z)}{1-\lambda\beta\sum_{i=1}^{N} q(i) z_{i,0} f_k^N(i,\omega,z)}\right]^{K+1}, \quad \mathrm{Re}\,\omega_j \geqslant 0,\ |z_{i,j}| \leqslant 1,\ i=1,...,N,\ j=0,...,k.$$

Using Theorem 5.2 most of the sojourn time characteristics can be immediately obtained from the results given in Section 2.4. Here we shall restrict ourself to a summary of the most important characteristics.

- The j-th sojourn time $\mathbf{S}_j$ of a Poisson customer has a $(K+1)$-stage Erlang distribution (E_{K+1}) with mean $(K+1)\beta/(1-\rho)$: (cf. (2.20))

$$E\{e^{-\omega_j \mathbf{S}_j}\} = \left[\frac{1-\rho}{1-\rho+\beta\omega_j}\right]^{K+1}, \quad j=1,...,k. \tag{5.53}$$

- The correlation coefficient, $corr(\mathbf{S}_i,\mathbf{S}_j)$, of the i-th and the j-th sojourn time of a Poisson customer is independent of the number of permanent customers in the system: (cf. (2.24))

$$corr(\mathbf{S}_i,\mathbf{S}_j) = 1-(1-\rho)C_{j-i}, \quad 1\leqslant i<j\leqslant k, \tag{5.54}$$

with C_n, $n=1,...,k-1$, determined by (2.22).

- The variance of the total sojourn time after k services, $var(\mathbf{S}^{(k)})$, is given by: (cf. (2.26))

$$var(\mathbf{S}^{(k)}) = (K+1)\left[\frac{\beta}{1-\rho}\right]^2 \left[k^2-2(1-\rho)\sum_{j=1}^{k-1} C_{k-j}\right], \quad k=1,2,.... \tag{5.55}$$

REMARK 5.8
Noting that in the present product form model a departing (and hence arriving) permanent customer sees the system in equilibrium with one less customer of his own type (see e.g. Walrand [1988, Section 3.4]) the characteristics of the successive cycle times of a particular permanent customer can be immediately obtained from the above sojourn time results for the Poisson customers. For example, the cycle times have a K-stage Erlang distribution (E_K) with mean $K\beta/(1-\rho)$, cf. (5.53).

5.6.3 The M/G/1 PS queue with additional permanent customers

In Section 3.3 it has been shown how queue length and sojourn time results for the M/G/1 processor sharing queue can be obtained from queue length and sojourn time results for the M/M/1 queue with general feedback. We applied a limiting procedure in which the mean service time $\beta\rightarrow 0$ while the feedback probabilities approach one in such a way that a customer's total required service time remains constant, see Subsection 3.3.1. It is easily seen that application of the same limiting procedure to the present M/M/1 feedback model with K permanent customers leads to the M/G/1 PS queue with K permanent customers. Note that the behaviour of the latter model is independent of the service time distribution(s) of the permanent customers (the permanent customers are *always* in service). From (5.50) it follows immediately that *for the M/G/1 PS queue with K permanent customers, the distribution of the queue length* $\mathbf{X}^{PS}$ *at an arbitrary epoch is the* $(K+1)$*-fold convolution of the queue length distribution in the same model without permanent customers:* (cf. (3.1))

$$E\{z^{\mathbf{X}^{PS}}\} = \left[\frac{1-\rho}{1-\rho z}\right]^{K+1}, \quad |z| \leqslant 1, \tag{5.56}$$

i.e.

$$Pr\{\mathbf{X}^{PS} = n\} = (1-\rho)^{K+1} \binom{n+K}{K} \rho^n, \quad n = 0, 1, \ldots, \tag{5.57}$$

with ρ the offered load to the system per unit of time due to the Poisson customers. From Theorem 5.2 we obtain the following remarkable sojourn time result:

THEOREM 5.3
For the M/G/1 PS queue with K permanent customers the distribution of the conditional sojourn time $\mathbf{S}^{PS}(x)$ *of a Poisson customer with given service demand* x *is the* $(K+1)$*-fold convolution of the conditional sojourn time in the same model without permanent customers. This also holds for the unconditional sojourn time* $\mathbf{S}^{PS}$ *of an arbitrary Poisson customer.*

Theorem 5.3 implies: (cf. (3.19))

$$E\{\mathbf{S}^{PS}(x)\} = (K+1)\frac{x}{1-\rho}, \quad x \geqslant 0. \tag{5.58}$$

REMARK 5.9
For the present PS model it is interesting to study the influence of the presence of the Poisson customers on the 'speed' with which the permanent customers are served. For $x \geqslant 0$ let $\mathbf{C}^{PS}(x)$ be the time required to give the permanent customers an amount x of service. From the discussion in Remark 5.8 and application of the limiting procedure it follows that $\mathbf{C}^{PS}(x)$ is distributed as the conditional sojourn time of a tagged Poisson customer with service demand x in the same model but with one less permanent customer. For example, from (5.58),

$$E\{\mathbf{C}^{PS}(x)\} = K\frac{x}{1-\rho}, \quad x \geqslant 0. \tag{5.59}$$

This formula shows that the influence of the Poisson customer stream on

$E\{\mathbf{C}^{PS}(x)\}$ is simply a reduction of the capacity of the server by an amount ρ. Moreover, (5.59) implies that the mean *total* amount of service obtained by the permanent customers per unit of time (given by $Kx / E\{\mathbf{C}^{PS}(x)\}$) is independent of K.

REMARK 5.10
In Remark 3.7 we concluded that for the M/G/1 PS queue (without permanent customers) the queue length distribution just after the departure of a tagged customer who has received an amount x of service is the same as at an arbitrary epoch, *independent* of x. From (5.56) it follows that for the M/G/1 PS queue with one permanent customer the queue length distribution at an arbitrary epoch is the two-fold convolution of the queue length distribution in the PS queue without permanent customers. Since one would expect that, when the required service time x of a tagged customer becomes very large, the behaviour of the M/G/1 PS queue approaches that of the corresponding PS queue with one permanent customer, it seems paradoxical that both statements are true. However, viewing the M/G/1 PS queue as the limiting case of the M/M/1 queue with general feedback this is immediately clear (the departure of a tagged customer in the PS model corresponds to the (last) service completion of a tagged customer in the feedback model which is more likely to occur when there are fewer customers in the system). A similar 'paradox' for queue lengths in PS queues is discussed in Foley and Klutke [1989].

REMARK 5.11
It should be noted that the M/G/1 PS queue with K permanent customers can also be viewed as a special case of generalized processor sharing (GPS), cf. Section 3.4: if there are j customers present in the system then the service rate for each of these customers is $f(j)=1/(j+K)$, $j=1,2,\ldots$. It is easily verified that (5.57) and (5.58) coincide with the results for the queue length distribution and the mean sojourn time in the M/G/1 GPS queue given by (3.59) and (3.60) respectively, that have already been obtained by Cohen [1979]. Theorem 5.3 is a new result.

REMARK 5.12
Using Theorem 5.3 the approximations for second moment characteristics of the sojourn time in the ordinary M/G/1 PS queue (developed in Chapter 4) can be easily extended to approximations for the corresponding quantities in the PS model with permanent customers.

The above results for the queue length and the sojourn time in the M/G/1 PS queue with permanent customers are interesting both from a theoretical and a practical point of view. One example where this queueing model may arise is provided by a 'Stored Program Controlled' (SPC) telephone exchange that is offered two types of jobs: (i) call requests, and (ii) operator tasks (see De Waal [1989]). To guarantee a certain quality of service of the call requests only a limited number (K) of operator tasks is allowed to be in service at the same time. It is clear that under heavy traffic conditions of the operator tasks and for appropriate assumptions about the system parameters the above formulas (5.56)-(5.58) (approximately) reflect the influence of the choice of the control parameter K on the queue length and the delay of the call requests. From the discussion in Remark 5.9 it follows that under certain conditions the *maximum* throughput of the operator tasks is independent of K. So, if the objective is to minimize the delay of the call requests and to maximize the throughput of the operator tasks one should take K as small as possible, i.e. $K=1$.

References

Adiri, I. (1972). Queueing models for multiprogrammed computers. In: *Proc. Symp. on Computer-Communications Networks and Teletraffic, Polytechnic Institute of Brooklyn,* 441-448.

Adiri, I., Avi-Itzhak, B. (1969A). A time-sharing queue. *Management Science* **15**, 639-657.

Adiri, I., Avi-Itzhak, B. (1969B). A time-sharing model with many queues. *Oper. Res.* **17**, 1077-1089.

Ali, O.M.E., Neuts, M.F. (1984). A service system with two stages of waiting and feedback of customers. *J. Appl. Prob.* **21**, 404-413.

Asare, B.K., Foster, F.G. (1983). Conditional response times in the M/G/1 processor-sharing system. *J. Appl. Prob.* **20**, 910-915.

Avi-Itzhak, B., Halfin, S. (1988). Response times in M/M/1 time sharing schemes with a limited number of service positions. *J. Appl. Prob.* **25**, 579-595.

Avi-Itzhak, B., Halfin, S. (1989A). Expected response times in a non-symmetric time sharing queue with limited number of service positions. In: *Teletraffic Science for New Cost-Effective Systems, Networks and Services,* ed. M. Bonatti. North-Holland Publ. Cy., Amsterdam, 1485-1493.

Avi-Itzhak, B., Halfin, S. (1989B). Response times in gated M/G/1 queues: the processor-sharing case. *Queueing Systems* **4**, 263-279.

Baskett, F., Chandy, K.M., Muntz, R.R., Palacios, F.G. (1975). Open, closed, and mixed networks of queues with different classes of customers. *J. ACM* **22**, 248-260.

Van den Berg, J.L. (1988). *Unpublished work.*

Van den Berg, J.L. (1989). Simple approximations for second moment characteristics of the sojourn time in the M/G/1 processor sharing queue. In: *Messung, Modellierung und Bewertung von Rechensystemen und Netzen,* eds. G. Stiege, J.S. Lie. Springer-Verlag, Berlin, 105-120.

Van den Berg, J.L. (1990). Queueing models with additional permanent customers. *Report Centre for Mathematics and Computer Science, Amsterdam* (to appear).

Van den Berg, J.L., Boxma, O.J. (1989A). Sojourn times in feedback queues. In: *Operations Research Proceedings 1988,* eds. D. Pressmar et al.. Springer-Verlag, Berlin, 247-257.

Van den Berg, J.L., Boxma, O.J. (1989B). Sojourn times in feedback and processor sharing queues. In: *Teletraffic Science for New Cost-Effective Systems, Networks and Services,* ed. M. Bonatti. North-Holland Publ. Cy., Amsterdam, 1467-1475.

Van den Berg, J.L., Boxma, O.J., Groenendijk, W.P. (1989). Sojourn times in the M/G/1 queue with deterministic feedback, *Stochastic Models* **5**, 115-129.

Boxma, O.J. (1986A). A queueing model of finite and infinite source interaction. *Oper. Res. Letters* **5**, 245-254.

Boxma, O.J. (1986B). Models of two queues - a few new views. In: *Teletraffic Analysis and Computer Performance Evaluation,* eds. O.J. Boxma, J.W. Cohen, H.C. Tijms. North-Holland Publ. Cy., Amsterdam.

Boxma, O.J., Cohen, J.W. (1989). The M/G/1 queue with permanent customers. *Report BS-R8919, Centre for Mathematics and Computer Science, Amsterdam.*

Boxma, O.J., Daduna, H. (1989). Sojourn times in queueing networks. *Report BS-R8916, Centre for Mathematics and Computer Science, Amsterdam.* To appear in: *Stochastic Analysis of Computer and Communication Systems,* ed. H. Takagi. North-Holland Publ. Cy., Amsterdam, 1990.

Çinlar, E. (1975). *Introduction to Stochastic Processes.* Prentice-Hall, Englewood Cliffs (NJ).

Coffman, E.G., Kleinrock, L. (1968). Feedback queueing models for time-shared systems. *J. ACM* **15**, 549-576.

Coffman, E.G., Muntz, R.R., Trotter, H. (1970). Waiting time distributions for processor-sharing systems. *J. ACM* **17**, 123-130.

Cohen, J.W. (1976). *On Regenerative Processes in Queueing Theory.* Springer-Verlag, Berlin.

Cohen, J.W. (1979). The multiple phase service network with generalized processor sharing. *Acta Informatica* **12**, 245-284.

Cohen, J.W. (1982). *The Single Server Queue,* 2nd ed.. North-Holland Publ. Cy., Amsterdam.

Disney, R.L. (1981). A note on sojourn times in M/G/1 queues with instantaneous Bernoulli feedback. *Nav. Res. Log. Quart.* **28**, 679-684.

Disney, R.L., König, D. (1985). Queueing networks: a survey of their random processes. *SIAM Review* **27**, 335-403.

Disney, R.L., König, D., Schmidt, V. (1984). Stationary queue-length and waiting-time distributions in single-server feedback queues. *Adv. Appl. Prob.* **16**, 437-446.

Disney, R.L., McNickle, D.C., Simon, B. (1980). The M/G/1 queue with instantaneous Bernoulli feedback. *Nav. Res. Log. Quart.* **27**, 635-644.

Doshi, B.T. (1986). Queueing systems with vacations - a survey. *Queueing Systems* **1**, 29-66.

Doshi, B.T., Kaufman, J.S. (1988). Sojourn time in an M/G/1 queue with Bernoulli feedback. In: *Queueing Theory and its Applications - Liber Amicorum for J.W. Cohen,* eds. O.J. Boxma and R. Syski. North-Holland Publ. Cy., Amsterdam, 207-233.

Doshi, B.T., Rege, K.M. (1985). Analysis of a multistage queue. *AT&T Tech. J.* **64**, 369-390.

Doshi, B.T., Wong, W.S. (1987). Exact solution of a simple finite infinite source interaction model. *Queueing Systems* **2**, 67-82.

Feller, W. (1950). *An Introduction to Probability Theory and Its Applications, Vol. I.* Wiley, New York.

Foley, R.D., Klutke, G.A. (1989). Stationary increments in the accumulated work process in processor-sharing queues. *J. Appl. Prob.* **26**, 671-677.

Fontana, B., Diaz Berzosa, C. (1984). Stationary queue-length distributions in an M/G/1 queue with two non-preemptive priorities and general feedback. In: *Performance of Computer-Communication Systems,* eds. H. Rudin and W. Bux. North-Holland Publ. Cy., Amsterdam, 333-347.

Fontana, B., Diaz Berzosa, C. (1985). M/G/1 queue with N-priorities and feedback: joint queue-length distributions and response time distribution for any particular sequence. In: *Teletraffic Issues in an Advanced Information Society,* ITC-11, ed. M. Akiyama. North-Holland Publ. Cy., Amsterdam, 452-458.

Fuhrmann, S.W., Cooper, R.B. (1985). Stochastic decomposition in the M/G/1 queue with generalized vacations. *Oper. Res.* **33**, 1117-1129.

Groenendijk, W.P. (1990). *Conservation Laws in Polling Systems,* Ph.D. Thesis. Centre for Mathematics and Computer Science, Amsterdam.

Hunter, J.J. (1989). Sojourn time problems in feedback queues. *Queueing Systems* **5**, 55-76.

Jackson, J.R. (1957). Networks of waiting lines. *Oper. Res.* **5**, 518-521.

Jackson, J.R. (1963). Jobshop-like queueing systems. *Management Science* **10**, 131-142.

Jaiswal, N.K. (1982). Performance evaluation studies for time-sharing computer systems. *Perfermance Evaluation* **2**, 223-236.

Kaufman, J.S. (1985). Finite and infinite source interactions. In: *Performance '84,* ed. E. Gelenbe. North-Holland Publ. Cy., Amsterdam.

Kelly, F.P. (1979). *Reversibility and Stochastic Networks.* Wiley, New York.

Kleinrock, L. (1964). Analysis of a time-shared processor. *Nav. Res. Log. Quart.* **11**, 59-73.

Kleinrock, L. (1967). Time-shared systems: a theoretical treatment. *J. ACM* **14**, 242-261.

Kleinrock, L. (1975). *Queueing Systems, Vol. I.* Wiley, New York.

Kleinrock, L. (1976). *Queueing Systems, Vol. II.* Wiley, New York.

Klutke, G.A., Kiessler, P.C., Disney, R.L. (1988). Interoutput times in processor sharing queues with feedback. *Queueing Systems* **3**, 363-376.

Lam, S.S., Shankar, A.U. (1981). A derivation of response time distributions for a multi-class feedback queueing system. *Performance Evaluation* **1**, 48-

61.

Lemoine, A.J. (1979). Total sojourn time in networks of queues. *Report TR 79-020-1, Systems Control, Inc., Palo Alto, CA.*

Muntz, R.R. (1972). Waiting time distribution for round-robin queueing systems. In: *Proc. Symp. on Computer-Communications Networks and Teletraffic, Polytechnic Institute of Brooklyn,* 429-439.

Nelson, R.D. (1987). Expected response time for a FCFS feedback queue with multiple classes. *Research Report RC 13221, IBM Research Division, T.J. Watson Research Center, Yorktown Heights, N.Y.*

O'Donovan, T.M. (1974). Direct solutions of M/G/1 processor-sharing models. *Oper. Res.* **22**, 1232-1235.

Ott, T.J. (1984). The sojourn-time distribution in the M/G/1 queue with processor sharing. *J. Appl. Prob.* **21**, 360-378.

Rege, K.M., Sengupta, B. (1989). A single server queue with gated processor-sharing discipline. *Queueing Systems* **4**, 249-261.

Resing, J.A.C., Hooghiemstra, G., Keane, M.S. (1989). The M/G/1 processor sharing queue as the almost sure limit of feedback queues. *Report no. 89-32, Faculty of Technical Mathematics and Informatics, Delft University of Technology.* To appear in: *J. Appl. Prob.* **27**.

Sakata, M., Noguchi, S., Oizumi, J. (1969). Analysis of a processor shared queueing model for time sharing systems. *Proc. 2nd Hawaii Int. Conf. on System Sciences,* 625-628.

Sakata, M., Noguchi, S., Oizumi, J. (1971). An analysis of the M/G/1 queue under round-robin scheduling. *Oper. Res.* **19**, 371-385.

Schassberger, R. (1981). On the response time distribution in a discrete round-robin queue. *Acta Informatica* **16**, 57-62.

Schassberger, R. (1984). A new approach to the M/G/1 processor-sharing queue. *Adv. Appl. Prob.* **16**, 202-213.

Schrage, L.E. (1967). The queue M/G/1 with feedback to lower priority queues. *Management Science* **13**, 466-474.

Shanthikumar, J.G. (1988). On stochastic decomposition in M/G/1 type queues with generalized server vacations. *Oper. Res.* **36**, 566-569.

Simon, B. (1984). Priority queues with feedback. *J. ACM* **31**, 134-149.

Stoyan, D. (1983). *Comparison Methods for Queues and Other Stochastic Models.* Wiley, New York.

Takàcs, L. (1963). A single-server queue with feedback. *Bell System Tech. J.* **42**, 505-519.

Takagi, H. (1986). *Analysis of Polling Systems.* The MIT Press, Cambridge, Massachusetts.

Tijms, H.C. (1986). *Stochasic Modeling and Analysis.* Wiley, New York.

Titchmarsh, E.C. (1968). *The Theory of Functions,* 2nd ed.. Oxford University

Press, London.

Voelker, D., Doetsch, G. (1950). *Die Zweidimensionale Laplace-Transformation.* Verlag Birkhäuser, Basel.

De Waal, P.R. (1989). An approximation method for a processor sharing queue with controlled arrivals and a waitbuffer. In: *Proc. of the 28th IEEE Conf. on Decision and Control, Tampa, FL.*

Walrand, J. (1988). *An Introduction to Queueing Networks.* Prentice-Hall, Englewood Cliffs (NJ).

Walrand, J., Varaiya, P. (1980). Sojourn times and the overtaking condition in Jacksonian networks. *Adv. Appl. Prob.* **12**, 1000-1018.

Wang, Y.T. (1981). An analysis of a round-robin schedule with preemptive priorities. In: *Performance '81,* ed. F.J. Kylstra. North-Holland Publ. Cy., Amsterdam, 147-157.

Whitt, W. (1982). Approximating a point process by a renewal process, I: Two basic methods. *Oper. Res.* **30**, 125-147.

Whitt, W. (1984). On approximations for queues, III: Mixtures of exponential distributions. *AT&T Bell Lab. Tech. J.* **63**, 163-175.

Wolff, R.W. (1982). Poisson arrivals see time averages. *Oper. Res.* **30**, 223-231.

Yashkov, S.F. (1983). A derivation of response time distribution for a M/G/1 processor sharing queue. *Problems Contr. & Info. Theory* **12**, 133-148.

Yashkov, S.F. (1986). A note on asymptotic estimates of the sojourn time variance in the M/G/1 queue with processor-sharing. *Syst. Anal. Model. Simul.* **3**, 267-269.

Yashkov, S.F. (1987). Processor-sharing queues: some progress in analysis. *Queueing Systems* **2**, 1-17.

SAMENVATTING

Wanneer meerdere gebruikers, op hetzelfde moment, de beschikking willen hebben over de diensten die door een systeem (bijvoorbeeld een computer) worden aangeboden zal vaak slechts een gedeelte van deze gebruikers direkt tot het systeem kunnen worden toegelaten en moeten de anderen op hun beurt *wachten.* De *wachtrijtheorie*, een tak van de toegepaste kansrekening, houdt zich bezig met het bestuderen van dit verschijnsel 'wachten' in systemen die diensten aanbieden voor collectief gebruik; in het bijzonder probeert men m.b.v. de wachtrijtheorie de *prestatie* van zulke systemen te bepalen. De wachtrijtheorie vindt zijn oorsprong in het onderzoek naar de vertraging en congestie bij het aanvragen van telefoongesprekken, rond het begin van deze eeuw. Tegenwoordig wordt de wachtrijtheorie op veel gebieden toegepast, in het bijzonder bij productieplanning en bij de prestatie analyse van computer- en communicatiesystemen.

Een wachtrijmodel wordt gewoonlijk beschreven in termen van klanten die een hoeveelheid bediening vragen, bedieningsfaciliteiten die bediening verlenen, en wachtrijen die op bediening wachtende klanten bevatten; de volgorde waarin de klanten worden bediend wordt bepaald door de bedieningsdiscipline. In dit proefschrift worden wachtrijmodellen bestudeerd waarin klanten meerdere malen naar een bepaalde bedieningsfaciliteit terug kunnen keren om zo verschillende bedieningsfasen te ontvangen voordat ze definitief het systeem verlaten. Zulke '*feedback*' verschijnselen treden o.a. op in processen die zich voordoen in computercommunicatie- en in productienetwerken. Een belangrijk voorbeeld is een computersysteem met *multiprogramming*. In zo'n systeem kunnen meerdere jobs 'tegelijkertijd' worden behandeld door elke job een kleine hoeveelheid tijd (een bedieningsquantum) toe te wijzen gedurende welke de job de beschikking over de CPU krijgt. Als de totale verwerkingstijd van een job groter is dan het hem toegewezen bedieningsquantum wordt hij in een wachtrij geplaatst; hier wacht de job tot hij aan de beurt is om opnieuw een bedieningsquantum te ontvangen, enzovoort; als de job uiteindelijk klaar is verlaat hij het systeem. Een ander voorbeeld van feedback vindt men in productieprocessen waarbij na het verrichten van een handeling gecontroleerd wordt of deze handeling eventueel opnieuw uitgevoerd dient te worden.

Het basis feedbackmodel dat in dit proefschrift wordt bestudeerd bestaat uit een bedieningsstation met één bediende en een wachtrij waarbij klanten arriveren volgens een Poisson proces. Nadat een klant een negatief exponentieel verdeelde hoeveelheid bediening (met gemiddelde β) heeft ontvangen keert hij terug naar het eind van de wachtrij (om later een volgende bediening te ontvangen) óf hij verlaat het systeem, respectievelijk met kans $p(i)$ en met kans $1-p(i)$; hierbij is $i=1,2,...$ het aantal keer dat de klant reeds bediend is. De klanten worden bediend in de volgorde waarin ze in de wachtrij staan. Dit model wordt het 'M/M/1 model met algemene feedback' of kortweg het

'M/M/1 feedbackmodel' genoemd. Het onderzoek in dit proefschrift is in de eerste plaats gericht op het bestuderen van de (kansverdeling van de) *verblijftijd* van een klant in zo'n systeem. De totale verblijftijd van een klant is de som van een stochastisch aantal partiële verblijftijden tussen de opeenvolgende bedieningen van die klant (de eerste partiële verblijftijd is de tijd tussen de aankomst van de klant en het eind van zijn eerste bediening). Het is duidelijk dat deze opeenvolgende (partiële) verblijftijden niet onafhankelijk zijn, hetgeen de analyse bemoeilijkt. Een ander aspect dat de bepaling van de verblijftijdverdeling in de meeste wachtrijmodellen met feedback zeer moeilijk of onmogelijk maakt is het optreden van het zogenaamde '*inhaal*' verschijnsel: de klanten verlaten het systeem niet noodzakelijk in de volgorde waarin ze het systeem binnenkomen. Een feedback wachtrijmodel is in feite het eenvoudigste voorbeeld van een *wachtrijnetwerk* waarin dit inhaal verschijnsel optreedt. Het bestuderen van verblijftijden in een één-bediende wachtrij met feedback is dan ook van belang voor het onderzoek naar verblijftijden in grotere wachtrijnetwerken waarin de klanten elkaar kunnen passeren; hiervoor zijn in de literatuur nog nauwelijks resultaten bekend.

In het proefschrift leiden we een uitdrukking af voor de samengestelde verdeling van de achtereenvolgende (partiële) verblijftijden van een klant in het hierboven beschreven M/M/1 feedbackmodel. Een belangrijk bijprodukt van de feedbackstudie is nieuw inzicht in het gedrag en de analyse van het bekende en veel gebruikte M/G/1 '*processor sharing*' (PS) model voor computersystemen met multiprogramming. In dit één-bediende model met algemeen verdeelde bedieningstijden worden alle aanwezige klanten tegelijkertijd en met gelijke snelheid bediend, zodanig dat de totale bedieningssnelheid constant blijft. In feite is het processor sharing model een model van een computer systeem met multiprogramming waarin de lengte van de bedieningsquanta naar nul gaat. Het is niet moeilijk in te zien dat wanneer in het M/M/1 feedbackmodel de gemiddelde bedieningstijd β naar nul gaat en de terugkeerkansen naar één zodanig dat de gemiddelde totale hoeveelheid bediening die een klant krijgt constant blijft, dit model zich precies hetzelfde gedraagt als het M/G/1 PS model. Verschillende keuzes van de terugkeerkansen in het M/M/1 feedbackmodel leiden tot verschillende bedieningsduurverdelingen in het PS model.

We geven nu een kort overzicht van de inhoud van de vijf hoofdstukken.

Hoofdstuk 1 is een algemene inleiding waarin de praktische en theoretische achtergronden van het onderzoek worden belicht. Dit hoofdstuk bevat ook een uitgebreid overzicht van de literatuur m.b.t. de in het proefschrift behandelde modellen. In de laatste paragraaf wordt een overzicht van de inhoud van de hoofdstukken 2-5 gegeven.

Hoofdstuk 2 bevat een fundamentele analyse van de verblijftijden in het M/M/1 feedbackmodel. We leiden eerst een recursieve uitdrukking af voor de samengestelde verdeling van de opeenvolgende (partiële) verblijftijden en van het aantal klanten in het systeem direkt na elke bediening van een klant die precies $k \geqslant 1$ keer bediend wordt. Met behulp van dit resultaat vinden we tamelijk eenvoudige uitdrukkingen voor de verschillende verblijftijdkarakteristieken, zoals de variantie van de totale verblijftijd na k bedieningen, de verdeling van de i-de (partiële) verblijftijd, $i = 1,...,k$, en de correlatiecoëfficient van de i-de en de j-de verblijftijd, $1 \leqslant i < j \leqslant k$. In het bijzonder wordt aangetoond dat deze laatste grootheid positief is en kleiner wordt als $j - i$ toeneemt, hetgeen op intuïtieve gronden ook verwacht mag worden. Hoofdstuk 2 wordt afgesloten met de analyse van een uitbreiding van het basismodel waarin de verdeling van de bedieningsduur van een klant algemeen is en afhangt van het aantal keren dat een klant reeds bediend is. De analyse van dit model is beperkt tot het bepalen van gemiddelde verblijftijden.

In hoofdstuk 3 wordt aangetoond hoe de verblijftijdresultaten voor het M/M/1 feedbackmodel, via de eerder geschetste limiet procedure, leiden tot resultaten voor de verblijftijd in het M/G/1 PS model. We geven o.a. de afleidingen voor het gemiddelde, de variantie en de Laplace-Stieltjes transformatie van de verblijftijd. Het blijkt dat deze nieuwe benadering van de analyse van het M/G/1 PS model veel inzicht geeft in een aantal bekende eigenschappen van de verblijftijd. De laatste paragraaf van hoofdstuk 3 is gewijd aan de analyse van een M/G/1 processor sharing model met feedback (PSFB) waarin de terugkeerkansen constant zijn. Uit de resultaten voor het M/M/1 feedbackmodel en m.b.v. de limiet procedure worden nieuwe resultaten afgeleid voor de correlatiecoëfficienten van de opeenvolgende (partiële) verblijftijden van een klant in dit PSFB model.

In hoofdstuk 4 worden enige benaderingsformules ontwikkeld voor het tweede moment van de verblijftijdverdeling in het M/G/1 PS model. De reden hiervoor is dat de in hoofdstuk 3 verkregen exacte uitdrukkingen in het algemeen alleen numeriek kunnen worden geëvalueerd en bovendien afhankelijk zijn van de *verdeling* van de bedieningsduur (die in de praktijk slechts zelden bekend is). De benaderingen hangen slechts af van de eerste twee momenten van de bedieningsduurverdeling. Ze zijn voornamelijk gebaseerd op enkele nieuwe asymptotische resultaten en op simpele exacte uitdrukkingen voor een aantal specifieke bedieningsduurverdelingen. De benaderingsresultaten worden uitgebreid getest aan de hand van (deels numeriek verkregen) exacte resultaten. Tenslotte wordt een verfijnder benadering afgeleid door ook het derde moment van de bedieningsduurverdeling erbij te betrekken.

In hoofdstuk 5 worden enige één-bediende wachtrijmodellen bestudeerd waarin naast de ‘gewone’ klanten een vast aantal permanente klanten in het systeem aanwezig is; de permanente klanten keren na iedere bediening terug naar het einde van de wachtrij. Centraal staat de vraag: wat is de invloed van de aanwezigheid van deze permanente klanten op de rijlengte en verblijftijd van de gewone klanten? Deze vraag wordt eerst beantwoord voor het M/G/1 (first-come-first-served) model met permanente klanten. Daarna wordt het M/M/1 feedbackmodel met permanente klanten geanalyseerd. Het blijkt dat voor dit laatste model de aanwezigheid van K permanente klanten leidt tot een verblijftijdverdeling die de $(K+1)$-voudige convolutie is van die voor het oorspronkelijke model (zonder permanente klanten). Een soortgelijk resultaat wordt afgeleid voor de verblijftijd in het M/G/1 PS model met permanente klanten.

CURRICULUM VITAE

De schrijver van dit proefschrift werd op 17 november 1961 geboren te Lexmond. Na het behalen van het Atheneum-B diploma aan het Chr. Lyceum te Gouda in 1980 begon hij de studie Wiskunde met bijvak Informatica aan de Rijksuniversiteit Utrecht. In januari 1986 studeerde hij af in de Toegepaste Wiskunde bij prof. dr. ir. J.W. Cohen. Het afstudeerwerk, een wachtrijstudie m.b.t. de prestatie analyse van een communicatienetwerk met flow control, werd verricht tijdens een stage bij het Centrum voor Wiskunde en Informatica (CWI) te Amsterdam. Van februari 1986 tot februari 1990 was hij als wetenschappelijk medewerker aan dit instituut verbonden. Het onderzoek dat hij gedurende die periode verrichtte onder begeleiding van prof. dr. ir. O.J. Boxma en prof. dr. ir. J.W. Cohen heeft geleid tot de totstandkoming van dit proefschrift.