CWI Tracts

Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

The CWI is a research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a nonprofit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

CWI Tract 8

Algorithms and approximations for queueing systems

M.H. van Hoorn

Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

1980 Mathematics Subject Classification: 60K25
ISBN 90 6196 274 9

Aknowledgements

I wish to thank all who have contributed to this monograph in one way or another during the research phase, the writing phase and the production phase. In particular I should mention Henk Tijms, Luuk Seelen, Paul Kühn, Phuoc Tran Gia, Egbert Kunst and my wife Beatrijs for their continuing support and interest.

Also, I thank the Institute for Actuarial Sciences and Econometrics at the Vrije Universiteit, Amsterdam, for the stimulating working environment wherein I have carried out the research and the Mathematical Centre in Amsterdam for the opportunity to publish this monograph in their series CWI Tracts.

Contents

0. INTRODUCTION AND SUMMARY

This monograph deals with a number of rather basic models from queueing theory. In particular, the attention is focused on the development of computational algorithms.

The question may arise whether after about 75 years of research in queueing, it is still possible to make a substantial contribution to the theory and to come up with some new results. To account for our positive answer to this question, and to place the present work in its proper context, we first give a short historical survey of the origin and the development of queueing theory and subsequently discuss the current state of affairs.

It is the Danish mathematician A.K. Erlang who is considered to have founded queueing theory in the beginning of this century. He was involved in queueing problems that occurred in automatic telephone exchanges. An example is the problem of finding the loss probability, i.e. the fraction of the incoming calls finding all trunks busy, as function of the offered traffic in a telephone switching system with a certain number of trunks. These systems were typical loss queueing systems with no waiting room. Erlang's name is still attached to some important concepts in queueing theory such as the Erlang delay probability and the phase method of Erlang.

Until 1940, the majority of the contributions to queueing theory was made by people active in the field of telephone traffic problems; cf. the book of Kosten[73]. After the second world war, the field of operations research rapidly developed and queueing applications were also found in production planning, inventory control and maintenance problems. In this period, much theoretically oriented research on queueing problems was done. The emphasis on theoretical aspects is, however, not surprising in view of the lack of computing facilities at that time.

In the fifties and sixties, the theory reached a very high mathematical level; see e.g. Cohen[69] and Takacs[62]. Advanced mathematical techniques like transform methods, Wiener Hopf decomposition and function theoretic tools were developed and refined. This research resulted in a number of elegant mathematical solutions.

Simultaneously with the progress in the mathematical theory of queueing, another important development was taking place in the fifties. The rise of the computer and other technically advanced digital equipment created a new field for applied queueing theory. In fact, on the one hand the computer enabled numerical work to be done, while on the other hand quantitative results of queueing models were needed for the design and performance evaluation of computers and telecommunication installations.

Unfortunately, the gap between theory and practical applications had grown too wide. Many solutions were given in a form resulting in numerically unstable solution procedures or in a form wholly unsuitable for computations. A famous example is, as Kendall[64] stated, ' the Laplacian curtain which has hitherto obscured much of the detail of the queue-theoretic scene'. The fact that many theorists neglected the numerical aspects of their work and the lack of time of practitioners to explore the numerical possibilities of new analytical results have prevented the application of potentially useful methods.

An important impulse to computational queueing analysis was given by research workers with an electrical engineering and computer science background; cf. Kleinrock[75] and Kühn[72]. Their aim was primarily to develop algorithms that were fast and easy to implement.

In the last ten years, much effort has been spent on numerical work for queueing models. In the course of this work many approaches have been explored besides the continuing attempts to implement the existing theoretical results. Rather than on the search for explicit results, the attention has been focused on algorithmic solution procedures. The most frequently used technique, which is essentially based on Erlang's phase method, is to model a queueing system as a continuous time Markov chain. By writing down the equilibrium equations, a system of linear equations for the state probabilities is obtained. These linear equations can effectively be solved by a proper application of iterative methods such as successive overrelaxation and aggregation/disaggregation; cf. Takahashi and Takami[76] and Groenevelt, van Hoorn and Tijms[82]. Obviously, this is a purely numerical approach not providing any qualitative insight. Such insight is indeed provided by the matrix geometric method developed by Neuts[81]. However, we feel that for numerical purposes alone the iterative methods mentioned earlier are in general to be preferred because of their ease of implementation and their effectiveness.

In view of the computational problems associated with exact solutions of complex queueing models, a lot of effort is being put into the development of approximations, using intuitive reasoning and heuristic methods. In particular, for the multiserver queue with Poisson input several good quality approximations have been found. These approximations will be reviewed later.

The effort put into the derivation of approximations for queueing models is justified by the fact that the models investigated are often themselves approximations of real world applications. In such applications, modeling and measurement errors may reduce the need for exact solutions.

In this monograph, we give for a wide class of queueing models recursive computational schemes for the state probabilities and other performance measures. The ultimate goal is all the time to obtain practical useful results and therefore the analysis is exact whenever possible and approximate whenever exact methods lead to intractable results. We want to emphasize the power of recursive methods and the need to think 'recursively' when developing numerical methods. We have not avoided the job of doing the actual numerical calculations in order to find out whether the proposed methods are indeed useful for practical purposes. Many numerical results and illustrations will be given.

As method of analysis we use the regenerative method first put forward by Hordijk and Tijms[76] and later applied to the M/G/1 queue with variable service rate in Federgruen and Tijms[80]. This is a unifying and intuitively appealing approach which uses results from the theory of regenerative processes. In the analysis we use up and down crossing properties of the queue length process. These up and down crossing properties are in fact the counterpart of the continuity theorem in physics concerning the principle of conservation of flow. A well known relationship derived with these properties is the equality of arriving customer and departing custo-

mer distributions in systems where customers arrive one at a time and leave one at a time. Also, throughout this monograph we shall in our analysis frequently use the fundamental property that 'Poisson arrivals see time averages'. We devote an appendix to the essence of this property, since by doing so we can streamline many of our derivations. The main organization of this monograph is as follows.

In Chapter 1, we introduce the regenerative method on the basis of the standard M/G/1 queue, i.e. the single server queue with Poisson input and general service times. We derive a recursive algorithm for the steady state probabilities of the number of customers in the system. We show that our algorithm is to be preferred above the classical algorithm obtained from the embedded Markov chain approach.

In Chapter 2, we consider an extension of the M/G/1 queue. We assume that the arrival rate of customers depends on the number of customers in the system. As special cases, we discuss the M/G/1 queue with finite capacity and the machine repair problem with a single repairman. Also, we give an algorithm to compute the waiting time distribution.

Chapter 3 is devoted to an approximate analysis for the intrinsically difficult M/G/c queue, i.e. the multiserver queue with Poisson input and general service times. Based on an approximation assumption concerning the residual service times of services in progress at a service completion epoch, we give a complete analysis for this model. In addition to algorithms for the state probabilities and the moments of the queue length and waiting time, we deal with the waiting time distribution function and the output process. Also asymptotic results for the state probabilities and the waiting time distribution will be given. At the end of Chapter 3 we give extensive numerical results and discuss the quality of the approximations. We not only validate our own general purpose approximations, but also review the various other approximation formulae for the mean queue length given in the literature.

In Chapter 4, we discuss briefly the extension of the results of Chapter 3 to an M/G/c queue with a state dependent Poisson arrival process. Again, we give special attention to the finite capacity model and the machine repair model with multiple repairmen.

In Chapter 5, we consider the $M^X/G/1$ queue. In this model the customers arrive in batches rather than singly and the batch size distribution depends on the state of the system. We give an algorithm for the state probabilities. In this model the arriving customer and the departing customer distributions clearly are not equal, since customers arrive in batches and leave singly. Using a batch arrival queueing model, we discuss as application an approximate method to compute the waiting time distribution in the M/G/1 queue with bounded sojourn time. We conclude this chapter with another application, namely the modeling of customer behaviour in telephone switching systems. In particular, we focus on the call completion rate, i.e. the fraction of customers who complete their calls successfully. Essentially, we consider here a discrete version of a queueing model in which the service time of a customer depends on his waiting time in the queue.

In Chapter 6, we analyze the SPP/G/1 queue. The arrival process in this model is a switched Poisson process (SPP), i.e. the rate of the arrival process alternates between two values. The limiting distribution of the interarrival intervals is hyperexponential, but the arrival process is in general not a renewal process. We end Chapter 6 with several numerical examples.

To avoid unnecessary interruptions of the text, we have shifted some details to appendices. In Appendix A, we discuss the 'Poisson Lemma', which forms the essence of the property 'Poisson arrivals see time averages', and review a number of basic results from probability theory and renewal theory. In Appendix B, we present some numerical auxiliary routines required for the implementation of the algorithms to be presented. Appendix C is devoted to the details of the algorithms given in Chapter 3. In Appendix D, we display the exact results for multiserver queues we have used for the validation of the approximations in Chapter 3.

1. INTRODUCTION TO THE REGENERATIVE METHOD

In this chapter, we present the regenerative method. By this technique a wide class of queueing systems can be systematically and thoroughly studied. The method is based on probabilistic arguments and leads in many cases to recursive and numerically stable algorithms for the steady state probabilities of the number of customers in the system. Moreover, the basic elements of the method allow of a clear and intuitive interpretation, thus increasing the insight in the system.

The origin of the regenerative method can be found in Hordijk and Tijms[76], but there the method has not yet crystallized into its present form. See also Federgruen and Tijms[80] where the method has been applied to the M/G/1 queue with variable service rates.

As the name of the method suggests, we consider regeneration cycles of the queueing process. In a cycle, we define quantities strongly related to the state probabilities. We derive for these quantities two sets of recurrence relations which can be solved efficiently. One of these sets of relations is obtained using an up and down crossing argument. For clarity of presentation, we explain the regenerative method on the basis of the standard M/G/1 queue. We review step by step all components of the method. The extension of the results of this chapter to e.g. the M/G/1 queue with server vacations is straightforward.

At the end of the chapter, we compare the regenerative method to two other techniques, the embedded Markov chain approach and a method based on up and down crossing properties of the queue length process.

1.1. The M/G/1 queue

We consider a single server queueing system with an infinite waiting capacity where customers arrive according to a Poisson process at rate λ. The service time $\mathbf{S}$ of a customer has a general probability distribution function $F(t)=\Pr\{\mathbf{S}\leq t\}$. the traffic intensity $\rho=\lambda E\mathbf{S}$ is less than one, i.e. the queue is stable

We are interested in the characteristics of the system when it is in statistical equilibrium. Unless stated otherwise we assume for the sake of convenience that at epoch 0 the system becomes empty after a service completion. We focus on the following steady state probabilities.

$p_n = \lim_{t\to\infty} \Pr\{$ at time t there are n customers in the system $\}$, $n \geq 0$

$q_n = \lim_{k\to\infty} \Pr\{$ the k^{th} customer leaves behind n customers in the system upon service completion $\}$, $n \geq 0$

$\pi_n = \lim_{k\to\infty} \Pr\{$ the k^{th} customer sees upon arrival n customers in the system $\}$, $n \geq 0$

These limits are well defined and are independent of the initial state of the queue length process; cf. Stidham[72]. At an arbitrary epoch (or seen by an outside observer) the distribution of the number of customers in the system is given by (p_n), at an arrival epoch by (π_n) and at a departure epoch by (q_n).

Although the probabilities (p_n), (q_n) and (π_n) are defined as limiting probabilities, we can interpret them without referring to the M/G/1 queue at a remote time when the system has reached the steady state. We can relate these steady state

probabilities to the behaviour of the system in a busy cycle. Therefore, we exploit the property that for the M/G/1 queue the process describing the queue length is a regenerative process. As regeneration points we can take the epochs at which the server becomes idle, i.e. the epochs at which a customer completes service and leaves no customers behind in the system. Indeed at these epochs the future behaviour of the queueing process is independent of its past. Moreover, the continuation of the process beyond a regeneration epoch is a probabilistic replica of the whole process starting at epoch 0. The time period between two successive regeneration points is called a busy cycle. Now, we study the behaviour of the queueing system during a fixed busy cycle; see Figure 1.1.

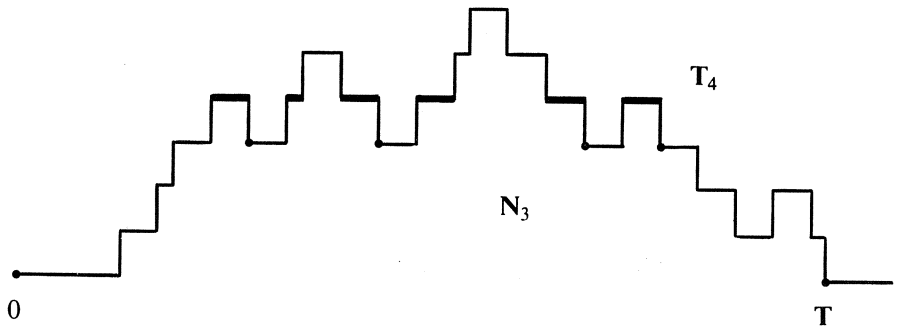

Figure 1.1 A sample path of the queueing process.

Recalling that we have assumed that at epoch 0 a customer has completed service and the system is empty, we define the random variables

$\mathbf{T}$ = the next time the system becomes empty

$\mathbf{T}_n$ = the amount of time in (0,**T**] that n customers are present, $n \geqslant 0$

$\mathbf{N}$ = the number of customers served in (0,**T**]

$\mathbf{N}_n$ = the number of service completion epochs in (0,**T**] at which n customers are left behind by the customer just served, $n \geqslant 0$

The following theorem supplies the justification for focusing on a busy cycle and the random variables associated with it.

Theorem 1.1a

$$p_n = \frac{\mathbf{ET}_n}{\mathbf{ET}}, \quad q_n = \frac{\mathbf{EN}_n}{\mathbf{EN}}, \quad n \geqslant 0 \tag{1.1}$$

Proof The theorem can be proved by the theory of the regenerative processes; cf. Stidham[72] and Ross[70]. See also Appendix A.

□

Theorem 1.1a gives another, more intuitive interpretation of (p_n) and (q_n), namely p_n equals the fraction of time the system is in state n and q_n is the fraction of customers who upon departure leave n customers behind in the system.

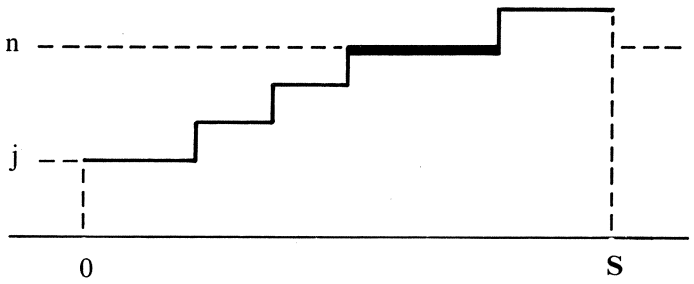

Figure 1.2 A service starting while j customers are present.

The next step in our analysis for the state probabilities is to partition the busy cycle (0,**T**] by means of the service completion epochs, i.e. to look at an embedded process within the busy cycle. $\mathbf{N}_j$ counts in a busy cycle the number of departures leaving the system in state j. Consequently, in a busy cycle, the number of new services starting with j customers present is equal to $\mathbf{N}_j$. Noting that the number of customers present during a particular service time is always larger than or equal to the number of customers present at the beginning of that service time, define for $0 \leqslant j \leqslant n$ (see Figure 1.2)

A_{jn} = the expected amount of time during which n customers are in the system until the next service completion epoch, given that at epoch 0 a service is completed with j customers left behind in the system.

Then, using Wald's equation (cf. Ross[70] and Appendix A), $\mathbf{E}\mathbf{N}_j A_{jn}$ is the expected amount of time in (0,**T**] that n customers are present, when we restrict ourselves to the services starting with j customers present.

In every busy cycle only the first service starts not immediately after the completion of another service, but this first service starts when a customer arrives at the empty system. Hence $\mathbf{E}\mathbf{N}_0 = 1$. Also note that because of the single arrivals $A_{0n} = A_{1n}$, $n \geqslant 1$. We summarize the above observations in the basic relation of the regenerative method.

Theorem 1.1b

$$\mathbf{E}\mathbf{T}_n = \sum_{j=0}^{n} \mathbf{E}\mathbf{N}_j A_{jn} \ , \quad n \geqslant 1 \tag{1.2}$$

□

A second set of relations between $\mathbf{E}\mathbf{T}_n$ and $\mathbf{E}\mathbf{N}_n$ is found by using an up and down crossing argument for 'level' n; see Figure 1.3. We shall show that

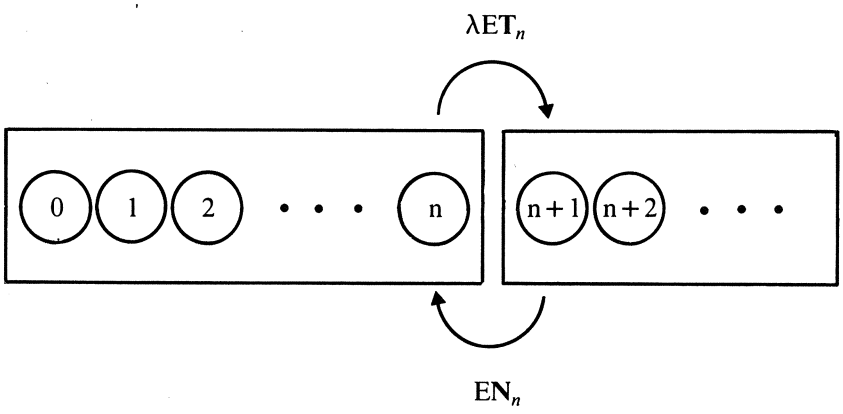

Figure 1.3 Up and downcrossings of level n.

in the busy cycle (0,T] the number of transitions from the set of states $\{0,1,\ldots,n\}$ to the set of states $\{n+1,n+2,\ldots\}$ is equal to the number of transitions from $\{n+1,n+2,\ldots\}$ to $\{0,1,\ldots,n\}$.

Theorem 1.1c

$$\mathbf{EN}_n = \lambda \mathbf{ET}_n \ , \quad n \geq 0 \tag{1.3}$$

Proof Direct transitions between $\{0,1,\ldots,n\}$ and $\{n+1,n+2,\ldots\}$ occur only between the two neighbouring states n and $n+1$. A downcrossing from $n+1$ to n can only occur at service completion epochs at which n customers are left behind in the system. Hence, $\mathbf{EN}_n$ equals the average number of downcrossings from $n+1$ to n in a busy cycle, and consequently from $\{n+1,n+2,\ldots\}$ to $\{0,1,\ldots,n\}$.

The second part of the proof follows by applying the Poisson Lemma (cf. Appendix A) to the busy cycle (0,T] and by noting that upcrossings of level n are generated by arriving customers finding the system in state n. Hence the expected number of transitions from $\{0,1,\ldots,n\}$ to $\{n+1,n+2,\ldots\}$ in (0,T] equals the expected number $\lambda \mathbf{ET}_n$ of arrivals finding the system in state n. Note that $\mathbf{ET}_n$ can be written as

$$\mathbf{ET}_n = \mathrm{E}\int_0^{\mathrm{T}} \Pr\{\text{ system in state } n \text{ at epoch s }\}\, ds$$

Since the number of downcrossing $\mathbf{EN}_n$ equals the number of upcrossings $\lambda \mathbf{ET}_n$, the theorem follows.

□

In this stage of the analysis, we have sufficient material to formulate a preliminary version of the algorithm to compute the steady state probabilities (p_n) and (q_n). For the numbers $(\mathbf{ET}_n)$ and $(\mathbf{EN}_n)$ two sets of linear equations have been derived, cf. Theorems 1.1b and 1.1c, while in Theorem 1.1a the connection to the desired characteristics of the model has been made.

Algorithm 1.2

1. Evaluate the constants A_{jn}, $0 \leqslant j \leqslant n$
2. Put $\mathbf{EN}_0 = 1$ and $\mathbf{ET}_0 = 1/\lambda$
3. Assume $\mathbf{EN}_0, \ldots, \mathbf{EN}_{n-1}$, $\mathbf{ET}_0, \ldots, \mathbf{ET}_{n-1}$ have been computed, solve for $\mathbf{EN}_n$ and $\mathbf{ET}_n$

$$\mathbf{ET}_n = \mathbf{EN}_n A_{nn} + \sum_{j=0}^{n-1} \mathbf{EN}_j A_{jn}$$

$$\mathbf{EN}_n = \lambda \mathbf{ET}_n$$

4. Return to step 3 if necessary.
5. Normalize $\mathbf{ET}_n$ by $\Sigma_{n=0}^{\infty} \mathbf{ET}_n$ to obtain p_n and normalize $\mathbf{EN}_n$ by $\Sigma_{n=0}^{\infty} \mathbf{EN}_n$ to find q_n.

□

We here note that the algorithms for the various queueing models to be presented have in essence the same simple structure. The evaluation of the constants A_{jn} is in general the difficult step and will be extensively discussed later. Depending on the specific properties of the model, further simplifications are possible.

Now, we proceed with the analysis of the M/G/1 queue, exploiting its specific structure.

Theorem 1.1d

$$\mathbf{ET} - \mathbf{ET}_0 = \mathbf{ENES} \tag{1.4}$$

$$\mathbf{EN} = \lambda \mathbf{ET} \tag{1.5}$$

Proof The first part follows by summing Equation (1.2) for $n \geqslant 1$ and by noting that $\Sigma_{n=j}^{\infty} A_{jn} = \mathbf{ES}$, $j \geqslant 0$ where $A_{00} = 0$.

$$\mathbf{ET} - \mathbf{ET}_0 = \sum_{n=1}^{\infty} \mathbf{ET}_n = \sum_{n=1}^{\infty} \sum_{j=0}^{n} \mathbf{EN}_j A_{jn} = \sum_{j=0}^{\infty} \mathbf{EN}_j \sum_{n=j}^{\infty} A_{jn} = \mathbf{ENES}$$

The second part follows trivially by summing (1.3) for $n \geqslant 0$.

□

Theorem 1.1d has the following intuitive interpretation. $\mathbf{ET} - \mathbf{ET}_0$ is the expected amount of time in (0,T] the server is busy and **ENES** is the total expected service time of all customers served in (0,T]. Clearly,these two quantities are equal. The second equality states that the expected number of customers served in (0,T] is equal to the expected number of customers arriving in (0,T]. By Theorem 1.1d, we have

$$\mathbf{ET} = \frac{\mathbf{ET}_0}{1-\rho}$$

Hence, $p_0 = \mathbf{ET}_0 / \mathbf{ET} = 1 - \rho$. Together with $\mathbf{ET}_0 = 1/\lambda$ this yields the well known formula for the expected length of a busy cycle.

$$\mathbf{ET} = \frac{1}{\lambda(1-\rho)}$$

By (1.1), (1.3) and (1.5), we have $p_n = q_n$ for all $n \geq 0$. Since in this model customers arrive one at a time and are served one at a time, we also have $(q_n) = (\pi_n)$. This yields for the M/G/1 queue the well known (cf. for example Hordijk and Tijms[76])

Corollary 1.3

$$\pi_n = p_n = q_n \ , \quad n \geq 0 \tag{1.6}$$

Next we turn to the characterization of the constants A_{jn}. Let

$$\alpha_k = \int_0^\infty (1 - F(t)) e^{-\lambda t} \frac{(\lambda t)^k}{k!} \mathrm{d}t \ , \quad k \geq 0 \tag{1.7}$$

Lemma 1.4

$$A_{jn} = \alpha_{n-j} \ , \quad 1 \leq j \leq n \tag{1.8}$$

$$A_{0n} = A_{1n} = \alpha_{n-1} \ , \quad n \geq 1 \tag{1.9}$$

Proof Assuming that at epoch 0 a service starts with j customers present, define the indicator function $\chi_{jn}(t)$, $1 \leq j \leq n$

$$\chi_{jn}(t) = \begin{cases} 1 \ , & \text{if at time } t \text{ there are } n \text{ customers present and} \\ & \text{the service started at epoch 0 is still in progress} \\ 0 \ , & \text{otherwise} \end{cases}$$

Then, by the definition of A_{jn}

$$A_{jn} = \int_0^\infty \mathrm{E}\chi_{jn}(t)\, \mathrm{d}t = \int_0^\infty \Pr\{\chi_{jn}(t) = 1\}\, \mathrm{d}t$$

Also,

$$\begin{aligned} \Pr\{\chi_{jn}(t) = 1\} &= \Pr\{\mathbf{S} > t\} \Pr\{n - j \text{ customers arrive in } (0,t)\} \\ &= (1 - F(t)) e^{-\lambda t} \frac{(\lambda t)^{n-j}}{(n-j)!} \ , \quad 1 \leq j \leq n \end{aligned}$$

This proves the first part of the lemma. From the definition of A_{jn} and the fact that the customers arrive singly, it follows that $A_{0n} = A_{1n}$.

□

By putting together all pieces of information obtained in Theorem 1.1, Corollary 1.3 and Lemma 1.4, we get

$$p_0 = 1 - \rho \tag{1.10}$$

$$p_n = \lambda p_0 \alpha_{n-1} + \lambda \sum_{j=1}^{n} p_j \alpha_{n-j} \ , \quad n \geq 1 \tag{1.11}$$

which leads to the simplified

Algorithm 1.5

1. Evaluate the constants α_n, $n \geqslant 0$
2. Assume $p_0 = 1-\rho, p_1, \ldots, p_{n-1}$ have been found, compute

$$p_n = (\lambda p_0 \alpha_{n-1} + \lambda \sum_{j=1}^{n-1} p_j \alpha_{n-j}) / (1-\lambda\alpha_0)$$

3. Return to step 2 if desired.

□

The actual computation of the constants α_n, $n \geqslant 0$ depends on the choice of the service time distribution function $F(t)$. To compute the numbers α_n , we refer to the algorithms given in Appendix C for the approximate M/G/c queue. These algorithms are exact for the case of a single server. See also Section 2.4.

1.2. Other approaches for the M/G/1 queue

In the literature, the analysis of the M/G/1 queue is usually done by the embedded Markov chain approach; cf. Cooper[72] and Allen[76]. We shall show how the computational scheme obtained by this method can be rewritten in the numerically stable scheme (1.11). Next, we discuss briefly another probabilistic approach directly resulting in 1.6.

The embedded Markov chain approach

It is easily seen that the queue length process embedded at the service completion epochs is a Markov chain. To know the state of the system at an embedded point, it is sufficient to know the state of the system at the previous embedded point.

Define

$$\beta_k = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^k}{k!} \mathrm{d}F(t) , \quad k \geqslant 0 \tag{1.12}$$

i.e. β_k is the probability of k arrivals during a service time. Then, it is well known that the steady state probabilities (q_n) satisfy the system of linear equations

$$q_n = q_0 \beta_n + \sum_{j=1}^{n+1} q_j \beta_{n+1-j} , \quad n \geqslant 0 \tag{1.13}$$

Rewriting (1.13) in a recursive form leads for $n \geqslant 0$ to

$$q_{n+1} = (q_n - q_0\beta_n - \sum_{j=1}^{n} q_j \beta_{n+1-j}) / \beta_0 \tag{1.14}$$

To demonstrate that (1.13) is equivalent to (1.11), note first by partial integration that

$$\beta_k = \lambda(\alpha_{k-1} - \alpha_k) , \quad k \geqslant 0 \text{ with } \alpha_{-1} = 1/\lambda \tag{1.15}$$

and hence $\sum_{j=0}^{k} \beta_j = 1 - \lambda\alpha_k$. Then, by summing (1.13) for $0 \leqslant k \leqslant n$ we get

$$\sum_{k=0}^{n} q_k = q_0 \sum_{k=0}^{n} \beta_k + \sum_{k=0}^{n} \sum_{j=1}^{n+1} q_j \beta_{k+1-j} = q_0(1-\lambda\alpha_n) + \sum_{j=1}^{n+1} q_j (1-\lambda\alpha_{n+1-j})$$

from which (1.11) follows with p_n replaced by q_n.

For numerical purposes (1.11) is better suited than (1.14) because it avoids taking differences repeatedly.

n	explicit	regenerative method	embedded Markov chain approach
$\rho=0.2$			
0	$8.0000000000000\,10^{-1}$	$8.0000000000000\,10^{-1}$	$8.0000000000000\,10^{-1}$
5	$2.5600000000000\,10^{-4}$	$2.5599999999999\,10^{-4}$	$2.5600000000460\,10^{-4}$
10	$8.1919999999999\,10^{-8}$	$8.1919999999996\,10^{-8}$	$8.1920004606391\,10^{-8}$
15	$2.6214400000000\,10^{-11}$	$2.6214399999998\,10^{-11}$	$2.6219006395036\,10^{-11}$
20	$8.3886079999999\,10^{-15}$	$8.3886079999992\,10^{-15}$	$1.2995003037904\,10^{-14}$
25	$2.6843545600000\,10^{-18}$	$2.6843545599997\,10^{-18}$	$4.6090793924652\,10^{-15}$
$\rho=0.5$			
0	$5.0000000000000\,10^{-1}$	$5.0000000000000\,10^{-1}$	$5.0000000000000\,10^{-1}$
10	$4.8828125000000\,10^{-4}$	$4.8828125000000\,10^{-4}$	$4.8828124999144\,10^{-4}$
20	$4.7683715820312\,10^{-7}$	$4.7683715820312\,10^{-7}$	$4.7683714961218\,10^{-7}$
30	$4.6566128730774\,10^{-10}$	$4.6566128730774\,10^{-10}$	$4.6565269631002\,10^{-10}$
40	$4.5474735088646\,10^{-13}$	$4.5474735088646\,10^{-13}$	$4.4615635311426\,10^{-13}$
50	$4.4408920985006\,10^{-16}$	$4.4408920985006\,10^{-16}$	$-8.14690856243\,10^{-15}$
$\rho=0.9$			
0	$1.0000000000000\,10^{-1}$	$1.0000000000000\,10^{-1}$	$1.0000000000000\,10^{-1}$
50	$5.1537752073195\,10^{-4}$	$5.1537752073192\,10^{-4}$	$5.1537752073246\,10^{-4}$
100	$2.6561398887581\,10^{-6}$	$2.6561398887578\,10^{-6}$	$2.6561398892295\,10^{-6}$
150	$1.3689147905854\,10^{-8}$	$1.3689147905852\,10^{-8}$	$1.3689148379367\,10^{-8}$
200	$7.0550791086517\,10^{-11}$	$7.0550791086506\,10^{-11}$	$7.0551264633078\,10^{-11}$
250	$3.6360291795847\,10^{-13}$	$3.6360291795839\,10^{-13}$	$3.6407646482688\,10^{-13}$

Table 1.4 Comparison of computation schemes.

For comparison of both schemes, we have computed in Table 1.4 the state probabilities in the M/M/1 queue in three different ways. In the first column, the explicit formula $p_n=(1-\rho)\rho^n$ has been used and in the second and third columns the schemes (1.11) and (1.14) have been employed respectively. It appears that the probabilities computed with (1.14) have an absolute error as big as the machine precision (here 10^{-14}), whereas with scheme (1.11) only a relative error of the machine precision is incurred.

The up and down crossing approach

The recursive relation (1.11) can directly be obtained from the following up and down crossing argument. For each $n \geqslant 1$, we have

> the fraction of services at the end of which n customers are left behind =
>
> the fraction of services having the property that at its beginning at most n customers are present and during its execution the number of customers present exceeds n (1.16)

To prove this statement, let n be fixed and let ω be any realization of the queue length process. For $m \geqslant 1$ define

$d(m,\omega)$ = { the number out of the first m services at the end of which n customers are left behind | ω }

$u(m,\omega)$ = { the number out of the first m services having the property that at their beginning at most n customers are present and during their execution the number of customers present exceeds n | ω }

Then, $d(m,\omega)$ denotes the number of downcrossings of level n in the realization ω of the queue length process restricted to the first m services. Similarly, $u(m,\omega)$ is the number of upcrossings of level n.

It is obvious that for any m and ω and independently of the initial state of the queue length process

$$|d(m,\omega)-u(m,\omega)| \leqslant 1 \tag{1.17}$$

Now, divide both sides of (1.17) by m and let $m \to \infty$. Then $d(m,\omega)/m$ and $u(m,\omega)/m$ converge to the corresponding fractions in (1.16) and also (1.16) follows.

Noting that by (1.15) $\sum_{k=n-j+1}^{\infty} \beta_k = \lambda \alpha_{n-j}$ is the probability that level n is exceeded in a service starting with j customers present, the up and down crossing relation (1.16) gives for $n \geqslant 1$

$$q_n = q_0 \sum_{k=n}^{\infty} \beta_k + \sum_{j=1}^{n} q_j \sum_{k=n-j+1}^{\infty} \beta_k = \lambda q_0 \alpha_{n-1} + \lambda \sum_{j=1}^{n} q_j \alpha_{n-j}$$

Finally (1.11) is found by replacing q_n by p_n.

□

Remark 1.6

Both alternative methods lead to computational schemes in terms of (q_n) but do not provide automatically the relation between (p_n) and (q_n).

2. THE M/G/1 QUEUE WITH STATE DEPENDENT ARRIVAL RATE

In this chapter we consider a rather wide class of single server queues with state dependent arrival processes. This class covers a number of standard queueing systems, including the M/G/1 queue. We will give a unifying treatment of the class of queueing systems with state dependent arrival processes; cf. also Tijms and van Hoorn[81b,81c].

We assume that the arrival rate of customers depends on the state of the system. New customers arrive according to a Poisson process at rate λ_j when j customers are in the system. The service time $\mathbf{S}$ of a customer has a general probability distribution function $F(t)=\Pr\{\mathbf{S}\leqslant t\}$. A sufficient condition for a stable queue is $\limsup_{n\to\infty}\lambda_n \mathbf{ES}<1$. It is no restriction to assume an infinite waiting capacity, i.e. every arriving customer actually enters the system.

This class of queueing systems contains a number of important models as special cases. For the M/G/1 queue having only place for K customers, the arrival process of entering customers can be modeled as a state dependent arrival process with $\lambda_j=\lambda$ for $0\leqslant j<K$ and $\lambda_j=0$, $j\geqslant K$.

Another example of a queueing system with state dependent arrivals is the finite source model which is variously called the machine repair model or the cyclic queue model. This model is one of the most useful queueing models for practical applications. The population of potential customers for this system consists of K identical machines and there is a single repairman. For each machine the operating time between breakdowns is exponentially distributed with mean $1/\lambda$. The repair time has a general probability distribution function F. The arrival process of the broken down machines can be modeled as a state dependent arrival process with $\lambda_j=(K-j)\lambda$, $j<K$ and $\lambda_j=0$, $j\geqslant K$, where state j denotes the number of machines broken down.

In the literature, this class of single server queues has received relatively little attention, except for the finite capacity M/G/1 queue. The state probabilities in this latter model are closely related to those in the M/G/1 queue with an infinite waiting capacity; cf. Keilson[65] and Cooper[72].

The machine repair problem has mainly been studied for exponential repair time; see e.g. the studies of Ferdinand[71] and Shum[76]. With the assumption of exponential repair time, the model essentially reduces to a birth and death queueing model.

Though the general model may seem rather complicated for an exact analysis, it has some pleasant characteristics. In the first place, the arrival process has a Markovian nature. Secondly, it is easily seen that the queue length process at departure epochs is a Markov chain. These two features make the model very suitable for applying the regenerative method.

After having derived the basic results in Section 2.1, we consider in the Sections 2.2 and 2.3 the two special models discussed above. The two models are special cases in the sense that we make assumptions on the arrival rates (λ_j).

In the Sections 2.4 and 2.5 we make assumptions on the service time distribution, namely that it is of phase type. Moreover, in Section 2.5 we focus on computational methods for the waiting time distribution.

2.1. The basic theorem

For the analysis we adopt the same notation and definitions as in Chapter 1 and using the regenerative method we derive

Theorem 2.1

$$p_n = \frac{\mathbf{ET}_n}{\mathbf{ET}}, \quad q_n = \frac{\mathbf{EN}_n}{\mathbf{EN}}, \quad n \geqslant 0 \tag{2.1}$$

$$\mathbf{ET}_n = \sum_{j=0}^{n} \mathbf{EN}_j A_{jn}, \quad n \geqslant 1 \tag{2.2}$$

$$\mathbf{EN}_n = \lambda_n \mathbf{ET}_n, \quad n \geqslant 0 \tag{2.3}$$

$$\mathbf{ET} - \mathbf{ET}_0 = \mathbf{ENES} \tag{2.4}$$

$$\mathbf{EN} = \sum_{n=0}^{\infty} \lambda_n \mathbf{ET}_n$$

Proof In this proof, we focus on relation (2.3). The other relations can be proved similarly as in Theorem 1.1.

Equation (2.3) expresses for this model the up and down crossing property of the queue length process in the busy cycle (0,**T**]. Since the expected number of down-crossings of level n is (trivially) equal to $\mathbf{EN}_n$, it is sufficient to prove that $\lambda_n \mathbf{ET}_n$ is the expected number of arrivals finding the system in state n in (0,**T**].

The most simple and intuitively appealing way to show this is to note that we may associate with each state n an interrupted Poisson process having intensity λ_n if the system is in state n and intensity 0 otherwise. In fact, the overall arrival process is decomposed into infinitely many processes of which only one at the time is 'active', i.e. has positive intensity. Next, for each n, we apply the Poisson Lemma to the arrival process associated with state n; cf. Remark A.3 in Appendix A. Hence, $\lambda_n \mathbf{ET}_n$ is the expected number of arrivals in (0,**T**] finding the system in state n.

□

An algorithm derived from Theorem 2.1 looks very much like Algorithm 1.1 in Chapter 1. Replacing in Algorithm 1.1 $\mathbf{ET}_0 = 1/\lambda$ by $\mathbf{ET}_0 = 1/\lambda_0$ and $\lambda \mathbf{ET}_n$ by $\lambda_n \mathbf{ET}_n$ gives a recursive and numerically stable scheme to compute the distributions (p_n) and (q_n).

In applications of this model, the distribution (π_n) is of interest too. This distribution describes the system from the point of view of an arriving (and by assumption also entering) customer. For example, $1 - \pi_0$ is the delay probability. Note that also in this model $(\pi_n) = (q_n)$. By Theorem 2.1 we have the

Corollary 2.2

$$q_n = \pi_n = \lambda_n p_n \Big/ \sum_{k=0}^{\infty} \lambda_k p_k, \quad n \geqslant 0. \tag{2.5}$$

□

Remark 2.3 The normalization factors ET and EN cannot be computed immediately as in Chapter 1. However, for models with finite capacity or for models where $\lambda_n = \lambda_{n_0}, n \geq n_0$, we find ET after having computed $ET_0 (= 1/\lambda_0)$, $ET_1, \ldots, ET_{n_0-1}$. By (2.3)

$$EN = \sum_{n=0}^{\infty} \lambda_n ET_n = \sum_{n=0}^{n_0-1} (\lambda_n - \lambda_{n_0}) ET_n + \lambda_{n_0} ET$$

and so, using (2.4)

$$ET = (ET_0 + ES \sum_{n=0}^{n_0-1} (\lambda_n - \lambda_{n_0}) ET_n) / (1 - \lambda_{n_0} ES) \tag{2.6}$$

□

2.2. The M/G/1 queue with finite capacity

In the literature on queueing theory, the analysis of the M/G/1 queue is naturally followed by that of the M/G/1 queue with only K waiting places; cf. Cooper[72]. Here we relate the finite M/G/1 queue to a queueing model with state dependent Poisson arrivals.

Let the M/G/1 queue with capacity K be defined as an ordinary M/G/1 queue where arriving customers finding K customers in the system are rejected. These customers are lost and do not return. For this model, we denote the steady state distributions at arrival, arbitrary and departure epochs by (Π_n),(P_n) and (Q_n) respectively.

The modified model is defined as the M/G/1 queue with state dependent arrival rate with $\lambda_j = \lambda$, $j < K$ and $\lambda_j = 0$, $j \geq K$. Since the blocked customers in the original model do not affect the system, both models have the same arrival process of entering customers. Hence, $(P_n) = (p_n)$ and $(Q_n) = (q_n)$, where (p_n) and (q_n) are the steady state distributions in the modified model.

The distributions (Π_n) and (π_n) are not equal. Indeed, in the original model the distribution (Π_n) is related to all customers, namely both accepted and rejected customers. Moreover, since the arrival process is Poisson, $(\Pi_n) = (P_n)$. In the modified model we have by the corollary of Theorem 2.1 that $(\pi_n) = (q_n)$. Note that the above also holds for the multiserver M/G/c queue with capacity K.

Next, we proceed with the analysis of the modified model. The special form of the sequence (λ_n) allows a number of simplifications in Theorem 2.1. Define as in Chapter 1

$$\alpha_k = \int_0^{\infty} (1 - F(t)) e^{-\lambda t} \frac{(\lambda t)^k}{k!} dt \ , \quad k \geq 0$$

Then, for $1 \leq j \leq n \leq K-1$ we have $A_{jn} = \alpha_{n-j}$ and also $A_{0n} = \alpha_{n-1}$, $1 \leq n \leq K-1$, since the 'boundary' K has no influence in this range of indices. After inserting (2.3) in (2.2), we get

$$ET_n = \lambda ET_0 \alpha_{n-1} + \lambda \sum_{j=1}^{n} ET_j \alpha_{n-j} \ , \quad 1 \leq n \leq K-1 \tag{2.7}$$

Starting with $ET_0 = 1/\lambda$, $ET_1, \ldots, ET_{K-1}$ are computed recursively using (2.7). Next ET is found from (2.4) (also cf.(2.6))

$$\mathbf{ET}=\mathbf{ET}_0+\lambda\mathbf{ES}\sum_{n=0}^{K-1}\mathbf{ET}_n \tag{2.8}$$

By dividing $\mathbf{ET}_0, \ldots, \mathbf{ET}_{K-1}$ by $\mathbf{ET}$ the state probabilities $p_0, \ldots, p_{K-1}$ are computed and $p_K = 1-\sum_{n=0}^{K-1} p_n$. Note that $q_n = p_n / (1-p_K)$, $0 \leq n \leq K-1$ and $q_K = 0$.

□

For $\lambda\mathbf{ES}<1$, there exists a simple relationship between the steady state distributions in the M/G/1 queue with capacity K and the infinite capacity M/G/1 queue; cf. Cooper[72]. Indeed, the computational schemes for the numbers $\mathbf{ET}_0, \ldots, \mathbf{ET}_{K-1}$ are identical in both models. Only the normalization factor $\mathbf{ET}$ is differently computed.

For sake of convenience, we mark the random variable $\mathbf{T}$ and the distributions (p_n) and (q_n) with a superscript (K) or (∞) to indicate the system they refer to. From the above it follows that $(p_n^{(K)})$ and $(p_n^{(\infty)})$ are proportional for $n=0, \ldots, K-1$, i.e. for some c_K

$$p_n^{(K)} = c_K p_n^{(\infty)}, \quad 0 \leq n \leq K-1 \tag{2.9}$$

Obviously, $c_K = \mathbf{ET}^{(\infty)} / \mathbf{ET}^{(K)}$ and using expression (2.8) for $\mathbf{ET}^{(K)}$, we find

$$c_K = 1 / (p_0^{(\infty)} + \rho \sum_{n=0}^{K-1} p_n^{(\infty)}) \tag{2.10}$$

The blocking probability $p_K^{(K)}$, expressed in terms of $(p_n^{(\infty)})$ is given by

$$p_K^{(K)} = (p_0^{(\infty)} + (\rho-1) \sum_{n=0}^{K-1} p_n^{(\infty)}) / (p_0^{(\infty)} + \rho \sum_{n=0}^{K-1} p_n^{(\infty)}) \tag{2.11}$$

Finally, by the proportionality of $p_n^{(K)}$ and $p_n^{(\infty)}$ for $n=0, \ldots, K-1$ and the equality of the arriving customer and departing customer distributions $(\pi_n^{(K)})$ and $(q_n^{(K)})$, we have for $0 \leq n \leq K-1$

$$\pi_n^{(K)} = q_n^{(K)} = p_n^{(\infty)} / \sum_{n=0}^{K-1} p_n^{(\infty)} \tag{2.12}$$

The interpretation of (2.12) regarding $(\pi_n^{(K)})$ is the following. The probability that an arriving customer sees n customers upon entering a M/G/1 system with capacity K equals the conditional probability that an arriving customer sees n customers in an infinite capacity M/G/1 queue, given that he sees less than K customers.

2.3. The machine repair model with a single repairman

In this section we discuss the finite source model known as the machine repair model , the machine interference model or the cyclic queue model. We consider a closed queueing system consisting of K identical machines and a single repairman (cf. Figure 2.1). A machine operates between breakdowns during an exponential time with mean $1/\lambda$. When a machine breaks down, it joins the queue for repair. If the repairman is free, he immediately begins to repair the machine, otherwise the machine must wait for repair. The repair time $\mathbf{S}$ has a general probability distribution function $F(t)$. We define the state of the system as the number of machines that are not working. Thus the arrival rate of broken down machines to the repair facility is $\lambda_j = (K-j)\lambda$, $0 \leq j \leq K$.

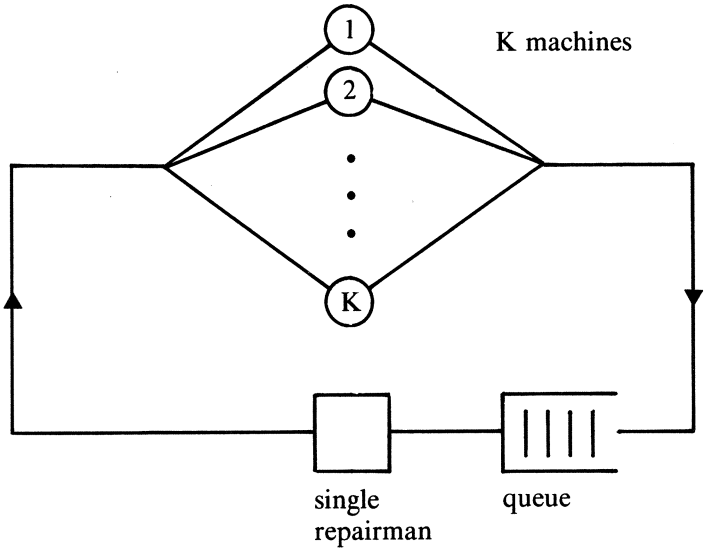

Figure 2.1 The machine repair model.

The machine repair model is also known in computer science. In that context, there are K jobs cycling in a system consisting of K terminals and a CPU (central processor unit) with a work queue. A job is sent from a terminal to the CPU after an exponentially distributed 'think time' and after being processed by the CPU the job enters another think phase at a terminal; cf. Shum[76].

For the constants A_{jn} in Theorem 2.1 we prove

Lemma 2.4

$$A_{jn} = \int_0^\infty (1-F(t))\phi_{jn}(t)\,dt \ , \quad 1 \leqslant j \leqslant n \leqslant K. \tag{2.13}$$

where

$$\phi_{jn}(t) = \binom{K-j}{n-j}(1-e^{-\lambda t(n-j)})\,e^{-\lambda t(K-n)}$$

Proof The lemma follows by noting that

$$A_{jn} = \int_0^\infty \mathrm{E}\chi_{jn}(t)\,dt$$

where, conditionally that at epoch 0 a new service starts with j machines broken down, $\chi_{jn}(t)=1$ if at time t this service is still in progress and n machines are broken down and $\chi_{jn}(t)=0$ otherwise. Then, $\phi_{jn}(t)$ is the binomial probability that $n-j$ machines fail in $(0,t)$ given that at epoch 0 $K-j$ machines are working.

□

We define the system response time **R** of a machine as the time interval between the instant when the machine breaks down and the instant when the machine is put into operation again. Hence, **R** is the sum of the waiting time in the queue and

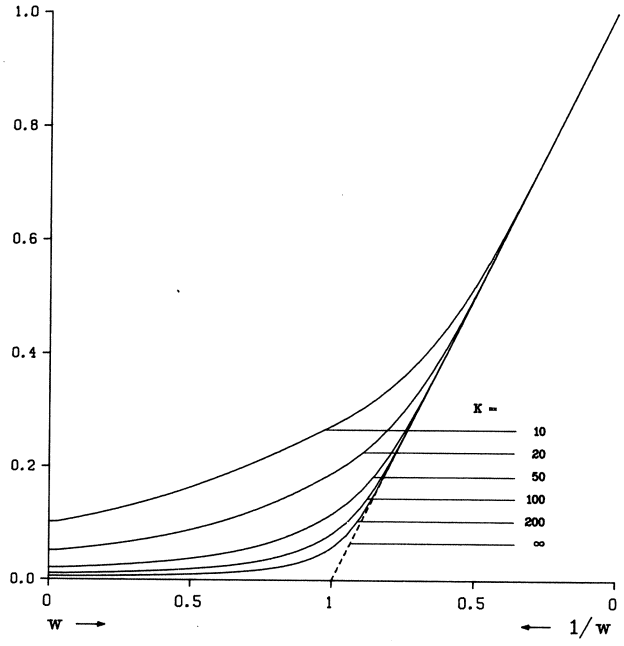

Figure 2.2 The doubly normalized response time $\overline{\mathbf{ER}}$.

the service time of a machine. By Little's law, we have

$$\lambda^*(\mathbf{ER}+1/\lambda)=K \tag{2.14}$$

where $\lambda^* = \sum_{n=0}^{K}\lambda_n p_n$ is the throughput i.e. the average arrival rate of machines to the queue. By combining (2.3) and (2.4) in Theorem 2.1, it follows that

$$\lambda^* = \frac{1-p_0}{\mathbf{ES}}$$

In Figure 2.2 and Table 2.3, we give some numerical results for the expected doubly normalized response time $\overline{\mathbf{ER}}$, defined as

$$\overline{\mathbf{ER}} = \frac{\mathbf{ER}}{K\,\mathbf{ES}} = \frac{1}{(1-p_0)} - \frac{1}{w},$$

where $w = K\lambda\mathbf{ES}$. See also Ferdinand[71] where asymptotic properties of this performance measure have been derived for the case of an exponential repair time distribution. Incidentally, the formula for $\overline{\mathbf{ER}}$ in his paper is incorrect.

In Figure 2.2, we have displayed $\overline{\mathbf{ER}}$ for several values of K. Here the repair time is exponentially distributed. In Table 2.3, $\overline{\mathbf{ER}}$ is computed for various other service distributions, namely

1: deterministic repair time ($C_S^2=0.0$)

2: Erlang - 2 repair time ($C_S^2=0.5$)

W	c_S^2	K = 10	K = 20	K = 30	K = 40
0.25	0.0	0.1138	0.0576	0.0386	0.0290
	0.5	0.1205	0.0613	0.0411	0.0309
	1.0	0.1269	0.0649	0.0437	0.0329
	1.5	0.1330	0.0685	0.0462	0.0348
	2.0	0.1389	0.0720	0.0486	0.0367
0.50	0.0	0.1350	0.0706	0.0479	0.0363
	0.5	0.1503	0.0801	0.0548	0.0417
	1.0	0.1643	0.0890	0.0613	0.0469
	1.5	0.1766	0.0972	0.0675	0.0518
	2.0	0.1878	0.1049	0.0733	0.0565
0.75	0.0	0.1681	0.0956	0.0678	0.0528
	0.5	0.1925	0.1130	0.0815	0.0642
	1.0	0.2136	0.1284	0.0938	0.0745
	1.5	0.2302	0.1411	0.1043	0.0835
	2.0	0.2448	0.1526	0.1139	0.0918
1.00	0.0	0.2180	0.1458	0.1161	0.0990
	0.5	0.2484	0.1694	0.1361	0.1167
	1.0	0.2732	0.1889	0.1527	0.1314
	1.5	0.2911	0.2036	0.1655	0.1430
	2.0	0.3064	0.2164	0.1767	0.1531
0.75	0.0	0.3105	0.2677	0.2564	0.2525
	0.5	0.3378	0.2833	0.2656	0.2579
	1.0	0.3605	0.2978	0.2753	0.2646
	1.5	0.3759	0.3085	0.2832	0.2705
	2.0	0.3888	0.3181	0.2905	0.2761
0.50	0.0	0.5024	0.5000	0.5000	0.5000
	0.5	0.5094	0.5005	0.5000	0.5000
	1.0	0.5187	0.5019	0.5002	0.5000
	1.5	0.5253	0.5034	0.5005	0.5001
	2.0	0.5313	0.5052	0.5010	0.5002
0.25	0.0	0.7500	0.7500	0.7500	0.7500
	1.5	0.7504	0.7500	0.7500	0.7500
	1.0	0.7502	0.7500	0.7500	0.7500
	1.5	0.7504	0.7500	0.7500	0.7500
	2.0	0.7507	0.7500	0.7500	0.7500

Table 2.3 Numerical results for $\overline{\mathrm{ER}}$.

3: exponential repair time ($C_S^2 = 1.0$)
4: hyperexponential repair time ($C_S^2 = 1.5$) and ($C_S^2 = 2.0$), where $F(t) = 1 - p_1 e^{-\mu_1 t} - p_2 e^{-\mu_2 t}$ with $p_1 / \mu_1 = p_2 / \mu_2$.

2.4. Algorithms for the quantities A_{jn}

In this section, we give computational schemes for the quantities A_{jn} when the service times have a phase type distribution. By exploiting the memoryless property of the exponential distribution and the property that $min(X_1,X_2)$ has an exponential distribution with mean $1/(\mu_1+\mu_2)$ if X_1 and X_2 are independent exponential random variables with mean values $1/\mu_1$ and $1/\mu_2$, we obtain a recursive scheme to compute the quantities A_{jn}. We demonstrate the ideas on the hand of three basic cases. Next the extension to general distributions of phase type is obvious.

For clarity of presentation, we repeat the definition of A_{jn},

A_{jn} = the expected amount of time during which n customers are in the system until the next service completion, given that at epoch 0 a service is completed with j customers left behind in the system

We consider only the range of indices $1 \leqslant j \leqslant n$ since $A_{0n} = A_{1n}$, $n \geqslant 1$. Clearly, in the definition of A_{jn} with $n \geqslant 1$, the completion of a service at epoch 0 coincides with the start of a new service.

Case 1 : $F(t) = 1 - e^{-\mu t}$

Let $j \geqslant 1$ be fixed and suppose that at epoch 0 a new service starts with j customers present. With probability $\lambda_j/(\lambda_j+\mu)$ an arrival occurs before the completion of the service. Further, by the memoryless property of the exponential distribution, the service in progress can be considered to start afresh at the arrival epoch. Hence,

$$A_{jn} = \frac{\lambda_j}{\lambda_j+\mu} A_{j+1,n} \ , \quad 1 \leqslant j < n$$

Also, since the expected time until either the next arrival or the next service completion occurs is equal to $1/(\lambda_j+\mu)$, it follows that

$$A_{nn} = \frac{1}{\lambda_n+\mu}$$

Hence, starting with A_{nn}, the numbers $A_{n-1,n}, \ldots, A_{1n}$ can recursively be computed.

Case 2 : $F(t) = p_1(1-e^{-\mu_1 t}) + p_2(1-e^{-\mu_2 t})$

In this case, the service time is with probability p_i exponentially distributed with mean $1/\mu_i$, $i = 1,2$. Thus, using Case 1 twice, we can compute $A_{jn}^{(1)}$ and $A_{jn}^{(2)}$ where in Case 1 the parameter μ is replaced by μ_1 and μ_2 respectively. Next, A_{jn} is obtained as the weighted sum of $A_{jn}^{(1)}$ and $A_{jn}^{(2)}$, namely

$$A_{jn} = p_1 A_{jn}^{(1)} + p_2 A_{jn}^{(2)} \ , \quad 1 \leqslant j \leqslant n$$

□

Case 3 : $F(t) = \Pr\{\mathbf{X}_1 + \mathbf{X}_2 \leqslant t\}$, where $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent exponential random variables with means $1/\mu_1$ and $1/\mu_2$.

In this case, the service consists of two phases. First an exponential phase $\mathbf{X}_1$ and next an exponential phase $\mathbf{X}_2$. The computation of A_{jn} is done in two steps. We first compute the auxiliary quantities B_{jn}, defined as

B_{jn} = the expected amount of time that n customers are in the system in the second phase of the service, given that the second phase starts with j customers present.

By this definition, B_{jn} is related to a service period consisting of a single exponential phase with mean $1/\mu_2$ and is computed according to the scheme of Case 1 with μ replaced by μ_2. Next turning to A_{jn}, note that with probability $\lambda_j/(\lambda_j+\mu_1)$ an arrival occurs before the end of the first phase of the service while with probability $\mu_1/(\lambda_j+\mu_1)$ the second phase of the service starts before an arrival occurs. Hence, we get the scheme

$$A_{jn}=\frac{\lambda_j}{\lambda_j+\mu_1}A_{j+1,n}+\frac{\mu_1}{\lambda_j+\mu_1}B_{jn}\ ,\quad 1\leqslant j<n$$

$$A_{nn}=\frac{1}{\lambda_n+\mu_1}+\frac{\mu_1}{\lambda_n+\mu_1}B_{nn}$$

from which the A_{jn} are computed recursively, starting with A_{nn}.

2.5. The waiting time distribution and its moments

In this section we focus on computational methods for the waiting time distribution. We assume that the service discipline is FIFO (first in first out), that is, service is in order of arrival. Define the random variable

$\mathbf{W}_q$ = the waiting time of an arbitrary arriving customer in the queue, excluding his service time

and the waiting time distribution function

$$W_q(t)=\Pr\{\mathbf{W}_q\leqslant t\}$$

For practical purposes it is often not enough to know the steady state distributions of the number of customers in a queueing system, but also the moments of the waiting time distribution or the distribution itself are required. Unfortunately, for models like the M/G/1 queue with state dependent arrival rate it is in general not possible to relate directly these performance measures to the distributions (p_n) or (q_n). An exception is the mean waiting time, which follows from the mean queue length by using Little's law:

$$\mathrm{E}\mathbf{W}_q=\sum_{n=1}^{\infty}(n-1)p_n\ /\ \sum_{n=0}^{\infty}\lambda_n p_n$$

Moreover, this equation holds for any work conserving service discipline; cf. Kleinrock[75]. To obtain results for higher moments of $\mathbf{W}_q$, it is unavoidable to make assumptions on F(t).

In the analysis for the waiting time the crucial point is to know the remaining service time of the service in progress (if any) at an arrival epoch. For exponential service times we know that the remaining service time has the same exponential distribution as the original service time. For general distributions of the service time, we must put more information in the state description in order to be able to describe when a service expires. However, we can elegantly handle this problem by considering phase type distributions and by defining properly a continuous time Markov chain.

We consider the following phase type distributions

$$F(t)=\sum_{l=1}^{s} r_l E_{m_l,\mu_l}(t) \tag{2.15}$$

with

$$E_{m_l,\mu_l}(t)=1-\sum_{i=0}^{m_l-1} e^{-\mu_l t}\frac{(\mu_l t)^i}{i!},$$

i.e. $F(t)$ is a finite mixture of Erlang distributions with different scale parameters μ_l. Hence, a customer entering service has with probability r_l a service time consisting of l consecutive phases, each having an exponential distribution with mean $1/\mu_l$, $1\leqslant l\leqslant s$. We note that any probability distribution function $G(t)$ concentrated on $[0,\infty)$ can be approximated arbitrarily closely by a function as (2.15); cf. Schassberger[73]. In fact, we can even find a mixture of Erlang distributions with identical scale parameters; cf. Bux and Herzog[77] for an algorithm to determine such a mixture.

A continuous time Markov chain representation for the queue length process in a system with a service distribution as (2.15) follows by defining the state (n,i,l) of the system by

n = the number of customers present,
i = the number of remaining phases of the service in progress, and
l = the index of the scale parameter μ_l for the phases of the service in progress, with the convention that state 0 describes the situation of an empty system.

If an arriving customer finds the system in state (n,i,l) his waiting time is the sum of i phases with mean $1/\mu_l$ and $n-1$ new services. Define for $n\geqslant 1$, $1\leqslant l\leqslant s$, $1\leqslant i\leqslant m_l$

$p_{ni}^{(l)}$ = the steady state probability that at an arbitrary epoch the system is in state (n,i,l)

$\pi_{ni}^{(l)}$ = the steady state probability that an arriving customer finds the system in state (n,i,l).

Then, the waiting time distribution $W_q(t)$ follows from the probabilities $\pi_{ni}^{(l)}$

$$W_q(t)=\pi_0+\sum_{n,i,l}\pi_{ni}^{(l)}W(t\mid n,i,l) \tag{2.16}$$

where the function $W(t\mid n,i,l)$ is the conditional waiting time distribution function of a customer finding upon arrival the system in state (n,i,l). Hence, $W(t\mid n,i,l)$ is the convolution of a residual service time and $n-1$ new service times

$$W(t\mid n,i,l)=E_{i,\mu_l}(t)*F^{(n-1)*}(t).$$

The moments of $\mathbf{W}_q$ can easily be computed from 2.16, e.g.

$$\mathrm{E}\mathbf{W}_q{}^2=\sum_{n,l,i}\pi_{ni}^{(l)}\left[\frac{i(i+1)}{\mu_l{}^2}+(n-1)\left(\frac{2i}{\mu_l}\mathrm{ES}+\mathrm{ES}^2+(n-2)(\mathrm{ES})^2\right)\right] \tag{2.17}$$

Analogously to (2.5) it follows that

$$\pi_{ni}^{(l)}=\lambda_n p_{ni}^{(l)}\Big/\sum_{j=0}^{\infty}\lambda_j p_j \quad \text{for all } n,i,l \tag{2.18}$$

i.e. $\pi_{ni}^{(l)}$ equals the average number of arrivals per unit time who see the system in state (n,i,l) divided by the average number of arrivals per unit time.

By writing down the equilibrium equations of the continuous time Markov chain, we get a system of linear equations for the probabilities $p_{ni}^{(l)}$. For $n \geqslant 1$ and $1 \leqslant l \leqslant r$ we have

$$(\lambda_n + \mu_l)p_{ni}^{(l)} = \lambda_{n-1}p_{n-1,i}^{(l)} + \mu_l p_{n,i+1}^{(l)} + r_l \delta_{i,m_l} \sum_{k=1}^{s} \mu_k p_{n+1,1}^{(k)}, \quad 1 \leqslant i \leqslant m_l \tag{2.19}$$

where $p_{ni}^{(l)} = 0$ for $i = m_l + 1$ and $p_{0i}^{(l)} = r_l \delta_{i,m_l}$ and $\delta_{i,j} = 1$ if $i = j$ and $\delta_{i,j} = 0$ if $i \neq j$.

This equation is obtained by equating the rate at which the system leaves the microstate (n,i,l) to the rate at which the system enters this state. Note that state (n,i,l) is left by an arrival at rate λ_n or by the completion of a service phase at rate μ_l, while state (n,i,l) is entered at rate λ_{n-1} by an arrival from $(n-1,i,l)$ or at rate μ_l by the completion of a phase from $(n,i+1,l)$ unless $i = m_l$ or by the start of a new service if $i = m_l$ after the completion of a service phase in one of the states $(n+1,1,k)$, $1 \leqslant k \leqslant s$.

Equation (2.19) is simplified by noting that

$$\sum_{k=1}^{s} \mu_k p_{n+1,1}^{(k)} = \lambda_n p_n$$

which expresses that the rate at which the system leaves the set of states with $0, \ldots, n$ customers present is equal to the rate at which the system enters that set. Using this relation, we can write (2.19) as

$$(\lambda_n + \mu_l)p_{ni}^{(l)} = \lambda_{n-1}p_{n-1,i}^{(l)} + \mu_l p_{n,i+1}^{(l)} + r_l \delta_{i,m_l} \lambda_n p_n, \quad 1 \leqslant i \leqslant m_l \tag{2.20}$$

To compute the state probabilities $p_{ni}^{(l)}$, we can do much better than solving the large system (2.19) of linear equations. If we first compute recursively the state probabilities (p_n) with the algorithm discussed in Section 2.4, we can next compute recursively the state probabilities $p_{ni}^{(l)}$ from (2.20). For fixed n and l, first compute $p_{n,m_l}^{(l)}$ and next $p_{ni}^{(l)}, i = m_l - 1, \ldots, 1$.

Theoretically, it is possible to obtain from the probabilities $\pi_{ni}^{(l)}$ the distribution $W_q(t)$ using (2.16). However, the computation of the conditional distribution functions $W(t \mid n,i,l)$ is a tedious and laborious task. Also, in practice it is not recommended to compute $W_q(t)$ in such a detailed way in view of the fact that the model itself is typically an approximate description of a queueing system in reality.

A good alternative is to approximate $W_q(t)$ or the functions $W(t \mid n,i,l)$ by fitting distribution functions based on the first few moments, assuming that the coefficient of variation is not too large. For example, in Kühn[76], the (two parameter) Weibull distribution function $1 - e^{-(at)^b}$, $a,b > 0$ is suggested to approximate the waiting time distribution function in various queueing models, based on the exactly computed first two moments. In Kühn[72] it is also suggested to use mixtures of exponential distributions as the parametric family of functions. Note that the moments of $W(t \mid n,i,l)$ can easily be computed from the $\pi_{ni}^{(l)}$.

If $F(t)$ is a mixture of Erlang distributions with the same scale parameter μ, the computational effort can greatly be reduced. In this case we drop the upper index of $\pi_{ni}^{(l)}$ and we take $m_l = l$. Define for $n \geqslant 0$

z_n = steady state probability that n service phases are in the system at an arrival epoch

Once the distribution (z_n) has been computed from (π_{ni}), $W_q(t)$ can be written as

$$W_q(t)=z_0+\sum_{n=1}^{\infty} z_n E_{n,\mu}(t) \tag{2.21}$$

For the computation of (z_n) from (π_{ni}) we define for $k, l \geqslant 1$

$r_l^{(k)}$ = the probability that k successive service times consist together of l exponential phases with mean $1/\mu$.

With the convention that $r_l^{(1)}=r_l$ and $r_l^{(k)}=0$ for $l<k$ or $l>sk$, we have

$$z_n=\sum_{k,i}\pi_{ki} r_{n-i}^{(k)}, \quad n \geqslant 1$$

and

$$z_0=\pi_0$$

For $r_l^{(k)}$ we have the recurrence relation $r_l^{(k)}=\sum_i r_i r_{l-i}^{(k-1)}$, $k \geqslant 2$.

t	V(t)	$V_1(t)$	$V_2(t)$	$V_3(t)$	$V_4(t)$
$\rho=0.8$, $C_S^2=2.0$					
0.0	1.0000	1.0000	1.0000	1.0000	1.0000
1.0	0.8405	0.8267	0.8453	0.8832	0.7656
2.0	0.7332	0.7262	0.7247	0.7568	0.6077
3.0	0.6337	0.6325	0.6214	0.6395	0.4955
4.0	0.5394	0.5416	0.5311	0.5352	0.4120
5.0	0.4549	0.4580	0.4518	0.4445	0.3477
6.0	0.3814	0.3848	0.3821	0.3669	0.2966
7.0	0.3181	0.3219	0.3211	0.3011	0.2552
8.0	0.2634	0.2678	0.2679	0.2460	0.2211
9.0	0.2164	0.2212	0.2217	0.2002	0.1926
10.0	0.1763	0.1811	0.1818	0.1622	0.1686
$\rho=1.5$, $C_S^2=2.0$					
0.0	1.0000	1.0000	1.0000	1.0000	1.0000
2.0	0.9712	0.9637	0.9648	0.9814	0.8203
4.0	0.9186	0.9047	0.9047	0.9212	0.6812
6.0	0.8288	0.8170	0.8169	0.8229	0.5695
8.0	0.7079	0.7055	0.7052	0.6977	0.4781
10.0	0.5684	0.5784	0.5777	0.5603	0.4026
12.0	0.4286	0.4468	0.4459	0.4255	0.3399
14.0	0.3038	0.3226	0.3217	0.3052	0.2875
16.0	0.2027	0.2158	0.2152	0.2064	0.2437
18.0	0.1277	0.1326	0.1323	0.1316	0.2068
20.0	0.0762	0.0740	0.0741	0.0790	0.1758

Table 2.4a Several approximations for $V(t)$.

t	V(t)	$V_1(t)$	$V_2(t)$	$V_3(t)$	$V_4(t)$
$\rho=0.8$, $C_S^2=0.5$					
0.0	1.0000	1.0000	1.0000	1.0000	1.0000
1.0	0.7975	0.8001	0.7909	0.7997	0.6355
2.0	0.5973	0.5920	0.5927	0.6102	0.4480
3.0	0.4471	0.4461	0.4461	0.4558	0.3363
4.0	0.3337	0.3340	0.3343	0.3356	0.2621
5.0	0.2481	0.2490	0.2492	0.2443	0.2094
6.0	0.1834	0.1844	0.1846	0.1761	0.1702
7.0	0.1346	0.1355	0.1356	0.1260	0.1403
8.0	0.0978	0.0985	0.0986	0.0895	0.1168
9.0	0.0700	0.0706	0.0706	0.0632	0.0982
10.0	0.0491	0.0496	0.0496	0.0443	0.0831
$\rho=1.5$, $C_S^2=0.5$					
0.0	1.0000	1.0000	1.0000	1.0000	1.0000
2.0	0.9984	0.9983	0.9983	0.9998	0.8460
4.0	0.9929	0.9917	0.9916	0.9961	0.7172
6.0	0.9746	0.9697	0.9696	0.9755	0.6088
8.0	0.9194	0.9108	0.9106	0.9125	0.5174
10.0	0.7866	0.7852	0.7851	0.7771	0.4401
12.0	0.5595	0.5763	0.5768	0.5615	0.3747
14.0	0.3022	0.3200	0.3206	0.3128	0.3192
16.0	0.1150	0.1118	0.1117	0.1187	0.2721
18.0	0.0292	0.0185	0.0181	0.0263	0.2320
20.0	0.0048	0.0009	0.0009	0.0028	0.1980

Table 2.4b Several approximations for $V(t)$.

In Table 2.4, we give some numerical results for the complementary waiting time distribution for the delayed customers, $V(t)$, defined as

$$V(t)=\Pr\{\mathbf{W}_q>t \mid \mathbf{W}_q>0\}$$

We have computed $V(t)$ for the M/G/1 queue with capacity $K=15$. The service time distribution is the following mixture of Erlang distributions

$$F(t)=pE_{1,\mu}(t)+(1-p)E_{s,\mu}(t)\ ,$$

where $s=8$ and $\mathbf{ES}=p/\mu+s(1-p)/\mu=1$. We consider the following four cases

1: $C_S^2=0.5$, $\rho=0.8$ $(p=(14-\sqrt{52})/21\ ,\ \mu=(10+\sqrt{52})/3)$

2: $C_S^2=0.5$, $\rho=1.5$ $(p=(14-\sqrt{52})/21\ ,\ \mu=(10+\sqrt{52})/3)$

3: $C_S^2=2.0$, $\rho=0.8$ $(p=6/7,\mu=2.0)$

4: $C_S^2=2.0$, $\rho=1.5$ $(p=6/7,\mu=2.0)$

We have computed $V(t)$ according to (2.21), i.e.

$$V(t)=\frac{1}{1-\pi_0}\sum_{j=1}^{(K-1)s} z_j(1-E_{j,\mu}(t))$$

Note that $\Pr\{\mathbf{W}_q>0\}=1-\pi_0$. Further, we have computed four approximations for $V(t)$ (denoted by $V_i(t), i=1, \ldots, 4$), namely

$$V_1(t)=\frac{1}{1-\pi_0}\sum_{n=1}^{K-1}\sum_{i=1}^{s}\pi_{ni}\, e^{-(a_{ni}t)^{b_{ni}}}$$

where a_{ni} and b_{ni} are chosen to fit a Weibull distribution to the conditional waiting time distribution $\Pr\{W_q\leqslant t \mid n,i\}$ by matching $\mathrm{E}\{\mathbf{W}_q \mid n,i\}$ and $\mathrm{E}\{\mathbf{W}_q{}^2 \mid n,i\}$.

The second approximation is given by

$$V_2(t)=\frac{1}{1-\pi_0}\sum_{n=1}^{K-1}\pi_n\, e^{-(a_n t)^{b_n}},$$

where a_n and b_n are chosen to fit a Weibull distribution to the conditional waiting time distribution $\Pr\{\mathbf{W}_q\leqslant t \mid n\}$ by matching $\mathrm{E}\{\mathbf{W}_q \mid n\}$ and $\mathrm{E}\{\mathbf{W}_q{}^2 \mid n\}$. Note that

$$\mathrm{E}\{\mathbf{W}_q{}^k \mid n\}=\sum_{i=1}^{s}\frac{\pi_{ni}}{\pi_n}\mathrm{E}\{\mathbf{W}_q{}^k \mid n,i\}, \quad k=1,2$$

The third approximation is

$$V_3(t)=e^{-(at)^b},$$

where a and b are chosen to fit a Weibull distribution to the conditional waiting time distribution $\Pr\{\mathbf{W}_q\leqslant t \mid \mathbf{W}_q>0\}$ by matching $\mathrm{E}\{\mathbf{W}_q \mid \mathbf{W}_q>0\}$ and $\mathrm{E}\{\mathbf{W}_q{}^2 \mid \mathbf{W}_q>0\}$ Finally,

$$V_4(t)=\frac{1}{1-\pi_0}\sum_{n=1}^{K-1}\pi_n\, e^{-\frac{t}{\mathrm{E}\{\mathbf{W}_q \mid n\}}}$$

Note that here only the first moments $\mathrm{E}\{\mathbf{W}_q \mid n\}$ are matched.

In Table 2.4, we see that V_1, V_2 and V_3 are good approximations for V and that V_4 is plainly a very bad approximation. The difference between V_1 and V_2 is remarkably small, though V_1 is a more detailed approximation than V_2. Further, V_3 requires much less computing time than V, V_1 and V_2 and is for that reason a very useful approximation.

Remark Let the random variable $\mathbf{X}$ have a Weibull distribution function $\Pr\{\mathbf{X}\leqslant t\}=1-e^{-(at)^b}$. Then

$$\mathrm{E}\mathbf{X}^k=\int_0^\infty t^k a^b t^{b-1} e^{-(at)^b}\,\mathrm{d}t=\frac{1}{a^k}\Gamma(1+k/b)$$

Hence, to find a and b when $\mathrm{E}\mathbf{X}$ and $\mathrm{E}\mathbf{X}^2$ are given, we need a numerical procedure to compute the gamma function and a numerical procedure to find b, the zero of

$$\frac{\Gamma(1+2/x)}{(\Gamma(1+1/x))^2}=\frac{\mathrm{E}\mathbf{X}^2}{(\mathrm{E}\mathbf{X})^2}$$

Next, $a=\dfrac{1}{\mathrm{E}\mathbf{X}}\Gamma(1+1/b)$.

□

3. APPROXIMATIONS FOR THE M/G/C QUEUE

The subject of this chapter is the standard M/G/c queue. We shall present an approximate analysis in order to obtain practically useful results for this multi-server queueing system. See also Tijms, van Hoorn and Federgruen[81a], Tijms and van Hoorn[81c] and van Hoorn and Tijms[82].

We consider a queueing system with $c>1$ servers and an infinite waiting capacity. Customers arrive according to a Poisson process with intensity λ and the service time $\mathbf{S}$ of a customer has a general probability distribution function $F(t)=\Pr\{\mathbf{S}\leqslant t\}$. The utilization factor $\rho=\lambda E\mathbf{S}/c$ is less than one.

In the literature much effort has been spent on exact solution methods for the M/G/c queue. So far, however, practically useful exact results have been obtained only for special cases such as the M/M/c queue and the M/D/c queue. The M/M/c queue can be solved as a birth and death process, while the M/D/c queue can be solved using the embedded Markov chain approach; cf. Crommelin[32]. In general we may not expect that an exact analysis of the M/G/c queue will ever yield computationally tractable results. To explain this, note that in order to set up the analysis of a queueing system, it is necessary to define the state of the system. The state description should contain sufficient information to describe the future probabilistic development of the system, given the state of the system. In particular, for the M/G/c queue this implies that the state description of the system at an arbitrary service completion epoch should contain the information about the residual service times of the other services in progress at that epoch.

In the supplementary variable technique, which was first introduced by Kosten (cf. Kosten[73]) the state of the system is described by one discrete variable for the number of customers present and by a continuous variable for each server representing the elapsed service times of the services in progress (if any). In Hokstad[80] and Ishikawa[79], this technique is employed to the $M/K_2/c$ queue and the $GI/E_k/c$ queue respectively (K_2 indicates a distribution with rational Laplace transform, where the denominator has degree 2). However, the algorithms obtained in this manner are numerically unstable and not reliable for higher values of c and k.

In de Smit[81], the solution of the $GI/H_m/c$ queue is given in the form of a Wiener Hopf type equation which is subsequently solved by using a factorization method.

Another exact approach is to consider service distributions of phase type. Then the queueing process can be represented as a continuous time Markov chain with a discrete valued state description which gives the number of customers present and the status of the phases of the services in progress. See Heffer[69] for a first analysis using this approach; cf. also the table books by Hillier and Yu[82] and by Sakasegawa[78].

In Takahashi and Takami[76], an efficient decomposition method is introduced which solves iteratively the equilibrium state equations of the Markov chain representation of the queueing process. This algorithm is also applicable to GI/G/c queues where both interarrival and service time distribution are of phase type. However, the computing time for this exact method increases rapidly with the number of servers and the dimensionality of the state space representation.

The above mentioned exact methods for special cases of the M/G/c queue have in common that they require a very skillful implementation to control the numerical reliability and that the computing times go up very rapidly with increasing number of servers.

As an alternative for an exact analysis, in recent years considerable attention has been paid to the development of approximations for various operating characteristics of the M/G/c queue. In Hokstad[78], an approximate formula for the generating function of the queue length distribution is derived using the supplementary variable technique. Our approximations deal both with the queue length distribution and the waiting time distribution. As a by-product we obtain an approximation for the mean waiting time. For this important performance measure, there have been obtained several other approximations in the literature. We mention here the good quality approximations in Boxma, Cohen and Huffels[80], Cosmetatos[76] and Takahashi[77]. The various approximations for the mean waiting time will be discussed at the end of this chapter.

The chapter is built up in the following way. In Section 3.1 we formulate the approximation assumptions which form the basis of the analysis. Using the regenerative method we derive a basic recursion relation for the steady state probabilities. In Section 3.2, this basic result is transformed into a tractable form, which is very suitable for practical computations.

The next four sections build on the recursion relation in Section 3.2 and are devoted to the derivation of approximations for other performance measures of the M/G/c queue, including the waiting time distribution and the output process. Also, asymptotic results for the queue length and waiting time distributions are obtained.

In Section 3.7, the validation of the approximations is done using a large number of exact results taken from Kühn[76] and Groenevelt, van Hoorn and Tijms[82]; cf. also Appendix D. The computational aspects of the algorithms derived in this chapter are extensively discussed in Appendix C.

3.1. The approximation assumption and the basic result

As method of analysis for the state probabilities p_n and q_n, we use the regenerative method. As regeneration points we choose the epochs at which the system becomes empty. Assuming that the system is empty at epoch 0, define

$\mathbf{T}$ = the next epoch at which the system becomes empty

$\mathbf{T}_n$ = the amount of time in (0,T] during which n customers are in the system, $n \geq 0$

$\mathbf{N}$ = the number of customers served in (0,T]

$\mathbf{N}_n$ = the number of service completion epochs in (0,T] at which n customers are left behind by the customer just served, $n \geq 0$.

Like in Chapter 1, we focus on the embedded process within the busy cycle induced by the service completion epochs. For the case of a single server, as in Chapter 1, this process is a Markov chain. However this is in general not true for multiple servers. To get implementable results, we have to compromise between mathematical and practical standpoints. Therefore we make an approximation assumption in order to construct a simple embedded Markov chain on the service completion epochs. We wish to choose this approximation assumption in such a way

that at each service completion epoch the future behaviour of the queueing process depends only on the number of customers left behind at that service completion epoch and not on the elapsed service times of services in progress, if any.

We distinguish between the case in which one or more servers are free and the case in which all servers are busy at the service completion epoch.

> Consider first the case in which j customers are left behind at a departure epoch where $1 \leqslant j \leqslant c-1$. Then all of these j customers are in service. *We now assume that their residual service times are independent random variables having the same distribution as a generic random variable with distribution function* $F_e(t)$. In this case the next service completion may be generated by a newly arriving customer who completes service before any of the j services in progress is completed.
>
> Next consider the case in which j customers are left behind at a service completion epoch where $j \geqslant c$. Then at this epoch one new service starts while $c-1$ other services are in progress. *We now make the assumption that the time until the next service completion has distribution function* $F(ct)$. Clearly, in this case, future arrivals do not influence the next service completion.

For the distribution function $F_e(t)$, we take

$$F_e(t) = \frac{1}{\mathbf{ES}} \int_0^t (1 - F(x))\, dx \tag{3.1}$$

i.e. the equilibrium or excess lifetime distribution of $F(t)$. This distribution is well known in renewal theory.

To motivate the specification of the approximation assumption, note that if not all c servers are busy, the M/G/c queue can be treated as a M/G/∞ queue for which the renewal theoretic result is valid that at an arbitrary epoch the remaining service times of services in progress (if any) are independent random variables with common probability distribution function $F_e(t)$ given by (3.1); cf. Takacs[62]. If all c servers are busy we treat the M/G/c queue with service time $\mathbf{S}$ as a M/G/1 queue with service time $\mathbf{S}/c$; cf. also Newell[73].

Note that the approximation assumption is exactly satisfied both for the M/G/c queue with exponential service time and the M/G/c queue with $c = 1$ or $c = \infty$.

Next we define for $0 \leqslant j \leqslant n$,

A_{jn} = the expected amount of time during which n customers are present until the next service completion epoch, given that at epoch 0 a service is completed with j customers left behind in the system.

The quantities A_{jn} are well defined owing to the approximation assumption.

Then, by partitioning the busy cycle (0,**T**] by means of the service completion epochs and using Wald's Theorem, we obtain a recurrence relation between $\mathbf{ET}_n$ and $\mathbf{EN}_n$. This is stated in

Theorem 3.1

$$p_n = \frac{\mathrm{ET}_n}{\mathrm{ET}}, \quad q_n = \frac{\mathrm{EN}_n}{\mathrm{EN}}, \quad n \geqslant 0 \tag{3.2}$$

$$\mathrm{ET}_n = \sum_{j=0}^{n} \mathrm{EN}_j A_{jn}, \quad n \geqslant 1 \tag{3.3}$$

$$\mathrm{EN}_n = \lambda \mathrm{ET}_n, \quad n \geqslant 1 \tag{3.4}$$

$$\mathrm{EN} = \lambda \mathrm{ET} \tag{3.5}$$

Proof The proof is identical to that of Theorem 1.1.

□

Corollary 3.2 The distributions (p_n) and (q_n) are equal.

□

Remark 3.3 Note that (3.2), (3.4) and (3.5) are exact and do not involve the approximation assumption.

□

In its present form Theorem 3.1 is not very useful for numerical purposes. The quantities A_{jn} hide a number of computational difficulties. Note e.g. that for $j < n \leqslant c-1$ an explicit expression for A_{jn} involves a $(n-j+1)$-dimensional integral because of the phenomenon that any of $n-j$ newly started services may be completed before each of the services in progress.

Fortunately, by the special form of the approximation assumption, we can succeed in eliminating the multidimensional integrals so that the ultimate recursive scheme is well suited for numerical purposes in practice.

3.2. The algorithm

Theorem 3.4

$$p_n = \frac{(\lambda \mathrm{ES})^n}{n!} p_0, \quad 0 \leqslant n \leqslant c-1 \tag{3.6}$$

$$p_n = \lambda p_{c-1} \alpha_{n-c} + \lambda \sum_{j=c}^{n} p_j \beta_{n-j}, \quad n \geqslant c \tag{3.7}$$

where

$$\alpha_k = \int_0^\infty (1-F_c(t))^{c-1} (1-F(t)) e^{-\lambda t} \frac{(\lambda t)^k}{k!} \mathrm{d}t, \quad k \geqslant 0 \tag{3.8}$$

$$\beta_k = \int_0^\infty (1-F(ct)) e^{-\lambda t} \frac{(\lambda t)^k}{k!} \mathrm{d}t, \quad k \geqslant 0 \tag{3.9}$$

$$p_0 = \frac{1}{\sum_{n=0}^{c-1} \frac{(\lambda \mathrm{ES})^n}{n!} + \frac{(\lambda \mathrm{ES})^c}{c!(1-\rho)}} \tag{3.10}$$

Proof We first derive an expression for the A_{jn} and then use its special structure to prove the theorem by induction. Define

$$M_{jn}(t)=\Pr\{\ n-j \text{ customers arrive in } (0,t) \text{ and each service started beyond time 0 is still in progress at time } t, \text{ given at epoch 0 there are } j \text{ customers are present}\} \tag{3.11}$$

Recalling the approximation assumption, it now follows that

$$A_{jn}=\int_0^\infty (1-F_e(t))^j M_{jn}(t)\,\mathrm{d}t\ ,\quad 1\leqslant j<c\ ,\quad n\geqslant j \tag{3.12}$$

$$A_{jn}=\int_0^\infty (1-F(ct))M_{jn}(t)\,\mathrm{d}t\ ,\quad c\leqslant j\leqslant n \tag{3.13}$$

$$A_{0n}=\int_0^\infty (1-F(t))M_{1n}(t)\,\mathrm{d}t\ ,\quad n\geqslant 1 \tag{3.14}$$

The first part of the integrand expresses the fact that the services in progress at a service completion epoch expire after t, the second part is the probability that n customers are present before the next service completion (cf. Lemma 1.4).
By considering what may happen in the interval $(0,\delta t)$, we find for $0\leqslant j\leqslant c-1$, $n>j$

$$M_{jn}(t+\delta t)=(1-\lambda\delta t)M_{jn}(t)+\lambda\delta t\,(1-F(t))M_{j+1,n}(t)+o(\delta t)$$

Note that if an arrival occurs in $(0,\delta t)$ a new service starts when $j<c$. Letting $\delta t\to 0$, we get

$$\frac{\mathrm{d}}{\mathrm{d}t}M_{jn}(t)=-\lambda M_{jn}(t)+\lambda(1-F(t))M_{j+1,n}(t)\ ,\quad 0\leqslant j\leqslant c-1\ ,\quad n>j \tag{3.15}$$

For $c\leqslant j\leqslant n$, an arriving customer joins the queue and $M_{jn}(t)$ equals the probability of $n-j$ arrivals in $(0,t)$, i.e.

$$M_{jn}(t)=e^{-\lambda t}\frac{(\lambda t)^{n-j}}{n-j!}\ ,\quad c\leqslant j\leqslant n \tag{3.16}$$

Clearly, for $n\geqslant 1$

$$M_{nn}(t)=e^{-\lambda t} \tag{3.17}$$

Define for $1\leqslant j\leqslant c-1$ the auxiliary quantities

$$B_{jn}=\int_0^\infty (1-F_e(t))^{j-1}(1-F(t))M_{jn}(t)\,\mathrm{d}t \tag{3.18}$$

Then

$$A_{jn}=B_{j+1,n}-\frac{j}{\lambda\mathbf{ES}}B_{jn}\ ,\quad 1\leqslant j\leqslant c-1\ ,\quad n>j \tag{3.19}$$

as easily follows by multiplying (3.15) with $(1-F_e(t))^j$ and applying partial integration to the left hand side.

To prove the theorem, assume that $p_0,\ldots,p_{n-1}$ satisfy (3.6), $n\leqslant c-1$, then

$$p_n(1-\lambda A_{nn}) = \sum_{j=0}^{n-1} \lambda p_j A_{jn} = \lambda p_0 A_{0n} + \sum_{j=1}^{n-1} \lambda p_j (B_{j+1,n} - \frac{j}{\lambda \mathbf{ES}} B_{jn}) \tag{3.20}$$

$$= \sum_{j=0}^{n-1} \lambda p_j B_{j+1,n} - \sum_{j=1}^{n-1} \lambda p_{j-1} B_{jn} = \lambda p_{n-1} B_{nn} \tag{3.21}$$

Note that (3.20) is equivalent to (3.3) by inserting $\mathbf{ET}_n = p_n \mathbf{ET}$ and $\mathbf{EN}_n = \lambda p_n \mathbf{ET}$. In (3.21) the induction assumption is used and A_{0n} is replaced by B_{1n}. With partial integration, it follows that

$$1-\lambda A_{nn} = \frac{\mathbf{ES}}{n} B_{nn}.$$

Now, p_n also satisfies (3.6).

The second part of Theorem 3.4 follows by deriving in a similar way that $\sum_{j=0}^{c-1} \lambda p_j A_{jn} = \lambda p_{c-1} B_{cn}$. By substituting (3.16) in (3.13), we find $A_{jn} = \beta_{n-j}$, $c \leqslant j \leqslant n$ and with $B_{cn} = \alpha_{n-c}$ the recurrence relation follows.

Finally, after summing (3.7) over $n \geqslant c$, it follows that

$$\sum_{n=c}^{\infty} p_n = \frac{\rho p_{c-1}}{1-\rho} \tag{3.22}$$

since $\sum_{k=0}^{\infty} \alpha_k = \sum_{k=0}^{\infty} \beta_k = \mathbf{ES}/c$. The expression for p_0 is now obvious. This ends the proof.

□

The computational scheme for the state probabilities is very simple and the computational complexity depends only on the evaluation of the constants α_k and β_k. For the evaluation of the integrals α_k and β_k several methods are available in numerical analysis. The selection of a method should be based on the properties of $F(t)$ and the associated equilibrium distribution $F_e(t)$. In some cases, numerical integration can be avoided. In Appendix C we give details for a number of important distributions.

A closer look at Theorem 3.4 learns that $p_0, \ldots, p_{c-1}$ and hence also the delay probability are equal to the corresponding probabilities in the M/M/c queue with the same utilization factor ρ. This approximate result also appears at other places in the literature. In particular, the resulting Erlang formula for the delay probability is successfully used as approximation in practice. A plausible explanation for this result is the observation that the approximation assumption in fact imposes a memoryless property to the system at service completion epochs. A service surviving k times (say) other services starts k times afresh with distribution function $F_e(t)$.

In the next four sections, we proceed with the analysis of the approximate M/G/c model. With Theorem 3.4 as starting point we focus successively on the generating function of the distribution (p_n), the waiting time distribution, the departure process and asymptotic properties of the queue length distribution and the waiting time distribution.

3.3. The generating function and the moments of the queue length

In this section, we determine the generating function of the probabilities p_n for $n \geqslant c$. From this generating function, we can easily compute the moments of the queue length and the waiting time distributions. Also, in Section 3.4, we use the generating function to derive an integral equation for the waiting time distribution function.

Assuming that the system is in the steady state, we define the random variables

$\mathbf{L}_q$ = the queue size at an arbitrary epoch (excluding the customers in service)

$\mathbf{W}_q$ = the waiting time of an arbitrary customer in the queue (excluding his service time)

Further, define for $|z| \leqslant 1$ the generating functions $P(z)=\sum_{n=c}^{\infty} p_n z^{n-c}$, $\alpha(z)=\sum_{k=0}^{\infty}\alpha_k z^k$, $\beta(z)=\sum_{k=0}^{\infty}\beta_k z^k$ and let

P_W = the probability that an arbitrary arriving customer has to wait in the queue

From (3.22) it follows that

$$\mathrm{P}_W = \frac{\rho}{1-\rho} p_{c-1} \tag{3.23}$$

The distribution of $\mathbf{L}_q$ follows directly from (p_n), namely $\Pr\{\mathbf{L}_q = n\} = p_{n+c}$, $n>0$ and $\Pr\{\mathbf{L}_q = 0\} = p_0 + \cdots + p_c$. Hence the generating function of $\mathbf{L}_q$ is equal to $P(z)$ up to an additive constant.

From (3.8) and (3.9), we obtain directly that

$$\alpha(z) = \int_0^{\infty} (1-F_e(t))^{c-1}(1-F(t))\, e^{-\lambda t(1-z)}\,\mathrm{d}t$$

$$\beta(z) = \int_0^{\infty} (1-F(ct))\, e^{-\lambda t(1-z)}\,\mathrm{d}t$$

To find $P(z)$, note that (3.7) is an equation of convolution type, and multiplying it with z^{n-c} and summing over $n \geqslant c$ yields

$$P(z) = \lambda p_{c-1}\alpha(z) + \lambda\beta(z)P(z) \tag{3.24}$$

and so

$$P(z) = \lambda p_{c-1}\frac{\alpha(z)}{1-\lambda\beta(z)} \tag{3.25}$$

By differentiation of (3.25) and after some algebra, we get

$$\mathrm{EL}_q = P'(1) = \mathrm{L}_q(exp)\{(1-\rho)\frac{c\gamma_1}{\mathrm{ES}} + \rho\frac{\mathrm{ES}^2}{2(\mathrm{ES})^2}\} \tag{3.26}$$

$$\mathrm{EL}_q(\mathbf{L}_q - 1) = P''(1) = \frac{\rho^2 \mathrm{P}_W}{1-\rho}\{(1-\rho)\frac{c^2\gamma_2}{(\mathrm{ES})^2} + \rho\frac{\mathrm{ES}^3}{3(\mathrm{ES})^3}\} + \frac{\rho^2}{1-\rho}\frac{\mathrm{ES}^2}{(\mathrm{ES})^2}\mathrm{EL}_q \tag{3.27}$$

where

$$\gamma_k = \int_0^{\infty} k t^{k-1}(1-F_e(t))^c\,\mathrm{d}t\ ,\quad k = 1,2. \tag{3.28}$$

In (3.26) $L_q(exp)$ denotes the mean queue length in the M/M/c queue with traffic intensity ρ and is given by

$$L_q(exp) = \frac{\rho}{1-\rho} P_W(exp) \tag{3.29}$$

where

$$P_W(exp) = \frac{\frac{(\lambda ES)^c}{c!(1-\rho)}}{\sum_{n=0}^{c-1} \frac{(\lambda ES)^n}{n!} + \frac{(\lambda ES)^c}{c!(1-\rho)}}$$

In addition to (3.26) and (3.27), higher moments of $\mathbf{L}_q$ can also be obtained. The moments of $\mathbf{W}_q$ are for FIFO service discipline related to those of $\mathbf{L}_q$ by the relation (cf. Marshall[71], and the derivation in Section 3.4).

$$E\mathbf{L}_q(\mathbf{L}_q - 1) \cdots (\mathbf{L}_q - k + 1) = \lambda^k E\mathbf{W}_q{}^k \ , \quad k \geqslant 1 \tag{3.30}$$

Remark 3.5 The approximation for the mean queue length is a linear interpolation of the exact light traffic value and the exact heavy traffic value of the mean queue length for the M/G/c queue; cf. Burman and Smith[81] and Köllerström[79]. Hence, the approximation assumption leads to results that are very close to the exact results in light traffic and heavy traffic situations.

3.4. The waiting time distribution

In this section we derive a relation between the distributions of $\mathbf{W}_q$ and $\mathbf{L}_q$ and an integral equation for the waiting time distribution function. As a by-product we get relation (3.30). We assume that service is in order of arrival. Define for $t \geqslant 0$

$$W_q(t) = \Pr\{\mathbf{W}_q \leqslant t\}$$

and

$$V(t) = \Pr\{\mathbf{W}_q \leqslant t \mid \mathbf{W}_q > 0\}$$

i.e. $V(t)$ is the distribution function of the waiting time of customers who have to wait. Using that $\Pr\{\mathbf{W}_q > 0\} = P_W$, it follows that

$$V(t) = \frac{\Pr\{0 < \mathbf{W}_q \leqslant t\}}{P_W} = 1 - \frac{1 - W_q(t)}{P_W} \tag{3.31}$$

For an arbitrary customer C_1, let C_2 be the first customer entering service after the service completion of customer C_1. The waiting time spent in the queue by customer C_2 is $\mathbf{W}_q$. By definition, with probability q_n, the customer C_1 leaves n customers behind in the system. If $0 \leqslant n \leqslant c-1$, then it is obvious that C_2 is not among these n customers since they all started their service before the departure of C_1. Hence, if $0 \leqslant n \leqslant c-1$ the customer C_2 immediately enters service upon arrival and his waiting time $\mathbf{W}_q$ is zero. Hence, the event that customer C_1 leaves behind n customers with $n \geqslant c$ occurs only if the waiting time $\mathbf{W}_q$ of customer C_2 is positive and $n-c$ arrivals occur during this time. This yields

$$q_n = \Pr\{\mathbf{W}_q > 0 \text{ and } n - c \text{ arrivals occur during } \mathbf{W}_q\} \ , \quad n \geqslant c \tag{3.32}$$

So far, the reasoning has only used the fact that customers singly arrive and are singly served under a FIFO discipline. Now, using the equality of (p_n) and (q_n) and the fact that the arrival process is Poisson, we get from (3.32)

$$p_n = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^{n-c}}{n-c!} \mathrm{dPr}\{0<\mathbf{W}_q \leqslant t\}\ ,\quad n \geqslant c \tag{3.33}$$

Taking generating functions in (3.33) and noting that $\Pr\{0<\mathbf{W}_q \leqslant t\} = \mathrm{P}_W V(t)$, leads to

$$P(z)/\mathrm{P}_W = \int_0^\infty e^{-\lambda t(1-z)} \mathrm{d}V(t).$$

Using the change of variable $s = \lambda(1-z)$ and using (3.25) we get for the Laplace stieltjes transform of $V(t)$

$$\int_0^\infty e^{-st}\,\mathrm{d}V(t) = \frac{p_{c-1}}{\mathrm{P}_W} \frac{\lambda \int_0^\infty (1-F_e(t))^{c-1}(1-F(t))e^{-st}\,\mathrm{d}t}{1-\lambda\int_0^\infty (1-F(ct))e^{-st}\,\mathrm{d}t} \tag{3.34}$$

To derive an integral equation of renewal type for $V(t)$, we introduce

$$\omega(s) = \int_0^\infty e^{-st}\,\mathrm{d}V(t),$$

$$\phi(s) = \int_0^\infty e^{-st}\,\mathrm{d}\{1-(1-F_e(t))^c\},$$

and

$$\psi(s) = \int_0^\infty e^{-st}\,\mathrm{d}F_e(ct)$$

Using $p_{c-1} = \frac{1-\rho}{\rho}\mathrm{P}_W$ (cf. (3.23)) it easily follows that

$$\omega(s) = \frac{(1-\rho)\phi(s)}{1-\rho\psi(s)}$$

Consequently

$$\omega(s) = (1-\rho)\phi(s) + \rho\omega(s)\psi(s) \tag{3.35}$$

Equation (3.35) is the Laplace-Stieltjes transform of the integral equation

$$\begin{aligned} V(t) &= (1-\rho)\{1-(1-F_e(t))^c\} + \rho\int_0^t V(t-x)\,\mathrm{d}F_e(cx) \\ &= (1-\rho)\{1-(1-F_e(t))^c\} + \lambda\int_0^t V(t-x)(1-F(cx))\,\mathrm{d}x \end{aligned} \tag{3.36}$$

This Volterra integral equation of the second kind is useful for numerical purposes. In Appendix B, we show how this equation is efficiently solved by discretizing the time parameter.

3.5. The departure process

An important characteristic of a queueing system is the departure process of customers from the system. To illustrate this, in a network of queues the input to a particular queue in the network is the output of one or more other queues plus traffic from external sources. Except for some special cases (see e.g. Pack[78], Heffes[76]) no tractable results are available for non-Markovian multiserver queues.

In this section, we give for the approximate M/G/c queue in steady state a derivation of the stationary interdeparture distribution, i.e. the distribution of the time between two successive departures from the system. From this distribution, the moments of the interdeparture intervals easily can be computed. Note that the interdeparture distribution does not completely specify the output process since the departure process is not renewal.

Under the assumption that epoch 0 is a service completion epoch, define the random variable

$\mathbf{T}_D$ = the time until the next departure

and the complementary distribution function

$$Q(t)=\Pr\{\mathbf{T}_D>t\}$$

The distribution of the number of customers left behind in the system at epoch 0 is (q_n). Also, by the approximation assumption, the knowledge of the number of customers left behind at epoch 0 gives enough information to describe the time until the next service completion. Hence, by conditioning on the number of customers left behind at epoch 0 and by using the equality $(q_n)=(p_n)$, we find

$$Q(t)=\sum_{i=0}^{c-1}p_i(1-F_e(t))^i\sum_{n=0}^{\infty}M_{i,n+i}(t)+\sum_{i=c}^{\infty}p_i(1-F(ct)),$$

where the functions $M_{jn}(t)$ are defined by (3.11). The factor $(1-F_e(t))^i\sum_{n=0}^{\infty}M_{i,n+i}(t)$ for example is the probability that the i services in progress at epoch 0 and any service started after epoch 0 expire beyond time t. Let

$$R(t)=\sum_{i=0}^{c-1}p_i(1-F_e(t))^i\sum_{n=0}^{\infty}M_{i,n+i}(t)$$

After some algebra, the following differential equation can be derived using the properties of p_n, $0\leqslant n\leqslant c$, $F_e(t)$ and $M_{jn}(t)$.

$$\frac{\mathrm{d}}{\mathrm{d}t}R(t)=-\lambda R(t)+\lambda p_{c-1}(1-F_e(t))^{c-1}(1-F(t)) \tag{3.37}$$

Finally, integration of (3.37), $R(0)=1-\mathrm{P}_W$ and $Q(t)=R(t)+\mathrm{P}_W(1-F(ct))$ yield

$$Q(t)=(1-\mathrm{P}_W)e^{-\lambda t}-\lambda p_{c-1}\int_0^t e^{-\lambda(t-u)}(1-F_e(u))^{c-1}(1-F(u))\,\mathrm{d}u$$

$$+\mathrm{P}_W(1-F(ct)) \tag{3.38}$$

The moments of $\mathbf{T}_D$ are given by, $m \geqslant 1$

$$\mathrm{ET}_D{}^m = \frac{m!}{\lambda^m}\{1-\rho \mathrm{P}_W + (1-\rho)\mathrm{P}_W \sum_{i=1}^{m-1} \frac{\lambda^i}{i!}\gamma_i + \mathrm{P}_W \frac{\lambda^m \mathrm{ES}^m}{c^m m!}\} \tag{3.39}$$

with γ_i defined by (3.28). In particular, we have $\mathrm{ET}_D = 1/\lambda$, in agreement with the fact that customers enter and leave the system at the same rate. After some algebra, the second moment of $\mathbf{T}_D$ can be rewritten as

$$\mathrm{ET}_D{}^2 = \frac{2}{\lambda^2}\{1-\rho \mathrm{P}_W + (1-\rho)\mathrm{EL}_q\} \tag{3.40}$$

3.6. Asymptotic properties of the state probabilities and the waiting time

The recurrence relation (3.7) for the state probabilities and the integral equation (3.36) for $V(t)$ are both defective renewal equations. In the discrete renewal Equation (3.7) the numbers $\lambda\beta_k$, $k \geqslant 0$ sum to $\lambda \mathrm{ES}/c = \rho < 1$ and in (3.36) the function $\rho F_e(cx)$ is not a proper probability distribution, since $\lim_{x\to\infty} \rho F_e(cx) = \rho < 1$. In order to apply the limiting theorems from renewal theory, we reduce the Equations (3.7) and (3.36) to proper renewal equations using a 'standard trick' (cf. Feller[66,68]). Next we apply the key renewal theorem to obtain the limiting behaviour of p_n for $n \to \infty$ and $V(t)$ for $t \to \infty$. Define

$$V^*(t) = 1 - V(t)$$

Then $V^*(t)$ satisfies (cf. 3.36)

$$V^*(t) = G(t) + \int_0^t V^*(t-x)\,\mathrm{d}H(x) \tag{3.41}$$

with $G(t) = 1 - \rho F_e(ct) - (1-\rho)\{1-(1-F_e(t))^c\}$ and and $H(t) = \rho F_e(ct)$. Note that $H(t)$ has density $\lambda(1-F(ct))$. The following lemma forms the basis of the standard trick to transform (3.41) into a proper renewal equation.

Lemma 3.6 There exists at most one $\xi > 0$ satisfying

$$\int_0^\infty e^{\xi t}\,\mathrm{d}H(t) = 1 \tag{3.42}$$

where $H(t)$ is defined as above.

Proof Let

$$k(x) = \int_0^\infty e^{xt}\,\mathrm{d}H(t) - 1\,, \quad x \geqslant 0$$

The function $k(x)$ is a monotone increasing function with $k(0) = \rho - 1 < 0$ and $\lim_{x\to\infty} k(x) = \infty$. Hence, $k(x)$ has at most one positive root. However, a root need not to exist. For example, when $F(t)$ is a log-normal distribution function, then there exists no ξ satisfying (3.42).

□

Suppose that a number ξ exists satisfying (3.42), then

$$e^{\xi t}\,\mathrm{d}H(t) = \lambda e^{\xi t}(1-F(ct))\,\mathrm{d}t$$

and apparently $\lambda e^{\xi t}(1-F(ct))$ is the density of a (proper) probability distribution. Define

$$V^{\#}(t)=e^{\xi t}V^{*}(t)$$

and multiply (3.41) with $e^{\xi t}$, then $V^{\#}(t)$ satisfies

$$V^{\#}(t)=e^{\xi t}G(t)+\int_0^t V^{\#}(t-x)e^{\xi x}\,\mathrm{d}H(x). \tag{3.43}$$

By applying the key renewal theorem (cf. Appendix A) to (3.43) we get the asymptotic result for $t\to\infty$

$$V^{\#}(t)\approx\frac{\int_0^\infty e^{\xi x}G(x)\,\mathrm{d}x}{\int_0^\infty x\,e^{\xi x}\,\mathrm{d}H(x)}$$

and finally, for $t\to\infty$

$$1-V(t)\approx e^{-\xi t}\,\frac{\int_0^\infty e^{\xi x}\{1-\rho F_c(cx)-(1-\rho)(1-(1-F_c(x))^c)\}\,\mathrm{d}x}{\int_0^\infty \lambda x\,e^{\xi x}(1-F(cx))\,\mathrm{d}x} \tag{3.44}$$

The asymptotic results for the state probabilities are obtained in a similar way by applying the discrete version of the renewal theorem; cf. Feller[68]. This results for $n\to\infty$ in

$$p_{n+c}\approx\eta^{-n-1}p_{c-1}\,\frac{\int_0^\infty e^{-\lambda x(1-\eta)}(1-F_c(x))^{c-1}(1-F(x))\,\mathrm{d}x}{\int_0^\infty x\,e^{-\lambda x(1-\eta)}(1-F(cx))\,\mathrm{d}x} \tag{3.45}$$

with $\eta=1+\frac{\xi}{\lambda}$.

For the M/G/c queue with phase type services, Takahashi[81] has shown that the distribution (p_n) has a geometric tail and that the waiting time distribution has an exponential tail. In support of our approximation, the same coefficients ξ and η are involved in Takahashi's asymptotic expansion. However, our approximate results (3.44) and (3.45) also hold for general service distributions for which a ξ satisfying (3.42) can be found.

Not only for reasons of completeness we have analyzed the asymptotic behaviour of the model. Also, for practical purposes these results are valuable. The computational effort to compute the tail probabilities of (p_n) or the tail of the waiting time distribution $V(t)$ increases with n and t respectively. Hence it is worthwhile to compute p_n or $V(t)$ from the asymptotic formulae (3.45) and (3.44) for n and t sufficiently large.

3.7. Numerical results

In this section, we present some numerical results for the M/G/c queue and discuss the quality of the approximations. We give results for a wide range of the parameters c , $\rho=\lambda ES/c$ and C_S^2, the number of servers, the utilization factor and the squared coefficient of variation of the service time respectively. We consider deterministic service times (D), hyperexponential service times (H_2) and service times which are a mixture of Erlang distributions with the same scale parameter ($E_{1,2}$) and ($E_{1,3}$). The mean service time is normalized to one. For a complete specification of these distributions we refer to Appendix C.

The exact results for the M/D/c queue have been taken from Kühn[76]. We have computed the exact results for the M/G/c queue with phase type service distributions using a specialization of the method of Takahashi and Takami[76]; cf. Groenevelt, van Hoorn and Tijms[82] and Appendix D.

The Tables 3.1, 3.2 and 3.3 concern the delay of probability P_W, the mean queue length EL_q and the coefficient of variation of the queue length $cv(L_q)$

$$cv(L_q)=\sqrt{EL_q^{\,2}/(EL_q)^2-1}$$

The top numbers are the exact values and the second top numbers are the approximations. It is clearly demonstrated that the approximations are very close to the exact results for the whole range of parameters covered in these tables. In Tijms and Van Hoorn[81c], another approximation is given for the M/D/c queue which improves in almost all cases the current approximation. This other approximation is obtained with the same technique as used in this chapter, but is based on a modification of the approximation assumption made in Section 3.1. The modified formulae for the delay probability and the mean queue length, denoted by $\overline{P_W}$ and $\overline{EL_q}$ are given by

$$\overline{P_W}=P_W(exp)\{1-(\frac{\eta_1}{\eta_2}-1)\frac{1-\rho}{\rho}\}$$

$$\overline{EL_q}=P_W(exp)\{\frac{\rho}{2(1-\rho)}+(\frac{\eta_1}{\eta_2}-1)\frac{1-\rho}{\rho}\}$$

where (with $ES=D$)

$$\eta_1=\frac{c-1}{D}\int_0^D(1-\frac{t}{D})^{c-2}e^{-\lambda t}\,dt \text{ and } \eta_2=e^{-\rho}$$

In Table 3.4, we give some results for the coefficient of variation c_D of the interdeparture time T_D. The top numbers in this table give the simulated actual values of c_D with a 95 % confidence interval and the second top numbers the approximate value of c_D corresponding to (3.40).

In Figure (3.5) and Table (3.6), we give some results for $V(t)$, the complementary waiting time distribution for the delayed customers. Our approximation for $V(t)$, (Appr.) has been obtained by solving the integral equation (3.36) using the numerical procedures given in Appendix B. For the case of deterministic service times we compare our results with the exact results (Exact) given in Kühn[76]. For the other cases, we compare our approximate results both with the asymptotic results (Asym.) using Takahashi[81] and with simulation results (Sim.). For each example, we have simulated one million customers. In the table, the notation 0.77(1) means that the 95 % confidence interval of the simulated value is 0.76-0.78. It can clearly be seen

	$\rho=0.5$			$\rho=0.8$			$\rho=0.9$			$\rho=0.95$		
c	P_W	EL_q	$cv(EL_q)$	P_W	EL_q	$cv(EL_q)$	P_W	EL_q	$cv(EL_q)$	P_W	EL_q	$cv(EL_q)$
2	0.3233	0.1767	3.1099	0.7019	1.4454	1.5149	0.8471	3.8654	1.2305	0.9227	8.8230	1.1098
	0.3333	0.1944	2.9864	0.7111	1.5170	1.4711	0.8526	3.9647	1.2068	0.9256	8.9401	1.0973
3	0.2253	0.1308	3.7165	0.6325	1.3294	1.6137	0.8077	3.7204	1.2712	0.9018	8.6638	1.1285
	0.2368	0.1480	3.5307	0.6472	1.4238	1.5512	0.8171	3.8606	1.2371	0.9070	8.8320	1.1103
5	0.1213	0.0766	5.0503	0.5336	1.1562	1.7872	0.7478	3.4965	1.3392	0.8692	8.4106	1.1593
	0.1304	0.0869	4.7784	0.5541	1.2560	1.7063	0.7625	3.6600	1.2960	0.8778	8.6171	1.1362
10	0.0331	0.0237	9.6101	0.3847	0.8787	2.1627	0.6469	3.1013	1.4744	0.8116	7.9496	1.2184
	0.0361	0.0254	9.1516	0.4092	0.9523	2.0574	0.6687	3.2555	1.4234	0.8256	8.1639	1.1921
15	0.0104	0.0080	17.072	0.2955	0.7011	2.5095	0.5771	2.8196	1.5868	0.7695	7.6098	1.2657
	0.0113	0.0081	16.401	0.3192	0.7501	2.3841	0.6026	2.9491	1.5320	0.7870	7.8032	1.2387
25	0.0012	0.0010	50.183	0.1900	0.4773	3.1957	0.4793	2.4116	1.7817	0.7063	7.0870	1.3437
	0.0013	0.0009	48.899	0.2091	0.4954	3.0291	0.5079	2.4966	1.7206	0.7284	7.2388	1.3166
50				0.0776	0.2143	5.1230	0.3355	1.7787	2.2034	0.6012	6.1944	1.4972
				0.0870	0.2073	4.8400	0.3639	1.7947	2.1254	0.6291	6.2635	1.4698
100				0.0176	0.0540	10.941	0.1953	1.1094	2.9888	0.4751	5.0778	1.7404
				0.0196	0.0470	10.345	0.2169	1.0719	2.8714	0.5065	5.0471	1.7096
200				0.0013	0.0043	41.131	0.0837	0.5196	4.7067	0.3351	3.7631	2.1488
				0.0014	0.0033	39.237	0.0945	0.4672	4.4966	0.3653	3.6418	2.1065

Table 3.1 Exact and approximate results for the M/D/c queue.

	$\rho=0.5$			$\rho=0.8$			$\rho=0.9$			$\rho=0.95$		
c	P_W	EL_q	$cv(EL_q)$	P_W	EL_q	$cv(EL_q)$	P_W	EL_q	$cv(EL_q)$	P_W	EL_q	$cv(EL_q)$
2	0.3308	0.2556	2.9251	0.7087	2.1478	1.4846	0.8512	5.7732	1.2194	0.9248	13.210	1.1052
	0.3333	0.2604	2.8997	0.7111	2.1689	1.4751	0.8526	5.8032	1.2145	0.9256	13.245	1.1026
3	0.2338	0.1844	3.5176	0.6432	1.9638	1.5849	0.8145	5.5441	1.2615	0.9056	12.957	1.1248
	0.2368	0.1891	3.4739	0.6472	1.9919	1.5702	0.8171	5.5867	1.2539	0.9070	13.009	1.1209
4	0.1710	0.1371	4.1411	0.5914	1.8165	1.6747	0.7844	5.3542	1.2980	0.8895	12.745	1.1416
	0.1739	0.1409	4.0813	0.5964	1.8466	1.6566	0.7878	5.4025	1.2889	0.8914	12.804	1.1369
5	0.1279	0.1038	4.8113	0.5484	1.6929	1.7580	0.7584	5.1899	1.3309	0.8754	12.559	1.1566
	0.1304	0.1067	4.7361	0.5541	1.7229	1.7372	0.7625	5.2406	1.3207	0.8778	12.623	1.1515
8	0.0576	0.0481	7.2147	0.4508	1.4089	1.9859	0.6960	4.7914	1.4166	0.8407	12.097	1.1950
	0.0590	0.0492	7.0910	0.4576	1.4341	1.9593	0.7015	4.8408	1.4046	0.8442	12.163	1.1892
10	0.0352	0.0298	9.2491	0.4021	1.2650	2.1274	0.6625	4.5761	1.4668	0.8216	11.841	1.2170
	0.0361	0.0303	9.0883	0.4092	1.2863	2.0979	0.6687	4.6219	1.4542	0.8256	11.905	1.2111
15	0.0110	0.0095	16.564	0.3122	0.9955	2.4657	0.5952	4.1402	1.5787	0.7819	11.309	1.2649
	0.0113	0.0096	16.282	0.3192	1.0082	2.4296	0.6026	4.1753	1.5652	0.7870	11.362	1.2589
20	0.0036	0.0032	28.798	0.2497	0.8046	2.7973	0.5427	3.7965	1.6787	0.7496	10.872	1.3062
	0.0037	0.0032	28.330	0.2561	0.8111	2.7550	0.5508	3.8214	1.6645	0.7554	10.915	1.3004
25	0.0012	0.0011	49.267	0.2033	0.6613	3.1322	0.4995	3.5114	1.7714	0.7219	10.497	1.3434
	0.0013	0.0011	48.512	0.2091	0.6635	3.0837	0.5079	3.5273	1.7566	0.7284	10.529	1.3379
30				0.1678	0.5499	3.4760	0.4628	3.2676	1.8592	0.6976	10.166	1.3777
				0.1729	0.5492	3.4210	0.4714	3.2758	1.8439	0.7045	10.188	1.3724
40				0.1173	0.3895	4.2047	0.4029	2.8662	2.0256	0.6560	9.5971	1.4401
				0.1212	0.3856	4.1357	0.4116	2.8624	2.0091	0.6636	9.6003	1.4353
50				0.0840	0.2820	5.0055	0.3554	2.5446	2.1844	0.6209	9.1152	1.4968
				0.0870	0.2770	4.9210	0.3639	2.5320	2.1667	0.6291	9.1031	1.4924

Table 3.2 Exact and approximate results for the $M/E_2/c$ queue.

	$\rho=0.5$			$\rho=0.8$			$\rho=0.9$			$\rho=0.95$		
c	P_W	EL_q	$cv(EL_q)$	P_W	EL_q	$cv(EL_q)$	P_W	EL_q	$cv(EL_q)$	P_W	EL_q	$cv(EL_q)$
2	0.3378	0.5056	2.8595	0.7152	4.5290	1.5103	0.8551	12.356	1.2382	0.9269	28.455	1.1161
	0.3333	0.4896	2.9194	0.7111	4.4444	1.5317	0.8526	12.230	1.2489	0.9256	28.304	1.1214
3	0.2418	0.3436	3.4586	0.6536	4.0635	1.6206	0.8211	11.763	1.2864	0.9093	27.795	1.1391
	0.2368	0.3333	3.5575	0.6472	3.9814	1.6492	0.8171	11.629	1.2996	0.9070	27.629	1.1455
4	0.1783	0.2439	4.0796	0.6044	3.7015	1.7176	0.7931	11.281	1.3276	0.8945	27.248	1.1586
	0.1739	0.2397	4.2144	0.5964	3.6413	1.7491	0.7878	11.171	1.3409	0.8914	27.104	1.1648
5	0.1341	0.1779	4.7422	0.5630	3.4045	1.8064	0.7689	10.869	1.3643	0.8815	26.772	1.1759
	0.1304	0.1776	4.9151	0.5541	3.3686	1.8394	0.7625	10.791	1.3769	0.8778	26.663	1.1813
8	0.0609	0.0758	7.1037	0.4680	2.7435	2.0455	0.7101	9.8859	1.4586	0.8494	25.607	1.2196
	0.0590	0.0792	7.4160	0.4576	2.7664	2.0812	0.7015	9.9008	1.4681	0.8442	25.608	1.2223
10	0.0373	0.0450	9.0981	0.4198	2.4201	2.1921	0.6783	9.3637	1.5132	0.8317	24.967	1.2444
	0.0361	0.0482	9.5301	0.4092	2.4693	2.2296	0.6687	9.4302	1.5207	0.8256	25.034	1.2455
15	0.0116	0.0134	16.272	0.3293	1.8357	2.5386	0.6138	8.3252	1.6335	0.7946	23.645	1.2979
	0.0113	0.0150	17.137	0.3192	1.9222	2.5823	0.6026	8.4895	1.6368	0.7870	23.851	1.2952
20	0.0038	0.0043	28.288	0.2652	1.4396	2.8749	0.5628	7.5233	1.7396	0.7642	22.575	1.3436
	0.0037	0.0049	29.874	0.2561	1.5406	2.9268	0.5508	7.7553	1.7399	0.7554	22.889	1.3380
25	0.0013	0.0014	48.425	0.2171	1.1530	3.2124	0.5204	6.8697	1.8371	0.7381	21.663	1.3845
	0.0013	0.0017	51.215	0.2091	1.2573	3.2743	0.5079	7.1500	1.8353	0.7284	22.067	1.3763

Table 3.3 Exact and approximate results for the $M/H_2/c$ queue ($C_S^2 = 2.25$).

c	D(C_S^2=0.0)	E_2(C_S^2=0.5)	$E_{1,3}$(C_S^2=0.5)	H_2(C_S^2=2.0)
2	0.7438 (±.0072) 0.6849	0.8836 (±.0058) 0.8543	0.8979 (±.0064) 0.8455	1.065 (±.0077) 1.121
3	0.8074 (±.0073) 0.7308	0.9136 (±.0069) 0.8725	0.9294 (±.0064) 0.8632	1.043 (±.0077) 1.106
4	0.8418 (±.0080) 0.7617	0.9321 (±.0063) 0.8856	0.9502 (±.0046) 0.8760	1.030 (±.0072) 1.096
5	0.8644 (±.0068) 0.7847	0.9474 (±.0051) 0.8959	0.9635 (±.0068) 0.8861	1.021 (±.0064) 1.089
10	0.9303 (±.0040) 0.8522	0.9734 (±.0029) 0.9273	0.9859 (±.0041) 0.9182	1.007 (±.0038) 1.065
15	0.9527 (±.0054) 0.8884	0.9849 (±.0047) 0.9273	0.9921 (±.0041) 0.9370	1.006 (±.0039) 1.051

Table 3.4 The coefficient of variation of the output process.

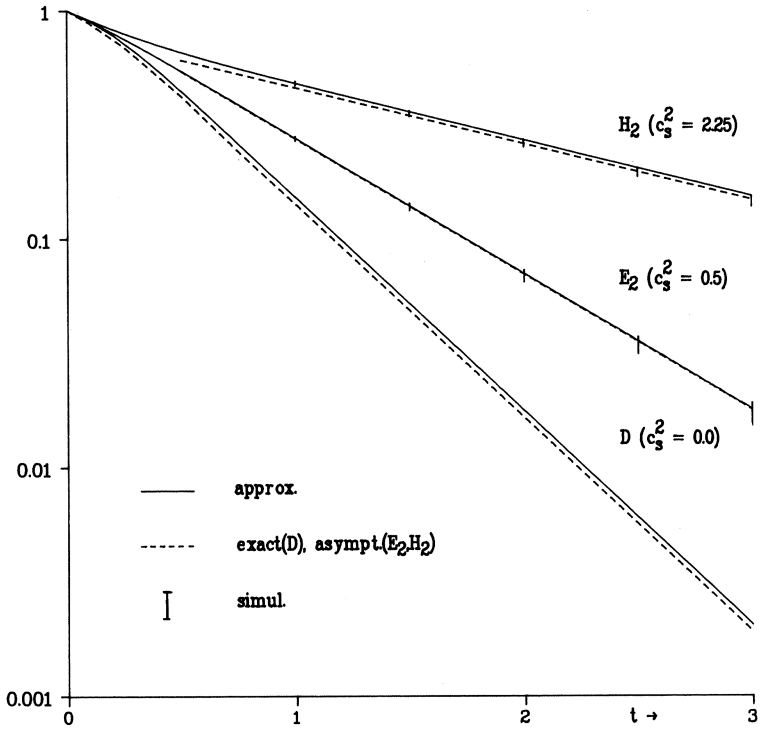

Figure 3. The complementary waiting time distribution $1-V(t)$, $c=5$ $\rho=0.8$.

	t	0.1	0.25	0.5	0.75	1.0	1.5	2.0	3.0
	$\rho=0.5$								
D	Appr.	0.7653	0.4475	0.1311	0.0293	0.0061			
	Exact	0.7366	0.4231	0.1192	0.0215	0.0046			
E_2	Appr.	0.7682	0.4927	0.2160	0.0902	0.0370	0.0062		
	Asym.	0.9668	0.5638	0.2294	0.0934	0.0380	0.0063		
	Sim.	0.77(1)	0.49(1)	0.216(7)	0.092(5)	0.038(4)	-		
$E_{1,3}$	Appr.	0.7730	0.5043	0.2240	0.0928	0.0374	0.0060		
	Asym.	1.060	0.6096	0.2425	0.0965	0.0384	0.0061		
	Sim.	0.76(1)	0.49(1)	0.22(1)	0.096(5)	0.040(3)	-		
H_2	Appr.	0.7834	0.5586	0.3450	0.2308	0.1619	0.0847	0.0454	
	Asym.	0.3351	0.2788	0.2052	0.1511	0.1112	0.0602	0.0326	
	Sim.	0.80(1)	0.58(2)	0.36(1)	0.23(1)	0.15(1)	0.073(7)	0.037(5)	
	$\rho=0.7$								
D	Appr.	0.8511	0.6115	0.2932	0.1277	0.0549	0.0101	0.0019	
	Exact	0.8244	0.5809	0.2722	0.1122	0.0489	0.0090	0.0017	
E_2	Appr.	0.8534	0.6527	0.3988	0.2387	0.1421	0.0502	0.0177	0.0022
	Asym.	0.9213	0.6743	0.4008	0.2383	0.1416	0.0500	0.0177	0.0022
	Sim.	0.85(1)	0.65(1)	0.40(1)	0.24(1)	0.144(7)	0.052(4)	0.019(2)	-
$E_{1,3}$	Appr.	0.8566	0.6616	0.4069	0.2429	0.1438	0.0502	0.0175	0.0021
	Asym.	0.9526	0.6946	0.4102	0.2423	0.1431	0.0499	0.0174	0.0021
	Sim.	0.85(1)	0.65(1)	0.40(1)	0.242(5)	0.143(4)	0.050(3)	0.017(2)	-
H_2	Appr.	0.8639	0.7063	0.5302	0.4153	0.3327	0.2190	0.1456	0.0646
	Asym.	0.6164	0.5457	0.4456	0.3638	0.2970	0.1979	0.1319	0.0586
	Sim.	0.87(1)	0.73(1)	0.55(1)	0.42(1)	0.33(1)	0.21(1)	0.14(1)	0.059(7)
	$\rho=0.9$								
D	Appr.	0.9475	0.8474	0.6673	0.5159	0.3982	0.2372	0.1413	0.0502
	Exact	0.9354	0.8297	0.6498	0.4985	0.3854	0.2297	0.1369	0.0486
E_2	Appr.	0.9484	0.8671	0.7368	0.6231	0.5265	0.3758	0.2682	0.1366
	Asym.	0.9621	0.8695	0.7346	0.6206	0.5243	0.3742	0.2671	0.1360
	Sim.	0.94(1)	0.86(1)	0.73(1)	0.62(1)	0.52(2)	0.37(2)	0.26(1)	0.13(1)
$E_{1,3}$	Appr.	0.9496	0.8709	0.7417	0.6272	0.5297	0.3776	0.2692	0.1368
	Asym.	0.9690	0.8754	0.7391	0.6240	0.5269	0.3756	0.2678	0.1361
	Sim.	0.95(1)	0.87(1)	0.74(1)	0.63(1)	0.53(1)	0.38(1)	0.27(1)	0.14(1)
H_2	Appr.	0.9525	0.8909	0.8096	0.7446	0.6888	0.5927	0.5111	0.3803
	Asym.	0.8893	0.8507	0.7901	0.7339	0.6816	0.5879	0.5071	0.3774
	Sim.	0.95(1)	0.89(1)	0.81(1)	0.74(1)	0.68(2)	0.58(2)	0.49(2)	0.36(2)

Table 3.6 The complementary waiting time distribution $1-V(t)$, $c=5$.

	$P_W<0.1$			$0.1\leqslant P_W<0.7$			$P_W\geqslant 0.7$	
error<	1%	3%	5%	1%	3%	5%	1%	3%
$L_q(Ap.)$	34	85	100	45	90	100	92	100
$L_q(Box)$	75	98	100	100	100	100	100	100
$L_q(Cos)$	47	100	100	79	100	100	100	100
$L_q(Tak)$	70	100	100	89	100	100	100	100
$P_W(Ap.)$	11	70	99	40	92	100	97	100
#Cases		88			230			198

Table 3.7 The quality of various approximations ($C_S^2<1$).

	$P_W<0.1$			$0.1\leqslant P_W<0.7$			$P_W\geqslant 0.7$	
error<	3%	5%	10%	1%	3%	5%	1%	3%
$L_q(Ap.)$	35	45	75	25	75	94	88	100
$L_q(Box)$	95	100	100	38	100	100	100	100
$L_q(Cos)$	0	15	55	0	29	60	73	100
$L_q(Tak)$	85	95	100	19	83	100	92	96
$P_W(Ap.)$	80	100	100	23	90	100	94	100
#Cases		20			48			48

Table 3.8 The quality of various approximations ($C_S^2=1.5625,2.25$).

	$P_W<0.1$			$0.1\leqslant P_W<0.7$			$P_W\geqslant 0.7$		
error<	3%	5%	10%	3%	5%	10%	1%	3%	5%
$L_q(Ap.)$	17	28	28	34	56	70	65	92	98
$L_q(Box)$	39	39	61	54	62	72	56	85	90
$L_q(Cos)$	0	0	0	0	4	24	33	69	90
$L_q(Tak)$	33	50	78	10	32	84	46	83	88
$P_W(Ap.)$	0	33	100	20	58	100	58	96	100
#Cases		18			50			48	

Table 3.9 The quality of various approximations ($C_S^2=4.0,9.0$).

seen, in particular in Figure 3.5, that the approximations are accurate enough for practical purposes and at least as accurate as the results obtained by time consuming computer simulation. The computation time for the approximate results was about 2 seconds CPU time for each example and practically independent of the values of c, ρ and C_S^2. The asymptotic results required per example between 2 and 15 seconds CPU time whereas the simulation of one example with one million customers took on the average 180 seconds CPU time (on a Cyber 175).

In the Tables 3.7, 3.8 and 3.9 we have summarized the conclusions regarding the quality of various approximations for the mean queue length in the M/G/c queue. Beside our approximation $L_q(Ap.)$ we have tested the approximations given in Boxma, Cohen and Huffels[80], in Cosmetatos[76], and in Takahashi[77], denoted by $L_q(Box)$, $L_q(Cos)$ and $L_q(Tak)$ respectively. The latter three approximations are special purpose approximations for EL_q (and EW_q) only; see Appendix C for a brief description. The approximation given in Nozaki and Ross[78], which is also found by Hokstad[78], and the various diffusion approximations (see Halachmi and Franta[78] and Kimura[81]) turned out to be inferior to the above mentioned approximations and will not further be discussed.

The validation is based on the exact results for the 748 different M/G/c systems displayed in Appendix D. We have classified the cases by distinguishing between three ranges for C_S^2 and three traffic levels. The traffic levels reflect light traffic ($P_W<0.1$), moderate traffic ($0.1\leqslant P_W<0.7$) and heavy traffic ($P_W\geqslant 0.7$) situations. We found the classification based on P_W better than the usual one, which depends only on ρ, since it takes into account the effect of the number of servers. Indeed, it makes quite a difference for a customer to enter a 2-server or a 20-server system with the same server utilization. More importantly, the classification being based on P_W, we can aggregate the information on the different systems over the values of c and ρ. A disadvantage is that P_W is not a priori known.

The results in the Tables 3.7, 3.8 and 3.9 are self-explanatory. As overall conclusion we may state that for $C_S^2<1$ all approximations are very accurate. For C_S^2 larger than 1 but not too large $L_q(Box)$ and $L_q(Tak)$ are superior, whereas the approximation $L_q(Cos)$ deteriorates. $L_q(Ap.)$ seems to be the best approximation in heavy traffic when C_S^2 becomes larger.

The practical applicability of $L_q(Ap.)$ and $L_q(Box)$ is increased by replacing the factor γ_1 (cf.(3.28)) appearing in both approximations by $\bar{\gamma}_1$ when $C_S^2<1.5$, where

$$\bar{\gamma}_1=C_S^2\frac{\mathbf{ES}}{c+1}+(1-C_S^2)\frac{\mathbf{ES}}{c}$$

Note that $\bar{\gamma}_1$ is a linear interpolation between the values of γ_1 for deterministic and exponential service times respectively. It was numerically verified that $\bar{\gamma}_1$ approximates very well γ_1 provided $C_S^2<1.5$.

In the Tables 3.10 and 3.11, we demonstrate the sensitivity of the various approximations for service distributions, having the same first two moments but different higher moments. We give results for the mean queue length EL_q. In Table 3.10, we consider hyperexponential service time distributions with the same squared coefficient of variation C_S^2 but different ratios p_1/μ_1. Note that in the extreme case $p_1/\mu_1=1$, the distribution becomes an exponential distribution whereas in the other extreme case $p_1/\mu_1=0$, the distribution has a positive mass in zero. Then, the service time is zero with probability $(C_S^2-1)/(C_S^2+1)$ and exponentially distributed with

probability $2/(C_S^2+1)$. In the latter case, the queueing systems may be interpreted as a batch service model, where the batch size is a truncated geometric distribution. Table 3.10 indicates that the approximation of $L_q(Tak)$ performs best for the skew distributions with $p_1/\mu_1 \to 1$.

	$\rho=0.5$			$\rho=0.9$		
$p_1/\mu_1=$	0.05	0.5	0.95	0.05	0.5	0.95
$C_S^2=1.5625$						
$L_q(Ex.)$	0.1620	0.1525	0.1395	8.7562	8.6737	8.4910
$L_q(Ap.)$	0.1592	0.1520	0.1493	8.7095	8.6338	8.6062
$L_q(Box)$	0.1610	0.1547	0.1522	8.7504	8.7040	8.6847
$L_q(Tak)$	0.1625	0.1550	0.1450	8.7698	8.7116	8.5873
$C_S^2=2.25$						
$L_q(Ex.)$	0.2033	0.1779	0.1435	11.090	10.869	10.311
$L_q(Ap.)$	0.1982	0.1776	0.1719	11.008	10.791	10.731
$L_q(Box)$	0.2011	0.1806	0.1737	11.076	10.910	10.848
$L_q(Tak)$	0.2039	0.1835	0.1560	11.115	10.965	10.587
$C_S^2=4.0$						
$L_q(Ex.)$	0.3099	0.2375	0.1483	17.039	16.400	14.672
$L_q(Ap.)$	0.3004	0.2409	0.2290	16.887	16.261	16.135
$L_q(Box)$	0.3054	0.2336	0.2128	17.010	16.333	16.067
$L_q(Tak)$	0.3104	0.2512	0.1752	17.091	16.679	15.506

Table 3.10 Sensitivity of EL_q for hyperexponential service times with different ratios p_1/μ_1, $c=5$.

	$C_S^2=0.5$		$C_S^2=1.0$		$C_S^2=13/12$	
	$E_{1,3}$	E_2	$E_{1,3}$	M	$E_{1,3}$	H_2
$\rho=0.5$						
$L_q(Ex.)$	0.1052	0.1038	0.1318	0.1304	0.1346	0.1337
$L_q(Ap.)$	0.1085	0.1067	0.1314	0.1304	0.1340	0.1336
$L_q(Box)$	0.1046	0.1036	0.1311	0.1304	0.1345	0.1342
$L_q(Tak)$	0.1042	0.1030	0.1315	0.1304	0.1347	0.1341
$\rho=0.9$						
$L_q(Ex.)$	5.1999	5.1899	6.8745	6.8624	7.1395	7.1319
$L_q(Ap.)$	5.2599	5.2406	6.8730	6.8624	7.1296	7.1254
$L_q(Box)$	5.1900	5.1837	6.8672	6.8624	7.1395	7.1374
$L_q(Tak)$	5.1853	5.1773	6.8714	6.8624	7.1423	7.1368

Table 3.11 Sensitivity of EL_q for the shape of the service time distribution, $c=5$.

In Table 3.11 we demonstrate the insensitivity of EL_q for the shape of the service time distribution for a few cases with $C_S^2 = 0.5, 1.0, 13/12$. The pairs of service time distributions in this table have the same first two moments but a different shape.

4. THE M/G/C QUEUE WITH STATE DEPENDENT ARRIVAL RATE

In this chapter, we give an extension of the results of Chapter 3. Using the same approximation assumption as in Chapter 3, we derive approximations for the steady state probabilities in the M/G/c queue with a state dependent Poisson arrival process. The arrival rate is λ_j when j customers are in the system, and we assume that $\limsup_{n\to\infty}\lambda_n \mathbf{E}S / c < 1$. It is no restriction to assume an infinite waiting capacity, i.e. each arriving customers actually enters the system.

After having derived an algorithm for the steady state probabilities, we discuss two important special models, namely the finite capacity M/G/c queue and the machine repair model with multiple repairmen. The validation of the approximations in this chapter has not yet been completed and will be reported later.

4.1. The basic theorem

For the analysis of the M/G/c queue with state dependent arrival rate, we adopt the same notation and definitions as in Section 3.1. Notably, we make the same approximation assumption. As generalization of Theorem 3.4, we formulate

Theorem 4.1

$$p_n = \frac{(\mathbf{E}S)^n}{n!} p_0 \prod_{j=0}^{n-1} \lambda_j \ , \quad 0 \leqslant n \leqslant c-1 \tag{4.1}$$

$$p_n = \lambda_{c-1} p_{c-1} \alpha_{cn} + \sum_{j=c}^{n} \lambda_j p_j \beta_{jn} \ , \quad n \geqslant c \tag{4.2}$$

$$\pi_n = q_n = \lambda_n p_n \Big/ \sum_{k=0}^{\infty} \lambda_k p_k \ , \quad n \geqslant 0 \tag{4.3}$$

where

$$\alpha_{cn} = \int_0^{\infty} (1-F_c(t))^{c-1} (1-F(t)) M_{cn}(t)\,\mathrm{d}t \ , \quad n \geqslant c$$

$$\beta_{jn} = \int_0^{\infty} (1-F(ct)) M_{jn}(t)\,\mathrm{d}t \ , \quad c \leqslant j \leqslant n$$

$M_{jn}(t) = \Pr\{\ n-j$ customers arrive in $(0,t)$ | there are j customers present at epoch 0 $\}$, $c \leqslant j \leqslant n$

Proof To prove the theorem, we duplicate the whole analysis in the Sections 3.1 and 3.2, except that we replace λ by the appropriate λ_j, notably in (3.4), (3.5) and (3.15). For the functions $M_{jn}(t)$, $c \leqslant j \leqslant n$, we cannot obtain explicit expressions in general and also for p_0 a formula cannot be found. Equation (4.3) is the analogon of (2.5) for the multiserver case.

□

Remark 4.2 The quotients p_n / p_0, $0 \leqslant n \leqslant c-1$ are equal to the corresponding quotients in the M/M/c queue with state dependent arrival rate λ_j.

□

4.2. The M/G/c/ queue with finite capacity

In this section, we consider the M/G/c queue with finite capacity K. As we have argued in Section 2.2, the distributions (p_n) and (q_n) in the M/G/c queue with capacity K are identical to the corresponding distributions in the M/G/c queue with state dependent arrival rate λ_j, where $\lambda_j = \lambda$, $j = 0, \ldots, K-1$ and $\lambda_j = 0$, $j \geqslant K$. For the latter model, we can simplify the results of Theorem 4.1 considerably. Equation (4.1) reduces to

$$p_n = \frac{(\lambda \mathrm{ES})^n}{n!} p_0, \quad 0 \leqslant n \leqslant c-1 \tag{4.4}$$

and $M_{jn}(t)$ is explicitly given as

$$M_{jn}(t) = e^{-\lambda t} \frac{(\lambda t)^{n-j}}{n-j!}, \quad c \leqslant j \leqslant n \leqslant K-1$$

$$M_{jK}(t) = \sum_{n=K}^{\infty} e^{-\lambda t} \frac{(\lambda t)^{n-j}}{n-j!}, \quad c \leqslant j \leqslant K$$

Hence, $\alpha_{cn} = \alpha_{n-c}$, $\beta_{jn} = \beta_{n-j}$, $c \leqslant j \leqslant n \leqslant K-1$ where α_k and β_k are defined by (3.8) and (3.9) respectively. Also, we have

$$\alpha_{cK} = \frac{\mathrm{ES}}{c} - \sum_{n=c}^{K-1} \alpha_{cn} \text{ and } \beta_{jK} = \frac{\mathrm{ES}}{c} - \sum_{n=j}^{K-1} \beta_{jn}$$

Using the above in (4.2) with $n = K$ and also using (4.2) for $n = c, \ldots, K-1$, we find after some algebra

$$p_K = \rho p_{c-1} - (1-\rho) \sum_{j=c}^{K-1} p_j \tag{4.5}$$

Hence, by using Equation (4.1) for $n = 0, \ldots, c-1$, Equation (4.2) for $n = c, \ldots, K-1$ and Equation (4.5), we can compute recursively the quotients p_n / p_0, $0 \leqslant n \leqslant K$ and next we find p_0 by noting that $\sum_{n=0}^{K} p_n / p_0 = 1 / p_0$.

For the case $\rho = \lambda \mathrm{ES} / c < 1$, it is straightforward to derive a relationship between the steady state distributions $(p_n^{(K)})$ and $(p_n^{(\infty)})$ in the M/G/c queue with finite and infinite capacity respectively. In analogy with the analysis in Section 2.2, we find

$$p_n^{(K)} = c_K p_n^{(\infty)}, \quad 0 \leqslant n \leqslant K-1 \tag{4.6}$$

where $c_K = 1 / \{1 - \rho + \rho \sum_{j=0}^{K-1} p_j^{(\infty)}\}$ and

$$p_K^{(K)} = c_K \{p_{c-1}^{(\infty)} - (1-\rho) \sum_{j=c-1}^{K-1} p_j^{(\infty)}\} \tag{4.7}$$

Note that the formulae for c_K in the single server case and the multiserver case are identical.

Letting the random variable $\mathbf{W}_q$ be the waiting time in the queue for an arbitrary arriving customer, we find with the same arguments as used in Section 3.4

$$q_n = \Pr\{\mathbf{W}_q > 0 \text{ and } n-c \text{ arrivals during } \mathbf{W}_q\}, \quad c \leqslant n \leqslant K-1$$

Hence, we have

$$q_n = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^{n-c}}{n-c!} \mathrm{dPr}\{0<\mathbf{W}_q \leqslant t\}\ ,\ c \leqslant n \leqslant K-2 \tag{4.8}$$

$$q_{K-1} = \int_0^\infty \sum_{n=K-1}^{\infty} e^{-\lambda t} \frac{(\lambda t)^{n-c}}{n-c!} \mathrm{dPr}\{0<\mathbf{W}_q \leqslant t\} \tag{4.9}$$

In (4.9), the sum accounts for the blocked customers. The Equations (4.8) and (4.9) define the waiting time distribution implicitly once the distribution (q_n) is known. However, by prescribing a representation for $\Pr\{0<\mathbf{W}_q \leqslant t\}$ containing several degrees of freedom, we can find an approximation for the waiting time distribution function. We have not yet done this, but it will be the subject of further research.

4.3. The machine repair model with multiple repairmen

In this section, we discuss the machine repair model with K machines and c repairmen. For a description of this model, we refer to Section 2.3.

Let p_n be the steady state probability that at an arbitrary epoch n machines are broken down, $0 \leqslant n \leqslant K$. Then, the sequence (λ_j) is specified by $\lambda_j = (K-j)\lambda$, $0 \leqslant j \leqslant K-1$ and $\lambda_j = 0$, $j \geqslant K$. Theorem 4.1 can now be simplified to

Theorem 4.3

$$p_n = \binom{K}{n} (\lambda \mathbf{ES})^n p_0\ ,\ 0 \leqslant n \leqslant c-1$$

$$p_n = (K-c+1)\lambda p_{c-1}\alpha_{cn} + \sum_{j=c}^{n} (K-j)\lambda p_j \beta_{jn}\ ,\ c \leqslant j \leqslant K$$

where α_{cn} and β_{jn} are defined as in Theorem 4.1. Now, $M_{jn}(t)$ is explicitly given as

$$M_{jn}(t) = \binom{K-j}{n-j} \left(1-e^{-\lambda t}\right)^{n-j} e^{-\lambda t(K-n)}$$

Proof The theorem follows directly by noting that $M_{jn}(t)$ is the binomial probability that in the interval $(0,t)$ $n-j$ machines fail, given that $K-j$ machines are working at epoch 0.

□

5. THE $M^X / G / 1$ QUEUE.

In this chapter we give an analysis for the steady state probabilities in a general class of single server queues with batch arrivals. See also van Hoorn[81].

In many practical queueing situations customers arrive in batches rather than singly. For example, at airports or train stations, passengers often arrive in groups at the check-in counters.

In communication networks the batch arrival model is also important. For example, messages sent through a data network are not handled as a whole, but only piecewise. At the entrance of the network, a message is split up into unit packets of a fixed number of bits and the packets are switched separately. Thus, the service time or processing time of a unit packet is likely to be deterministic. Applications of this kind can be found in Manfield and Tran Gia[81] and Manfield and Tran Gia[82], where the $GI^X/M/1$ queue and the $M^X/M/1$ queue are investigated respectively.

Batch arrival queues are not only interesting in their own right, but are also very useful to model practical situations which are difficult to solve analytically or allow of no analytical treatment at all. We demonstrate this on the M/G/1 queue with bounded sojourn time and a queueing system in which the service time of customers depends on their waiting time. Incidentally, the model defined in Chapter 2 can also be seen as a special case of the batch arrival model where a batch consists of 0 or 1 customer.

In Section 5.4 we define a batch arrival model that is a discrete version of the M/G/1 queue with bounded sojourn time. The latter model is a M/G/1-like queueing system where a customer is only accepted if the sum of his estimated waiting time $\mathbf{W}_q$ in the queue and his service time $\mathbf{S}$ is less than a number K. Otherwise the customer is not admitted to the system. The analytical solution for the waiting time distribution in this system is given as an integro-differential equation, which can be solved numerically by discretization. Hence, there are two ways to solve the M/G/1 queue with bounded sojourn time, namely to discretize the model, as we suggest in Section 5.4, or to discretize the exact solution of the model.

In Section 5.5, we use the batch arrival model to describe a number of peculiarities of customer behaviour in telephone switching systems. It appears that a queueing system where the service time of a customer depends on his waiting time in the queue, is appropriate to model 'impatience characteristics' of customers. In this application, an analytical approach is hardly possible.

Further, the chapter is organized as follows. In Section 5.1, we define the general model and give its solution using the regenerative method. In Section 5.2. we focus on some special choices for the service time distribution, whereas in Section 5.3 we treat the case of a uniform batch size distribution, i.e. here the standard $M^X/G/1$ queue is analyzed.

5.1. The model and the regenerative analysis

The model

We consider a single server queueing system at which batches of customers arrive according to a Poisson process at rate λ. The number of customers in a batch is distributed as a random variable $\mathbf{G}^{(j)}$ when j customers are in the system. We allow that $\Pr\{\mathbf{G}^{(j)}=0\}>0$ and hence it is no restriction to assume that the queueing system has an infinite capacity. We can model various finite capacity queueing situations by taking $\Pr\{\mathbf{G}^{(j)}=0\}=1$, $j\geqslant K$ for some K. The customers of a batch are served individually and have a service time $\mathbf{S}$ with a general probability distribution function $F(t)=\Pr\{\mathbf{S}\leqslant t\}$. To have a stable system we assume that $\limsup_{n\to\infty} \lambda \mathbf{E}\mathbf{G}^{(n)}\mathbf{E}\mathbf{S}<1$.

It is no restriction to assume that all customers in a batch actually enter the system. Indeed, for the analysis of the system 'server + queue' only the entering customers are relevant since they interact with the system. In practice, this assumption is usually not fulfilled and it may occur that batches are partially or totally blocked. Then, it is of interest to know the blocking probability. However, by a proper choice of the distribution of the batch size $\mathbf{G}^{(j)}$, it is always possible to define a modified model having the same characteristics as the original model restricted to the entering customers. After having analyzed the modified model, we can find the characteristics of the original model using the results of the modified model. We illustrate this with the following example.

Example 5.1.

Let model 1 be a single server queueing model with batch arrivals having a finite capacity K. The arrival rate is λ and the batch size is distributed as a random variable $\mathbf{H}$. When upon arrival there are not enough waiting places for all customers in a batch, the whole batch is not admitted to the system. We define model 2, the modification of model 1, as follows. Model 2 has the same service process as model 1, a uniform arrival rate λ of batches of customers and a state dependent batch size $\mathbf{G}^{(j)}$ when j customers are in the system, where (see Figure 5.1)

$$\Pr\{\mathbf{G}^{(j)}=k\}=\Pr\{\mathbf{H}=k\}\,,\quad 0\leqslant j\leqslant K-1\,,\quad 1\leqslant k\leqslant K-j$$

$$\Pr\{\mathbf{G}^{(j)}=0\}=\Pr\{\mathbf{H}=0\}+\Pr\{\mathbf{H}>K-j\}\,,\quad 0\leqslant j\leqslant K-1$$

$$\Pr\{\mathbf{G}^{(j)}=0\}=1\,,\quad j\geqslant K$$

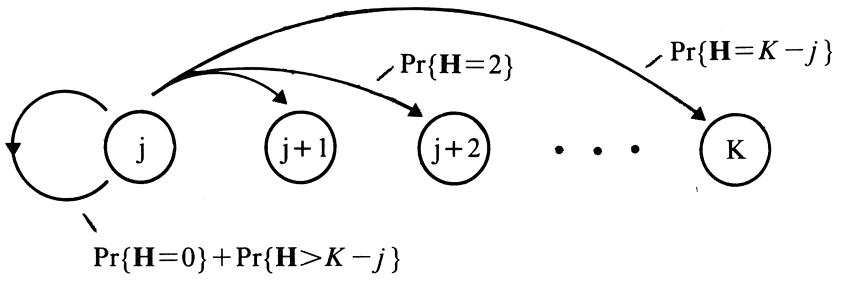

Figure 5.1 In model 2 the batch size is state dependent.

The analysis

The regenerative method is also very suitable to derive an algorithm for the steady state probabilities in a batch arrival queueing system with a Markovian arrival process. For clarity of presentation, we repeat the necessary definitions. For $n \geqslant 0$ we define

p_n = the steady state probability that n customers are in the system at an arbitrary epoch

q_n = the steady state probability that at an arbitrary service completion epoch n customers are left behind in the system by the customer just served.

We assume that at epoch 0 the system has become empty after a service completion and define the random variables

$\mathbf{T}$ = the next time the system becomes empty

$\mathbf{T}_n$ = the amount of time in (0,**T**] that n customers are present, $n \geqslant 0$

$\mathbf{N}$ = the number of customers served in (0,**T**]

$\mathbf{N}_n$ = the number of service completion epochs in (0,**T**] at which n customers are left behind by the customer just served

Finally, let for $0 \leqslant j \leqslant n$

A_{jn} = the expected amount of time during which n customers are in the system until the next service completion epoch, given that at epoch 0 a service is completed with j customers left behind in the system

The following theorem supplies the basic relations for the model in its general form.

Theorem 5.3

$$p_n = \mathrm{ET}_n / \mathrm{ET}\,, \quad q_n = \mathrm{EN}_n / \mathrm{EN}\,, \quad n \geqslant 0 \tag{5.1}$$

$$\mathrm{ET}_n = \sum_{j=0}^{n} \mathrm{EN}_j A_{jn}\,, \quad n \geqslant 1 \tag{5.2}$$

$$\mathrm{EN}_n = \lambda \sum_{k=0}^{n} \mathrm{ET}_k \Pr\{\mathbf{G}^{(k)} \geqslant n+1-k\}\,, \quad n \geqslant 0 \tag{5.3}$$

$$\mathrm{ET} - \mathrm{ET}_0 = \mathrm{ENES} \tag{5.4}$$

$$\mathrm{EN} = \lambda \sum_{k=0}^{\infty} \mathrm{ET}_k \mathrm{EG}^{(k)} \tag{5.5}$$

Proof Equation (5.1) is as before a general result and follows from the theory of the regenerative processes. Relation (5.2) is obtained by partitioning the busy cycle (0,**T**] by means of the service completion epochs and (5.4) is an immediate consequence of it.

To prove (5.3) we resort again on an up and down crossing property of the queue length process. Consider the macro states $M_1 = \{0,1,\ldots,n\}$ and $M_2 = \{n+1, n+2, \ldots\}$. Then, in a busy cycle the average number of transitions from M_1 to M_2 is equal to the average number of transitions from M_2 to M_1; see Figure 5.2. The latter number is simply equal to EN_n, since transitions from M_2 to M_1 occur at departure epochs when the system is left behind in state n.

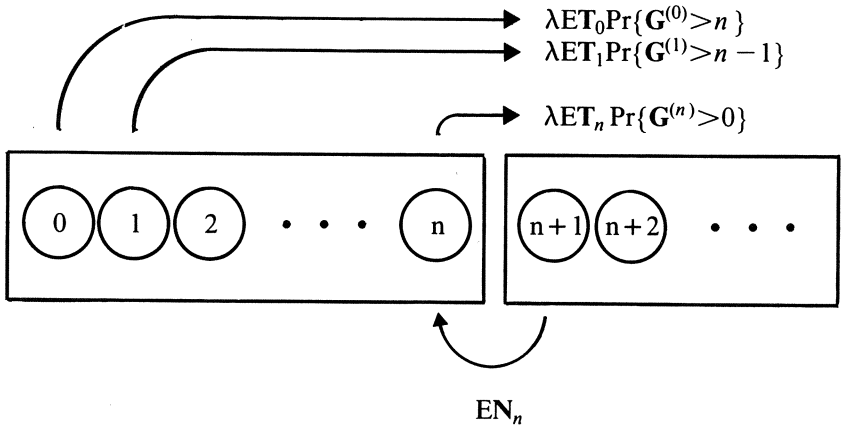

Figure 5.2 Up and downcrossings of level n.

Direct transitions from M_1 to M_2 occur when a batch arrives, seeing n or less customers in the system, that is large enough to bring the number of customers in the system above n. Let $0 \leq k \leq n$, then according to the Poisson Lemma , λET_k is the average number of batches arriving while the system is in state k in (0,T] and $\lambda ET_k \Pr\{G^{(k)} \geq n+1-k\}$ is the average number of batches generating a transition from state k to M_2 in (0,T]. By conditioning on k, we have

the average number of transitions from M_1 to M_2 in (0,T]

$$= \lambda ET_0 \Pr\{G^{(0)} \geq n+1\} + \cdots + \lambda ET_k \Pr\{G^{(k)} \geq n+1-k\} + \cdots + \lambda ET_n \Pr\{G^{(n)} \geq 1\}$$

Note that in this derivation we use the assumption that any customer is admitted to the system. This completes the proof of (5.3) and (5.5) follows by summing (5.3) over $n \geq 0$.

□

Remark 5.4 From Theorem 5.1 it is clear that the distributions (p_n) and (q_n) are not equal in general.

□

Remark 5.5 For practical purposes it is of interest to know the distribution of the number of customers in the system at an arrival epoch. Since we allow that $\Pr\{G^{(j)}=0\}>0$, we distinguish between zero and non-zero batches. We are primarily interested in non-zero batches, since these batches typically interact with the system. Therefore, we define for $n \geq 0$

π_n = the steady state probability that an arbitrary non-zero batch sees upon arrival n customers in the system

π_n^* = the steady state probability that an arbitrary batch sees upon arrival n customers in the system.

Using the property 'Poisson arrivals see time averages' (cf. Wolff[81] or Appendix A) we have $\pi_n^* = p_n$, and

$$\pi_n = p_n \Pr\{\mathbf{G}^{(n)} \geqslant 1\} \Big/ \sum_{k=0}^{\infty} p_k \Pr\{\mathbf{G}^{(k)} \geqslant 1\} \tag{5.6}$$

Note that $(\pi_n) \neq (q_n)$ as is intuitively clear since customers arrive in batches and leave singly.

□

To compute the distributions (p_n) and (q_n) we formulate the

Algorithm 5.6

1. Evaluate the constants A_{jn}, $0 \leqslant j \leqslant n$
2. Put $\mathbf{EN}_0 = 1$ and $\mathbf{ET}_0 = 1 / \lambda \Pr\{\mathbf{G}^{(0)} \geqslant 1\}$
3. Assume that $\mathbf{EN}_0, \ldots, \mathbf{EN}_{n-1}$, $\mathbf{ET}_0, \ldots, \mathbf{ET}_{n-1}$ have been computed, solve for $\mathbf{EN}_n$ and $\mathbf{ET}_n$

$$\mathbf{ET}_n = \mathbf{EN}_n A_{nn} + \sum_{j=0}^{n-1} \mathbf{EN}_j A_{jn}$$

$$\mathbf{EN}_n = \lambda \mathbf{ET}_n \Pr\{\mathbf{G}^{(n)} \geqslant 1\} + \lambda \sum_{k=0}^{n-1} \mathbf{ET}_k \Pr\{\mathbf{G}^{(k)} \geqslant n+1-k\}$$

4. Return to step 3 if necessary
5. Normalize $\mathbf{ET}_n$ by $\mathbf{ET} = \sum_{k=0}^{\infty} \mathbf{ET}_k$ to find p_n and normalize $\mathbf{EN} = \sum_{k=0}^{\infty} \mathbf{EN}_k$ to obtain q_n.

The complexity of step 1 in the algorithm depends on the distribution of the batch size $\mathbf{G}^{(j)}$, $j \geqslant 0$ and on the distribution $F(t)$ of the service time $\mathbf{S}$. Step 5 can be simplified if the batch size distribution is independent of the state of the system. In the next section, we treat a number of important cases and give schemes to compute the quantities A_{jn} .

For the constants A_{jn} the following integral representation holds.

Lemma 5.7

$$A_{jn} = \int_0^{\infty} (1 - F(t)) a_{jn}(t)\, \mathrm{d}t \ , \quad 1 \leqslant j \leqslant n \tag{5.7}$$

$$A_{0n} = \sum_{k=1}^{n} \frac{\Pr\{\mathbf{G}^{(0)} = k\}}{\Pr\{\mathbf{G}^{(0)} \geqslant 1\}} A_{kn} \ , \quad n \geqslant 1 \tag{5.8}$$

where $a_{jn}(t) = \Pr\{\ n - j$ customers arrive in $(0,t) \mid j$ customers present at epoch 0 $\}$

Proof The proof of (5.7) goes analogously to the proof of Lemma 1.4. The second part of the lemma follows by noting that $\Pr\{\mathbf{G}^{(0)} = k\} / \Pr\{\mathbf{G}^{(0)} \geqslant 1\}$ is the probability that the first non-zero batch entering the system in $(0,\mathbf{T}]$ consists of k customers and hence is the probability that the first service in $(0,\mathbf{T}]$ starts with k customers present.

□

5.2. Algorithms for the quantities A_{jn}

In this section, we pay some attention to the evaluation of the constants A_{jn}, the difficult step in Algorithm 5.6. To show in which way to extend the algorithms of Section 2.4 to batch arrival processes, we consider the case of an exponential service time distribution. Next it will be obvious how to deal with other phase type service distributions. Finally, for a general service distribution, we discuss a representation of the numbers A_{jn} which could be used for numerical purposes in some cases.

Case 1 $F(t)=1-e^{-\mu t}$

By exploiting the memoryless property of the exponential distribution and the property that $min(\mathbf{X}_1,\mathbf{X}_2)$ has an exponential distribution with mean $1/(\mu_1+\mu_2)$ if $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent, exponential random variables with mean values $1/\mu_1$ and $1/\mu_2$, we obtain a recursive scheme for the A_{jn}. Put for abbreviation

$$g_k^{(j)}=\Pr\{\mathbf{G}^{(j)}=k\} \text{ for all } j,k$$

Then, for any fixed $n \geqslant 1$

$$A_{jn}=\frac{\lambda}{\lambda+\mu}\sum_{k=0}^{n-j} g_k^{(j)}A_{j+k,n}\ ,\quad 1\leqslant j<n$$

$$A_{nn}=\frac{1}{\lambda+\mu}+\frac{\lambda\, g_0^{(n)}}{\lambda+\mu}A_{nn}$$

Here $\lambda g_k^{(j)}/(\lambda+\mu)$ is the joint probability that a batch arrives before the completion of the service and that the size of this batch is k given j. After some rewriting, we get

$$A_{jn}=\frac{\lambda}{\lambda(1-g_0^{(j)})+\mu}\sum_{k=1}^{n-j} g_k^{(j)}A_{j+k,n}\ ,\quad 1\leqslant j<n$$

$$A_{nn}=\frac{1}{\lambda(1-g_0^{(n)})+\mu}.$$

From these relations we can recursively compute A_{jn} for $j=n,\ldots,1$.

Case 2 General service time distribution

Though we handle here the case of a general service distribution, we do not give recommendations how or when to use the representation below for the A_{jn}. Nevertheless, from a theoretical point of view the derivation is interesting.

The integral representation (5.7) suggests to focus on the arrival process to get an expression for $a_{jn}(t)$. Consider a Markov chain with countable state space $\{0,1,2,\ldots\}$, where state n denotes the number of customers present, and with transition matrix T, given by

$$T=\begin{pmatrix} g_0^{(0)} & g_1^{(0)} & g_2^{(0)} & g_3^{(0)} & \cdots \\ & g_0^{(1)} & g_1^{(1)} & g_2^{(1)} & \cdots \\ & & g_0^{(2)} & g_1^{(2)} & \cdots \\ & & & \cdot & \cdots \\ & & & & \cdot \end{pmatrix}$$

Then, by conditioning on the number of arriving batches in $(0,t)$, $a_{jn}(t)$ is given by the (j,n)-element of the matrix

$$[a](t)=\sum_{k=0}^{\infty} e^{-\lambda t}\frac{(\lambda t)^k}{k!}T^k \tag{5.9}$$

Next, after integration of (5.9), A_{jn} is given by the (j,n)-element of the matrix

$$[A]=\sum_{k=0}^{\infty}\int_0^{\infty}(1-F(t))e^{-\lambda t}\frac{(\lambda t)^k}{k!}\mathrm{d}tT^k \tag{5.10}$$

The coefficients of the powers of T in (5.10) are identical to the numbers α_k in the M/G/1 queue; cf. (1.7). As said before, the value of (5.10) for computational purposes is dubious. The representation (5.10) can be used for numerical purposes if the matrix T is small or if T is diagonizable. The numerical results in Section 5.4 have been obtained by using (5.10). We omit further details.

5.3. The $M^X/G/1$ queue with uniform batch size.

In this section, we give an analytical treatment of the standard $M^X/G/1$ queue with uniform batch size. Let the random variable **G** denote the batch size and let $g_k=\Pr\{\mathbf{G}=k\}$, $k\geqslant 0$. The capacity of the system is infinite. It is no restriction to assume that $g_0=0$, otherwise we take $\lambda^*=\lambda(1-g_0)$ and $g_k^*=g_k/(1-g_0)$. The assumption of a uniform batch size permits a number of simplifications in the algorithm. The numbers A_{jn} depend now only on the difference $n-j$, $1\leqslant j\leqslant n$. Also from (5.5) it follows that $\mathbf{EN}=\lambda\mathbf{EGET}$ and hence by (5.4) $p_0=1-\lambda\mathbf{ESEG}$. Define for $k\geqslant 0$

$$\alpha_k=\int_0^{\infty}(1-F(t))\Pr\{\ k \text{ customers arrive in } (0,t)\}\,\mathrm{d}t$$

Also, define for $|z|\leqslant 1$ the generating functions $P(z)=\sum_{n=0}^{\infty}p_nz^n$, $Q(z)=\sum_{n=0}^{\infty}q_nz^n$, $G(z)=\sum_{n=1}^{\infty}g_nz^n$, and $\alpha(z)=\sum_{n=0}^{\infty}\alpha_nz^n$. We have

$$A_{jn}=\alpha_{n-j}\ ,\quad 1\leqslant j\leqslant n$$

and

$$\alpha(z)=\int_0^{\infty}(1-F(t))e^{-\lambda t(1-G(z))}\,\mathrm{d}t$$

Note that $e^{-\lambda t(1-G(z))}$ is the generating function of the number of arrivals in $(0,t)$ according to the compound Poisson arrival process.

By taking generating functions in (5.2) and (5.3) and by noting that $\mathbf{ET}_n=p_n\mathbf{ET}$ and $\mathbf{EN}_n=q_n\lambda\mathbf{EGET}$, we easily get after some algebra

$$P(z)-p_0=\lambda\mathbf{EG}q_0G(z)\alpha(z)+(Q(z)-q_0)\lambda\mathbf{EG}\alpha(z)$$

and

$$\lambda\mathbf{EG}Q(z)=\lambda P(z)\frac{1-G(z)}{1-z}$$

Finally, with $\rho=\lambda\mathbf{ESEG}$, we find for $P(z)$

$$P(z)=(1-\rho)\frac{1-\lambda\alpha(z)(1-G(z))}{1-\lambda\alpha(z)(1-G(z))/(1-z)} \tag{5.11}$$

From (5.11) we obtain the following well known expansion for the mean queue size $\mathbf{EL}_q$.

$$\mathbf{EL}_q = \frac{\rho^2}{1-\rho}\,\frac{1+C_S^2}{2} + \frac{\rho}{2(1-\rho)}\left(\frac{\mathbf{EG}^2}{\mathbf{EG}} - 1\right)$$

The formula for $\mathbf{EL}_q$ consists of two parts, the first of which is equal to the mean queue length in a M/G/1 queue with traffic intensity ρ. The second part reflects the additional effects of variability of the batch size to the mean queue length; cf. also Cosmetatos[78].

The computation of the numbers α_j, $j \geqslant 0$, is essentially identical to the computation of A_{jn} in Section 5.2. For the important special case of constant service time D, we prove a very simple computational scheme for the α_j. Define for $j \geqslant 0$

$$a_j(t) = \Pr\{\ j \text{ customers arrive in } (0,t)\}$$

Then

$$\alpha_j = \int_0^D a_j(t)\,\mathrm{d}t \tag{5.12}$$

In Adelson[66] a simple recurrence relation for $a_j(t)$ is derived.

$$a_j(t) = \frac{\lambda t}{j}\sum_{i=1}^{j} i\, g_i\, a_{j-i}(t)\ ,\quad j \geqslant 1 \tag{5.13}$$

$$a_0(t) = e^{-\lambda t}$$

Also, by conditioning on the first arrival in $(0,t)$ we get for $a_j(t)$ the identity

$$a_j(t) = \int_0^t \sum_{i=1}^{j} g_i \lambda\, e^{-\lambda(t-u)} a_{j-i}(u)\,\mathrm{d}u \tag{5.14}$$

After integration of 5.14 and using 5.12 and 5.13, we obtain

$$\alpha_j = \sum_{i=0}^{j-1} \alpha_i g_{j-i} - \frac{1}{\lambda} a_j(D) \tag{5.15}$$

$$\alpha_0 = \frac{1}{\lambda}(1 - e^{-\lambda D})$$

The numbers α_j are computed very efficiently by applying first 5.13 to compute $a_j(D)$ and next 5.15. However, the scheme 5.15 need in general not be numerically stable, since loss of accuracy can occur by taking successive differences. To circumvent this difficulty, note that 5.15 is in fact a discrete renewal equation. By applying the discrete renewal theorem (cf. Appendix A) it easily follows that

$$\alpha_j = \frac{1}{\lambda} m_j - \frac{1}{\lambda}\sum_{i=0}^{j} m_i\, a_{j-i}(D) \tag{5.16}$$

where the renewal quantity m_j is recursively computed by

$$m_j = \sum_{k=1}^{j} g_k\, m_{j-k}\ ,\quad j \geqslant 1 \text{ with } m_0 = 1$$

In relation 5.16 the number α_j is expressed as the difference of two expressions which are both computed in a stable way.

5.4. The M/G/1 queue with bounded sojourn time

In this section, we present a method to approximate the waiting time distribution in the M/G/1 queue with bounded sojourn time. For that purpose we use the batch arrival model analyzed in this chapter with a suitable choice of the batch size distribution.

The M/G/1 queue with bounded sojourn time is a queueing model of M/G/1 type where arriving customers are refused if their waiting time plus service time exceeds a fixed amount K. Let $\mathbf{W}_q$ be the waiting time which an arbitrary customer faces upon arrival and let $\mathbf{S}$ be the service time which this customer would like to receive. Then this customer is admitted to the system if $\mathbf{W}_q + \mathbf{S} \leqslant K$ and is rejected if $\mathbf{W}_q + \mathbf{S} > K$. In a telecommunication application, the model describes a storage buffer emptying at constant rate and receiving messages from a high-speed data channel. A message is rejected if its length plus the length of all messages in storage would overflow the buffer capacity K.

The M/G/1 queue with bounded sojourn time has been studied by Cohen[69] and Gavish and Schweitzer[77], who arrived at an analytical solution for the waiting time distribution in the form of an integro-differential equation. An explicit solution of this equation can be given for the case of an exponential service time distribution. For a general service distribution the equation can only be solved numerically.

Our approach is to approximate the M/G/1 queue with bounded sojourn time by a finite capacity $M^X/D/1$ queue, where the whole batch is not admitted to the system if there are not sufficient waiting places for all members of the batch. Therefore, we approximate the service time distribution $F(t)$ by a discrete distribution which is characterized by the probabilities (g_i) where

$$g_i = F(iD) - F((i-1)D), \quad i \geqslant 1$$

and D is the grid size of the discretization. Then, with probability g_i an arriving customer brings i unit packets of work into the system, where each packet has a fixed length D. A customer is admitted to the system if his number of packets plus the number of packets he sees in the system upon arrival does not exceed N with $N = K/D$. Note that the approximate model for the M/G/1 queue with bounded sojourn time now satisfies the conditions in Example 5.1 in Section 5.1. By following the same procedure as in Example 5.1 we fit the above model in the framework of the model discussed in Section 5.1.

After having computed the distributions (p_n) and (π_n) in the modified model (cf. Relation (5.6)), we can easily compute some interesting performance measures for the original model. Note that (π_n) gives an approximation for the actual waiting time distribution of an entering customer in the original system. π_n is an approximation for the probability that an entering customer has to wait an amount of time between $(n-1)D$ and nD. Similarly, (p_n) gives an approximation for the virtual waiting time distribution. Define for the original system

P_{idle} = the fraction of time the server is idle

P_{join} = the probability that an arbitrary customer enters the system

E_{wait} = the expected waiting time of an entering customer (excluding service time).

These performance measures are approximated by

$$P_{idle} = p_0$$

$$P_{join} = \sum_{k=0}^{N-1} \Pr\{G^{(k)} > 0\} p_k$$

$$E_{wait} = \sum_{k=1}^{N-1} \pi_k (k - \tfrac{1}{2}) D$$

We have computed the approximations using Equation (5.10). In Table 5.3 we compare for a number of values of the buffer capacity K and the offered load $\rho = \lambda ES$ the numerical results of our method to the exact analytical results according to Gavish and Schweitzer[77] for several values of D, the grid size of the discretization of the exponential service time distribution. The mean service time is taken equal to 1. It turns out that for rather crude discretizations the approximations are close to the exact results.

In Table 5.4 we demonstrate that the quantities P_{idle}, P_{join} and E_{wait} are very sensitive for the distribution of the service time. For several values of K and ρ we consider Erlang-3, exponential and hyperexponential ($C_S^2 = 3$) service time distributions. In the results in Table 5.4 we have taken $N = 40$. This sensitivity is intuitively clear by noting that , roughly said, customers with a short service requirement are more likely to enter the system than customers with a longer service requirement. When the service time distribution is hyperexponential, the majority of the customers has a very short service time, while in the case of an Erlang service time distribution the services are longer and more concentrated around the mean service time.

	P_{idle}			P_{join}			E_{wait}		
$\rho =$	0.8	1.0	4.0	0.8	1.0	4.0	0.8	1.0	4.0
$K = 0.6$									
$N = 20$	0.9030	0.8804	0.6098	0.4379	0.4347	0.3938	0.0190	0.0235	0.0820
$N = 40$	0.9047	0.8824	0.6079	0.4387	0.4357	0.3954	0.0186	0.0231	0.0831
exact	0.9066	0.8844	0.6059	0.4396	0.4367	0.3970	0.0182	0.0226	0.0840
$K = 2.0$									
$N = 20$	0.5838	0.5060	0.0445	0.7943	0.7773	0.5780	0.2683	0.3306	0.9793
$N = 40$	0.5884	0.5095	0.0397	0.7977	0.7807	0.5694	0.2675	0.3320	1.024
exact	0.5932	0.5131	0.0347	0.8009	0.7839	0.5594	0.2661	0.3327	1.073
$K = 6.0$									
$N = 20$	0.2705	0.1643	-	0.9580	0.9341	0.6198	1.430	1.941	4.743
$N = 40$	0.2830	0.1717	-	0.9589	0.9341	0.5967	1.429	1.967	4.880
exact	0.2956	0.1790	-	0.9599	0.9340	0.5652	1.427	1.993	5.041

Table 5.3 The performance measures for various grid sizes D $(= K / N)$.

	P_{idle}			P_{join}			E_{wait}		
$\rho=$	0.8	1.0	4.0	0.8	1.0	4.0	0.8	1.0	4.0
$K=0.6$									
Erl.	0.9168	0.8976	0.6671	0.2565	0.2535	0.2164	0.0189	0.0233	0.0781
Exp.	0.9047	0.8824	0.6079	0.4387	0.4357	0.3954	0.0186	0.0231	0.0831
Hyp.	0.8881	0.8620	0.5474	0.5546	0.5508	0.4976	0.0212	0.0264	0.0959
$K=2.0$									
Erl.	0.4652	0.3778	0.0136	0.7926	0.7567	0.4195	0.3857	0.4719	1.154
Exp.	0.5884	0.5095	0.0397	0.7977	0.7807	0.5694	0.2675	0.3320	1.024
Hyp.	0.6375	0.5616	0.0445	0.8547	0.8443	0.6633	0.2103	0.2666	1.013
$K=6.0$									
Erl.	0.2295	0.1123	-	0.9698	0.9289	0.4543	1.549	2.332	5.041
Exp.	0.2830	0.1717	-	0.9589	0.9341	0.5967	1.429	1.967	4.880
Hyp.	0.4138	0.3014	-	0.9570	0.9482	0.6893	0.9246	1.274	4.841

Table 5.4 The sensitivity of the performance measures for the service time distribution ($N=40$).

5.5. Dependency of service time on waiting time in switching systems: a queueing analysis with aspects of overload control†

In this section, we use the batch arrival model defined in this chapter for a case study concerning a telephone switching system. We model certain aspects of customer behaviour in such a system and in particular we give curves for the call completion rate. It turns out that we have developed an approximation for a queueing system where the service time of a customer depends on his waiting time in the queue. Posner[73] gives an exact analytical solution for the latter model, but, unfortunately, this solution is not suitable for computational purposes.

Problem formulation

A telephone switching system is a medium that connects a customer (subscriber) initiating a telephone call with another customer. For that purpose, a network of wires, exchanges and processor units is available. A call set-up is a sequence of actions processed by the system. First the dial tone is given to the customer, next digit for digit links are selected and the connection is built up piecewise and finally a talk phase is entered by the customers. In the mean time, control messages are travelling within the system, not recognizable by the customers and also actions are taken for the accounting of the call.

In overload situations, i.e. when more traffic is offered to the system than it is dimensioned for, the behaviour of the customers very strongly affects the system performance. Reactions of customers can influence the system in various ways. On the one hand, a customer may abandon his call with a certain probability when he is

† This section is the result of joint research with dr. Phuoc Tran Gia, University of Siegen/Stuttgart; cf. also Tran Gia and van Hoorn[82].

confronted with large delays during the call set-up phase (e.g. waiting for dial tone, post dialling delay). In this case, an ineffective amount of work has been offered to the processor and hence the call completion rate of the system decreases. On the other hand, rejected customers may reattempt their call after a certain time. The repeated attempts will further inflate the overload.

In this study, we determine the performance limitations of switching systems in overload situations, where we do not take into account repeated attempts. To measure the performance of the system, we use the call completion rate, which is the number of successfully completed calls per unit time.

Modelling approach

We model the processor of a telephone switching system as a single server queue with an infinite waiting capacity where calls (customers) arrive according to a Poisson process.

We observe a test call entering a switching system. The call sees an amount of work waiting for processing. Concretely this work may stand for the number of subcalls or telephonic events buffered in the processor queue. Based on this observation and in order to simplify the analysis without losing essential effects, we consider the amount of work in the processor queue as a discrete number of phases which are assumed to be independent and identically distributed random variables having a distribution function $F(t)$.

The number of phases the test call sees upon arrival corresponds to its waiting time before entering service. Depending on the duration of its waiting time, the test call decides to bring a number of phases into the system. These phases can be interpreted as the number of subcalls and the corresponding call handling effort the switching system must spend for the call attempt. From the point of view of analysis we can consider the decision to be taken at the arrival epoch of the call, although in reality it is taken at the instant the customer enters service.

Calls with incompleted dialling or abandoned calls usually offer a small number of phases to the system while successful calls with completed dialling often have offered a larger number of phases to the system. Therefore, according to the number of phases chosen by a call we define the probability that it will become a bad call or a successful call.

Considering all arguments discussed above, we have modelled the system as a single server queueing system of type $M^X/G/1$ with state dependent batch arrivals. In this model the following assumptions are made:

- Call arrivals follow a Poisson process at rate λ.
- A call that sees k phases in the system (including the phase in service) will offer j phases to the system with probability $g_j^{(k)}$.
- A call having chosen j phases becomes a successful call (completed call) with the conditional completion probability c_j.
- The service time of an arbitrary phase has distribution function $F(t)=\Pr\{\mathbf{S}\leqslant t\}$.

We define the performance measures

P_{compl} = the probability that an arbitrary call is completed successfully

Y = the fraction of calls that are completed successfully

EL = the mean number of phases in the system at an arbitrary epoch.

These performance measures are computed from the distribution (p_n), the steady state distribution of the number of phases in the system at an arbitrary epoch.

$$P_{compl} = \sum_{k=0}^{\infty} p_k \sum_{j=0}^{\infty} g_j^{(k)} c_j$$

$$Y = \lambda P_{compl}$$

$$EL = \sum_{k=1}^{\infty} k p_k$$

Specification of c_j and $g_j^{(k)}$

From practical considerations, the number of service phases chosen by a call may vary between fixed numbers N_0 and N_1. For the probabilities c_j we have made the following choice, containing the parameters γ and N_{lim} as degrees of freedom; cf. Figure 5.5.

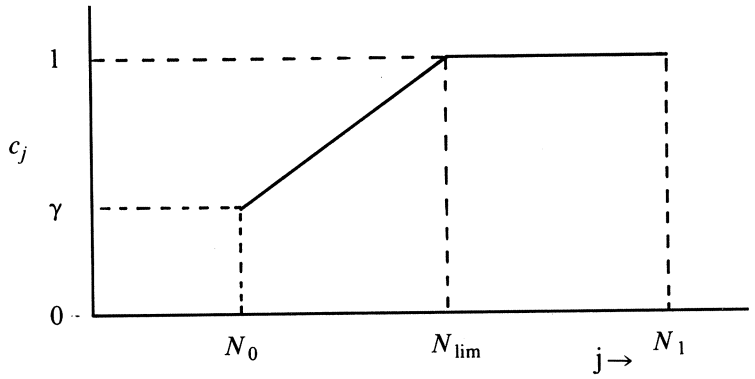

Figure 5.5 The conditional completion probability.

$$c_j = \gamma + (1-\gamma)(j - N_0)/(N_{lim} - N_0), \quad N_0 \leq j \leq N_{lim}$$
$$c_j = 1, \quad N_{lim} \leq j \leq N_1$$
$$c_j = 0, \quad \text{otherwise}$$

In practical situations, the number of subcalls produced by the majority of the successful calls varies between certain limits, here represented by N_{lim} and N_1. If the number of subcalls generated by a call is less than N_{lim}, the probability for it to be completed successfully decreases but need not be zero.

The batch size distribution is the factor that takes into account the dependency between the service time of a customer and his waiting time. When a customer sees upon arrival k phases in the system, his waiting time is approximately the convolution of k service times. He is supposed to have a certain patience, i.e. he is willing to wait a reasonable time, say τ, before entering service. If his waiting time is short, he will choose a service time consisting of a relatively large number of phases, corresponding to a large number of subcalls. If his waiting time is longer than τ he will tend to bring a smaller number of phases into the system because he abandons his call sooner. The length of the patience τ could be obtained by measurement in a real system. Here we choose $\tau = 3N_1ES$. In Figure 5.6 a typical example is given for the average number of phases chosen by a customer. For details on the specification of $g_j^{(k)}$ we refer to Tran Gia and van Hoorn[82].

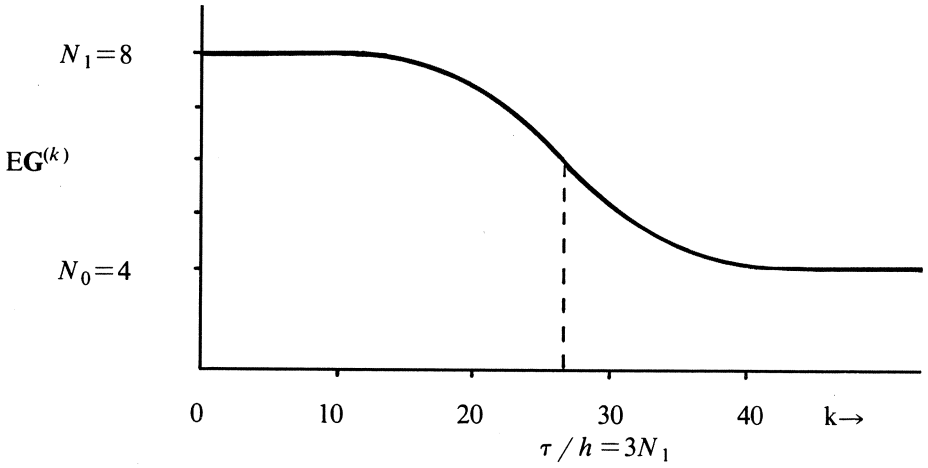

Figure 5.6 The mean number of phases chosen by a customer.

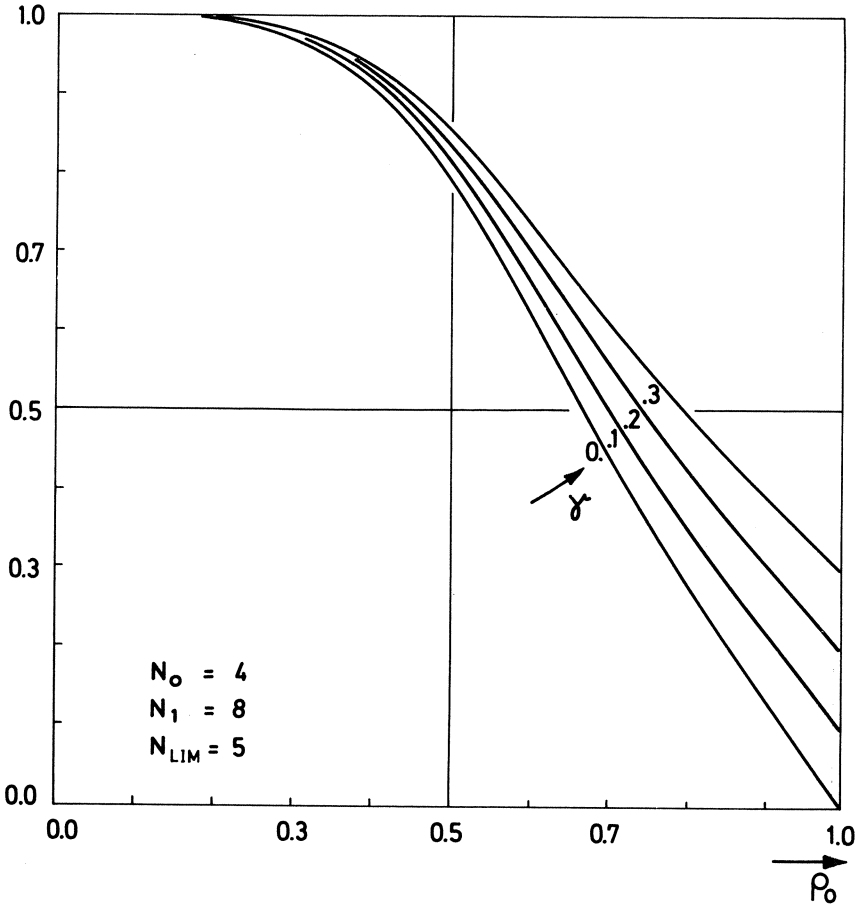

Figure 5.7 The completion probability for calls.

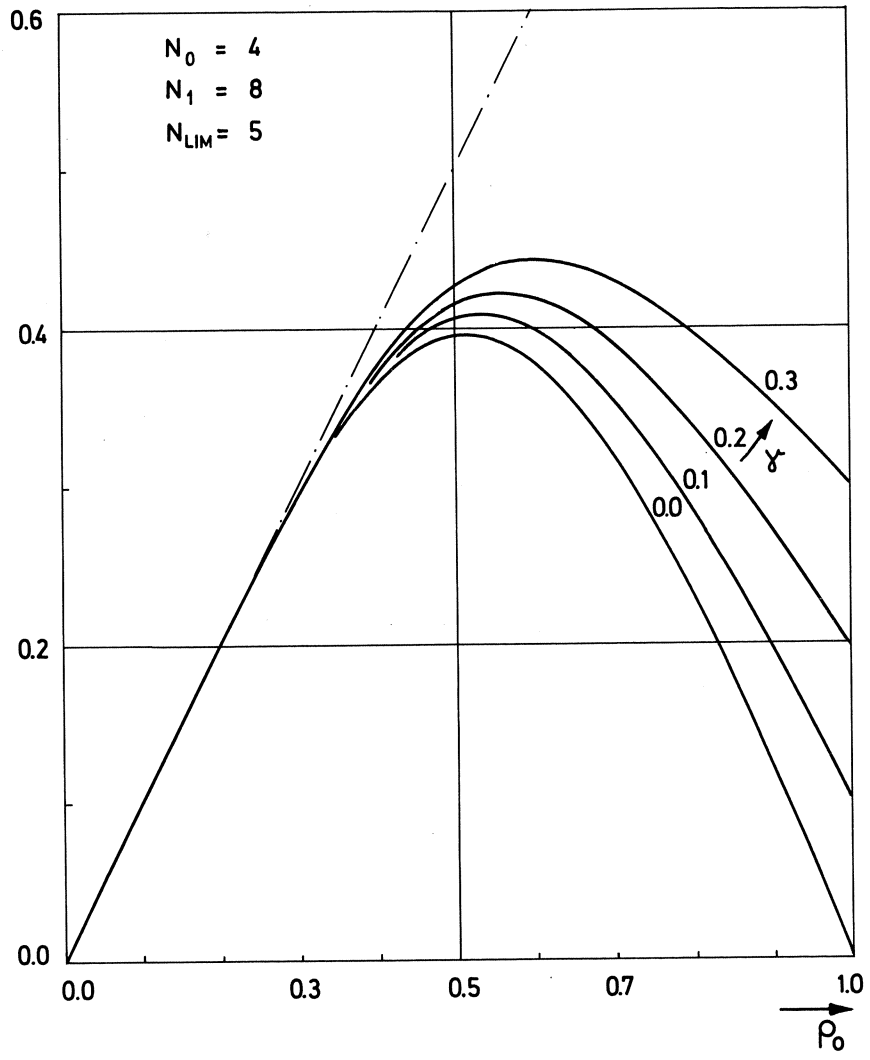

Figure 5.8 The normalized call completion rate.

Numerical results

For the numerical results, except for those in table 5.13, we have taken an exponential distribution of the service phases. In the next 3 figures, we have depicted some numerical results for the performance measures as function of the normalized traffic intensity $\rho_0=\lambda N_0 ES$.

Figure 5.7 shows the completion probability for an arbitrary call. The curves are drawn for different values of γ. It should be recalled that γ represents the completion probability for calls that have a relatively long waiting time and choose the minimum number N_0 of phases. It can be seen here that the call completion probability decreases rapidly above a certain level of the offered traffic. A degradation of

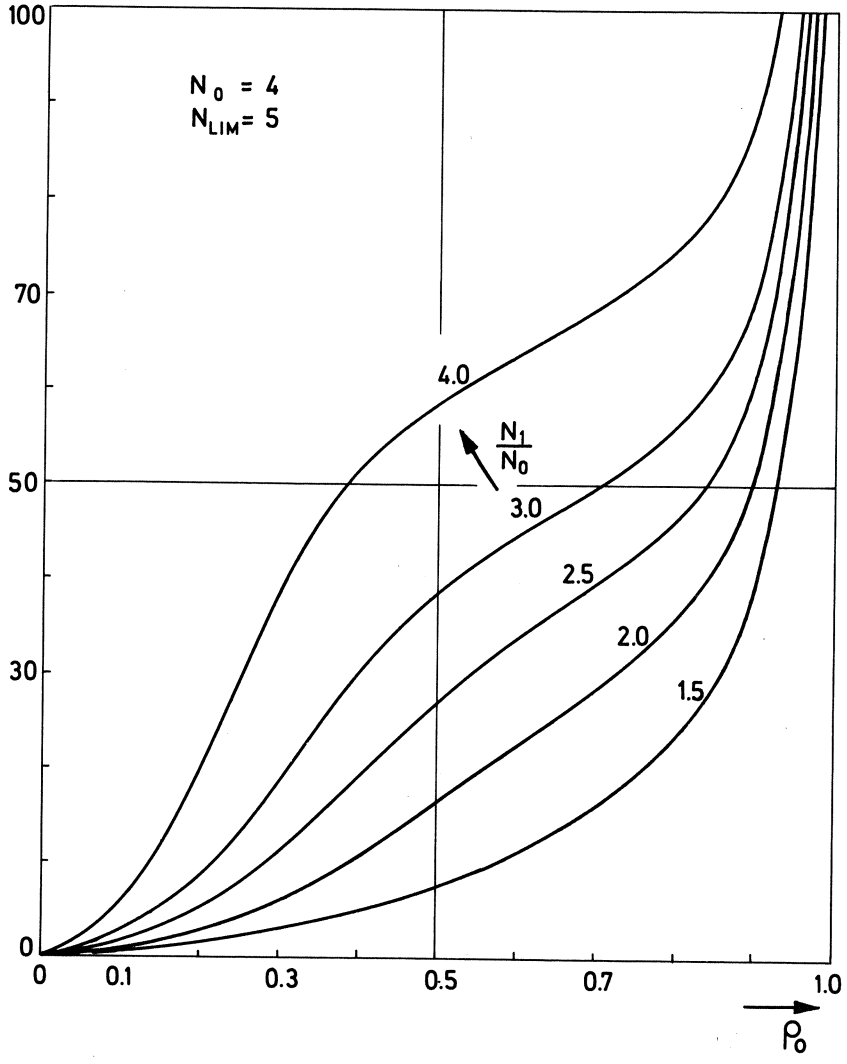

Figure 5.9 The mean number of phases in the system.

the system performance is said to have occurred. This effect is shown more clearly in Figure 5.8. where the normalized call completion rate YN_0ES is depicted.

The mean number of phases in the system is shown in Figure 5.9, where different values of the ratio N_1 / N_0 are considered. For higher values of N_1 / N_0 the curve can clearly be recognized as a superposition of two segments. The first segment of the curve corresponds to lower traffic levels where the batch size is almost always equal to N_1. The second segment corresponds to higher traffic intensities where the majority of customers chooses N_0 phases.

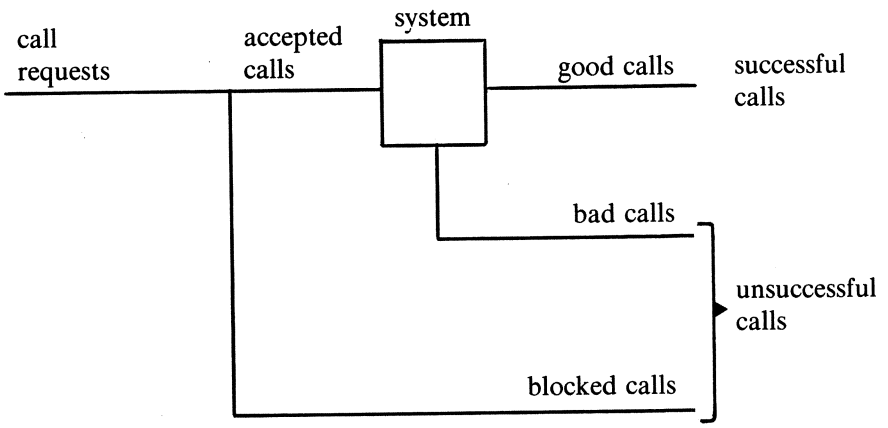

Figure 5.10 On the call completion rate in a switching system.

In Figure 5.8 it is shown that the system performance, say the normalized call completion rate, decreases rapidly above a critical level of the offered load. Above this critical level, the queue becomes large and customers must wait a long time before they enter service. Then they loose patience, tend to abandon their call and as a result, the call completion rate decreases.

In order to avoid this effect, the system may stop accepting all calls at a certain load level. The idea behind it is that if the switching system accepts fewer calls it is able to handle them well. As illustrated in Figure 5.10 we can save processor time and increase the amount of good calls if we allow the system to reject calls according to a scheme which will be described in the following. It should be noted here that the phenomenon of repeated attempts of blocked calls is not taken into account. A simple but effective overload control scheme is to stop accepting calls when the number of phases in the systems is larger than L_1 (say). Then the capacity of the system is limited to L_1+N_1 phases.

In Figure 5.11, it can be clearly seen that for a suitable choice of L_1 the system performance is improved considerably.

Table 5.12 compares the call completion rate in a system with overload control for different service time distributions of phases. For the given batch statistics the difference caused by the phase distributions is not essential. This argument justifies a Markovian phase modelling approach, which entails a simpler analysis and less computing effort.

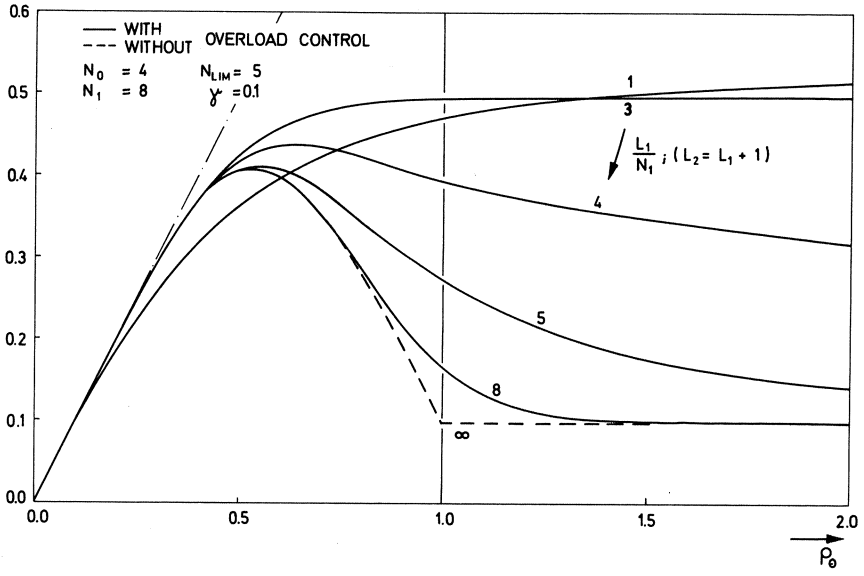

Figure 5.11 The performance of the overload control scheme. The normalized call completion rate.

offered traffic ρ_0	Phase service time distribution		
	E_3	M	H_2 ($C_S^2=3.0$)
0.1	0.099980	0.099974	0.099949
0.3	0.294192	0.293271	0.290380
0.5	0.425019	0.421388	0.412192
0.7	0.445664	0.443856	0.439294
1.0	0.405236	0.407232	0.412044
1.2	0.375811	0.379266	0.387623
1.5	0.338031	0.343039	0.354784
2.0	0.290323	0.297062	0.312029

Table 5.12 The call completion rate with overload control for different service time distributions of phases. ($N_0=4$, $N_1=8$, $N_{lim}=5$, $\gamma=0.1$, $L_1=5N_1$).

6. THE SPP/G/1 QUEUE: A SINGLE SERVER QUEUE WITH A SWITCHED POISSON PROCESS AS INPUT PROCESS

* This sector has been published in OR spectrum (1983), 5, 207-218.

In this chapter, we give an analysis of a non-standard queueing model, namely the SPP/G/1 queue. The arrival process of customers is a switched Poisson process (SPP), i.e. the intensity of the arrival stream is alternately λ_1 and λ_2, governed by some random mechanism. The model is interesting since this arrival process covers both renewal and non-renewal processes with coefficients of variation larger than one. For the renewal case, the interarrival distribution is in fact hyperexponential; cf. Kuczura[73]. With our model, the consequence of a 'renewal assumption' can be investigated . In practice, one often assumes that the arrival process is renewal and fits a hyperexponential distribution to the interarrival time by matching the first two moments, provided that the coefficient of variation of the interarrival time is at least one. In Section 6.6 we discuss sensitivity questions with respect to the renewal assumption and the influence of the third moment of the interarrival time.

A special case of the switched Poisson process is the interrupted Poisson process (IPP), which has alternately a positive and a zero intensity; cf. Kuczura[73]. In Heffes[73], an analysis is given for the IPP/M/c queue, which is in fact a special case of the GI/M/c queue. In Yechiali and Naor[71], the queue length distribution is studied in a Markovian queueing model with two possible arrival rates. The model is a generalization of the SPP/M/1 queue in that with each arrival rate a service rate is associated.

In Section 6.1, we describe the model in detail and give the analysis using the regenerative method. The analysis presented here demonstrates the power of the regenerative method to investigate systematically rather complex queueing models. The properties of the arrival process are derived in Section 6.2. Also, in Section 6.2 some preparatory work is done for the generating function analysis in Section 6.3. Section 6.4 concerns some computational aspects of the algorithms formulated in the Sections 6.1 and 6.5. Section 6.5 deals with the SPP/G/1 queue having only place for a finite number of customers. Finally, in Section 6.6 we present a few numerical results for the delay probability and the mean queue length. We investigate the sensitivity of these performance measures for the service time distribution, the third moment of the interarrival intervals and the degree of correlation between successive arrivals. Also, we discuss the quality of the two moment approximations for the delay probability and the mean waiting time given in Krämer and Langenbach-Belz[76].

6.1. The model and the regenerative analysis

We consider a single server queueing system with an infinite capacity. The arrival process of customers is a switched Poisson process. That is, the arrival process is alternately in the phases 1 and 2 during exponentially distributed times with means $1/\omega_1$ and $1/\omega_2$. If the arrival process is in phase i, then independently of the phase process, customers arrive according to a Poisson process with intensity λ_i , $i=1,2$. See Figure 6.1. To avoid trivialities we assume that ω_1 and ω_2 are strictly positive and that at least one arrival intensity is non-zero. The service time $\mathbf{S}$ of a customer has a general probability distribution function $F(t)=\Pr\{\mathbf{S}\leq t\}$. We assume that the traffic intensity $\rho=\lambda^* \mathbf{ES}$ is less than 1, where λ^* is the average arrival rate and is given by

$$\lambda^* = (\lambda_1\omega_2+\lambda_2\omega_1)/(\omega_1+\omega_2)$$

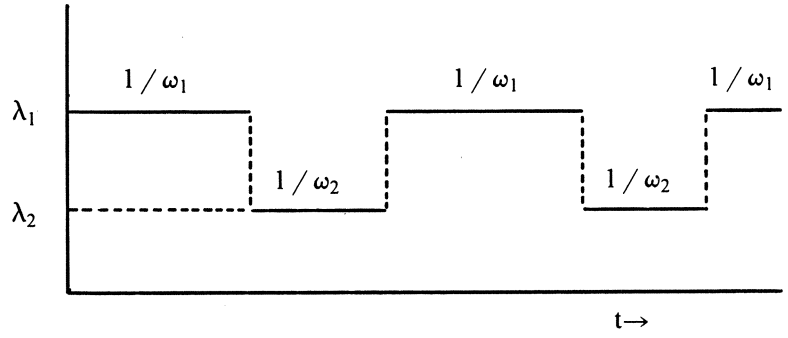

Figure 6.1 The arrival rate

We consider the non-Markovian process whose state is given by the pair (n,i), where n denotes the number of customers present and i the phase (or level) of the arrival process, $n \geq 0$, $i = 1,2$.

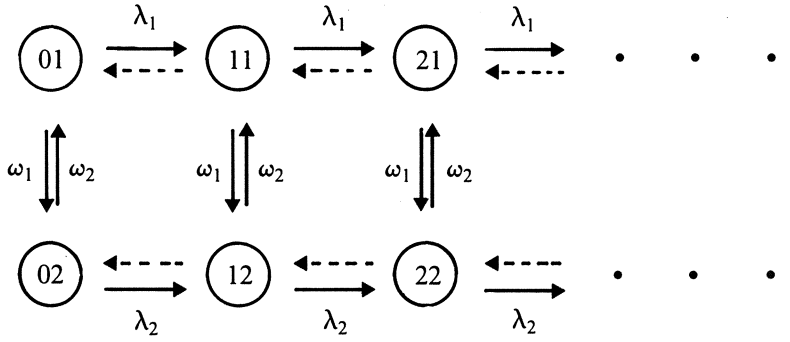

Figure 6.2 The state transition diagram

This process considered on suitable embedded epochs becomes Markovian. In Figure 6.2 the state transition diagram is depicted. The straight arrows and the corresponding transition rates represent transitions which may occur at 'random' epochs (i.e. at epochs at which an arrival occurs or the phase of the arrival process changes) and the dotted arrows represent transitions at service completion epochs. As method of analysis for this queueing model with a two dimensional state, we use the regenerative method and up and down crossing arguments. We focus on the following steady state probabilities. For $n \geq 0$, $i = 1,2$ let

$$p_{ni} = \lim_{t \to \infty} \Pr\{ \text{ at epoch } t \text{ the system is in state } (n,i) \}$$

$$q_{ni} = \lim_{k \to \infty} \Pr\{ \text{ the } k^{\text{th}} \text{ customer leaves behind the system in state } (n,i) \text{ upon service completion } \}$$

$$\pi_{ni} = \lim_{k \to \infty} \Pr\{ \text{ the } k^{\text{th}} \text{ customer sees upon arrival the system in state } (n,i) \}$$

Using standard results from the theory of the regenerative processes (cf. Stidham[72]), it can easily be verified that these limits are well defined and are independent of the initial state. After having derived a recursion scheme for the probabilities (p_{ni}) and (q_{ni}), we give a direct relation between (π_{ni}) and (p_{ni}).

We assume that at epoch 0 a customer has completed service and has left behind the system in state (0,1) and define the random variables

$\mathbf{T}$ = the next time the system is left behind by a customer in state (0,1)

$\mathbf{T}_{ni}$ = the amount of time the system is in state (n,i) in the busy cycle $(0,\mathbf{T}]$, $n \geqslant 0$, $i = 1,2$

$\mathbf{N}$ = the number of customers served in $(0,\mathbf{T}]$

$\mathbf{N}_{ni}$ = the number of service completion epochs in $(0,\mathbf{T}]$ at which the customer served leaves behind the system in state (n,i), $n \geqslant 0$, $i = 1,2$

Note that there are other possibilities to define a busy cycle.

In the next theorem, which can be proved using the theory of the regenerative processes (cf. Ross[70], Stidham[72]), we relate the above defined random variables to the steady state probabilities (p_{ni}) and (q_{ni}).

Theorem 6.1

$$p_{ni} = \frac{\mathbf{ET}_{ni}}{\mathbf{ET}}, \quad q_{ni} = \frac{\mathbf{EN}_{ni}}{\mathbf{EN}}, \quad n \geqslant 0, \quad i = 1,2 \tag{6.1}$$

□

For $0 \leqslant j \leqslant n$, $k,l = 1,2$ we define the quantities

A_{jn}^{kl} = the expected amount of time during which the system is in state (n,l) until the next service completion epoch, given that at epoch 0 a service is completed and the system is left behind in state (j,k).

In the multi-indexed quantity A_{jn}^{kl}, the superscripts concern the arrival intensity and the subscripts the number of customers. Then, by partitioning the busy cycle $(0,\mathbf{T}]$ by means of the service completion epochs and using Wald's Theorem, it readily follows that

Theorem 6.2

$$\mathbf{ET}_{ni} = \sum_{j=0}^{n} \mathbf{EN}_{j1} A_{jn}^{1i} + \sum_{j=0}^{n} \mathbf{EN}_{j2} A_{jn}^{2i}, \quad n \geqslant 1, \quad i = 1,2 \tag{6.2}$$

□

By using an up and down crossing argument, we derive a second set of relations between the numbers $(\mathbf{ET}_{ni})$ and $(\mathbf{EN}_{ni})$. Let S be an arbitrary subset of the state space, then the following property holds.

in the busy cycle $(0,\mathbf{T}]$, the number of transitions out of S is equal to the number of transitions into S.

We apply this property to the set of states $S_{n1} = \{(0,1),(1,1),...,(n,1)\}$. Note that a transition from $(n,1)$ to $(n+1,1)$ can only be caused by an arrival, while a transition

from $(n,1)$ to $(n,2)$ can only be caused by a switching of the arrival rate; see also Figure 6.3. By the Poisson Lemma we have that in the busy cycle (0,T] the expected number of transitions from $(n,1)$ to $(n+1,1)$ is equal to $\lambda_1 \mathbf{ET}_{n1}$ and the expected number of transitions from $(k,1)$ to $(k,2)$ is equal to $\omega_1 \mathbf{ET}_{k1}$, $0 \leqslant k \leqslant n$. Also, the expected number of transitions in (0,T] from $(n+1,1)$ to $(n,1)$ is equal to $\mathbf{EN}_{n1}$. Summarizing, we get

the expected number of transitions out of the set S_{n1} in the busy cycle (0,T]

$= \lambda_1 \mathbf{ET}_{n1} + \omega_1(\mathbf{ET}_{01} + \cdots + \mathbf{ET}_{n1})$

= the expected number of transitions into the set S_{n1} in (0,T]

$= \mathbf{EN}_{n1} + \omega_2(\mathbf{ET}_{02} + \cdots + \mathbf{ET}_{n2})$

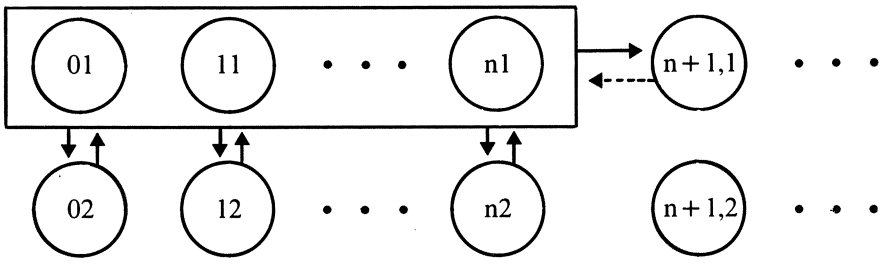

Figure 6.3 The rate out of S_{n1} equals the rate into S_{n1}.

A similar relation applies to the set $S_{n2} = \{(0,2),(1,2), \ldots, (n,2)\}$. Hence we have the following

Theorem 6.3

$$\lambda_1 \mathbf{ET}_{n1} + \omega_1 \sum_{k=0}^{n} \mathbf{ET}_{k1} = \mathbf{EN}_{n1} + \omega_2 \sum_{k=0}^{n} \mathbf{ET}_{k2}, \quad n \geqslant 0 \tag{6.3}$$

$$\lambda_2 \mathbf{ET}_{n2} + \omega_2 \sum_{k=0}^{n} \mathbf{ET}_{k2} = \mathbf{EN}_{n2} + \omega_1 \sum_{k=0}^{n} \mathbf{ET}_{k1}, \quad n \geqslant 0 \tag{6.4}$$

□

Now, we can easily obtain some useful relations which we formulate in

Theorem 6.4

$$\sum_{n=0}^{\infty} p_{n1} = \frac{\omega_2}{\omega_1 + \omega_2} = 1 - \sum_{n=0}^{\infty} p_{n2} \tag{6.5}$$

$$\mathbf{EN} = \lambda^* \mathbf{ET}, \quad \text{where } \lambda^* = (\lambda_1\omega_2 + \lambda_2\omega_1)/(\omega_1 + \omega_2). \tag{6.6}$$

$$p_{01} + p_{02} = 1 - \lambda^* \mathbf{ES} \tag{6.7}$$

Proof The first equation follows by an up and down crossing argument for the set of states $S_1 = \{(0,1),(1,1),(2,1),\ldots\}$. S_1 is left at rate $\omega_1 \Sigma_{n=0}^{\infty} p_{n1}$ and entered at rate

$\omega_2 \Sigma_{n=0}^{\infty} p_{n2}$. Clearly these rates are equal and since the probabilities sum up to one, (6.5) follows. By adding (6.3) and (6.4) and summing over $n \geqslant 0$, we get

$$\sum_{n=0}^{\infty} (\lambda_1 \mathrm{ET}_{n1} + \lambda_2 \mathrm{ET}_{n2}) = \mathrm{EN}$$

Next, using (6.5) and the first part of Theorem 6.1, it follows that $\mathrm{EN}/\mathrm{ET} = (\lambda_1\omega_2 + \lambda_2\omega_1)/(\omega_1 + \omega_2)$ $(= \lambda^*)$. In the same way, by adding (6.2) for $i = 1,2$ and summing over $n \geqslant 1$, we find

$$\mathrm{ET} - \mathrm{ET}_{01} - \mathrm{ET}_{02} = \sum_{k=1}^{2} \sum_{n=1}^{\infty} \sum_{j=0}^{n} \mathrm{EN}_{jk} (A_{jn}^{k1} + A_{jn}^{k2})$$

Note that $\Sigma_{n=j}^{\infty} (A_{jn}^{i1} + A_{jn}^{i2}) = \mathrm{ES}$, $j \geqslant 0$, $i = 1,2$, where $A_{00}^{kl} = 0$, $k, l = 1,2$. By changing the order of summations we get $\mathrm{ET} - \mathrm{ET}_{01} - \mathrm{ET}_{02} = \mathrm{ENES}$. This expression implies (6.7). □

Next we combine the results obtained in the Theorems 6.1. - 6.4. This yields the following algorithm for the probabilities (p_{ni}) and (q_{ni}).

Algorithm 6.5

1. Evaluate the constants A_{jn}^{kl}, $0 \leqslant j \leqslant n$, $k, l = 1,2$
2. Compute p_{01} and p_{02} using Relation (6.7) and Relation (6.24) to be derived hereafter. Next, using (6.3) and (6.4) with $n = 0$ and using (6.6), compute q_{01} and q_{02}.
3. Assume that $p_{0i}, \ldots, p_{n-1,i}$, $q_{0i}, \ldots, q_{n-1,i}$, $i = 1,2$ have been computed, compute p_{n1}, p_{n2}, q_{n1}, and q_{n2} by solving the four linear equations

$$p_{n1} = \lambda^* q_{n1} A_{nn}^{11} + \lambda^* q_{n2} A_{nn}^{21} + \lambda^* \sum_{j=0}^{n-1} (q_{j1} A_{jn}^{11} + q_{j2} A_{jn}^{21})$$

$$p_{n2} = \lambda^* q_{n1} A_{nn}^{12} + \lambda^* q_{n2} A_{nn}^{22} + \lambda^* \sum_{j=0}^{n-1} (q_{j1} A_{jn}^{12} + q_{j2} A_{jn}^{22})$$

$$\lambda^* q_{n1} = (\lambda_1 + \omega_1) p_{n1} - \omega_2 p_{n2} + \sum_{k=0}^{n-1} (\omega_1 p_{k1} - \omega_2 p_{k2})$$

$$\lambda^* q_{n2} = -\omega_1 p_{n1} + (\lambda_2 + \omega_2) p_{n2} + \sum_{k=0}^{n-1} (\omega_2 p_{k2} - \omega_1 p_{k1})$$

4. Return to step 3 if necessary. □

The computational effort to compute the distributions (p_{ni}) and (q_{ni}) lies mainly in step 1 and step 2. For the evaluation of the constants A_{jn}^{kl}, we refer to Section 6.4. To start the algorithm, the initial values p_{01} and p_{02} are required, but so far we have only one relationship between p_{01} and p_{02}, namely Equation (6.7). A second relation is provided by a standard method in queueing theory, namely by using the generating functions of the state probabilities; see Section 6.3. Once we have computed (p_{ni}), we can directly find the arriving customer distributions (π_{ni}), $i = 1,2$.

Theorem 6.6

$$\pi_{ni} = \frac{\lambda_i}{\lambda^*} p_{ni} \ , \ n \geqslant 0 \ , \ i = 1,2 \tag{6.8}$$

Proof The steady state probability π_{ni} at arrival epochs can be interpreted as the long run fraction of customers who find the system in state (n,i). By the theory of the regenerative processes, π_{ni} is also equal to the expected number of customers who find the system in state (n,i) in the busy cycle $(0,\mathbf{T}]$ divided by the total expected number of arrivals in $(0,\mathbf{T}]$. Hence, $\pi_{ni} = \lambda_i \mathbf{ET}_{ni} / \lambda^* \mathbf{ET}$ and the theorem follows

□

6.2. Properties of the arrival process

The way in which the arrival process has been introduced in the previous section does not give a clear understanding what this process really looks like. Important characteristics are its coefficient of variation and its autocorrelation coefficient, which contains information about the dependence between successive arrivals.

In this section, we investigate the switched Poisson process we introduced in Section 6.1. We give the stationary distribution of the interarrival time and explicit expressions for its coefficient of variation and its correlation coefficient. As a preparation for the generating function analysis in the next section, we first focus on the probabilities

$$a_j^{kl}(t) = \Pr\{ \ j \text{ customers arrive in } (0,t) \text{ and the arrival process is in phase } l \text{ at epoch } t \mid \text{the arrival process is in phase } k \text{ at epoch } 0 \ \}, \quad j \geqslant 0, \ k,l = 1,2.$$

Note that the arrival process is Markovian since all distribution functions involved are exponential. By standard arguments, we get the following systems of Chapman-Kolmogorov differential equations, $i = 1,2$

$$\frac{\mathrm{d}}{\mathrm{d}t} a_j^{1i}(t) = -(\lambda_1 + \omega_1) a_j^{1i}(t) + \omega_1 a_j^{2i}(t) + \lambda_1 a_{j-1}^{1i}(t) \ , \ j \geqslant 0 \tag{6.9}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} a_j^{2i}(t) = \omega_2 a_j^{1i}(t) - (\lambda_2 + \omega_2) a_j^{2i}(t) + \lambda_2 a_{j-1}^{2i}(t) \ , \ j \geqslant 0 \tag{6.10}$$

with $a_{-1}^{kl}(t) \equiv 0$ and $a_0^{kl}(0) = \delta_{kl}$, $k,l = 1,2$. For $|z| \leqslant 1$, $k,l = 1,2$ we define

$$a^{kl}(t,z) = \sum_{j=0}^{\infty} a_j^{kl} z^j$$

Multiplying (6.9) and (6.10) with z^k and summing over $k \geqslant 0$ yields for $i = 1,2$

$$\frac{\partial}{\partial t} a^{1i}(t,z) = -(\lambda_1(1-z) + \omega_1) a^{1i}(t,z) + \omega_1 a^{2i}(t,z)$$

$$\frac{\partial}{\partial t} a^{2i}(t,z) = \omega_2 a^{1i}(t,z) - (\lambda_2(1-z) + \omega_2) a^{2i}(t,z)$$

The solution of this system of linear differential equations is

$$a^{ii}(t,z)=\frac{1}{r_2(z)-r_1(z)}\Big\{[r_2(z)-(\lambda_i(1-z)+\omega_i)]e^{-r_1(z)t} - [r_1(z)-(\lambda_i(1-z)+\omega_i)]e^{-r_2(z)t}\Big\}\,,\quad i=1,2 \tag{6.11}$$

$$a^{kl}(t,z)=\omega_k\,\frac{e^{-r_1(z)t}-e^{-r_2(z)t}}{r_2(z)-r_1(z)}\,,\quad (k,l)=(1,2),(2,1) \tag{6.12}$$

where

$$r_{1,2}(z)=\tfrac{1}{2}\{\lambda_1(1-z)+\omega_1+\lambda_2(1-z)+\omega_2 \pm \sqrt{(\lambda_1(1-z)+\omega_1+\lambda_2(1-z)+\omega_2)^2-4[(\lambda_1(1-z)+\omega_1)(\lambda_2(1-z)+\omega_2)-\omega_1\omega_2]}\}$$

The functions $a^{kl}(t,z)$, $k,l=1,2$ contain the information of the arrival process and will be used in Section 6.3 to derive the generating functions of (p_{ni}) and (q_{ni}), $i=1,2$. In particular, we note that $a_0^{kl}(t)=a_{kl}(t,0)$.

Next, we turn to the determination of the limiting distribution of the interarrival time. Let τ_1 be the epoch of the first arrival and let τ_{n+1} be the time elapsed between the n^{th} and the $n+1^{\text{th}}$ arrival, $n\geqslant 1$. Define now

$$G(t)=\lim_{n\to\infty}\Pr\{\tau_n\leqslant t\}$$

It is easily verified that $G(t)$ is independent of the initial state. Let the random variable $\mathbf{X}(t)$ be the state of the arrival process at epoch t, where $\mathbf{X}(t)=i$ if the current arrival rate at epoch t is λ_i, $i=1,2$. Further, define

$$t_n=\tau_1+\cdots+\tau_n\,,\quad n\geqslant 1$$

Then t_n is the epoch of the n^{th} arrival. By conditioning on the state of the arrival process at epoch t and by noting that $\lim_{n\to\infty}\Pr\{\mathbf{X}(t_{n-1})=1\}=\sum_{n=0}^{\infty}\pi_{n1}=\lambda_1\omega_2/(\lambda_1\omega_2+\lambda_2\omega_1)$, we get

$$G(t)=\lim_{n\to\infty}\sum_{i=1}^{2}\Pr\{\tau_n\leqslant t\mid\mathbf{X}(t_{n-1})=i\}\Pr\{\mathbf{X}(t_{n-1})=i\} \tag{6.13}$$

$$=\frac{\lambda_1\omega_2}{\lambda_1\omega_2+\lambda_2\omega_1}\{1-a_0^{11}(t)-a_0^{12}(t)\}+\frac{\lambda_2\omega_1}{\lambda_1\omega_2+\lambda_2\omega_1}\{1-a_0^{21}(t)-a_0^{22}(t)\}$$

$$=1-\frac{\lambda_1\omega_2}{\lambda_1\omega_2+\lambda_2\omega_1}\,\frac{(r_2-\lambda_1)e^{-r_1t}-(r_1-\lambda_1)e^{-r_2t}}{r_2-r_1} - \frac{\lambda_2\omega_1}{\lambda_1\omega_2+\lambda_2\omega_1}\,\frac{(r_2-\lambda_2)e^{-r_1t}-(r_1-\lambda_2)e^{-r_2t}}{r_2-r_1}$$

where $r_1=r_1(0)$ and $r_2=r_2(0)$. Recall that $a_0^{kl}(t)=a^{kl}(t,0)$. Hence the distribution $G(t)$ is hyperexponential with mean $1/\lambda^*$ (cf. 6.6) and squared coefficient of variation C_a^2, which is after some algebra found to be

$$C_a^2=1+\frac{2\omega_1\omega_2(\lambda_1-\lambda_2)^2}{(\lambda_1\lambda_2+\lambda_1\omega_2+\lambda_2\omega_1)(\omega_1+\omega_2)^2} \tag{6.14}$$

Though the function $G(t)$ already gives some insight in the arrival process, there is another aspect which influences strongly the queueing process, namely the dependence

between successive arrivals. The autocorrelation coefficient of the sequence of interarrival intervals measures this effect. Let θ be defined as

$$\theta = \lim_{n\to\infty} \frac{E\tau_n \tau_{n+1} - (E\tau_n)^2}{E\tau_n^2 - (E\tau_n)^2}$$

After some algebra (see below), it follows that

$$\theta = \frac{\lambda_1\lambda_2}{\lambda_1\lambda_2 + \lambda_1\omega_2 + \lambda_2\omega_1} \frac{C_a^2 - 1}{2C_a^2} \tag{6.15}$$

Note that the only cases where the arrival process is a renewal process are the cases $\lambda_1\lambda_2 = 0$ and $\lambda_1 = \lambda_2$.

Since we already know that (cf.(6.14))

$$\lim_{n\to\infty} E\tau_n = \frac{1}{\lambda^*} \text{ and } \lim_{n\to\infty} E\tau_n^2 - (E\tau_n)^2 = \frac{C_a^2}{\lambda^{*2}},$$

the difficult element in the computation of θ remains to compute $\lim_{n\to\infty} E\tau_n \tau_{n+1}$. Now, let n be fixed, then

$$E\tau_n \tau_{n+1} = \tfrac{1}{2}E(\tau_n + \tau_{n+1})^2 - \tfrac{1}{2}(E\tau_n)^2 - \tfrac{1}{2}(E\tau_{n+1})^2$$

and hence it is sufficient to focus on $E(\tau_n + \tau_{n+1})^2$. We get

$$\Pr\{\tau_n + \tau_{n+1} \leq x\} = \sum_{i=1}^{2} \Pr\{\tau_n + \tau_{n+1} \leq x \mid \mathbf{X}(t_n) = i\} \Pr\{\mathbf{X}(t_n) = i\}$$

$$= \sum_{i=1}^{2} \Pr\{\text{ in the interval } (0,x) \text{ at least two arrivals occur } \mid \mathbf{X}(0) = i\} \Pr\{\mathbf{X}(t_n) = i\}$$

$$= \sum_{i=1}^{2} \{1 - a_0^{i1}(t) - a_0^{i2}(t) - a_1^{i1}(t) - a_1^{i2}(t)\} \Pr\{\mathbf{X}(t_n) = i\}$$

Using

$$a_1^{kl}(t) = \frac{\partial}{\partial z} a^{kl}(t,z) \Big|_{z=0} \text{ and } \lim_{n\to\infty} \Pr\{\mathbf{X}(t_n) = 1\} = \frac{\lambda_1\omega_2}{\lambda_1\omega_2 + \lambda_2\omega_1},$$

it is only a matter of straightforward but tedious algebra to compute $\lim_{n\to\infty} E(\tau_n + \tau_{n+1})^2$ and next to compute θ.

6.3. The generating functions and the mean queue length

After the preparatory work in Section 6.2, we derive in this section the generating functions of the probabilities (p_{ni}) and (q_{ni}), $i = 1,2$. From these functions we deduce a second relation between the probabilities p_{01} and p_{02}, beside Relation (6.7). We conclude this section with a formula for EL_q, the mean number of customers in the queue at an arbitrary epoch (excluding the customer in service).

Before actually determining the generating functions we give a representation of the quantities A_{jn}^{kl} appearing in the basic Relation (6.2).

Lemma 6.7 For $1 \leqslant j \leqslant n$, $k,l = 1,2$

$$A_{jn}^{kl} = \int_0^\infty (1-F(t)) a_{n-j}^{kl}(t)\,\mathrm{d}t \tag{6.16}$$

Proof First, remark that for $1 \leqslant j \leqslant n$, A_{jn}^{kl} only depends on j and n through the difference $n-j$. Secondly, note that

$$A_{jn}^{kl} = \int_0^\infty \mathrm{E}\chi(t)\,\mathrm{d}t$$

(letting j,n,k,l fixed) where under the assumption that at epoch 0 a service starts while the system is in state (j,k), $\chi(t)=1$ if at epoch t this service is still in progress and the system is in state (n,l). Otherwise, $\chi(t)=0$. Now (6.16) follows, since

$$\mathrm{E}\chi(t) = \Pr\{\chi(t)=1\} = (1-F(t))\, a_{n-j}^{kl}(t)$$

□

The quantities A_{0n}^{kl} are computed in an other way than the A_{jn}^{kl} for $j \geqslant 1$. A_{0n}^{kl} does not satisfy (6.16), since in state $(0,k)$ first an arrival should occur before a service can start. If the system is in state $(0,1)$, then with probability $\lambda_1/(\lambda_1+\omega_1)$ an arrival occurs before a switching to the arrival rate λ_2. Hence, for $k=1,2$

$$A_{0n}^{1k} = \frac{\lambda_1}{\lambda_1+\omega_1} A_{1n}^{1k} + \frac{\omega_1}{\lambda_1+\omega_1} A_{0n}^{2k} \tag{6.17}$$

and

$$A_{0n}^{2k} = \frac{\lambda_2}{\lambda_2+\omega_2} A_{0n}^{1k} + \frac{\omega_2}{\lambda_2+\omega_2} A_{1n}^{2k} \tag{6.18}$$

Define for $|z| \leqslant 1$ the generating functions

$$\alpha^{kl}(z) = \sum_{n=j}^{\infty} A_{jn}^{kl} z^{n-j} \quad \text{and} \quad \alpha_0^{kl}(z) = \sum_{n=1}^{\infty} A_{0n}^{kl} z^{n-1}$$

It readily follows that

$$\alpha^{kl}(z) = \int_0^\infty (1-F(t)) a^{kl}(t,z)\,\mathrm{d}t\,, \quad k,l = 1,2 \tag{6.19}$$

$$\alpha_0^{1k}(z) = \frac{\lambda_1(\lambda_2+\omega_2)\alpha^{1k}(z) + \lambda_2\omega_1\alpha^{2k}(z)}{\lambda_1\lambda_2+\lambda_1\omega_2+\lambda_2\omega_1}\,, \quad k=1,2 \tag{6.20}$$

$$\alpha_0^{2k}(z) = \frac{\lambda_1\omega_2\alpha^{1k}(z) + \lambda_2(\lambda_1+\omega_1)\alpha^{2k}(z)}{\lambda_1\lambda_2+\lambda_1\omega_2+\lambda_2\omega_1}\,, \quad k=1,2 \tag{6.21}$$

Next, it is straightforward to derive a system of four equations for the generating functions of (p_{n1}), (p_{n2}), (q_{n1}) and (q_{n2}). Define for $|z| \leqslant 1$, $i=1,2$ the generating functions

$$P_i(z) = \sum_{n=0}^{\infty} p_{ni} z^n \quad \text{and} \quad Q_i(z) = \sum_{n=0}^{\infty} q_{ni} z^n$$

These generating functions essentially follow from the Equations (6.2), (6.3) and (6.4). By rewriting (6.2), (6.3) and (6.4) in terms of the probabilities (p_{ni}) and (q_{ni}) using

(6.1) and (6.6) and next by taking generating functions, we obtain

$$P_k(z)-p_{0k}=\sum_{i=1}^{2}\{\lambda^* q_{0i} z\,\alpha_0^{ik}(z)+\lambda^*(Q_i(z)-q_{0i})\alpha^{ik}(z)\}\ ,\quad k=1,2$$

$$\lambda_1 P_1(z)+\frac{\omega_1}{1-z}P_1(z)=\lambda^* Q_1(z)+\frac{\omega_2}{1-z}P_2(z)$$

$$\lambda_2 P_2(z)+\frac{\omega_2}{1-z}P_2(z)=\lambda^* Q_2(z)+\frac{\omega_1}{1-z}P_1(z)$$

After a cumbersome but straightforward derivation, we get from this system of equations

$$\{P_1(z)+P_2(z)\}g(z)=(p_{01}+p_{02})((1-z)g(z)-z^2) \tag{6.22}$$
$$+zp_{01}\int_0^\infty (a^{22}(t,z)-a^{12}(t,z))\,\mathrm{d}F(t)$$
$$+zp_{02}\int_0^\infty (a^{11}(t,z)-a^{21}(t,z))\,\mathrm{d}F(t)$$

where

$$g(z)=\frac{1}{1-z}\left(\int_0^\infty e^{-r_2(z)t}\,\mathrm{d}F(t)-z\right)\left(\int_0^\infty e^{-r_1(z)t}\,\mathrm{d}F(t)-z\right) \tag{6.23}$$

To find a second relation between p_{01} and p_{02} we proceed as follows. The function $P_1(z)+P_2(z)$ is regular for $|z|\leqslant 1$ since it is a probability generating function. Hence, a zero of $g(z)$ is also a zero of the right hand side of (6.22). If we can find a zero , say z_0, of $g(z)$ with $|z_0|\leqslant 1$, then (6.22) supplies a second equation for p_{01} and p_{02} beside (6.7). Now, consider the function $g_1(z)$ defined by

$$g_1(z)=\int_0^\infty e^{-r_1(z)t}\,\mathrm{d}F(t)-z$$

Since $g_1(0)>0$ and $g_1(1)<0$, $g_1(z)$ has a zero z_0 with $0<z_0<1$. Hence, z_0 is also a zero of $g(z)$. Substituting z_0 in (6.22) and after some rewriting, we obtain

$$\{r_1(z_0)-(\lambda_2(1-z_0)+\omega_1+\omega_2)\}p_{01}+\{r_1(z_0)-(\lambda_1(1-z_0)+\omega_1+\omega_2)\}p_{02}=0 \tag{6.24}$$

It can be easily verified that the Equations (6.7) and (6.24) are independent, unless $\lambda_1=\lambda_2$ or $\omega_1\omega_2=0$. We have excluded the trivial case $\omega_1\omega_2=0$ and if $\lambda_1=\lambda_2$, the arrival process reduces to an ordinary Poisson process.

Remark It is not necessary to prove that $g(z)$ has a unique root for $|z|\leqslant 1$, since we already know the existence and unicity of the distribution (p_{ni}), $i=1,2$.

□

The actual computation of z_0 may be done using a numerical standard procedure.

Finally, from (6.22) we obtain by differentiation a formula for EL_q, the mean number of customers in the queue (excluding any customer in service)

$$\mathrm{EL}_q = \frac{\rho}{2(1-\rho)}(C_a^2 - 1 + \rho(C_S^2+1)) + \frac{C_a^2-1}{2(1-\rho)}\frac{\lambda_1\lambda_2 \mathrm{ES}}{\omega_1+\omega_2} + \frac{\lambda^*}{\omega_1+\omega_2}\left(\frac{q_{01}+q_{02}}{1-\rho} - 1\right) \tag{6.25}$$

where C_a^2 is given in Section 6.2 and C_S^2 is the squared coefficient of variation of the service time. The first term in (6.25) is equal to the bound given in Marshall[68] (cf. also Whitt[82]) for the mean queue length in a GI/G/1 queue with a DFR (decreasing failure rate) interarrival distribution. The mean waiting time follows by using Little's Law.

6.4. The computation of the quantities A_{jn}^{kl}

The first step in Algorithm 6.5 involves the computation of the quantities A_{jn}^{kl}. In this section, we focus on an exponential service distribution, to be considered as a representant of the class of phase type service distributions. The computational scheme we obtain can easily be generalized to cover more interesting phase type distributions; cf. Chapter 2.4. For deterministic service times, a computation scheme can be derived from the differential Equations (6.9) and (6.10). We omit details.

Let $F(t) = 1 - e^{-\mu}$ and assume that at epoch 0 the system is in state $(n,1)$. Then, a state transition occurs by an arrival, a service completion or a switch of the arrival rate. By the memoryless property of the exponential distribution, the first transition from state $(n,1)$ is caused by an arrival with probability $\lambda_1/(\lambda_1+\omega_1+\mu)$ and by a switching of the arrival rate with probability $\omega_1/(\lambda_1+\omega_1+\mu)$. Hence, we have for $1 \leqslant j \leqslant n$, $k = 1,2$

$$A_{jn}^{1k} = \frac{\lambda_1}{\lambda_1+\omega_1+\mu} A_{j+1,n}^{1k} + \frac{\omega_1}{\lambda_1+\omega_1+\mu} A_{jn}^{2k}$$

$$A_{jn}^{2k} = \frac{\omega_2}{\lambda_2+\omega_2+\mu} A_{jn}^{1k} + \frac{\lambda_2}{\lambda_2+\omega_2+\mu} A_{j+1,n}^{2k}$$

and

$$A_{nn}^{1k} = \frac{\delta_{1k}}{\lambda_1+\omega_1+\mu} + \frac{\omega_1}{\lambda_1+\omega_1+\mu} A_{nn}^{2k}$$

$$A_{nn}^{2k} = \frac{\delta_{2k}}{\lambda_2+\omega_2+\mu} + \frac{\omega_2}{\lambda_2+\omega_2+\mu} A_{nn}^{1k}$$

where $\delta_{ij} = 0$ if $i \neq j$ and $\delta_{ij} = 1$ if $i = j$. For fixed n and k, we solve A_{jn}^{1k} and A_{jn}^{2k} recursively for $j = n, n-1, \ldots, 1$ starting with A_{nn}^{1k} and A_{nn}^{2k}. As we have remarked before, A_{jn}^{kl} depends on j and n only through the difference $n-j$. In particular, $A_{jn}^{kl} = A_{1,n-j+1}^{kl}$ for $1 \leqslant j \leqslant n$, $k,l = 1,2$ and hence it is sufficient to compute A_{1n}^{kl}, $n \geqslant 1$, $k,l = 1,2$. The numbers A_{0n}^{1k} and A_{0n}^{2k} follow from (6.17) and (6.18).

6.5. The finite capacity case

In this section we discuss the SPP/G/1 queue with a finite capacity K. There are only $K-1$ waiting places and customers who see upon arrival K customers in the system are lost and do not influence the system.

With some minor modifications, the analysis of the infinite capacity model carries over to the finite model. Since obviously $p_{ni} = 0$ for $n > K$, $i = 1,2$ and $q_{ni} = 0$ for

$n \geqslant K$, $i = 1,2$, Equation (6.2) holds now only for $1 \leqslant n \leqslant K$, $i = 1,2$ and the Equations (6.3) and (6.4) only hold for $0 \leqslant n \leqslant K$. The average arrival rate of entering customers is not longer equal to λ^*, hence in (6.6) and (6.7) we replace λ^* by λ'. It will appear that we do not need λ' explicitly. The quantity $\rho = \lambda^* \mathbf{ES}$ is now interpreted as the offered traffic, whereas $\lambda' \mathbf{ES}$ is the carried traffic. Note that Lemma 6.7 remains valid for $1 \leqslant j \leqslant n \leqslant K-1$, $k,l = 1,2$ but that for $n = K$, $k,l = 1,2$

$$A_{jK}^{kl} = \int_0^\infty (1 - F(t)) \sum_{n=K}^{\infty} a_{n-j}^{kl}(t)\, dt \tag{6.26}$$

Indeed, to have K customers present at epoch t at least $K-j$ customers should arrive in $(0,t)$, including customers who will be blocked. From (6.26), we derive for $1 \leqslant j \leqslant K$, $k,l = 1,2$

$$A_{jK}^{kl} = \alpha^{kl}(1) - \sum_{n=j}^{K-1} A_{jn}^{kl} \tag{6.27}$$

and similarly for $k,l = 1,2$

$$A_{0K}^{kl} = \alpha_0^{kl}(1) - \sum_{n=j}^{K-1} A_{0n}^{kl} \tag{6.28}$$

Inserting (6.27) and (6.28) in (6.2) (with obviously $\mathbf{EN}_{K1} = \mathbf{EN}_{K2} = 0$) yields after some algebra for $i = 1,2$

$$\mathbf{ET}_{Ki} = \sum_{k=1}^{2} \{\mathbf{EN}_{0k}\, \alpha_0^{ki}(1) + \sum_{n=1}^{K-1} \mathbf{EN}_{nk}\, \alpha^{ki}(1) - \sum_{n=1}^{K-1} \mathbf{ET}_{ni}\} \tag{6.29}$$

Below we present the modified version of Algorithm 6.5 for the finite capacity SPP/G/1 queue. Since λ' is not a priori known, we compute the numbers $\lambda' q_{ni}$ instead of q_{ni}.

Algorithm 6.8

1. Evaluate the constants A_{jn}^{kl}, $0 \leqslant n \leqslant K-1$, $k,l = 1,2$
2. Compute p_{01} and p_{02} and from these $\lambda' q_{01}$ and $\lambda' q_{02}$ using (6.3) and (6.4), $n := 1$.
3. Assume that $p_{0i}, \ldots, p_{n-1,i}$, $\lambda' q_{0i}, \ldots, \lambda' q_{n-1,i}$ have been computed, solve the system of 4 equations as in Algorithm 6.5 with λ^* replaced by λ' for p_{n1}, p_{n2}, $\lambda' q_{n1}$ and $\lambda' q_{n2}$.
4. $n := n + 1$, return to step 3 if $n < K - 1$.
5. Compute p_{K1} and p_{K2} from (6.29)

An unanswered question in Algorithm 6.8 is how to compute the initial values p_{01} and p_{02}. The generating function technique, which provided a second equation for p_{01} and p_{02} in the infinite capacity model, does not work here any more. Moreover, since λ' is not a priori known we cannot use an analogon of Equation (6.7). To overcome this difficulty, we propose a numerical procedure to identify p_{01} and p_{02}. This numerical method was first applied to similar queueing models by Herzog, Woo and Chandy[75].

Step 2 in Algorithm 6.8

2a. Apply Algorithm 6.8 to compute (x_{ni}) instead of (p_{ni}) and use the initial values $x_{01}=1$, $x_{02}=0$. Compute $x_{.1}=\sum_{n=0}^{K} x_{n1}$, $x_{.2}=\sum_{n=0}^{K} x_{n2}$.

2b. Analogously, compute (y_{ni}) with initial values $y_{01}=0$, $y_{02}=1$. Compute $y_{.1}=\sum_{n=0}^{K} y_{n1}$, $y_{.2}=\sum_{n=0}^{K} y_{n2}$.

2c. Solve p_{01} and p_{02} from the following system of linear equations.

$$\omega_1(p_{01}x_{.1}+p_{02}y_{.1})=\omega_2(p_{01}x_{.2}+p_{02}y_{.2})$$

$$p_{01}(x_{.1}+x_{.2})+p_{02}(y_{.1}+y_{.2})=1$$

The first equation is in fact the balance equation $\omega_1\sum_{n=0}^{K} p_{n1}=\omega_2\sum_{n=0}^{K} p_{n2}$ and the second is the normalization equation $\sum_{n=0}^{K}(p_{n1}+p_{n2})=1$.

□

To see why the above procedure works, note that Algorithm 6.8 is linear in (p_{ni}). Hence, p_{ni} can be written as a linear combination of the initial values p_{01} and p_{02}. In the steps 2a and 2b, we compute precisely the coefficients x_{ni} and y_{ni} in p_{n1} and p_{n2} of p_{01} and p_{02} respectively. Next, in step 2c, the balance relation and the normalization equation provide 2 equations for p_{01} and p_{02}. Obviously, p_{ni} is found to be $p_{ni}=p_{01}x_{ni}+p_{02}y_{ni}$. Finally, note that the carried traffic equals $\lambda' \mathbf{ES}=1-p_{01}-p_{02}$ by Little's formula.

In the next section we discuss our numerical experience with the Algorithms 6.5 and 6.8.

6.6. Numerical results

The Algorithms 6.5 and 6.8 are not numerically stable for all values of the parameters λ_1, λ_2, ω_1 and ω_2. The instability is introduced by the use of the Equations (6.3) and (6.4) in the algorithms. Unfortunately, we have not yet obtained a clear insight when the algorithm can be used safely.

In this section we present numerical results for the delay probability

$$\Pi_W=1-\pi_{01}-\pi_{02}$$

and the mean queue length $\mathbf{EL}_q$ as function of the traffic intensity. We consider deterministic service times ($C_S^2=0$) and hyperexponential service times ($C_S^2\geqslant 1$, $F(t)=1-p_1e^{-\mu_1 t}-p_2e^{-\mu_2 t}$ with $p_1/\mu_1=p_2/\mu_2$). We have normalized the mean service time to one. For sake of convenience we write the limiting distribution of the interarrival time $G(t)$ (cf.(6.13)) as

$$G(t)=1-u_1e^{-r_1 t}-u_2e^{-r_2 t}$$

In the Figures 6.4-6.7 the arrival process is a renewal process, i.e. $\lambda_2=0$. For fixed C_a^2 and $\rho=\lambda^* \mathbf{ES}$, together with the additional condition $u_1/r_1=u_2/r_2$, we find all parameters of the arrival process as follows

$$\lambda_1=\frac{2C_a^2}{C_a^2+1}\lambda^*\ ,\quad \omega_1=\frac{C_a^2-1}{C_a^2(C_a^2+1)}\lambda^*\ ,\quad \omega_2=\frac{\lambda^*}{C_a^2}$$

Note that Π_W is decreasing in C_S^2 for fixed C_a^2 and increasing in C_a^2 for fixed C_S^2, whereas $\mathbf{EL}_q$ is increasing in both cases.

In the Figures 6.8, 6.10 and the Tables 6.9 and 6.11 we show the consequence of a renewal assumption concerning the arrival process. We have fixed C_S^2 and C_a^2 and have taken the autocorrelation coefficient θ as parameter. To specify the parameters of the arrival process we use again the condition $u_1/r_1=u_2/r_2$. The parameters λ_1 , λ_2 , ω_1 and ω_2 are displayed below the Tables 6.9 and 6.11. Beside our (exact) results we show the approximate results for the GI/G/1 queue according to Krämer and Langenbach-Belz[76] given by (for $C_a^2 \geqslant 1$)

$$\Pi_W(KLB)=\rho+(C_a^2-1)\frac{4\rho^2(1-\rho)}{C_a^2+\rho^2(4C_a^2+C_S^2)}$$

and

$$\mathrm{EL}_q(KLB)=\frac{\rho^2}{2(1-\rho)}(C_a^2+C_S^2)e^{-(1-\rho)\frac{C_a^2-1}{C_a^2+4C_S^2}}$$

These approximations are designed for the renewal case $\theta=0$ and they indeed perform well. Note that in particular EL_q is very sensitive for θ.

In the Figures 6.12, 6.14 and the Tables 6.13 and 6.15 we investigate the influence of the higher moments of the interarrival time on Π_W and EL_q. As parameter we have taken k defined as

$$k=\frac{u_1/r_1}{u_1/r_1+u_2/r_2}$$

Further we have fixed C_S^2, C_a^2 and $\lambda_2=0$. The parameters λ_1, λ_2, ω_1 and ω_2 are given below the Tables 6.13 and 6.15. The limiting case $k=0$ corresponds to a batch Poisson process as arrival process with a geometric batch size distribution and the case $k=1$ corresponds to Poisson input. This can clearly be recognized in the Figures 6.12 and 6.14. Note that Π_W is very sensitive for k.

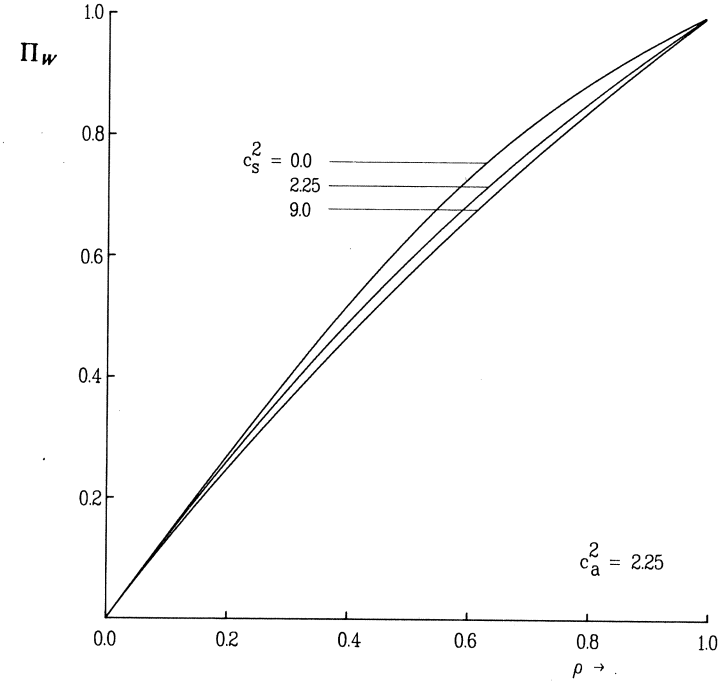

Figure 6.4

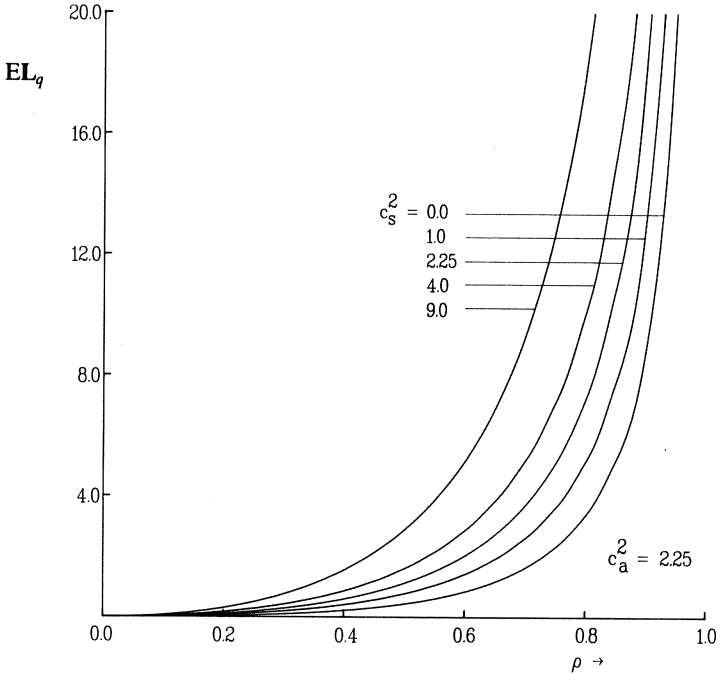

Figure 6.5

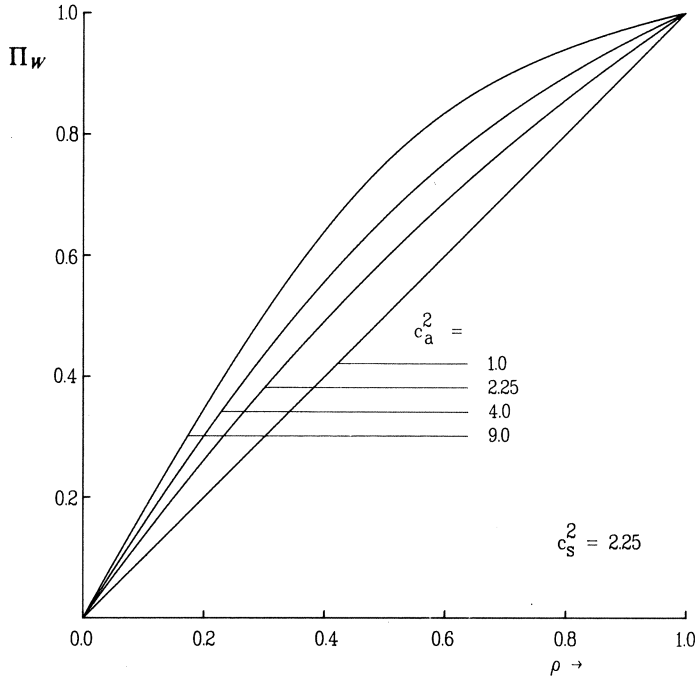

Figure 6.6

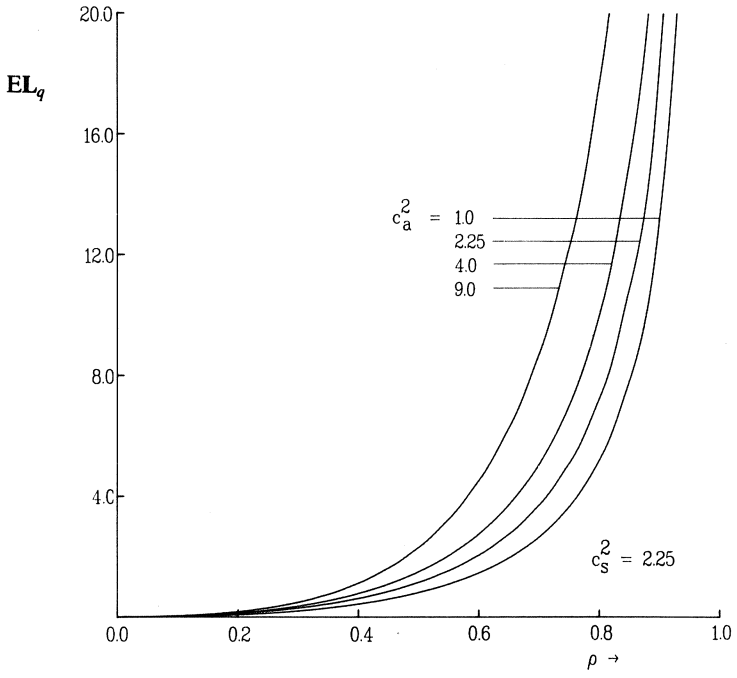

Figure 6.7

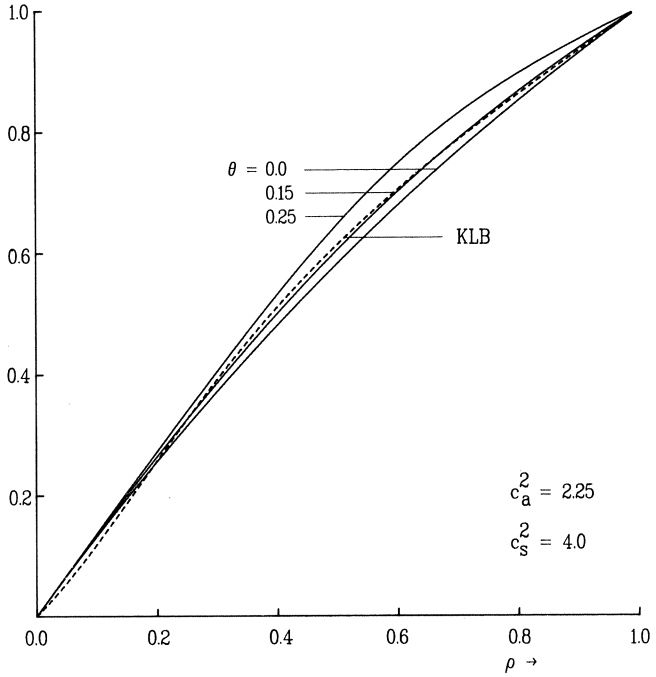

Figure 6.8 Sensitivity of Π_w in the case of correlation between successive arrivals

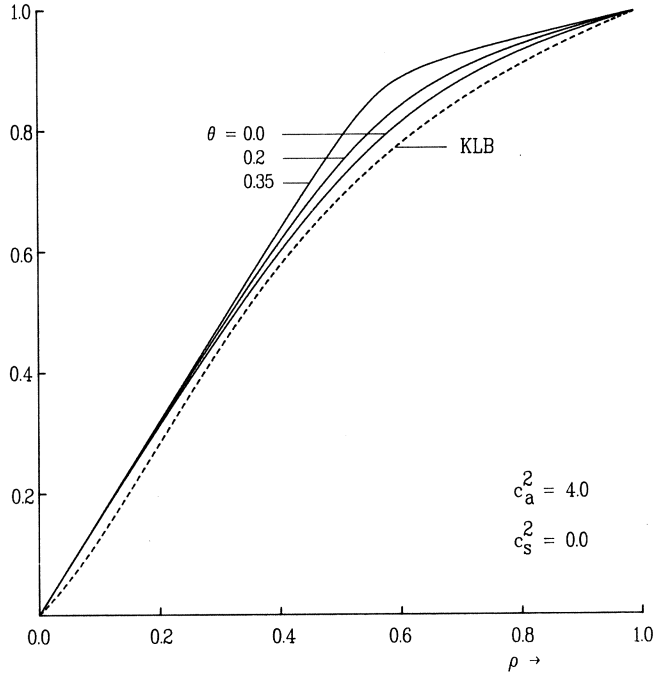

Figure 6.10 Sensitivity of Π_w in the case of correlation between successive arrivals

ρ / θ	0.00	0.05	0.10	0.15	0.20	0.25	KLB
0.10	0.0375	0.0380	0.0385	0.0391	0.0398	0.0405	0.0326
0.20	0.1657	0.1692	0.1734	0.1788	0.1856	0.1950	0.1479
0.30	0.4204	0.4318	0.4466	0.4669	0.4967	0.5462	0.3830
0.40	0.8629	0.8903	0.9278	0.9829	1.0736	1.2628	0.7998
0.50	1.6049	1.6616	1.7422	1.8679	2.0970	2.7036	1.5099
0.60	2.8696	2.9789	3.1395	3.4013	3.9193	5.6269	2.7365
0.70	5.1782	5.3867	5.7002	6.2298	7.3430	11.660	5.0004
0.80	10.093	10.517	11.164	12.285	14.743	25.249	9.8639
0.90	25.428	26.524	28.223	31.216	37.953	68.229	25.140
0.95	56.534	58.996	62.829	69.627	85.062	155.41	56.213
0.99	306.42	319.85	340.83	378.18	463.47	855.17	306.07
λ_1 / ρ	1.3846154	1.4172254	1.4537643	1.4945814	1.5399754	1.5901673	
λ_2 / ρ	0.0000000	0.0781592	0.1523895	0.2223416	0.2877169	0.3482943	
ω_1 / ρ	0.1709402	0.1572277	0.1373266	0.1100458	0.0742993	0.0292445	
ω_2 / ρ	0.4444444	0.3473876	0.2565196	0.1730312	0.0980084	0.0322939	

Table 6.9 Sensitivity of EL_q in the case of correlation between successive arrivals ($C_S^2=4.0, C_a^2=2.25$).

ρ / θ	0.00	0.05	0.15	0.20	0.30	0.35	KLB
0.10	0.0095	0.0095	0.0095	0.0096	0.0096	0.0097	0.0113
0.20	0.0459	0.0463	0.0472	0.0476	0.0486	0.0491	0.0549
0.30	0.1302	0.1324	0.1373	0.1402	0.1467	0.1506	0.1521
0.40	0.3051	0.3143	0.3371	0.3516	0.3907	0.4188	0.3401
0.50	0.6596	0.6932	0.7865	0.8552	1.1037	1.4033	0.6873
0.60	1.3720	1.4755	1.7948	2.0662	3.4515	7.3393	1.3335
0.70	2.8061	3.0735	3.9442	4.7343	9.3339	25.075	2.6085
0.80	5.9639	6.6020	8.7162	10.669	22.324	62.987	5.5085
0.90	15.830	17.616	23.560	29.075	62.132	177.78	15.029
0.95	35.775	39.866	53.499	66.154	142.06	407.70	34.771
0.99	195.73	218.29	293.45	363.25	782.01	2247.7	194.56
λ_1 / ρ	1.6000000	1.6204201	1.6638367	1.6868661	1.7356288	1.7613781	
λ_2 / ρ	0.0000000	0.0329133	0.0961633	0.1264673	0.1843712	0.2119553	
ω_1 / ρ	0.1500000	0.1354822	0.1016288	0.0821681	0.0379372	0.0131039	
ω_2 / ρ	0.2500000	0.2111844	0.1383712	0.1044986	0.0420628	0.0135628	

Table 6.11 Sensitivity of EL_q in the case of correlation between successive arrivals ($C_S^2=0.0, C_a^2=4.0$).

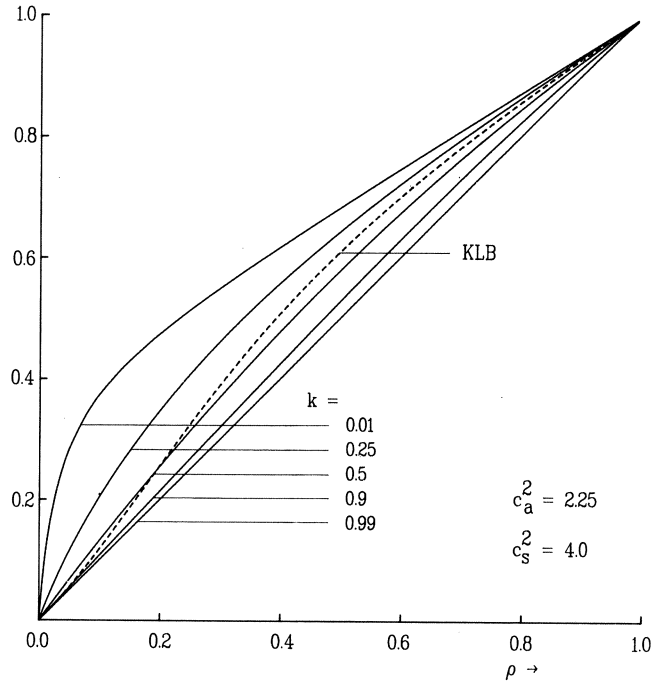

Figure 6.12 Sensitivity of Π_W for the third moment of the interarrival time.

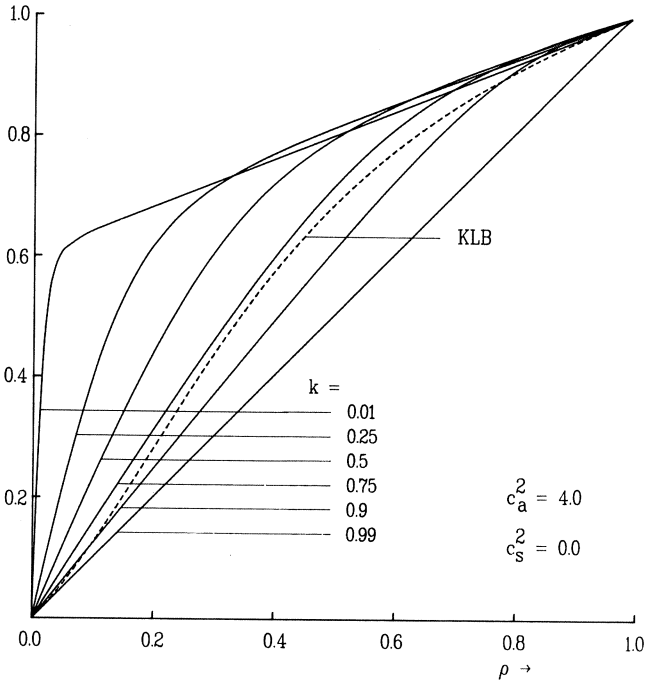

Figure 6.14 Sensitivity of Π_W for the third moment of the interarrival time.

ρ / k	0.01	0.25	0.50	0.75	0.90	0.99	KLB
0.10	0.0845	0.0444	0.0375	0.0331	0.0303	0.0281	0.0326
0.20	0.2670	0.1872	0.1657	0.1491	0.1370	0.1265	0.1479
0.30	0.5744	0.4607	0.4204	0.3844	0.3548	0.3258	0.3830
0.40	1.0681	0.9253	0.8629	0.8005	0.7419	0.6772	0.7998
0.50	1.8596	1.6916	1.6049	1.5082	1.4053	1.2735	1.5099
0.60	3.1720	2.9823	2.8696	2.7307	2.5615	2.3020	2.7365
0.70	5.5261	5.3177	5.1782	4.9889	4.7231	4.2060	5.0004
0.80	10.484	10.260	10.093	9.8459	9.4417	8.3431	9.8639
0.90	25.859	25.621	25.428	25.115	24.518	21.765	25.140
0.95	56.984	56.740	56.534	56.187	55.469	50.437	56.213
0.99	306.88	306.63	306.42	306.04	305.22	295.79	306.07
λ_1 / ρ	16.1100473	1.7757869	1.3846154	1.1906825	1.0852147	1.0097901	
λ_2 / ρ	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	
ω_1 / ρ	22.6753926	0.5422681	0.1709402	0.0488591	0.0107062	0.0001519	
ω_2 / ρ	1.5006831	0.6989910	0.4444444	0.2562329	0.1256374	0.0155123	

Table 6.13 Sensitivity of EL_q for the third moment of the interarrival time ($C_S^2=4.0, C_a^2=2.25$).

ρ / k	0.01	0.25	0.50	0.75	0.90	0.99	KLB
0.10	0.1472	0.0155	0.0095	0.0071	0.0062	0.0056	0.0113
0.20	0.3749	0.0818	0.0459	0.0333	0.0281	0.0253	0.0549
0.30	0.6821	0.2461	0.1302	0.0893	0.0736	0.0652	0.1521
0.40	1.1082	0.5703	0.3051	0.1965	0.1562	0.1356	0.3401
0.50	1.7249	1.1250	0.6596	0.3995	0.3027	0.2550	0.6873
0.60	2.6749	2.0388	1.3720	0.8083	0.5725	0.4613	1.3335
0.70	4.2916	3.6337	2.8061	1.7392	1.1279	0.8442	2.6085
0.80	7.5749	6.9035	5.9639	4.2706	2.5904	1.6820	5.5085
0.90	17.525	16.845	15.830	13.601	9.3665	4.4786	15.029
0.95	37.500	36.817	35.775	33.351	27.543	11.055	34.771
0.99	197.48	196.79	195.73	193.19	186.41	112.84	194.56
λ_1 / ρ	37.1136678	2.4278775	1.6000000	1.2521225	1.0983154	1.0099685	
λ_2 / ρ	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	
ω_1 / ρ	23.4270747	0.5598400	0.1500000	0.0338443	0.0058671	0.0000656	
ω_2 / ρ	0.6487038	0.3920784	0.2500000	0.1342374	0.0596765	0.0065801	

Table 6.15 Sensitivity of EL_q for the third moment of the interarrival time ($C_S^2=0.0, C_a^2=4.0$).

Appendix A

The Poisson Lemma

In this monograph we shall frequently use Poisson processes as components of the queueuing models to be analyzed. In the Chapters 1 and 3 we consider a Poisson arrival process of customers with uniform rate λ, whereas in the Chapters 2 and 4 the arrival intensity varies with the state of the system. That is, the intensity is λ_j when j customers are present. The batch arrival process studied in Chapter 5 is a compound Poisson process. Batches of customers arrive in a Poisson stream with uniform rate λ and the batch size distribution is state dependent. Finally, in Chapter 6 a so called switched Poisson process is introduced to describe the arrival process. Here the arrival rate is alternately λ_1 and λ_2 governed by some random mechanism, which is independent of the number of customers in the system. Note that all the above mentioned arrival processes are Markovian of nature.

One crucial step in the analysis of the queueing models defined in these chapters is due to a property of Poisson processes. We derive below the 'Poisson Lemma' which states this property. To prove the Poisson Lemma, we copy the derivation given in Wolff[82] with only some small modifications. Incidentally, the Poisson Lemma is the crucial step in Wolff's proof of 'Poisson arrivals see time averages' and is somewhat obscurely hidden in his paper.

Let $\mathbf{N}=\{\mathbf{N}(t)\,,\ t\geqslant 0\}$ be a stochastic process and $\mathbf{A}=\{\mathbf{A}(t)\,,\ t\geqslant 0\}$ be a Poisson process at rate λ, both defined on some probability space $(\Omega\,,\ \boldsymbol{F}\,,\ P)$. $\mathbf{N}(t)$ represents the status of a system at $t\geqslant 0$ and $\mathbf{A}$ an arrival process of customers to the system. For example, if $\mathbf{N}(t)$ is the number of customers in the system at epoch t, it will increase with unit jumps at customer arrival epochs. In general, we let $\mathbf{N}(t)$ take on values in an arbitrary measurable space and the interaction between $\mathbf{A}$ and $\mathbf{N}$ is unspecified.

For an arbitrary set $\mathbf{B}$ in the outcome space of $\mathbf{N}$ (called the state space) such that $\{\mathbf{N}(t)\in\mathbf{B}\}$ is measurable for every $t\geqslant 0$, define

$$\mathbf{U}(t)=\begin{cases}1\,, & \text{if } \mathbf{N}(t)\in\mathbf{B}\\ 0\,, & \text{otherwise}\end{cases}$$

$$\mathbf{Y}(t)=\int_0^t \mathbf{U}(s)\,\mathrm{d}\mathbf{A}(s)$$

We assume that the sample paths of $\mathbf{U}$ are left continuous and have right hand limits with probability one and that $(\omega,t)\in\Omega\times[0,\infty)$. With these definitions $\mathbf{Y}(t)$ is the number of arrivals in the interval $[0,t]$ while the process $\mathbf{N}$ is in state $\mathbf{B}$.

Note that with probability one $\mathbf{U}$ has only a finite number of discontinuities on any finite interval. Hence, for each ω in a set that has probability one, $\mathbf{Y}(t)$ can be approximated arbitrarily closely by a function of the form

$$\mathbf{Y}_n(t)=\sum_{k=0}^{n-1}\mathbf{U}\left(\frac{kt}{n}\right)\left\{\mathbf{A}\left(\frac{(k+1)t}{n}\right)-\mathbf{A}\left(\frac{kt}{n}\right)\right\} \tag{A.1}$$

for sufficienly large n. Although we did not specify the interaction between $\mathbf{A}$ and $\mathbf{N}$, these processes are typically dependent, especially in a queueing context. However,

we assume that the system has no anticipation, i.e. we do not want the future increments of **A** to depend on the past of **U**. We make the

Lack of Anticipation Assumption A.1

For each $t \geqslant 0$ $\{\mathbf{A}(t+u)-\mathbf{A}(t)\,,\ u \geqslant 0\}$ and $\{\mathbf{U}(s)\,,\ 0 \leqslant s \leqslant t\}$ are independent. □

We now give the lemma that is of crucial interest for this monograph.

The Poisson Lemma A.2

Under the lack of anticipation assumption

$$\mathrm{E}\mathbf{Y}(t)=\lambda \mathrm{E}\int_0^t \mathbf{U}(s)\,\mathrm{d}s \text{ for any } t \geqslant 0 \tag{A.2}$$

i.e. on any finite interval $[0,t]$, the expected number of arrivals who find the system in state B is equal to the arrival rate times the expected length of time the system stays in state B during $[0,t]$.

Proof Note that $\mathrm{E}\{\mathbf{A}(\frac{(k+1)t}{n})-\mathbf{A}(\frac{kt}{n})\}=\lambda t/n$ since **A** is a Poisson process. The use of the lack of anticipation assumption in Equation (A.1) implies

$$\mathrm{E}\mathbf{Y}_n(t)=\lambda \mathrm{E}\sum_{k=0}^{n-1}\frac{t}{n}\mathbf{U}(\frac{kt}{n})$$

Now, we let $n \to \infty$ and apply the dominated convergence theorem ($\mathbf{U}(kt/n) \leqslant 1$)

$$\mathrm{E}\mathbf{Y}(t)=\lim_{n\to\infty}\mathrm{E}\mathbf{Y}_n(t)=\lambda \mathrm{E}\int_0^t \mathbf{U}(s)\,\mathrm{d}s$$

□

Remark A.3 The assumptions made to prove the Poisson Lemma are rather general. Notably, the interaction between the processes **A** and **N** remained unspecified. Therefore the Poisson Lemma can be applied in more situations than a first glance at the conditions of the lemma suggests, and in particular the lemma can be applied in situations seemingly not satisfying these assumptions. To illustrate this we consider the following

□

Example A.4

Let **N**, **A** and B be defined as in the problem formulation of this section and define an 'interrupted' Poisson process $\mathbf{A}'(t)$ by

$\mathbf{A}'(t)$ is a Poisson process at rate $\lambda > 0$ if $\mathbf{N}(t) \in \mathrm{B}$

$\mathbf{A}'(t)$ is a Poisson process at rate 0 if $\mathbf{N}(t) \notin \mathrm{B}$ (if $\mathbf{N}(t) \notin \mathrm{B}$ then $\mathbf{A}'(t)$ produces no arrivals)

We compare the following two systems. The first system is described by the triple (**N**,**A**,B) with the specification that arriving customers finding $\mathbf{N}(t) \in \mathrm{B}$ enter the system and otherwise are lost. The second system is described by the triple (**N**,**A**′,B) with the specification that all arrivals of **A**′ enter the system. Obviously, both systems behave identically, though an outside observer sees the difference. Indeed, in the first system blocking may occur, in the second not. Hence,

$$E\mathbf{Y}'(t)=E\mathbf{Y}(t),$$

where $\mathbf{Y}'(t)=\int_0^t \mathbf{U}(s)\mathrm{d}\mathbf{A}'(s)$. Note that $\mathbf{A}'(t)$ satisfies the lack of anticipation assumption.

□

Some results from renewal theory

Below we review some results from renewal theory. For more details see e.g. Ross[70] and Feller[66,68].

Let $\{\mathbf{X}_n\ ,\ n\geqslant 1\}$ be a sequence of nonnegative, independent and identically distributed random variables, distributed as a random variable $\mathbf{X}$ with distribution function $F(x)=\Pr\{\mathbf{X}\leqslant x\}$. To avoid trivialities we assume that $F(0)<1$ and $E\mathbf{X}<\infty$. Let

$$\mathbf{S}_0=0\ ,\ \mathbf{S}_n=\sum_{i=1}^{n}\mathbf{X}_i\ ,\ n\geqslant 1 \text{ and define } \mathbf{N}(t)=sup\ \{n:\mathbf{S}_n\leqslant t\}$$

It follows from the strong law of large numbers that $\mathbf{S}_n/n\to E\mathbf{X}$ with probability one. Hence, $\mathbf{S}_n\leqslant t$ only finitely often and so $\mathbf{N}(t)<\infty$ with probability one. The process $\{\mathbf{N}(t)\ ,\ t\geqslant 0\}$ is said to be a renewal process. We say that a renewal occurs at t if $\mathbf{S}_n=t$ for some n. A particular important renewal process is the well known Poisson process; cf. Ross[70]. Define the renewal function $M(t)$ by

$$M(t)=E\mathbf{N}(t)$$

i.e. $M(t)$ is the expected number of renewals in $(0,t]$. The following theorems are well known; cf. Feller[66] and Ross[70].

Theorem A.5

$$M(t)=\sum_{n=1}^{\infty}F^{n*}(t)$$

where $F^{n*}(t)$ is the n-fold convolution of F with itself. Further,

$$M(x)=F(x)+\int_0^t M(x-y)\,\mathrm{d}F(y)\ ,\ y\geqslant 0$$

□

Theorem A.6 If $g(t)$ satisfies the renewal-type equation

$$g(t)=h(t)+\int_0^t g(t-x)\,\mathrm{d}F(x)$$

where $h(t)$ is a given function, which is bounded on finite intervals, then

$$g(t)=h(t)+\int_0^t h(t-x)\,\mathrm{d}M(x)$$

□

Theorem A.7 (Key Renewal Theorem)
If $h(t)$ is directly Riemann integrable and $\mathbf{X}$ is non lattice (i.e. there exists no number $d>0$ with $\Sigma_{n=0}^{\infty}\Pr\{\mathbf{X}=nd\}=1$), then

$$\lim_{t\to\infty}\int_0^t h(t-x)\,\mathrm{d}M(x)=\frac{1}{\mathrm{E}\mathbf{X}}\int_0^{\infty}h(t)\,\mathrm{d}t$$

□

We refer to Feller[68] for the discrete version of Theorem A.7 when the random variable $\mathbf{X}$ is lattice.

Theorem A.8 Suppose $(f_n)_{n=1}^{\infty}$ is a discrete probability distribution. Let the renewal quantities m_n be defined by

$$m_n=f_n+\sum_{i=0}^{n-1}m_i f_{n-i}\ ,\quad n\geqslant 0$$

and let $(g_n)_{n=0}^{\infty}$ satisfy the discrete renewal equations

$$g_n=h_n+\sum_{i=0}^{n-1}g_i f_{n-i}\ ,\quad n\geqslant 0$$

where (h_n) is a sequence of finite numbers, then

$$g_n=h_n+\sum_{i=0}^{n-1}h_i m_{n-i}$$

Regenerative processes

Consider a stochastic process $\{\mathbf{X}(t)\ ,\ t\geqslant 0\}$ with state space $\{0,1,2,...\}$ having the property that there exist time points at which the process (probabilistically) restarts itself. That is, suppose that with probability one there exist a time $\mathbf{T}_1$ such that the continuation of the process beyond $\mathbf{T}_1$ is a probabilistic replica of the whole process starting at 0. Note that this property implies the existence of further times $\mathbf{T}_2,\mathbf{T}_3,...$ having the same property as $\mathbf{T}_1$. Such a stochastic process is known as a regenerative process; see e.g. Ross[70], Stidham[72]. It follows that $\{\mathbf{T}_1,\mathbf{T}_2,...\}$ forms a renewal process and we say that a cycle is completed every time a renewal occurs.

Theorem A.9 If $\mathbf{T}_1$ has an absolutely continuous component and $\mathrm{E}\mathbf{T}_1<\infty$, then for all $j\geqslant 0$

$$\lim_{t\to\infty}\Pr\{\mathbf{X}(t)=j\}=\frac{\mathrm{E}\{\text{ amount of time in state } j \text{ during one cycle }\}}{\mathrm{E}\{\text{ length of one cycle }\}}$$

□

Theorem A.10 (Wald's equation)

If $\{\mathbf{X}_n\ ,\ n\geqslant 1\}$ is a sequence of independent and identically distributed random variables, distributed as a random variable $\mathbf{X}$ with $\mathrm{E}\mathbf{X}<\infty$ and if $\mathbf{N}$ is a stopping time for $\mathbf{X}_1,\mathbf{X}_2,...$, such that $\mathrm{E}\mathbf{N}<\infty$, then

$$\mathrm{E}\sum_{i=1}^{\mathbf{N}}\mathbf{X}_i=\mathrm{E}\mathbf{N}\mathrm{E}\mathbf{X}_1$$

$\mathbf{N}$ is a stopping time if the event $\{\mathbf{N}=n\}$ is independent of $\mathbf{X}_{n+1},\mathbf{X}_{n+2},...,\ n\geqslant 1$.

Appendix B

Appendix B Some numerical auxiliary routines

In this appendix, we present several numerical auxiliary routines we have used to obtain our numerical results. These routines are easy to implement and we recommend to use them for the range of applications covered in this monograph. Clearly, our recommendation does not hold beyond the scope of this monograph.

Two quadrature procedures

In numerical analysis, various procedures are available to compute the integral of a known function over a finite or infinite interval. Below we give details of two so-called 'quadrature rules of Gauss type', which are particularly suited for the evaluation of the integrals we are interested in.

Let $f(x)$ and $w(x)$ be functions defined on the interval $[a,b]$. A Gauss quadrature rule of order n has the following form

$$\int_a^b f(x)w(x)\,dx = \sum_{j=1}^{n} w_j f(x_j) + E_n(f,w) \tag{B.0}$$

where x_j, $1 \leqslant n$ are the zeros of the n^{th} orthonormal polynomial associated with $w(x)$ and w_j, $1 \leqslant j \leqslant n$ are appropriately chosen weights such that $E_n(f,w)=0$ if f is a polynomial of degree less than $2n$. We refer to Davis and Rabinovitz[67] for the theory behind these quadrature rules and for the determination of the orthonormal polynomials and the associated weights. The abscissae and weights, given in the Table B.1 have been computed using a numerical procedure from the Numal library; cf. Hemker[81]. By neglecting the error term $E_n(f,w)$ we obtain an approximate integration formula for the function $f(x)w(x)$ which is easy to implement. If $f(x)$ is smooth enough, the error incurred by neglecting $E_n(f,w)$ is very small.

Gauss Laguerre quadrature

For the special choice $w(x)=e^{-x}$, $a=0$ and $b=\infty$, we obtain the Gauss Laguerre quadrature rule. In Table B.1, we have displayed the abscissae x_j and the numbers $\ln(w_j)$ for $j=1,\ldots,n$ and $n=64$. Since for given j the situation may occur that w_j is very small while $f(x_j)$ is very large, it is better to rewrite (B.0) in the following form in order to prevent computer overflow.

$$\int_0^\infty f(x)e^{-x}\,dx \approx \sum_{j=1}^{n} \exp\{\ln(w_j)+\ln(f(x_j))\} \tag{B.1}$$

Using the numbers of Table B.1, the integration rule B.1 is exact for polynomial functions $f(x)$ of degree up to 127.

	Gauss-Laguerre weights		Gauss-Legendre weights	
n	x_n	$\ln(w_n)$	x_n	w_n
1	2.3480957917134 10^{2}	-2.3182437850894 10^{2}	4.8782485366824 10^{-1}	4.8690957009140 10^{-2}
2	2.1803185193534 10^{2}	-2.1534789782750 10^{2}	4.6350343910606 10^{-1}	4.8575467441507 10^{-2}
3	2.0467202848507 10^{2}	-2.0216260196083 10^{2}	4.3926859035193 10^{-1}	4.8344762234803 10^{-2}
4	1.9315113603708 10^{2}	-1.9076805878736 10^{2}	4.1517778978800 10^{-1}	4.7999388596462 10^{-2}
5	1.8285820469144 10^{2}	-1.8057598116541 10^{2}	3.9128817812998 10^{-1}	4.7540165714832 10^{-2}
6	1.7347494683646 10^{2}	-1.7127779920295 10^{2}	3.6765641889560 10^{-1}	4.6968182816212 10^{-2}
7	1.6480860265513 10^{2}	-1.6268582454498 10^{2}	3.4433856400488 10^{-1}	4.6284796581316 10^{-2}
8	1.5673107513271 10^{2}	-1.5467494840814 10^{2}	3.2138992083114 10^{-1}	4.5491627927418 10^{-2}
9	1.4915166590003 10^{2}	-1.4715638562595 10^{2}	2.9886492101798 10^{-1}	4.4590558163756 10^{-2}
10	1.4200312148997 10^{2}	-1.4006418445125 10^{2}	2.7681699137324 10^{-1}	4.3583724529325 10^{-2}
11	1.3523378794950 10^{2}	-1.3334761898845 10^{2}	2.5529842714645 10^{-1}	4.2473515123650 10^{-2}
12	1.2880287876921 10^{2}	-1.2696658953300 10^{2}	2.3436026799003 10^{-1}	4.1262563242624 10^{-2}
13	1.2267746026857 10^{2}	-1.2088868651852 10^{2}	3.4747913212030 10^{-4}	1.7832807217413 10^{-3}
14	1.1683044505128 10^{2}	-1.1508723302485 10^{2}	1.8299416140337 10^{-3}	4.1470332605916 10^{-3}
15	1.1123920752448 10^{2}	-1.0953993223793 10^{2}	4.4933142616337 10^{-3}	6.5044579689874 10^{-3}
16	1.0588459946883 10^{2}	-1.0422790456449 10^{2}	8.3318730576956 10^{-3}	8.8467598263756 10^{-3}
17	1.0075023196946 10^{2}	-9.9134984540436 10^{1}	1.3336586105051 10^{-2}	1.1168139460140 10^{-2}
18	9.5821940015556 10^{1}	-9.4247196120305 10^{1}	1.9495600173983 10^{-2}	1.3463047896727 10^{-2}
19	9.1087375613167 10^{1}	-8.9552353599709 10^{1}	2.6794312570804 10^{-2}	1.5726030476030 10^{-2}
20	8.6535693349431 10^{1}	-8.5039753007138 10^{1}	3.5215413934036 10^{-2}	1.7951715775701 10^{-2}
21	8.2157303778352 10^{1}	-8.0699929924172 10^{1}	4.4738931460753 10^{-2}	2.0134823153533 10^{-2}
22	7.7943677434410 10^{1}	-7.6524466936750 10^{1}	5.5342277002449 10^{-2}	2.2270173808387 10^{-2}
23	7.3887187232516 10^{1}	-7.2505838748843 10^{1}	2.1405217689868 10^{-1}	3.9953741132719 10^{-2}
24	6.9980980377179 10^{1}	-6.8637286273331 10^{1}	6.7000300922960 10^{-2}	2.4352702568714 10^{-2}
25	6.6218873251250 10^{1}	-6.4912713308123 10^{1}	1.9442232241380 10^{-1}	3.8550153178614 10^{-2}
26	6.2595264400126 10^{1}	-6.1326601014306 10^{1}	7.9685351873714 10^{-2}	2.6377469715058 10^{-2}
27	5.9105061918992 10^{1}	-5.7873936579321 10^{1}	9.3367342438604 10^{-2}	2.8339672614263 10^{-2}
28	5.5743622413340 10^{1}	-5.4550153293154 10^{1}	1.7551726437267 10^{-1}	3.7055128540241 10^{-2}
29	5.2506699341321 10^{1}	-5.1351079891438 10^{1}	1.0801382052833 10^{-1}	3.0234657072405 10^{-2}
30	4.9390399025600 10^{1}	-4.8272897489907 10^{1}	1.5738184347288 10^{-1}	3.5472213256884 10^{-2}
31	4.6391142978678 10^{1}	-4.5312102785146 10^{1}	1.4005907491419 10^{-1}	3.3805161837143 10^{-2}
32	4.3505635466396 10^{1}	-4.2465476473522 10^{1}	1.2359004636973 10^{-1}	3.2057928354854 10^{-2}
33	4.0730835444433 10^{1}	-3.9730056048317 10^{1}	8.7640995363027 10^{-1}	3.2057928354854 10^{-2}
34	3.8063932165649 10^{1}	-3.7103112297503 10^{1}	8.5994092508581 10^{-1}	3.3805161837143 10^{-2}
35	3.5502323891173 10^{1}	-3.4582128960368 10^{1}	8.4261815652712 10^{-1}	3.5472213256884 10^{-2}
36	3.3043599236411 10^{1}	-3.2164785101613 10^{1}	8.9198617947167 10^{-1}	3.0234657072405 10^{-2}
37	3.0685520767529 10^{1}	-2.9848939849485 10^{1}	8.2448273562733 10^{-1}	3.7055128540241 10^{-2}
38	2.8426010527532 10^{1}	-2.7632619213515 10^{1}	9.0663265756140 10^{-1}	2.8339672614263 10^{-2}
39	2.6263137227092 10^{1}	-2.5514004760939 10^{1}	9.2031464812629 10^{-1}	2.6377469715058 10^{-2}
40	2.4195104875935 10^{1}	-2.3491423982025 10^{1}	8.0557767758620 10^{-1}	3.8550153178614 10^{-2}

Table B.1 The abscissae and weights of two quadrature rules of Gauss type.

n	Gauss-Laguerre weights		Gauss-Legendre weights	
	x_n	$\ln(w_n)$	x_n	w_n
41	2.2220242665953 10^{1}	-2.1563342224120 10^{1}	9.3299969907704 10^{-1}	2.4352702568714 10^{-2}
42	2.0336995948761 10^{1}	-1.9728356125174 10^{1}	7.8594782310132 10^{-1}	3.9953741132719 10^{-2}
43	1.8543918170890 10^{1}	-1.7985188524692 10^{1}	9.4465772299755 10^{-1}	2.2270173808387 10^{-2}
44	1.6839663652650 10^{1}	-1.6332684889832 10^{1}	9.5526106853925 10^{-1}	2.0134823153533 10^{-2}
45	1.5222981111526 10^{1}	-1.4769811360914 10^{1}	9.6478458606596 10^{-1}	1.7951715775701 10^{-2}
46	1.3692707845520 10^{1}	-1.3295654608594 10^{1}	9.7320568742920 10^{-1}	1.5726030476030 10^{-2}
47	1.2247764504275 10^{1}	-1.1909423813103 10^{1}	9.8050439982602 10^{-1}	1.3463047896727 10^{-2}
48	1.0887150383888 10^{1}	-1.0610455239124 10^{1}	9.8666341389495 10^{-1}	1.1168139460140 10^{-2}
49	9.6099391927968 10^{0}	-9.3982201228247 10^{0}	9.9166812694230 10^{-1}	8.8467598263756 10^{-3}
50	8.4152752394547 10^{0}	-8.2723369446582 10^{0}	9.9550668573837 10^{-1}	6.5044579689874 10^{-3}
51	7.3023700026174 10^{0}	-7.2325897239877 10^{0}	9.9817005838597 10^{-1}	4.1470332605916 10^{-3}
52	6.2704990468662 10^{0}	-6.2789548611926 10^{0}	9.9965252086788 10^{-1}	1.7832807217413 10^{-3}
53	5.3189992545552 10^{0}	-5.4116405303448 10^{0}	7.6563973200997 10^{-1}	4.1262563242624 10^{-2}
54	4.4472663433132 10^{0}	-4.6311451432444 10^{0}	7.4470157285355 10^{-1}	4.2473515123650 10^{-2}
55	3.6547526501790 10^{0}	-3.9383459116074 10^{0}	7.2318300862676 10^{-1}	4.3583724529325 10^{-2}
56	2.9409651567547 10^{0}	-3.3346369368340 10^{0}	7.0113507898202 10^{-1}	4.4590558163756 10^{-2}
57	2.3054637393076 10^{0}	-2.8221528889635 10^{0}	6.7861007916886 10^{-1}	4.5491627927418 10^{-2}
58	1.7478596260595 10^{0}	-2.4041494787846 10^{0}	6.5566143599512 10^{-1}	4.6284796581316 10^{-2}
59	1.2678140407463 10^{0}	-2.0856926100797 10^{0}	6.3234358110440 10^{-1}	4.6968182816212 10^{-2}
60	8.6503700464830 10^{-1}	-1.8750137659721 10^{0}	6.0871182187002 10^{-1}	4.7540165714832 10^{-2}
61	2.2415874175622 10^{-2}	-2.8778987074406 10^{0}	5.8482221021200 10^{-1}	4.7999388596462 10^{-2}
62	1.1812251208346 10^{-1}	-2.1284302322291 10^{0}	5.6073140964807 10^{-1}	4.8344762234803 10^{-2}
63	2.9036574400281 10^{-1}	-1.8483526536354 10^{0}	5.3649656089394 10^{-1}	4.8575467441507 10^{-2}
64	5.3928622122800 10^{-1}	-1.7864910691569 10^{0}	5.1217514633176 10^{-1}	4.8690957009140 10^{-2}

Table B.1 (Continued)

Gauss Legendre quadrature

For the special choice $w(x)=1$, $a=0$ and $b=1$, we obtain the Gauss Legendre quadrature rule

$$\int_0^1 f(x)\,dx \approx \sum_{j=1}^{n} w_j f(x_j) \tag{B.2}$$

The abscissae x_j and the weights w_j for $j=1,...,n$ and $n=64$ are given in Table B.1. Hence, the Formula (B.2) is exact for polynomial functions $f(x)$ of degree up to 127.

The numerical solution of an integral equation

Consider the Volterra integral equation of the second kind

$$f(t)=g(t)+\int_0^t f(x)k(t-x)\,dx \tag{B.3}$$

to be solved for $f(t)$, where $g(t)$ and $k(t)$ are known differentiable functions. Choose a step length h and let f_n denote $f(nh)$, etc. In addition to $f_0(=g_0)$, a second initial

value f_1 can be obtained using Day's starting procedure. Letting $g_{1/2}=g(h/2)$ and $k_{1/2}=k(h/2)$, define the numbers

$$a_1=g_1+hg_0k_1$$
$$a_2=g_1+h(g_0k_1+a_1k_0)/2$$
$$a_3=g_{1/2}+h(g_0k_{1/2}+\tfrac{1}{2}g_0k_0+\tfrac{1}{2}a_2k_0)/4$$

then

$$f_1=g_1+h(g_0k_1+4a_3k_{1/2}+a_2k_0)/6$$

Next, to solve (B.3), we rewrite the integral as a finite sum using repeatedly Simpson's rule. Therefore, we distinguish between n even and n odd. For n even, we get

$$f_n=g_n+\frac{1}{3}h\sum_{j=0}^{n}d_{nj}f_jk_{n-j} \tag{B.4}$$

and for n odd

$$f_n=g_n+\frac{1}{3}h\sum_{j=0}^{n-3}d_{n-3,j}f_jk_{n-j} \tag{B.5}$$
$$+\frac{3}{8}h(f_{n-3}k_3+3f_{n-2}k_2+3f_{n-1}k_1+f_nk_0)$$

where $d_{nj}=3-(-1)^j$, $1\leqslant j\leqslant n-1$, $d_{n0}=d_{nn}=1$ are the weights of Simpson's integration rule. Once $f_0,\ldots,f_{n-1}$ have been found, f_n is computed from (B.4) if n is even and from (B.5) if n is odd. Hence, for a fixed step length h, the function values $f(nd)$ are computed by a simple, recursive scheme. By repeatedly halving the step length, $f(t)$ can in principle be computed in any desired accuracy. For more details, see Delves and Walsh[74].

The numerical solution of a differential equation

Consider the ordinary differential equation

$$\frac{dy}{dx}=f(x,y),\quad y(0)=y_0 \tag{B.6}$$

to be solved for y, where $f(x,y)$ is a known continuous function in x and y. Choose a step length h and let y_n denote $y(nh)$ and x_n denote nh. If y_n is known, we compute y_{n+1} using a Runge Kutta step. Define the numbers

$$k_1=hf(x_n,y_n)$$
$$k_2=hf(x_n+\tfrac{1}{2}h,y_n+\tfrac{1}{2}k_1)$$
$$k_3=hf(x_n+\tfrac{1}{2}h,y_n+\tfrac{1}{2}k_2)$$
$$k_4=hf(x_n+h,y_n+k_3)$$

then

$$y_{n+1}=y_n+h(k_1+2k_2+2k_3+k_4)/6 \tag{B.7}$$

Hence, for a fixed step length h, the function values $y(nh)$ are computed by this straightforward scheme. By successive halving of h, any prescribed accuracy can in principle be obtained; cf. e.g. Stoer and Boelirsch[80].

Appendix C

Numerical aspects concerning the M/G/c queue

In this appendix, we focus on the computational aspects of the algorithm and formulae presented in Chapter 3, which chapter deals with an approximate analysis for the M/G/c queue. The gap between application-oriented research and its actual execution in practice is in general larger than expected at first sight. The theorist is often only willing to do that minimum amount of numerical work needed to support his theoretical methods, partly due to the fact that numerical validation is not always rewarding in academic circles. The potential user is willing to spend some time to explore the computational possiblities of a research paper, but time constraints or lack of skill in numerical analysis often prevent him from applying the methods of the research paper if this paper leaves the numerical details to the reader.

In our opinion, when presenting an algorithm one should in principle give the details of each step and indicate for which parameter range the algorithm works well and is reliable. A statement like 'this integral equation can be numerically solved by standard methods' is not helpful to someone who has never solved an integral equation. A suggestion for the method to be used should be welcome, but also comments on the special cases in which numerical difficulties may arise.

Here, we follow the intermediate approach of giving the relevant numerical details to the reader. First, we discuss the numerical details of the algorithm for the state probabilities and for the formulae for the moments of the queue length. Next, we turn to the integral equation for the waiting time distribution.

The state probabilities and the moments of the queue length in the M/G/c queue

For completeness, we repeat the approximation formulae for the state probabilities (see Theorem 3.4).

$$p_n = \frac{(\lambda ES)^n}{n!} p_0 , \quad 0 \leqslant n \leqslant c-1$$

$$p_n = \lambda p_{c-1} \alpha_{n-c} + \lambda \sum_{j=c}^{n} p_j \beta_{n-j} , \quad n \geqslant c$$

where

$$\alpha_k = \int_0^\infty (1-F_e(t))^{c-1} (1-F(t)) e^{-\lambda t} \frac{(\lambda t)^k}{k!} \mathrm{d}t$$

$$\beta_k = \int_0^\infty (1-F(ct)) e^{-\lambda t} \frac{(\lambda t)^k}{k!} \mathrm{d}t$$

$$p_0 = \frac{1}{\sum_{i=0}^{c-1} \frac{(\lambda ES)^i}{i!} + \frac{(\lambda ES)^c}{c!(1-\rho)}}$$

Also, the delay probability and the first two moments of the queue size are given by

$$P_W = \frac{\rho}{1-\rho} p_{c-1} = \frac{(\lambda ES)^c}{c!(1-\rho)} p_0$$

$$EL_q = L_q(exp)\{(1-\rho)\frac{c\gamma_1}{ES} + \rho\frac{ES^2}{2(ES)^2}\}$$

$$EL_q(L_q - 1) = \frac{\rho^2 P_W}{1-\rho}\{(1-\rho)\frac{c^2\gamma_2}{(ES)^2} + \rho\frac{ES^3}{3(ES)^3}\} + \frac{\rho^2}{1-\rho}\frac{ES^2}{2(ES)^2} EL_q$$

where

$$\gamma_k = \int_0^\infty k t^{k-1}(1-F_e(t))^c \, dt \ , \quad k \geq 1 \tag{C.1}$$

$$L_q(exp) = \frac{\rho}{1-\rho} P_W = \frac{\rho(\lambda ES)^c}{c!(1-\rho)} p_0 \tag{C.2}$$

Once p_0 , α_k , β_k and γ_k have been obtained, the other calculations are trivial. Note that for $n \geq c$ p_n is recursively computed from $p_0, \ldots, p_{n-1}$ by writing

$$p_n = (\lambda p_{c-1}\alpha_{n-c} + \lambda \sum_{j=c}^{n-1} p_j \beta_{n-j}) / (1 - \lambda\beta_0)$$

To compute p_0, we propose the scheme (in PASCAL style written)

```
p0:=1 ; sum:=1 ;
for i:=1 to c-1 do
begin pi:=λES/i pi-1 ; sum := sum +pi ;
end ;
sum := sum +ρpc-1/(1-ρ) ;
for i:=0 to c-1 do pi:=pi / sum ;
```

The evaluation of α_k, β_k and γ_k depends on the type of the service time distribution. We shall specify the actual computations for deterministic, Erlangian and hyperexponential service times. Recall that $F_e(t)$ is given by (cf(3.1))

$$F_e(t) = \frac{1}{ES}\int_0^t (1 - F(x))\, dx$$

Deterministic service time

Assume

$$F(t) = \begin{cases} 0 \ , & t < D \\ 1 \ , & t \geq D \end{cases}$$

Thus

$$F_e(t) = \begin{cases} t/D \ , & t < D \\ 1 \ , & t \geq D \end{cases}$$

See Figure C.1. Note that $ES = D$ and $C_S^2 = 0$.

In the deterministic case, the expressions for α_k, β_k and γ_k reduce to integrals over a finite interval. We propose to use Gauss-Legendre quadrature (cf. Appendix B) to compute α_k, whereas the numbers β_k and γ_k are explicitly given as

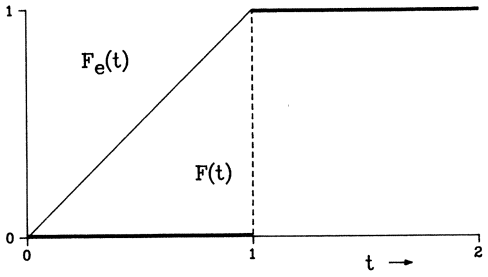

Figure C.1 $F(t)$ and $F_e(t)$ for deterministic service time.

$$\beta_k = \int_0^{D/c} e^{-\lambda t} \frac{(\lambda t)^k}{k!} dt = \frac{1}{\lambda}\{1 - \sum_{i=0}^{k} e^{-\rho} \frac{\rho^k}{k!}\}$$

$$\gamma_k = \int_0^{D} kt^{k-1}(1 - t/D)^c \, dt = \frac{k!c!}{(k+c)!} D^k$$

Erlangian service time

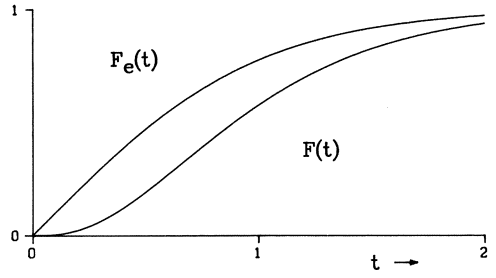

Figure C.2 $F(t)$ and $F_e(t)$ for Erlang-3 service times.

We consider a mixture of Erlang-1, Erlang-2 and Erlang-3 distributions with the same scale parameter μ. See Figure C.2. The extension to other mixtures is straightforward. Hence we assume that

$$F(t) = p_1(1 - e^{-\mu t}) + p_2(1 - (1+\mu t)e^{-\mu t}) + p_3(1 - (1 + \mu t + \frac{\mu^2 t^2}{2})e^{-\mu t})$$

$$= 1 - (1 + u_1 t + u_2 t^2)e^{-\mu t} \quad , \quad t \geq 0 \, ,$$

where $u_1 = (p_2 + p_3)\mu$, $u_2 = \frac{1}{2} p_3 \mu^2$. It easily follows that

$$F_e(t) = 1 - (1 + v_1 t + v_2 t^2)e^{-\mu t} \quad , \quad t \geq 0 \, ,$$

where $\nu_1=(p_2+2p_3)\mu/(p_1+2p_2+3p_3)$, $\nu_2=\frac{1}{2}p_3\mu^2/(p_1+2p_2+3p_3)$. Also, note that $\mathbf{ES}=(p_1+2p_2+3p_3)/\mu$ and $C_S^2=(2p_1+6p_2+12p_3)/\mu^2-1$. The above mixture of Erlang distributions has three degrees of freedom. As usual, we take $\mathbf{ES}=1$. Further, we consider the special choices with $p_3=0$ or $p_2=0$. Then, we can fix the parameters when C_S^2 is given.

For the case of $p_3=0$, we obtain from $\mathbf{ES}=1$

$$p_1=\frac{2C_S^2-\sqrt{2(1-C_S^2)}}{C_S^2+1}\,,\quad 1/2\leqslant C_S^2\leqslant 1$$

For the case of $p_2=0$, we get from $\mathbf{ES}=1$

$$p_1=\frac{1+6C_S^2-\sqrt{13-12C_S^2}}{4(C_S^2+1)}\,,\quad 1/3\leqslant C_S^2\leqslant 13/12$$

or

$$p_1=\frac{1+6C_S^2+\sqrt{13-12C_S^2}}{4(C_S^2+1)}\,,\quad 1\leqslant C_S^2\leqslant 13/12$$

Hence, by taking a mixture of Erlang-1, Erlang-2 and Erlang-3 distributions with the same scale parameter, we can obtain squared coefficients of variation between 1/3 and 13/12. For the evaluation of the α_k, we suggest to use Gauss-Laguerre quadrature. The polynomial in the integrand has degree $2c+k$, hence the usage of the numbers of Table B.1 will yield an exact evaluation of α_k for a wide range of values for c and k, while for practical purposes a good approximation will in general be obtained for parameter values outside this range.

For β_k, we need not use Gauss-Laguerre quadrature since we also have the explicit formula

$$\beta_k=\frac{\lambda^k}{(\lambda+c\mu)^{k+1}}\{1+\frac{(k+1)u_1c}{\lambda+c\mu}+\frac{(k+1)(k+2)u_2c^2}{(\lambda+c\mu)^2}\}$$

To compute γ_k again Gauss-Laguerre quadrature can be used.

Hyperexponential service time

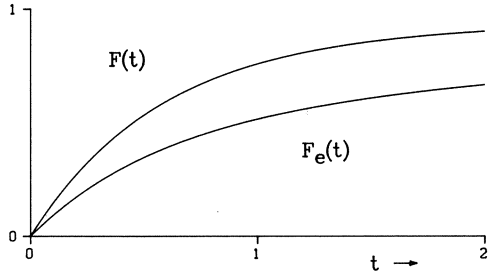

Figure C.3 $F(t)$ and $F_e(t)$ for hyperexponential service times ($C_S^2=4.0$).

For hyperexponential services, we have

$$F(t)=1-(p_1e^{-\mu_1 t}+p_2e^{-\mu_2 t}),\quad t\geqslant 0$$

where $0\leqslant p_1,p_2\leqslant 1$ and $p_1+p_2=1$. It follows that

$$F_e(t)=1-(q_1e^{-\mu_1 t}+q_2e^{-\mu_2 t}),\quad t\geqslant 0$$

where $q_1=(p_1/\mu_1)/(p_1/\mu_1+p_2/\mu_2)$ and $q_1+q_2=1$. We have $\mathbf{ES}=p_1/\mu_1+p_2/\mu_2$ and $C_S^2=2p_1/\mu_1^2+2p_2/\mu_2^2-1$. See Figure C.3 for an example. Note that $F_e(t)\geqslant F(t)$.

The hyperexponential distribution has the three parameters p_1, μ_1 and μ_2 as degrees of freedom. The coefficient of variation is always at least one, where for $C_S^2=1$ the hyperexponential distribution reduces to an exponential distribution. For a given value of C_S^2 and $\mathbf{ES}=1$ we can fix the parameters by the usual normalization $q_1=\frac{1}{2}\mathbf{ES}=\frac{1}{2}$. In fact q_1 measures the contribution to the mean of the 'branch' with parameter μ_1. For the choice $q_1=r$, we get for $\mathbf{ES}=1$

$$p_1=\frac{1}{2(C_S^2+1)}\{C_S^2+4r-1+\sqrt{(C_S^2-1)((C_S^2-1)+8r(1-r))}\}$$

In particular, if $q_1=\frac{1}{2}$, we have

$$p_1=\frac{1}{2}\{1+\sqrt{\frac{C_S^2-1}{C_S^2+1}}\}$$

For the numbers α_k, β_k and γ_k the following explicit expressions are easily obtained

$$\alpha_k=\sum_{i=0}^{c-1}\binom{c-1}{i}q_1^{\,i}q_2^{\,c-1-i}\{p_1\frac{\lambda^k}{(\lambda+(i+1)\mu_1+(c-1-i)\mu_2)^{k+1}}$$

$$+p_2\frac{\lambda^k}{(\lambda+i\mu_1+(c-i)\mu_2)^{k+1}}\}$$

$$\beta_k=p_1\frac{\lambda^k}{(\lambda+c\mu_1)^{k+1}}+p_2\frac{\lambda^k}{(\lambda+c\mu_2)^{k+1}}$$

$$\gamma_k=\sum_{i=0}^{c}\binom{c}{i}q_1^{\,i}q_2^{\,c-i}\frac{k}{(i\mu_1+(c-i)\mu_2)^k}$$

The implementation of these expressions is straightforward.

The waiting time distribution

To compute the waiting time distribution, the following Volterra integral equation has to be solved for $V(t)=\Pr\{\mathbf{W}_q<t\mid\mathbf{W}_q>0\}$.

$$V(t)=(1-\rho)\{1-(1-F_e(t))^c\}+\lambda\int_0^t V(t-x)(1-F(cx))\,dx$$

For Erlangian and hyperexponential service distributions, we suggest to solve the integral equation directly by using the scheme given in Appendix B. Obviously, for large t the computing effort increases. Therefore, it is worthwhile to use the asymptotic expansion of $V(t)$ (cf. Section 3.6) to save computing time by building in a test to check wether the tail of $V(t)$ has already an exponential form. It turns out from our numerical investigations that the tail behaviour is rather rapidly exponential.

For the deterministic case , $F(t)$ and $F_e(t)$ are not differentiable for all t, and hence the numerical procedure can not blindly be used. Then, the distribution function $V(t)$ has points of inflection at the values $t = k\,\mathrm{ES}/c$, $k \geq 1$. Hence, we should successively solve the integral equation in the intervals $[(k-1)\mathrm{ES}/c\ ,\ k\,\mathrm{ES}/c]$, using the value $V((k-1)\mathrm{ES}/c)$ as starting value in the corresponding interval. However, for the deterministic case there is an alternative way to compute $V(t)$. By differentiation of $V(t)$, we get the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}V(t) = \begin{cases} c(1-\rho)(1-t)^{c-1} + \lambda V(t) & , \ 0 \leq t \leq \mathrm{ES}/c \\ c(1-\rho)(1-t)^{c-1} + \lambda\{V(t) - V(t-\mathrm{ES}/c)\} & , \ \mathrm{ES}/c \leq t \leq \mathrm{ES} \\ \lambda\{V(t) - V(t-\mathrm{ES}/c)\} & , \ t \geq \mathrm{ES} \end{cases}$$

We proceed now as follows using the numerical procedure described in Appendix B.

1. Choose $N > 0$ and let $h = (\mathrm{ES}/c)/\mathrm{N}$ and $V_n = V(nh)$.
2. Compute $V_0, \ldots, V_\mathrm{N}$ using (B.7) with $V_0 = 0$ as initial value.
3. To compute $V_{kN}, \ldots, V_{(k+1)N}$, $k \geq 1$ consider $V_{(k-1)N}, \ldots, V_{kN}$ as belonging to the known part of the differential equation (taking when necessary $V((n+\tfrac{1}{2})h) = \tfrac{1}{2}(V_n + V_{n+1})$) and use (B.7) with V_{kN} as initial value.

Other approximations for the mean queue length in the M/G/c queue.

For completeness, we give here the approximations for the mean queue length EL_q in the M/G/c queue we have tested in Chapter 3. In the following, we assume that c and ρ are fixed. Let $\mathrm{L}_q(exp)$ and $\mathrm{L}_q(det)$ denote the exact values for EL_q in the M/M/c queue and the M/D/c queue respectively. $\mathrm{L}_q(exp)$ is explicitly given by (C.2), whereas $\mathrm{L}_q(det)$ has been tabulated for c up to 250 in Kühn[76]. In Cosmetatos[75], the approximation $\mathrm{L}_q(Cosd)$ for $\mathrm{L}_q(det)$ is suggested

$$\mathrm{L}_q(Cosd) = \tfrac{1}{2}\mathrm{L}_q(exp)\{1 + (1-\rho)(c-1)\frac{(\sqrt{4+5c}-2)}{16\rho c}\}$$

Cosmetatos[76] proposes as approximation for EL_q the following linear interpolation between $\mathrm{L}_q(exp)$ and $\mathrm{L}_q(det)$

$$\mathrm{L}_q(Cos) = (1-C_S^2)\mathrm{L}_q(det) + C_S^2\mathrm{L}_q(exp)$$

The approximation of Boxma, Cohen and Huffels[80] requires the evaluation of $\mathrm{L}_q(exp)$, $\mathrm{L}_q(det)$ and the constant γ_1 and is given by

$$\mathrm{L}_q(Box) = \frac{1+C_S^2}{2}\,\frac{2\mathrm{L}_q(exp)\mathrm{L}_q(det)}{2A\mathrm{L}_q(det) + (1-A)\mathrm{L}_q(exp)} ,$$

where $A = \{\dfrac{\mathrm{ES}^2}{\mathrm{ES}\gamma_1} - c - 1\}/(c-1)$ and γ_1 is given by (C.1).

In the approximation of Takahashi[77] we also need to evaluate $\mathrm{L}_q(exp)$, $\mathrm{L}_q(det)$ and in addition the root α of the equation

$$\mathrm{L}_q(exp) = (\Gamma(\alpha+1))^{\frac{1}{\alpha-1}}\mathrm{L}_q(det) , \quad \alpha \leq 2$$

Then Takahashi's approximation is given by

$$L_q(Tak) = \left(\frac{ES^\alpha}{(ES)^\alpha} \right)^{\frac{1}{\alpha - 1}} L_q(det)$$

Remark The practical applicability of the latter two approximations is increased by using $L_q(Cosd)$ in stead of the exact value $L_q(det)$. We have done so in the comparison of the approximations in Section 3.7.

□

Appendix D

In this appendix we display the exact results for a number of M/G/c systems with phase type service time distributions. These results have been computed with a specialization of the method described in Takahashi and Takami[76] and can also be found in Groenevelt, van Hoorn and Tijms[82]. We give numerical results for the following quantities.

$$P_W = \Pr\{W_q > 0\}$$

$$T_W = E\{W_q \mid W_q > 0\}$$

$$C_W = cv(W_q \mid W_q > 0) = \sqrt{E\{W_q^2 \mid W_q > 0\} / (E\{W_q^2 \mid W_q > 0\})^2 - 1}$$

$$= \sqrt{P_W EW_q^2 / (EW_q)^2 - 1}$$

We have considered hyperexponential service times with $p_1 / \mu_1 = \frac{1}{2}ES$ and service times which are a mixture of Erlang distributions. For given values of c, ρ, C_S^2 and with $ES = 1$ the cases are completely specified; see also Appendix C.

		$c=2$			$c=3$			$c=4$		
ρ	C_S^2	P_W	T_W	C_W	P_W	T_W	C_W	P_W	T_W	C_W
0.30	1.5625	0.1395	0.8537	1.1285	0.0708	0.5480	1.1094	0.0375	0.4010	1.0934
	2.2500	0.1404	1.0189	1.2498	0.0714	0.6308	1.2260	0.0379	0.4504	1.1991
	4.0000	0.1418	1.4264	1.4552	0.0724	0.8280	1.4570	0.0385	0.5641	1.4275
	9.0000	0.1431	2.5578	1.7207	0.0736	1.3470	1.8388	0.0392	0.8493	1.8668
0.50	1.5625	0.3357	1.2269	1.1055	0.2395	0.7959	1.1005	0.1763	0.5851	1.0945
	2.2500	0.3378	1.4970	1.1965	0.2418	0.9476	1.1959	0.1783	0.6838	1.1897
	4.0000	0.3410	2.1641	1.3396	0.2453	1.3152	1.3623	0.1816	0.9193	1.3678
	9.0000	0.3447	4.0067	1.5166	0.2497	2.3017	1.6004	0.1855	1.5384	1.6470
0.70	1.5625	0.5791	2.0883	1.0714	0.4962	1.3709	1.0745	0.4330	1.0156	1.0758
	2.2500	0.5814	2.5958	1.1293	0.4995	1.6809	1.1387	0.4368	1.2319	1.1439
	4.0000	0.5852	3.8634	1.2146	0.5047	2.4486	1.2395	0.4427	1.7633	1.2558
	9.0000	0.5901	7.3971	1.3137	0.5113	4.5660	1.3653	0.4500	3.2153	1.4037
0.80	1.5625	0.7133	3.1595	1.0503	0.6506	2.0857	1.0545	0.6007	1.5519	1.0573
	2.2500	0.7152	3.9577	1.0903	0.6536	2.5903	1.0997	0.6044	1.9139	1.1063
	4.0000	0.7184	5.9653	1.1476	0.6584	3.8521	1.1678	0.6102	2.8149	1.1826
	9.0000	0.7226	11.607	1.2119	0.6644	7.3771	1.2478	0.6174	5.3172	1.2762
0.90	1.5625	0.8539	6.3659	1.0265	0.8192	4.2243	1.0297	0.7906	3.1562	1.0321
	2.2500	0.8551	8.0278	1.0473	0.8211	5.3057	1.0536	0.7931	3.9510	1.0586
	4.0000	0.8569	12.234	1.0765	0.8242	8.0350	1.0884	0.7971	5.9525	1.0978
	9.0000	0.8595	24.159	1.1081	0.8281	15.750	1.1273	0.8020	11.594	1.1429
0.95	1.5625	0.9263	12.774	1.0136	0.9082	8.4967	1.0154	0.8930	6.3607	1.0170
	2.2500	0.9269	16.156	1.0242	0.9093	10.726	1.0278	0.8945	8.0166	1.0307
	4.0000	0.9280	24.744	1.0390	0.9110	16.377	1.0455	0.8967	12.210	1.0507
	9.0000	0.9293	49.186	1.0548	0.9131	32.438	1.0648	0.8995	24.111	1.0730
0.99	1.5625	0.9852	64.025	1.0028	0.9814	42.665	1.0032	0.9783	31.987	1.0035
	2.2500	0.9853	81.159	1.0049	0.9817	54.062	1.0057	0.9786	40.519	1.0064
	4.0000	0.9855	124.75	1.0079	0.9820	83.051	1.0093	0.9791	62.216	1.0104
	9.0000	0.9858	249.21	1.0111	0.9825	165.79	1.0132	0.9797	124.13	1.0149

Table D.1 Exact results for the $M/H_2/c$ queue.

		$c=5$			$c=8$			$c=10$		
ρ	C_S^2	P_W	T_W	C_W	P_W	T_W	C_W	P_W	T_W	C_W
0.50	1.5625	0.1324	0.4609	1.0884	0.0601	0.2789	1.0727	0.0367	0.2199	1.0644
	2.2500	0.1341	0.5306	1.1813	0.0609	0.3110	1.1543	0.0373	0.2416	1.1380
	4.0000	0.1368	0.6946	1.3646	0.0624	0.3834	1.3324	0.0382	0.2893	1.3047
	9.0000	0.1402	1.1169	1.6750	0.0641	0.5583	1.6916	0.0393	0.3997	1.6690
0.70	1.5625	0.3824	0.8042	1.0761	0.2749	0.4909	1.0744	0.2256	0.3880	1.0723
	2.2500	0.3863	0.9665	1.1467	0.2787	0.5773	1.1484	0.2291	0.4511	1.1466
	4.0000	0.3925	1.3626	1.2672	0.2846	0.7842	1.2852	0.2344	0.6005	1.2893
	9.0000	0.4002	2.4350	1.4343	0.2919	1.3305	1.4972	0.2411	0.9888	1.5240
0.80	1.5625	0.5589	1.2331	1.0593	0.4632	0.7582	1.0626	0.4148	0.6013	1.0636
	2.2500	0.5630	1.5116	1.1113	0.4680	0.9160	1.1205	0.4198	0.7207	1.1240
	4.0000	0.5696	2.2020	1.1942	0.4755	1.3029	1.2185	0.4274	1.0116	1.2292
	9.0000	0.5776	4.1090	1.2996	0.4847	2.3567	1.3529	0.4369	1.7971	1.3796
0.90	1.5625	0.7659	2.5166	1.0341	0.7061	1.5601	1.0386	0.6739	1.2425	1.0408
	2.2500	0.7689	3.1413	1.0627	0.7101	1.9336	1.0719	0.6783	1.5339	1.0766
	4.0000	0.7735	4.7114	1.1056	0.7163	2.8678	1.1239	0.6852	2.2607	1.1333
	9.0000	0.7793	9.1265	1.1561	0.7238	5.4787	1.1876	0.6935	4.2849	1.2046
0.95	1.5625	0.8798	5.0804	1.0182	0.8470	3.1625	1.0213	0.8289	2.5244	1.0229
	2.2500	0.8815	6.3940	1.0332	0.8494	3.9665	1.0391	0.8317	3.1600	1.0422
	4.0000	0.8842	9.7172	1.0551	0.8532	5.9957	1.0658	0.8360	4.7622	1.0716
	9.0000	0.8875	19.139	1.0800	0.8578	11.732	1.0971	0.8413	9.2841	1.1064
0.99	1.5625	0.9755	25.581	1.0038	0.9685	15.976	1.0046	0.9645	12.775	1.0050
	2.2500	0.9759	32.396	1.0069	0.9690	20.218	1.0083	0.9652	16.161	1.0091
	4.0000	0.9765	49.722	1.0114	0.9699	30.999	1.0138	0.9662	24.764	1.0152
	9.0000	0.9772	99.150	1.0164	0.9710	61.737	1.0200	0.9675	49.286	1.0220

Table D.2 Exact results for the $M/H_2/c$ queue.

		$c=15$			$c=20$			$c=25$		
ρ	C_S^2	P_W	T_W	C_W	P_W	T_W	C_W	P_W	T_W	C_W
0.50	1.5625	0.0115	0.1432	1.0494	0.0038	0.1059	1.0396	0.0013	0.0839	1.0329
	2.2500	0.0116	0.1535	1.1060	0.0038	0.1117	1.0840	0.0013	0.0876	1.0686
	4.0000	0.0119	0.1746	1.2384	0.0039	0.1232	1.1865	0.0013	0.0946	1.1484
	9.0000	0.0122	0.2188	1.5637	0.0040	0.1449	1.4459	0.0014	0.1067	1.3454
0.70	1.5625	0.1440	0.2529	1.0664	0.0955	0.1865	1.0607	0.0649	0.1474	1.0555
	2.2500	0.1465	0.2875	1.1381	0.0973	0.2086	1.1280	0.0661	0.1627	1.1179
	4.0000	0.1504	0.3673	1.2863	0.1000	0.2582	1.2738	0.0680	0.1963	1.2575
	9.0000	0.1554	0.5671	1.5600	0.1035	0.3777	1.5676	0.0705	0.2742	1.5563
0.80	1.5625	0.3246	0.3939	1.0642	0.2610	0.2914	1.0633	0.2134	0.2306	1.0617
	2.2500	0.3293	0.4645	1.1279	0.2652	0.3392	1.1283	0.2171	0.2655	1.1268
	4.0000	0.3367	0.6340	1.2458	0.2719	0.4525	1.2538	0.2231	0.3472	1.2568
	9.0000	0.3459	1.0829	1.4288	0.2803	0.7470	1.4618	0.2305	0.5560	1.4844
0.90	1.5625	0.6086	0.8206	1.0448	0.5572	0.6108	1.0476	0.5146	0.4854	1.0495
	2.2500	0.6138	1.0047	1.0853	0.5628	0.7426	1.0914	0.5204	0.5867	1.0960
	4.0000	0.6218	1.4611	1.1516	0.5715	1.0679	1.1653	0.5295	0.8354	1.1760
	9.0000	0.6316	2.7225	1.2388	0.5820	1.9607	1.2658	0.5405	1.5135	1.2882
0.95	1.5625	0.7911	1.6750	1.0261	0.7602	1.2513	1.0286	0.7336	0.9977	1.0307
	2.2500	0.7946	2.0882	1.0486	0.7642	1.5547	1.0536	0.7381	1.2358	1.0577
	4.0000	0.8001	3.1269	1.0834	0.7705	2.3155	1.0929	0.7450	1.8318	1.1009
	9.0000	0.8067	6.0480	1.1258	0.7781	4.4488	1.1416	0.7534	3.4985	1.1552
0.99	1.5625	0.9561	8.5087	1.0059	0.9491	6.3765	1.0066	0.9429	5.0977	1.0072
	2.2500	0.9570	10.755	1.0107	0.9502	8.0549	1.0121	0.9441	6.4356	1.0133
	4.0000	0.9584	16.461	1.0180	0.9518	12.315	1.0203	0.9460	9.8303	1.0224
	9.0000	0.9600	32.712	1.0263	0.9537	24.443	1.0299	0.9482	19.490	1.0330

Table D.3 Exact results for the $M/H_2/c$ queue.

		$c=2$			$c=3$			$c=4$		
ρ	C_S^2	P_W	T_W	C_W	P_W	T_W	C_W	P_W	T_W	C_W
0.30	0.5000	0.1373	0.5614	0.9116	0.0691	0.3856	0.9122	0.0365	0.2956	0.9149
	0.6400	0.1376	0.6070	0.9328	0.0694	0.4138	0.9350	0.0367	0.3153	0.9382
	0.7500	0.1379	0.6416	0.9509	0.0696	0.4346	0.9537	0.0368	0.3297	0.9568
	0.8100	0.1380	0.6600	0.9614	0.0697	0.4455	0.9642	0.0368	0.3370	0.9671
0.50	0.5000	0.3308	0.7727	0.9314	0.2338	0.5258	0.9270	0.1710	0.4006	0.9253
	0.6400	0.3315	0.8388	0.9473	0.2347	0.5678	0.9450	0.1719	0.4309	0.9446
	0.7500	0.3321	0.8896	0.9610	0.2354	0.5997	0.9602	0.1725	0.4535	0.9604
	0.8100	0.3324	0.9169	0.9691	0.2357	0.6166	0.9688	0.1729	0.4654	0.9693
0.70	0.5000	0.5736	1.2703	0.9561	0.4880	0.8566	0.9503	0.4235	0.6484	0.9466
	0.6400	0.5744	1.3833	0.9660	0.4893	0.9302	0.9621	0.4251	0.7024	0.9597
	0.7500	0.5751	1.4712	0.9746	0.4902	0.9869	0.9721	0.4262	0.7438	0.9706
	0.8100	0.5754	1.5188	0.9797	0.4908	1.0174	0.9780	0.4268	0.7660	0.9769
0.80	0.5000	0.7087	1.8942	0.9699	0.6432	1.2721	0.9651	0.5914	0.9598	0.9616
	0.6400	0.7094	2.0657	0.9766	0.6444	1.3847	0.9732	0.5929	1.0432	0.9708
	0.7500	0.7099	2.1996	0.9825	0.6453	1.4722	0.9802	0.5940	1.1077	0.9785
	0.8100	0.7102	2.2723	0.9860	0.6458	1.5196	0.9842	0.5946	1.1425	0.9830
0.90	0.5000	0.8512	3.7682	0.9846	0.8145	2.5210	0.9817	0.7844	1.8962	0.9794
	0.6400	0.8516	4.1149	0.9880	0.8153	2.7504	0.9859	0.7854	2.0672	0.9842
	0.7500	0.8519	4.3864	0.9910	0.8158	2.9298	0.9895	0.7861	2.2007	0.9883
	0.8100	0.8521	4.5342	0.9928	0.8161	3.0273	0.9916	0.7865	2.2732	0.9907
0.95	0.5000	0.9248	7.5177	0.9922	0.9056	5.0204	0.9907	0.8895	3.7706	0.9894
	0.6400	0.9251	8.2144	0.9939	0.9060	5.4833	0.9928	0.8901	4.1168	0.9918
	0.7500	0.9253	8.7611	0.9954	0.9063	5.8461	0.9946	0.8905	4.3879	0.9939
	0.8100	0.9254	9.0589	0.9963	0.9065	6.0436	0.9957	0.8907	4.5354	0.9952
0.99	0.5000	0.9849	37.517	0.9984	0.9809	25.020	0.9981	0.9775	18.770	0.9978
	0.6400	0.9849	41.014	0.9988	0.9810	27.350	0.9985	0.9776	20.516	0.9983
	0.7500	0.9849	43.761	0.9991	0.9810	29.179	0.9989	0.9777	21.888	0.9988
	0.8100	0.9850	45.259	0.9993	0.9811	30.177	0.9991	0.9778	22.635	0.9990

Table D.4 Exact results for the $M/E_{1,2}/c$ queue.

		$c=5$			$c=8$			$c=10$		
ρ	C_S^2	P_W	T_W	C_W	P_W	T_W	C_W	P_W	T_W	C_W
0.50	0.5000	0.1279	0.3247	0.9250	0.0576	0.2088	0.9272	0.0352	0.1694	0.9295
	0.6400	0.1286	0.3480	0.9450	0.0581	0.2221	0.9480	0.0355	0.1795	0.9503
	0.7500	0.1292	0.3653	0.9611	0.0584	0.2317	0.9641	0.0357	0.1866	0.9661
	0.8100	0.1295	0.3743	0.9701	0.0586	0.2366	0.9729	0.0358	0.1903	0.9746
0.70	0.5000	0.3724	0.5227	0.9440	0.2651	0.3327	0.9396	0.2167	0.2688	0.9382
	0.6400	0.3740	0.5652	0.9580	0.2668	0.3580	0.9555	0.2182	0.2885	0.9549
	0.7500	0.3752	0.5975	0.9696	0.2680	0.3771	0.9684	0.2194	0.3032	0.9682
	0.8100	0.3759	0.6148	0.9763	0.2687	0.3872	0.9755	0.2200	0.3109	0.9755
0.80	0.5000	0.5484	0.7717	0.9589	0.4508	0.4883	0.9533	0.4021	0.3932	0.9509
	0.6400	0.5501	0.8377	0.9689	0.4529	0.5283	0.9652	0.4043	0.4247	0.9636
	0.7500	0.5514	0.8886	0.9773	0.4544	0.5590	0.9749	0.4059	0.4488	0.9739
	0.8100	0.5521	0.9160	0.9821	0.4552	0.5754	0.9804	0.4067	0.4616	0.9797
0.90	0.5000	0.7584	1.5207	0.9776	0.6960	0.9562	0.9733	0.6625	0.7675	0.9711
	0.6400	0.7596	1.6568	0.9829	0.6976	1.0402	0.9798	0.6644	0.8342	0.9783
	0.7500	0.7605	1.7629	0.9874	0.6989	1.1054	0.9853	0.6658	0.8859	0.9842
	0.8100	0.7610	1.8204	0.9900	0.6996	1.1406	0.9884	0.6665	0.9138	0.9876
0.95	0.5000	0.8754	3.0202	0.9883	0.8407	1.8933	0.9858	0.8216	1.5171	0.9844
	0.6400	0.8761	3.2964	0.9911	0.8418	2.0648	0.9892	0.8228	1.6539	0.9882
	0.7500	0.8767	3.5126	0.9934	0.8425	2.1989	0.9921	0.8237	1.7607	0.9914
	0.8100	0.8769	3.6301	0.9948	0.8430	2.2717	0.9937	0.8242	1.8186	0.9932
0.99	0.5000	0.9745	15.020	0.9976	0.9670	9.3929	0.9970	0.9628	7.5168	0.9967
	0.6400	0.9747	16.416	0.9982	0.9672	10.265	0.9977	0.9631	8.2136	0.9975
	0.7500	0.9748	17.512	0.9986	0.9674	10.949	0.9983	0.9633	8.7605	0.9982
	0.8100	0.9748	18.110	0.9989	0.9675	11.322	0.9987	0.9634	9.0584	0.9985

Table D.5 Exact results for the $M/E_{1,2}/c$ queue.

		$c=15$			$c=20$			$c=25$		
ρ	C_S^2	P_W	T_W	C_W	P_W	T_W	C_W	P_W	T_W	C_W
0.50	0.5000	0.0110	0.1159	0.9356	0.0036	0.0885	0.9410	0.0012	0.0717	0.9457
	0.6400	0.0111	0.1218	0.9556	0.0037	0.0925	0.9601	0.0013	0.0747	0.9637
	0.7500	0.0112	0.1260	0.9705	0.0037	0.0953	0.9739	0.0013	0.0767	0.9766
	0.8100	0.0112	0.1281	0.9782	0.0037	0.0966	0.9809	0.0013	0.0777	0.9830
0.70	0.5000	0.1373	0.1826	0.9372	0.0907	0.1390	0.9377	0.0615	0.1125	0.9389
	0.6400	0.1385	0.1950	0.9548	0.0916	0.1478	0.9558	0.0622	0.1193	0.9570
	0.7500	0.1394	0.2042	0.9686	0.0923	0.1543	0.9696	0.0627	0.1242	0.9708
	0.8100	0.1398	0.2089	0.9761	0.0926	0.1576	0.9771	0.0629	0.1266	0.9781
0.80	0.5000	0.3122	0.2657	0.9470	0.2497	0.2014	0.9449	0.2033	0.1626	0.9438
	0.6400	0.3144	0.2860	0.9612	0.2516	0.2162	0.9600	0.2051	0.1741	0.9594
	0.7500	0.3160	0.3013	0.9725	0.2531	0.2273	0.9718	0.2065	0.1827	0.9716
	0.8100	0.3168	0.3094	0.9788	0.2539	0.2331	0.9784	0.2071	0.1872	0.9783
0.90	0.5000	0.5952	0.5152	0.9670	0.5427	0.3886	0.9640	0.4995	0.3124	0.9616
	0.6400	0.5975	0.5590	0.9754	0.5452	0.4210	0.9732	0.5021	0.3380	0.9716
	0.7500	0.5992	0.5928	0.9822	0.5470	0.4459	0.9808	0.5040	0.3577	0.9797
	0.8100	0.6000	0.6110	0.9861	0.5480	0.4593	0.9850	0.5050	0.3682	0.9842
0.95	0.5000	0.7819	1.0149	0.9817	0.7496	0.7634	0.9796	0.7219	0.6122	0.9778
	0.6400	0.7835	1.1054	0.9862	0.7514	0.8308	0.9847	0.7239	0.6659	0.9834
	0.7500	0.7846	1.1760	0.9900	0.7527	0.8833	0.9889	0.7253	0.7076	0.9880
	0.8100	0.7852	1.2142	0.9921	0.7534	0.9117	0.9913	0.7261	0.7301	0.9906
0.99	0.5000	0.9539	5.0146	0.9961	0.9464	3.7631	0.9955	0.9398	3.0120	0.9950
	0.6400	0.9542	5.4786	0.9970	0.9468	4.1107	0.9966	0.9403	3.2898	0.9963
	0.7500	0.9545	5.8425	0.9978	0.9472	4.3832	0.9975	0.9407	3.5075	0.9973
	0.8100	0.9547	6.0407	0.9983	0.9473	4.5316	0.9980	0.9409	3.6261	0.9978

Table D.6 Exact results for the $M/E_{1,2}/c$ queue.

		$c=30$			$c=40$			$c=50$		
ρ	C_S^2	P_W	T_W	C_W	P_W	T_W	C_W	P_W	T_W	C_W
0.70	0.5000	0.0425	0.0947	0.9404	0.0209	0.0722	0.9434	0.0107	0.0585	0.9464
	0.6400	0.0429	0.1001	0.9584	0.0212	0.0760	0.9611	0.0108	0.0613	0.9636
	0.7500	0.0433	0.1040	0.9719	0.0214	0.0786	0.9741	0.0109	0.0633	0.9729
	0.8100	0.0435	0.1059	0.9789	0.0215	0.0799	0.9799	0.0109	0.0643	0.9784
0.80	0.5000	0.1678	0.1366	0.9432	0.1173	0.1038	0.9430	0.0840	0.0840	0.9434
	0.6400	0.1694	0.1460	0.9592	0.1185	0.1105	0.9594	0.0849	0.0891	0.9601
	0.7500	0.1705	0.1529	0.9717	0.1194	0.1154	0.9721	0.0856	0.0929	0.9728
	0.8100	0.1711	0.1565	0.9785	0.1199	0.1180	0.9789	0.0860	0.0948	0.9796
0.90	0.5000	0.4628	0.2615	0.9597	0.4029	0.1976	0.9568	0.3554	0.1591	0.9547
	0.6400	0.4655	0.2826	0.9703	0.4056	0.2132	0.9683	0.3580	0.1714	0.9669
	0.7500	0.4674	0.2988	0.9788	0.4075	0.2250	0.9775	0.3599	0.1806	0.9766
	0.8100	0.4684	0.3074	0.9835	0.4086	0.2313	0.9826	0.3609	0.1855	0.9819
0.95	0.5000	0.6976	0.5113	0.9763	0.6560	0.3850	0.9737	0.6209	0.3091	0.9716
	0.6400	0.6997	0.5559	0.9823	0.6583	0.4181	0.9804	0.6234	0.3354	0.9789
	0.7500	0.7013	0.5904	0.9872	0.6600	0.4437	0.9859	0.6253	0.3557	0.9848
	0.8100	0.7021	0.6090	0.9900	0.6610	0.4575	0.9889	0.6262	0.3666	0.9881
0.99	0.5000	0.9339	2.5111	0.9946	0.9234	1.8849	0.9939	0.9143	1.5090	0.9932
	0.6400	0.9345	2.7424	0.9959	0.9241	2.0580	0.9954	0.9151	1.6474	0.9949
	0.7500	0.9349	2.9236	0.9970	0.9246	2.1937	0.9966	0.9157	1.7557	0.9963
	0.8100	0.9351	3.0223	0.9977	0.9249	2.2676	0.9973	0.9160	1.8146	0.9970

Table D.7 Exact results for the $M/E_{1.2}/c$ queue.

		$c=2$			$c=3$			$c=4$		
ρ	C_S^2	P_W	T_W	C_W	P_W	T_W	C_W	P_W	T_W	C_W
0.30	0.3333	0.1366	0.5113	0.8658	0.0686	0.3566	0.8686	0.0362	0.2763	0.8743
	0.4000	0.1368	0.5342	0.8755	0.0688	0.3712	0.8796	0.0363	0.2868	0.8861
	0.4500	0.1369	0.5512	0.8829	0.0689	0.3819	0.8880	0.0363	0.2944	0.8949
	0.5000	0.1371	0.5679	0.8906	0.0690	0.3924	0.8966	0.0364	0.3018	0.9038
0.50	0.3333	0.3294	0.6975	0.8958	0.2321	0.4795	0.8899	0.1695	0.3682	0.8883
	0.4000	0.3298	0.7300	0.9028	0.2326	0.5006	0.8984	0.1699	0.3837	0.8976
	0.4500	0.3300	0.7541	0.9083	0.2330	0.5162	0.9048	0.1703	0.3950	0.9047
	0.5000	0.3303	0.7781	0.9139	0.2333	0.5316	0.9114	0.1706	0.4062	0.9119
0.70	0.3333	0.5720	1.1385	0.9335	0.4856	0.7722	0.9250	0.4207	0.5871	0.9196
	0.4000	0.5724	1.1932	0.9377	0.4863	0.8082	0.9303	0.4216	0.6138	0.9257
	0.4500	0.5727	1.2341	0.9411	0.4868	0.8349	0.9344	0.4222	0.6336	0.9304
	0.5000	0.5731	1.2748	0.9445	0.4873	0.8616	0.9386	0.4228	0.6532	0.9351
0.80	0.3333	0.7073	1.6925	0.9545	0.6410	1.1409	0.9473	0.5887	0.8633	0.9422
	0.4000	0.7077	1.7750	0.9574	0.6417	1.1954	0.9509	0.5895	0.9039	0.9464
	0.4500	0.7080	1.8367	0.9596	0.6422	1.2361	0.9537	0.5901	0.9342	0.9496
	0.5000	0.7083	1.8983	0.9619	0.6426	1.2767	0.9566	0.5907	0.9643	0.9528
0.90	0.3333	0.8503	3.3578	0.9768	0.8131	2.2504	0.9724	0.7825	1.6951	0.9690
	0.4000	0.8506	3.5237	0.9782	0.8135	2.3605	0.9743	0.7831	1.7774	0.9712
	0.4500	0.8507	3.6479	0.9793	0.8138	2.4429	0.9757	0.7835	1.8389	0.9729
	0.5000	0.8509	3.7720	0.9805	0.8141	2.5252	0.9771	0.7839	1.9003	0.9746
0.95	0.3333	0.9244	6.6905	0.9883	0.9048	4.4719	0.9859	0.8884	3.3610	0.9840
	0.4000	0.9245	7.0230	0.9890	0.9050	4.6931	0.9868	0.8887	3.5267	0.9851
	0.4500	0.9246	7.2723	0.9896	0.9052	4.8589	0.9876	0.8890	3.6507	0.9860
	0.5000	0.9247	7.5214	0.9901	0.9054	5.0245	0.9883	0.8892	3.7746	0.9868
0.99	0.3333	0.9848	33.357	0.9976	0.9807	22.249	0.9971	0.9772	16.694	0.9967
	0.4000	0.9848	35.023	0.9978	0.9807	23.359	0.9973	0.9773	17.526	0.9970
	0.4500	0.9848	36.272	0.9979	0.9808	24.192	0.9975	0.9774	18.150	0.9971
	0.5000	0.9848	37.521	0.9980	0.9808	25.024	0.9976	0.9774	18.774	0.9973

Table D.8 Exact results for the $M/E_{1,3}/c$ queue.

		$c=5$			$c=8$			$c=10$		
ρ	C_S^2	P_W	T_W	C_W	P_W	T_W	C_W	P_W	T_W	C_W
0.50	0.3333	0.1265	0.3003	0.8887	0.0569	0.1958	0.8941	0.0347	0.1599	0.8986
	0.4000	0.1269	0.3124	0.8986	0.0571	0.2029	0.9048	0.0349	0.1654	0.9096
	0.4500	0.1272	0.3212	0.9061	0.0573	0.2080	0.9128	0.0350	0.1693	0.9177
	0.5000	0.1275	0.3299	0.9136	0.0575	0.2130	0.9208	0.0351	0.1731	0.9257
0.70	0.3333	0.3694	0.4752	0.9160	0.2622	0.3052	0.9103	0.2140	0.2476	0.9088
	0.4000	0.3703	0.4963	0.9227	0.2632	0.3180	0.9181	0.2149	0.2577	0.9171
	0.4500	0.3710	0.5119	0.9277	0.2639	0.3274	0.9240	0.2155	0.2651	0.9234
	0.5000	0.3716	0.5274	0.9329	0.2646	0.3367	0.9300	0.2162	0.2724	0.9296
0.80	0.3333	0.5453	0.6959	0.9382	0.4472	0.4429	0.9303	0.3984	0.3578	0.9269
	0.4000	0.5463	0.7282	0.9429	0.4484	0.4627	0.9360	0.3996	0.3735	0.9331
	0.4500	0.5470	0.7522	0.9464	0.4492	0.4775	0.9403	0.4005	0.3852	0.9377
	0.5000	0.5477	0.7760	0.9500	0.4501	0.4920	0.9446	0.4014	0.3967	0.9424
0.90	0.3333	0.7562	1.3611	0.9663	0.6930	0.8584	0.9600	0.6591	0.6901	0.9568
	0.4000	0.7569	1.4267	0.9687	0.6939	0.8991	0.9630	0.6602	0.7225	0.9602
	0.4500	0.7574	1.4758	0.9706	0.6946	0.9295	0.9654	0.6610	0.7467	0.9628
	0.5000	0.7579	1.5247	0.9725	0.6953	0.9597	0.9677	0.6618	0.7708	0.9654
0.95	0.3333	0.8741	2.6937	0.9824	0.8389	1.6911	0.9787	0.8195	1.3562	0.9767
	0.4000	0.8745	2.8260	0.9837	0.8395	1.7735	0.9803	0.8202	1.4220	0.9785
	0.4500	0.8748	2.9251	0.9846	0.8399	1.8351	0.9815	0.8207	1.4712	0.9798
	0.5000	0.8751	3.0241	0.9856	0.8403	1.8967	0.9827	0.8212	1.5203	0.9812
0.99	0.3333	0.9742	13.360	0.9964	0.9666	8.3572	0.9955	0.9623	6.6890	0.9951
	0.4000	0.9743	14.025	0.9966	0.9667	8.7730	0.9959	0.9624	7.0215	0.9954
	0.4500	0.9744	14.525	0.9968	0.9668	9.0847	0.9961	0.9626	7.2708	0.9957
	0.5000	0.9744	15.024	0.9970	0.9669	9.3962	0.9964	0.9627	7.5199	0.9960

Table D.9 Exact results for the $M/E_{1,3}/c$ queue.

		$c=15$			$c=20$			$c=25$		
ρ	C_S^2	P_W	T_W	C_W	P_W	T_W	C_W	P_W	T_W	C_W
0.50	0.3333	0.0108	0.1107	0.9096	0.0036	0.0852	0.9187	0.0012	0.0695	0.9264
	0.4000	0.0109	0.1140	0.9204	0.0036	0.0875	0.9284	0.0012	0.0713	0.9226
	0.4500	0.0109	0.1164	0.9284	0.0036	0.0892	0.9361	0.0012	0.0725	0.9244
	0.5000	0.0110	0.1187	0.9362	0.0036	0.0907	0.9424	0.0012	0.0737	0.9268
0.70	0.3333	0.1353	0.1698	0.9085	0.0893	0.1302	0.9102	0.0605	0.1060	0.9127
	0.4000	0.1360	0.1763	0.9175	0.0898	0.1348	0.9196	0.0609	0.1096	0.9223
	0.4500	0.1365	0.1810	0.9242	0.0902	0.1383	0.9265	0.0612	0.1122	0.9293
	0.5000	0.1370	0.1856	0.9309	0.0906	0.1416	0.9334	0.0615	0.1148	0.9362
0.80	0.3333	0.3086	0.2434	0.9218	0.2464	0.1854	0.9192	0.2004	0.1504	0.9180
	0.4000	0.3098	0.2536	0.9288	0.2475	0.1930	0.9267	0.2014	0.1563	0.9259
	0.4500	0.3107	0.2612	0.9340	0.2483	0.1986	0.9324	0.2022	0.1607	0.9318
	0.5000	0.3116	0.2686	0.9393	0.2491	0.2040	0.9380	0.2029	0.1649	0.9376
0.90	0.3333	0.5914	0.4648	0.9508	0.5386	0.3515	0.9465	0.4952	0.2833	0.9431
	0.4000	0.5926	0.4862	0.9548	0.5399	0.3675	0.9510	0.4966	0.2960	0.9480
	0.4500	0.5936	0.5022	0.9579	0.5410	0.3793	0.9543	0.4977	0.3054	0.9516
	0.5000	0.5945	0.5180	0.9609	0.5420	0.3911	0.9577	0.4987	0.3147	0.9553
0.95	0.3333	0.7793	0.9088	0.9727	0.7465	0.6845	0.9695	0.7186	0.5496	0.9669
	0.4000	0.7801	0.9525	0.9748	0.7475	0.7171	0.9720	0.7197	0.5757	0.9696
	0.4500	0.7808	0.9851	0.9764	0.7483	0.7415	0.9738	0.7205	0.5951	0.9716
	0.5000	0.7814	1.0176	0.9781	0.7490	0.7658	0.9756	0.7213	0.6145	0.9736
0.99	0.3333	0.9532	4.4639	0.9941	0.9456	3.3508	0.9933	0.9389	2.6826	0.9926
	0.4000	0.9534	4.6854	0.9945	0.9458	3.5168	0.9938	0.9392	2.8154	0.9932
	0.4500	0.9536	4.8514	0.9949	0.9460	3.6412	0.9942	0.9394	2.9148	0.9936
	0.5000	0.9537	5.0173	0.9952	0.9462	3.7655	0.9946	0.9396	3.0142	0.9940

Table D.10 Exact results for the $M/E_{1,3}/c$ queue.

References

Adelson, R.M. (1966), 'Compound Poisson distributions', *Operational Research Quarterly* **17**, 73-75.

Allen, A.O. (1978), *Probability, Statistics and Queueing Theory*, Academic Press, N.Y.

Boxma, O.J., Cohen, J.W. and Huffels, N. (1980), 'Approximations of the mean waiting time in an M/G/s queueing system', *Operations Research* **27**, 1115-1127.

Burke, P.J. (1975), 'Delays in single server queues with batch input', *Operations Research* **23**, 830-833.

Burman, D.Y. and Smith,D.R. (1981), 'A light traffic theorem for multiserver queues', Report Bell Labs, Holmdel, New Jersey.

Bux, W. (1979), 'Single server queues with general interarrival and phase type service time distribution - Computational algorithms', *Proceedings ITC 9*, Torremolinos, Spain.

Bux, W. and Herzog, U. (1977), 'The phase concept: Approximation of measured data and performance analysis', presented at: *International Symposium on Computer Performance Modelling, Measurement and Evaluation*, Yorktown Heights, New York.

Cohen, J.W. (1969), *The single server queue*, North Holland Publishing Company, Amsterdam.

Cohen, J.W. (1969), 'Single Server Queues with Restricted Accessibility', *J. Eng. Math.* **3**, 265-284,

Cohen, J.W. (1977), 'On up and down crossings', *Journal of Applied Probability* **14**, 405-410.

Cooper, R.B. (1981), *Introduction to queueing theory*, Edward Arnold, London.

Cosmetatos, G.P. (1975), 'Approximate explicit formulae for the average queueing time in the processes (M/D/r) and (D/M/r)', *INFOR* **13**, 328-331.

Cosmetatos, G.P. (1976), 'Some approximate equilibrium results for the multiserver queue M/G/r', *Operational Research Quarterly* **27**, 615-620

Cosmetatos, G.P. (1978), 'Some practical considerations on multi server queues with multiple Poisson arrivals', *Omega* **6**, 443-448.

Crommelin, C.D. (1932), 'Delay probability formulae when the holding times are constant', *P.O. Elect. Engr. J.* **25**, 41-50.

Davis, P.J. and Rabinowitz, P. (1967), *Numerical integration*, Blaisdell Publishing Company , Waltham, Massachusetts.

Delves, L.M. and Walsh, J. (1974), *Numerical solution of integral equations*, Clarendon Press, Oxford.

Feller, W. (1968, 1966), *An introduction to probability theory and its applications*, Volume I and II, Wiley, New York.

Ferdinand, A.E. (1971), 'An analysis of the machine inference model', *IBM System Journal* **10**, 129-142.

Gavish, B. and Schweitzer, P.J. (1977), 'The Markovian queue with bounded waiting time', *Management Science* **23**, 1349-1357.

Groenevelt. H, van Hoorn, M.H. and Tijms, H.C. (1982), 'Tables for M/G/c queueing systems with phase type service', Research Report 79 Vrije Universiteit, Amsterdam (to appear in *European Journal of Operational Research*).

Haji, R. and Newell, G.F. (1971), 'A relation between stationary queue and waiting time distributions', *Journal of Applied Probability* **8**, 617-620.

Halachmi, B. and Franta, W.R. (1978), 'A diffiusion approximation to the multiserver queue', *Management Science* **24**, 522-529.

Heffer, J.C. (1969), 'Steady state solution of the $M/E_k/c(\infty,FIFO)$ queueing system', *CORS Journal* **7**, 16-30.

Heffes, H. (1973), 'Analysis of First-Come First-Served queueing systems with peaked inputs', *Bell System Technical Journal* **52**, 1215-1228.

Heffes, H. (1976), 'On the output of a GI/M/N queueing system with interrupted Poisson input', *Operations Research* **24**, 530-542.

Hemker, P.W. (editor) (1981), 'Numal, numerical procedures in Algol', *MC Syllabus* **47**, Amsterdam.

Herzog, U., Woo, L and Chandy, K.M. (1975), 'Solution of queueing problems by a recursive technique', *IBM Journal of Research and Development* **19**, 295-300.

Hillier, F.S. and Yu, O.S. et al. (1982), *Queueing tables and graphs*, North Holland Publishing Company, New York.

Hokstad, P. (1978), 'Approximations for the M/G/m queue', *Operations Research* **26**, 511-523.

Hokstad, P, (1980), 'The steady state solution of the $M/K_2/m$ queue', *Advances in Applied Probability* **12**, 799-823.

van Hoorn, M.H. (1981), 'Algorithms for the state probabilities in a general class of single server queueing systems with group arrivals', *Management Science* **27**, 1178-1187.

van Hoorn, M.H. and Tijms, H.C. (1982), 'Approximations for the waiting time distribution of the M/G/c queue', *Performance Evaluation* **2**, 22-28.

Hordijk, A. and Tijms, H.C. (1976), 'A simple proof of the equivalence of the limiting distributions of the continuous time and the embedded process of the queue size in the M/G/1 Queue', *Statistica Neerlandica* **30**, 97-100.

Ishikawa, A. (1979), 'On the equilibrium solution for the queueing system $GI/E_k/m$', *TRU Math.* **15**, 47-66.

Keilson, J. (1965), *Green's function methods in probability theory*, Hafner, New York.

Kendall, D.G. (1964), 'Some recent work and further problems in the theory of queues', *Theory of probability and its applications* **9**, 1-13.

Kiefer,J. and Wolfowitz,J. (1955), 'On the theory of queues with many servers', *Transactions of the Americain Mathematical Society* **78**, 1-18.

Kimura, T. (1982), 'Diffusion approximation for an M/G/m queue', Research report Tokyo Institute of Technology.

Kleinrock, L. (1975, 1976), *Queueing sysytems*, Volume I and II, Wiley, New York.

Köllerström, J. (1974), 'Heavy traffic theory for queues with several servers', *Journal of Applied Probability* **11**, 544-552.

Kosten, L. (1973), *Stochastic theory of service systems*, Pergamon press, New York.

Krämer, W. and Langenbach-Belz, M. (1976), Approximate formulae for the delay in the queueing system GI/G/1', *Proceedings ITC 8*, Melbourne.

Kuczura, A. (1973), 'The interrupted Poisson process as an overflow process', *Bell System Technical Journal* **52**, 437-448.

Kühn, P. (1972), *On the calculation of waiting times in switching and computer systems*, Institute of Switching and Data Technics, University of Stuttgart.

Kühn, P. (1976), *Tables on Delay Systems*, Institute of Switching and Data Technics, University of Stuttgart.

Lavenberg, S.S. (1975), 'The steady state queueing time distribution for the M/G/1 finite capacity queue', *Management Science* **21**, 501-506.

Manfield, D.R. and Tran Gia, P. (1981), 'Queueing analysis of scheduled communication phases in distributed processing systems', *Proceedings of International Symposium on Computer Performance Modelling, Measurement and Evaluation*, 1981, Amsterdam.

Manfield, D.R. and Tran Gia, P. (1981), 'Analysis of a finite storage system with batch input arising out of message packetisation', Research Report University of Siegen, Siegen, FRG.

Marshall, K.T. (1968), 'Some Inequalities in Queuing', *Operations Research* **16**, 651-665

Marshall, K.T. and Wolff, R.W. (1971), 'Customer average and time average queue lengths and waiting times', *Journal of Applied Probability* **8**, 535-542.

Neuts, M.F. (1981), *Matrix geometric solutions in stochastic models - an algorithmic approach*, The John Hopkins University Press.

Newell, G. (1973), 'Approximate stochastic behaviour of n-server service systems with large n', *Lecture notes in Economics and Mathematical Systems* **87**, Springer Verlag, Berlin.

Nozaki, S.A. and Ross, S.M. (1978), 'Approximations in finite capacity multiserver queues with Poisson arrivals', *Journal of Applied Probability* **15**, 826-834.

Pack, C.D. (1978), 'The output of multiserver queueing systems', *Operations Research* **26**, 492-509.

Posner, M. (1973), 'Single server queues with service time dependent on waiting time', *Operations Research* **21**, 610-616.

Ross, S.M. (1970), *Applied probability models with optimization applications*, Holden-day, San Francisco.

Sakasegawa, H. (1978), 'Numerical tables of the queueing systems $E_k/E_2/s$', *Computer science monographs* **10**, The institute of statistical mathematics, Tokyo.

Schassberger, R. (1973), *Warteschlangen*, Springer-Verlag, Berlin.

Shum, A.W. (1976), *Queueing models for computer systems with general service time distributions*, Garland Publishing Company, New York.

Smit, J.H.A. de (1981), 'The queue GI/M/s with customers of different types or the queue $GI/H_m/s$', Research Report 364 Technical University Twente, Netherlands.

Stidham, S. Jr. (1972), 'Regenerative processes in the theory of queues with applications to the alternating priority queue', *Advances in Applied Probability* **4**, 542-557.

Stoer J. and Bulirsch, R. (1980), *Introduction to numerical analysis*, Springer Verlag, New York.

Takacs, L. (1962), *Introduction to the theory of queues*, Oxford University Press, New York.

Takahashi, Y. and Takami, Y. (1976), 'A numerical method for the steady state probabilities of a GI/G/c queueing system in a general class', *Journal of the Operations Research Society of Japan* **19,** 147-157.

Takahashi, Y. (1977), 'An approximation formula for the mean waiting time of an M/G/c queue', *Journal of the Operations Research Society of Japan* **20,** 150-163.

Takahashi, Y. (1981), 'Asymptotic exponentiality of the tail of the waiting time distribution in a Ph/Ph/c queue', *Advances in Applied Probability* **13**, 619-630.

Tijms, H.C. and Federgruen, A. (1980), 'Computation of the stationary distribution of the queue size in an M/G/1 queueing system with variable service rate', *Journal of Applied Probability* **17**, 515-522.

Tijms, H.C., van Hoorn, M.H. and Federgruen, A. (1981a), 'Approximations for the steady state probabilities in the M/G/c queue', *Advances in Applied Probability* **13**, 186-206.

Tijms, H.C., van Hoorn, M.H. (1981b), 'Algorithms for the state probabilities and waiting times in single server queueing systems with random and quasirandom input and phase-type service times', *OR Spektrum* **2**, 145-152.

Tijms, H.C. and van Hoorn, M.H. (1981c), 'Computational methods for Single Server and Multi Server queues with Markovian input and general service times', in: R.E. Disney (editor) *Proceedings of the special TIMS Meeting on Applied Probability - Computer Science*, Birkhauser, Boston.

Tran Gia, P. and van Hoorn, M.H. (1982), 'Dependency of service time on waiting time in switching systems - A queueing analysis with aspects of overload control', Research Report 83 Vrije Universiteit, Amsterdam.

Whitt, W. (1982), 'The Marshall and Stoyan bounds for IMRL/G/1 queues are tight', Report Bell Labs, Holmdel, New Jersey.

Wolff, R.W. (1982), 'Poisson arrivals see time averages', *Operations Research* **30**, 223-231.

Yechiali, U. and Naor, P, (1971), 'Queueing problems with heterogeneous arrivals and service', *Operations Research* **19**, 722-734.

CWI TRACTS

1 D.H.J. Epema. *Surfaces with canonical hyperplane sections.* 1984.

2 J.J. Dijkstra. *Fake topological Hilbert spaces and characterizations of dimension in terms of negligibility.* 1984.

3 A.J. van der Schaft. *System theoretic descriptions of physical systems.* 1984.

4 J. Koene. *Minimal cost flow in processing networks, a primal approach.* 1984.

5 B. Hoogenboom. *Intertwining functions on compact Lie groups.* 1984.

6 A.P.W. Böhm. *Dataflow computation.* 1984.

7 A. Blokhuis. *Few-distance sets.* 1984.

8 M.H. van Hoorn. *Algorithms and approximations for queueing systems.* 1984.

9 C.P.J. Koymans. *Models of the lambda calculus.* 1984.