Queueing Models with Dependence Structures

Marco Combé

Queueing Models with Dependence Structures

Proefschrift

ter verkrijging van de graad van doctor
aan de Katholieke Universiteit Brabant,
op gezag van de rector magnificus, prof.dr. L.F.W. de Klerk,
in het openbaar te verdedigen ten overstaan van
een door het college van dekanen aangewezen commissie
in de aula van de Universiteit
op vrijdag 27 januari 1995 te 16.15 uur

door

Marco Bastiaan Combé

geboren te Rotterdam

Promotor: Prof.dr.ir. O.J. Boxma

The Penrose pattern used on the cover is generated by a set of two tiles, named darts and kites, which tiles only non-periodically. By 'only' is meant that neither a single shape or subset nor the entire set tiles periodically, but by using all of them a non-periodic tiling is possible. The number of Penrose patterns consisting of darts and kites is uncountable. Although it is possible to construct Penrose patterns with a high degree of symmetry, most patterns are a mystifying mixture of order and unexpected deviation from order. A Penrose pattern is made by starting with darts and kites around a vertex and expanding radially. An alternative start can be made with a combination of ten identical isosceles triangles; a decapod. Ignoring rotations and reflections there are 62 different decapods. Surprisingly 60 decapods force a tiling of one specific type: an infinite cartwheel pattern. Only two do not. The decapod chosen for the cover also allows patterns of other kinds.

Voorwoord

Het is een genoegen een ieder te bedanken die heeft bijgedragen aan de totstandkoming van dit proefschrift. Enkelen wil ik graag in het bijzonder noemen. Dit proefschrift is het resultaat van het promotie-onderzoek dat ik gedurende de afgelopen vier jaar op het CWI verricht heb in het kader van het NFI-project "Prestatie-analyse en regeling van gespreide computersystemen". Ik ben NFI en het CWI erkentelijk voor de mij geboden ondersteuning.
Veel dank ben ik verschuldigd aan Onno Boxma; ik besef dat ik mij gelukkig mag prijzen met de zeldzame combinatie van zijn professionele begeleiding, wiskundige expertise, sociaal karakter en vooral humor.
Mijn OIO-schap zou niet hetzelfde zijn geweest zonder het gezelschap van kamergenoot en mede-OIO Sem Borst. Onze gesprekken vormden een verrijking van mijn periode op het CWI.
Van mijn collega's op het CWI wil ik vooral Adri Steenbeek en Peter de Waal bedanken voor het verzachten van de af en toe pijnlijke omgang met computers en software.
Aan Katja tenslotte ben ik zeer veel dank verschuldigd. Haar liefde en steun waren onmisbaar.

Amsterdam, oktober 1994
Marco Combé

Contents

Chapter 1

Introduction

1.1 Motivation

The explosive growth of the exchange of information in today's society places high demands on the carriers of the information, which are the communication and computer networks. The traffic must be controlled in such a way that an acceptable level of service can be guaranteed for the users of the networks. The requirements of the infrastructure not only concern hardware technology but also touch on the quality of network design and the efficiency of control mechanisms.

Queueing models form a natural paradigm for the study of the latter issues. Accordingly, queueing theory plays an important role in the design, dimensioning, fine-tuning, control, and performance evaluation of communication and computer systems.

Modern networks operate at high transmission speed and are intended for a high level of utilization, while the traffic itself has evolved to highly irregular processes with typical "non-Poisson" behaviour. The increasing intensity and complexity of communication traffic remains a driving force behind queueing research, urging for analysis and new modelling tools.

This monograph discusses queueing models and methods for the analysis and optimization of the behaviour of communication and computer networks.

In Sections 1.2 - 1.3 we present the main themes of this thesis: Section 1.2 considers dependence structures in traffic input processes, Section 1.3 reviews traffic controlling. Section 1.4 concludes this chapter with an outline of the thesis. Regarding the references in this chapter, we have restricted ourself to mentioning general contributions. More detailed overviews of related literature are presented in the introductions of the chapters.

1.2 Dependence structures in input processes of queueing systems

In the early days of queueing analysis the aim was to describe the behaviour of processes in elementary queueing systems in terms of the system parameters. The technical difficulties arising in the analysis were mainly due to the stochastic nature of the input processes, the latter being sequences of random variables for the interarrival times and the service requests of customers. The analysis helped to build up intuition and also revealed fundamental properties of queueing systems, one example being the phenomenon that mean waiting times and mean queue lengths behave proportional to $\frac{1}{1-\rho}$, where ρ denotes the traffic offered to the system. In the early queueing models the characteristics of the interarrival and the service time distributions remain constant in time. Moreover, customers behave more or less independent from one another, i.e. the interarrival and service times of customers are independent.

The elementary queueing models have been studied in great detail. Unfortunately, early queueing models have limited ability to adequately model the typical traffic characteristics of modern communication networks. This is due to the fact that in modern queueing systems the interarrival times and service requests of customers (messages, jobs) no longer are independent or identically distributed. Typical examples are streams of packets arriving in "bursty" arrival processes or varying arrival rates of the arrival process over periods of time. Unlike the traditional queueing models, the observation of the sequence of interarrival times, or the sequence of service requests, gives extra information on future interarrival and service times, e.g. observing relatively short interarrival intervals might indicate some kind of "rush-hour".

Fendick et al.[47] consider three types of correlation structures in the sequences of interarrival and service times: between consecutive interarrival times, between consecutive service times, and between the interarrival and service time of a customer. All three types arise in a natural way in queueing systems.

Dependence between interarrival times and between service times

In [47] it is argued that dependence between consecutive interarrival times is mainly caused by two factors. Firstly, as communication channels are usually designed for more than one user, various arrival processes are superpositioned. In general the superpositioning of independent renewal processes is *not* a renewal process. A nice illustration of this phenomenon is the ticking of clocks that run slightly out of phase. A second reason for correlated interarrival times is that messages are packetized, so a single message that is generated at user level results in a burst of a large number of small packets at cell level. A common way to model this is to view the arrival process as traffic generated by sources that alternate between periods of silence (Off periods) and periods of transmission (On periods). In this On/Off model the transmission period starts with the arrival of a message, and the relatively short interarrival times during the transmission period reflect the burst of packets triggered by this arrival. Superpositioning the arrival processes generated by such sources leads

to á model with periods of varying arrival rates.
The reasons for dependence between consecutive service times can be quite similar. A source that starts transmitting messages might have its typical characteristics of message length. In that situation, observing the sequence of service requests gives information about which source is transmitting.
A second reason for correlation between successive service times can be caused by the characteristics of the service facility. For example in a multi-server queueing model the number of operative servers might not be the same at all time, it might vary due to maintenance activities or breakdowns. We remark that in a model with varying system configurations the output process might indicate a status of the queuing system, for example the number of operative servers.

The two above-described dependence structures have received quite some attention over the years. The main approach to model and analyse queues with these structures is to consider the queueing model as in some way being directed by an exogenous process. The state of this process may contain information about the parameters of the current arrival process and the service time distributions of customers present in the system; or it may contain information about current system characteristics, like in the previous example the number of servers available. For reasons of mathematical tractability the exogenous process is usually a finite state Markov process. Queueing models that are directed in this way are usually referred to as Markov modulated queueing systems.
Over the past years numerous papers discussing these Markov modulated queueing systems have appeared. A brief survey of literature is presented in Chapter 7, in which a typical Markovian queue is analysed. One of the first Markov modulated queueing models to be introduced, and also the one that has received the most attention over the years, is the single-server queue with the Markov Modulated Poisson Process (MMPP) as the arrival process. In the MMPP the state of the exogenous Markov process determines the arrival rate of a Poisson process. This model has been extensively used to model correlations in the interarrival process, such as arise in the above-mentioned On/Off models. The potential of Markov models has already been recognized in 1962 by Loynes[71]; Naor & Yechiali[77] and Neuts[79] were the first to consider the MMPP queueing model, these papers appeared in 1971. An extensive overview of the MMPP and the MMPP/G/1 queue is presented in Fischer & Meier-Hellstern[48], while recent applications are discussed in Grünenfelder & Robert[53] and Bonomi et al.[17].

Dependence between interarrival and service times
Dependence between the interarrival and service time of a customer arises in a somewhat different way. In communication networks the existence of such dependence can occur in the following way. Messages pass through a number of switches before reaching their destination. At each switch a queueing situation may occur as messages wait in buffers for a clear path to the next link.

The processing of a message through a network can be viewed as a customer receiving service from a number of servers in series. Since propagation times are proportional to message lengths, successive service times in such a tandem configuration are correlated. As the interarrival time of a customer at a queue is strongly connected to its service time at the previous queue, this creates a positive correlation between interarrival and service times.
Dependence between interarrival and service times might also arise as a result of operational aspects. In particular we have communication networks with collection or reservation protocols in mind. An example is a bridge queue connecting a network with other networks, where at certain intervals the messages in the network that are destined for other networks are collected and delivered at the bridge queue. In this situation the number of messages collected, and hence the collected bulk message, is positively correlated to the time between collections.
In contrast with the dependence structures that we described earlier in this section, dependence between interarrival and service times of customers has received relatively little attention. With the results presented in Chapter 2,3 and 4 we hope to have filled up part of this void in queueing theory. For a further introduction and a survey on related literature on this subject we refer to Section 1 of Chapter 2.

1.3 Traffic control in queueing networks

The suppliers of communication facilities or services must offer an acceptable *Quality Of Service* (QOS) to their clients, the *service* being the transportation of information, the *quality* being defined in terms such as delay, loss, and reliability, with the observation that different clients can have different priorities regarding these aspects. For example, for data transfer the messages must be received unaltered, hence avoiding packet loss and data distortion receives highest priority; for real-time communication, such as voice or video, delay is the most important factor.
Given a certain level of technology, there are two instruments to obtain the desired level of service in communication networks. Firstly, one has options in the design of the network; this is mainly a question of resource management. We will not directly consider design issues in this thesis. The second means, to which we will restrict ourself in this section, is the operational management of the network; here the aim is to find effective mechanisms to control the communication traffic. Obviously, as both instruments are strongly connected, *decisions* regarding control mechanisms and design can not be made independently.
Concentrating on control mechanisms, one may make a distinction between access protocols and routing policies. The function of access protocols is to organize the acceptance of new messages (jobs). Routing policies prescribe by which servers (viz. communication channels, processors) a customer (message, job) is handled. The aim of control policies is to provide a level of service that

is acceptable for all users; note again that different users might have different priorities. Next we discuss access protocols and routing policies in some detail.

(i) Access protocols
Focusing on access protocols, the increasing momentum in the development of communication technology, network design and control protocols has led to an extensive literature on this subject; a large number of existing and new suggestions for access protocols have been discussed and evaluated in numerical and analytical performance studies. However, since we only indirectly touch on the subject of access protocols, it is not in the scope of the thesis to extensively discuss the developments in this area.
In this thesis we discuss two models that are related to access protocols:
-The dependence structure between interarrival and service times that we introduced in the previous section can be viewed in the light of access protocols with a reservation/collecting mechanism.
-In Chapter 5 we consider a queueing model in which for each customer at the moment of arrival it is decided whether this customer joins the queue or not, the decision being based on the amount of work present in the system at that time.

(ii) Routing policies
In communication networks routing policies control the path(s) a message follows from its origin to its destination. In computer networks in which the processing capacity is distributed, routing (job assignment) is applied to make efficient use of the processing capacity.
From a mathematical viewpoint there are two main aspects of assignment protocols:
Load sharing: distributing the load over the servers of the system in order to make efficient use of capacity.
Reducing variability: in systems that are not saturated, overflow losses and delay are caused by the stochastic nature of interarrival and service times. By reducing variability of these processes, the performance of the system can be drastically improved. In communication networks "regularizing" the arrival process is known as traffic shaping.

Besides mathematical aspects there are a number of considerations of an operational and economical nature, such as robustness, failure tolerance, implementation, and overhead.

Taking all these considerations into account, the decision for a routing policy by a network designer might be guided by considering two main characterizing features of routing policies (cf. Wang & Morris[100]):

Static vs. dynamic
An important aspect is the use of state information of the queueing network. The amount of detail may range from a complete knowledge of all events in the system, including the size of service requests and current queue lengths, to the situation in which only basic characteristics of the system are used. In this

spectrum the policies that make active use of state information are referred to as *dynamic policies*, whereas policies that only use basic characteristics are called *static policies*. Since the state information in dynamic policies is more accurate, in general the mathematical performance of dynamic policies is better than that of static policies. However, one has to make a trade-off between the better performance of dynamic policies and the costs associated with the additional exchange, processing and storage of information.

Source vs. server initiative
In a multiple-source multiple-queue environment both sources and servers can take the initiative in the allocation of the customers (messages, jobs). In source initiative policies the source of a customer decides to which server the customer is routed. In server initiative policies a server decides from which source to obtain a next customer. These are again extremes of a spectrum; an allocation policy might also be characterized as a a hybrid of source and server initiative policies.

In Chapter 6 we present a study on the optimization of static traffic allocation policies, concentrating on the mathematical aspects of routing policies. There we also present a survey of literature related to traffic allocation. Here we mention two general papers on control and design of queueing systems. Wang & Morris[100] present a systematic description of load sharing policies for distributed systems. Their aim is to categorize and compare these policies. In Towsley[97] the discussion concentrates around the earlier-introduced concept *quality of service*. Towsley[97] discusses many issues such as: call admission, monitoring and shaping of traffic, routing, and buffer management. Wang & Morris[100] approach the subject from a controller's viewpoint, starting by the *means* of network operators, in Towsley[97] the starting point is formed by the *objectives*. Both papers provide an extensive list of references.

1.4 Outline of the thesis and overview of the results

Chapters 2,3 and 4 are devoted to single-server queues in which the interarrival and service times of a customer are positively correlated. The dependence structure can be physically represented by a single-server queue in which customers are not immediately allowed to join the queue, but instead have to wait for a gate to be opened (gated admission); alternatively, the model can be regarded as a queue in which the customers have to wait at a bus-stop for a bus to collect them (customer collection). Both interpretations of the model indicate some kind of access control mechanism. Chapter 2 discusses the basic model, Chapters 3 and 4 generalize this model in two different ways.
In Chapter 2 we study the dependence structure from the perspective of a collecting procedure; individual customers arrive at a bus-stop where they have to wait for a bus to be collected and to be delivered as a batch customer at the queue of a single-server facility. In the basic model the intercollecting intervals are exponentially distributed, hence the arrival process of busses at the service facility is Poisson. We derive the Laplace-Stieltjes Transform (LST)

of the stationary joint distribution of the waiting and service time of a batch customer. This result leads to LST expressions for the waiting times, sojourn times and generating functions for the numbers of customers, both for batch customers and for individual customers. A main contribution of Chapter 2 is the derivation of a work decomposition law by viewing the model from the perspective of a queue with server vacations.
Chapter 3 generalizes the arrival process of work at the bus-stop. Here we connect the dependence structure to the situation where work arrives at a gate according to a process with non-negative independent increments, and where at exponential intervals the gate is opened and - after the addition of an independent component - the work is delivered to the server as a single customer. The work in Chapter 3 constitutes a unification of recently studied $M/G/1$ queues with dependence between interarrival and service times, including the customer collection procedure of Chapter 2. Similar to Chapter 2 we derive the LST of the joint stationary distribution of the waiting and service time of a batch customer. By viewing the queueing system as a dam model with server vacations we derive a work decomposition property.
As stated in the first sections of this introduction, the arrival processes in modern queueing systems are in general not Poisson processes. In Chapter 4 we generalize the basic collector model of Chapter 2 to a queue with customer collecting in which the arrival process of customers and the process of collecting are more general processes than the Poisson process. This Chapter concentrates on the modelling of dependence between interarrival and service times with the use of Markovian queueing systems (cf. Section 1.2), in particular we apply the framework of the Batch Markovian Arrival Process (BMAP) and the $BMAP/G/1$ queue. In the BMAP batch arrivals of customers are generated at the moments of transition in a finite state Markov process. The BMAP can be considered as a matrix generalization of the Poisson process and is a very general non-renewal batch arrival process. Throughout this thesis we encounter the BMAP or special variants of the BMAP, which motivates the inclusion of an appendix dedicated to the BMAP and the $BMAP/G/1$ queue (cf. Appendix A).

In Chapter 5 we discuss a queueing model with customer impatience; on arrival customers decide whether to join the queue or not, their decision being based on the work present at that moment, i.e. the actual waiting time of the arriving customer. Another way of looking at this mechanism is to consider the model as a queue with restricted access; at the moment of the arrival of a customer it is decided whether the customer is allowed to join the queue or not, and again the decision is based on the amount of work in the system present at that time. This mechanism is studied for a single-server queue with a Markovian Arrival Process (MAP) as the arrival process and with generally distributed service times. The MAP is a special case of the BMAP: the MAP is a Markovian arrival process of single customers. The amount of time an arriving customer is prepared to wait, or the amount of unfinished work in the system beyond which

a customer is rejected, is a random variable with an exponential distribution. In Chapter 5 we derive the LST for the stationary distribution of the amount of work in the system. From this we obtain expressions for the stationary waiting time of a customer and the steady state probability that a customer will be rejected.

In Chapter 6 we present a study on optimal static allocation policies for customers that arrive according to a Poisson process at a service facility and which have to be allocated over a group of servers in parallel. We consider two types of policies: probabilistic assignment and pattern allocation. Under probabilistic assignment customers are allocated over the queues according to fixed probabilities, under pattern allocation customers are assigned to the queues according to an allocation table. We present an optimization procedure for pattern allocation. By approaching static allocation from a theoretical viewpoint and by comparing various approaches, we develop insights for general allocation problems and clarify some reported, but hitherto unexplained properties.
In this chapter we again encounter the MAP, here as a result of the typical arrival process that occurs at a queue when customers are routed to that queue according to a fixed pattern.
Chapter 6 also discusses three extensions of the traffic allocation problem. The first extension considers a general arrival process of customers at the routing point. The second extension describes the case in which all queues receive a Poisson arrival stream, on top of which an external arrival stream has to be allocated. The third extension considers allocation to multiple-server stations. This last problem gives rise to an interesting queueing model which may be characterized as a multiple-server queue with BMAP input.

The latter queueing model is studied in Chapter 7 for the case of exponentially distributed service times. The study of this so-called $BMAP/M/s$ queue supports an overview of current methods for analysing Markov modulated queueing systems. In particular we concentrate on three methods that exploit homogeneity structures in the Markov process that describes the number of customers in the system. With each of the three methods we derive the generating function for the stationary probability vector of the number of customers in the $BMAP/M/s$ queue. The aim of the overview is to point out differences and similarities between these important methods.
To illustrate the results, the $BMAP/M/s$ queue is applied to the multi-server generalization of the dependence structure that is discussed in Chapter 2 by using the modelling technique that is presented in Chapter 4. In this generalization the batch of collected customers is delivered at a single-queue multi-server service facility.

The results to be presented in the next chapters were obtained during the course of the author's thesis research at CWI, Amsterdam. Some of the results have already appeared in the open literature. The results in Chapter 2 are based on Borst, Boxma, & Combé[18, 19, 20], and on Borst & Combé[21]. Chapter 3 is based on Boxma & Combé[25]. The results in Chapter 4 are based on Combé[37]. Chapter 5 is based on Combé[38]. In Chapter 6 we present the results obtained in Combé & Boxma[36]. The results of Chapter 7 are obtained from Combé[39].

Throughout the thesis we apply the following conventions for bibliographical references and numbering.
Stochastic variables are printed bold, stochastic variables that are not denoted by Greek symbols are in capital. References to the literature are presented as the name(s) of the author(s) followed by the corresponding index in the reference list or, in some cases of repeated occurrence, as the index in the reference list only, omitting the name(s) of the author(s). The chapters are each divided in a number of sections. Formulas, figures, and tables are numbered per chapter, e.g., formula (3.5) is the fifth numbered formula in Chapter 3. Assumptions, corollaries, conjectures, examples, lemma's, properties, remarks, theorems, etc. are numbered per section, e.g., Theorem 2.3.1 is the first theorem in Section 3 of Chapter 2. Appendices that have global reference are included after Chapter 7 and are indicated by capitals, e.g., Appendix A. Appendices only referring to a particular chapter are included directly after the chapter in question. They are numbered by chapter and are indicated by capitals, e.g., Appendix 6.B is the second appendix of Chapter 6.

Chapter 2

The $M/G/1$ queue with dependence between interarrival and service times

The next three chapters consider single-server queues in which the service times of arriving customers depend on the length of the interval between their arrival and the previous arrival. The dependence structure that is studied in Chapter 2 arises when individual customers arrive at pick-up points according to a Poisson process, while customer collectors are sent out according to a Poisson process to collect the customers and to bring them to the service facility. In this case the collected numbers of customers, and hence the total collected service requests, are positively correlated with the corresponding collecting intervals. Two generalizations of this collecting procedure are discussed in Chapter 3 and 4 respectively.

2.1 Introduction

Consider the following situation. Customers arrive at pick-up points ("bus-stops") according to independent Poisson processes. At these pick-up points they wait for a bus to bring them to a single-server service facility (e.g., the check-in counter of a hotel). Busses with unlimited customer capacity move according to a fixed route along the pick-up points, with fixed speed, collecting all waiting customers that they encounter and finally delivering all collected customers at the service facility. The intervals between the starts of successive bus tours are exponentially distributed. Because of the fixed tour length, the arrival process of busses at the service facility is a Poisson process. Viewing a batch of customers that is brought to the service facility by a bus as one supercustomer, the service facility very closely resembles an ordinary $M/G/1$

queue; the only difference is that *the service time of a supercustomer depends on the previous interarrival time.* Indeed, if two consecutive busses arrive at a relatively long (short) interval, then the second bus is likely to pick up relatively many (few) customers: the interarrival time and the size of the picked-up batch are positively correlated, and hence so are the interarrival time and the supercustomer service time.

In the present chapter we analyse the $M/G/1$ queue with the above-sketched correlation structure between interarrival and service times. In Chapter 3, the arrival process of work at the pick-up point is generalized. There the service time of a supercustomer consists of two components: a service request that is independent of the interarrival time, i.e. an 'ordinary' $M/G/1$ amount of work, plus a service request that *is* dependent on the interarrival time, the amount of work associated with this request being the result of a work accumulation process that is a generalization of the Poisson arrival process of single customers considered in this chapter. In the basic collector model both the arrival process of customers at the bus-stop and the process of collecting are Poisson processes. In Chapter 4 we discuss a generalization of the basic model in which these processes are more general than the Poisson process.

Our motivation for studying correlated interarrival and service times is twofold. Firstly, until a few years ago it has almost exclusively been assumed that there exists no dependence between arrival intervals, between service times and between interarrival and service times. In recent years dependence between consecutive arrival intervals and between consecutive service times has received quite some attention. However, dependence between the interarrival and service time of a customer was hardly explicitly studied. The main reason for ignoring this dependence structure seems the mathematical complexity that it gives rise to. However, the present model allows a very detailed analysis of most performance measures of interest, thus giving valuable insight into the effect that dependence between interarrival and service times may have on those performance measures.

Our second motive for the analysis is that the correlation structure under consideration seems quite natural in many queueing situations. For example, it arises when customers are collectively brought to a central service facility. Examples are mail pick-up and the pick-up of customers at airport terminals. In computer-communications, one might think of the collection of packets in a 'train' in a Local Area Network with interconnected rings, to be delivered at a bridge queue. Furthermore, modern reservation protocols for the use of transmission slots in high-speed Local and Metropolitan Area Networks also may give rise to customer collection.

Besides collecting mechanisms, dependence between interarrival and service times also arises in a natural way in queueing networks. Kleinrock[65] discusses the traffic characteristics of message flows in a message switching communication network. In the corresponding queueing network the service times of messages at each queue are proportional to the message length. In a two-stage tandem configuration this would lead to a strong positive correlation between

interarrival and service times at the second queue.
Next we return to the basic collecting model. In the remainder of this section we present a model description, a survey of related literature, and an overview of the chapter.

Model description
Individual customers require service from a service facility with a single server. Their service times are independent, identically distributed stochastic variables with distribution $B(\cdot)$, with mean β_1, second moment β_2 and Laplace-Stieltjes Transform (LST) $\beta(\cdot)$; $B(0+) = 0$. These individual customers arrive at a pick-up point according to a Poisson process with rate λ. They are collected by a collector and delivered in batches at the service facility at times $\mathbf{t}_1, \mathbf{t}_2, \ldots$. The collecting intervals $\boldsymbol{\sigma}_n := \mathbf{t}_n - \mathbf{t}_{n-1}, n = 1, 2, \ldots$, with $\mathbf{t}_0 = 0$, are independent, negative exponentially distributed stochastic variables with mean $1/\gamma$. A delivered batch of customers can be viewed as one supercustomer. Batches (supercustomers) apparently arrive at the service facility according to a Poisson process with rate γ. In the sequel we use both the terms supercustomers and batch customers to refer to the collected batches of customers.

Remark 2.1.1
Instead of assuming that there is a single Poisson arrival stream of individual customers with service time distribution $B(\cdot)$, we could also have allowed several independent Poisson arrival streams at various pick-up points, with different service time distributions. A distinction between the various arrival streams may be useful in certain applications, where e.g. waiting times of individual customers must be determined while taking into account the location of the pick-up point and the travel time of the collector to the service facility. Multiple pick-up points and non-zero deterministic travel times can easily be implemented into our model without seriously complicating the analysis. In the sequel the travel time of the collector from the pick-up point to the service facility is assumed to be zero. Thus the arrival of the batch customer at the service facility coincides with the arrival of the collector at the pick-up point. □

Define the total offered traffic load as $\rho := \lambda\beta_1$. In Appendix 2.A, following the lines of Laslett et al.[69], we prove that $\rho < 1$ is a sufficient condition for the waiting and the service times to have a proper joint limiting distribution. It follows that the same holds for the other quantities under consideration: sojourn times, queue lengths and busy periods. We remark that in a more general framework Loynes[71] showed that $\rho < 1$ is a *necessary* and sufficient condition for the waiting times to have a proper limiting distribution.

After this global model description we consider batch sizes, service times of batch customers and the correlation structure of the model in some more detail. The number of individual customers, $\mathbf{K}_n$, constituting the n-th batch customer, is distributed as the number of arrivals in a Poisson process during the collecting

interval $\boldsymbol{\sigma}_n$. It follows that $\mathbf{K}_n$ for a collecting interval $\boldsymbol{\sigma}_n$ of length u has conditional distribution

$$\Pr\{\mathbf{K}_n = k \,|\, \boldsymbol{\sigma}_n = u\} = e^{-\lambda u}\frac{(\lambda u)^k}{k!}, \qquad u > 0, k = 0, 1, \ldots, \tag{2.1}$$

so

$$\begin{aligned}\Pr\{\mathbf{K}_n = k\} &= \int_{u=0}^{\infty} \gamma e^{-\gamma u}\Pr\{\mathbf{K}_n = k \,|\, \boldsymbol{\sigma}_n = u\}du \\ &= \frac{\gamma}{\gamma+\lambda}\left(\frac{\lambda}{\gamma+\lambda}\right)^k, \qquad k = 0, 1, \ldots .\end{aligned} \tag{2.2}$$

The service time $\boldsymbol{\tau}_n$ of the n-th batch customer, being the set of service requests associated with the individual customers that are collected, in a collecting interval $\boldsymbol{\sigma}_n$ of length u has conditional distribution

$$\begin{aligned}\Pr\{\boldsymbol{\tau}_n < t \,|\, \boldsymbol{\sigma}_n = u\} &= \sum_{k=0}^{\infty}\Pr\{\mathbf{K}_n = k \,|\, \boldsymbol{\sigma}_n = u\}B^{k*}(t) \\ &= \sum_{k=0}^{\infty} e^{-\lambda u}\frac{(\lambda u)^k}{k!}B^{k*}(t), \quad t \geq 0, u > 0.\end{aligned} \tag{2.3}$$

Hence

$$\mathrm{E}(e^{-\omega\boldsymbol{\tau}_n} \,|\, \boldsymbol{\sigma}_n = u) = e^{-\lambda(1-\beta(\omega))u}, \qquad Re\,\omega \geq 0, u > 0. \tag{2.4}$$

It follows that

$$\Pr\{\boldsymbol{\tau}_n < t\} = \sum_{k=0}^{\infty}\frac{\gamma}{\gamma+\lambda}\left(\frac{\lambda}{\gamma+\lambda}\right)^k B^{k*}(t), \qquad t \geq 0, \tag{2.5}$$

$$\mathrm{E}(e^{-\omega\boldsymbol{\tau}_n}) = \frac{\gamma}{\gamma+\lambda(1-\beta(\omega))}, \qquad Re\,\omega \geq 0. \tag{2.6}$$

The service times $\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \ldots$ of batch customers are independent, identically distributed stochastic variables. It should be noted that a batch may be empty, and hence a supercustomer may have zero service time:

$$\Pr\{\mathbf{K}_n = 0 \,|\, \boldsymbol{\sigma}_n = u\} = \Pr\{\boldsymbol{\tau}_n = 0 \,|\, \boldsymbol{\sigma}_n = u\} = e^{-\lambda u}, \quad u > 0, \tag{2.7}$$

$$\Pr\{\mathbf{K}_n = 0\} = \Pr\{\boldsymbol{\tau}_n = 0\} = \frac{\gamma}{\gamma+\lambda}. \tag{2.8}$$

From (2.6) it follows that

$$\mathrm{E}\boldsymbol{\tau}_n = \frac{\lambda\beta_1}{\gamma}, \qquad \mathrm{E}\boldsymbol{\tau}_n^2 = \frac{\lambda\beta_2}{\gamma} + 2\left(\frac{\lambda\beta_1}{\gamma}\right)^2. \tag{2.9}$$

The bivariate LST of $\boldsymbol{\sigma}_n$ and $\boldsymbol{\tau}_n$ follows from (2.3):

$$\mathrm{E}(e^{-\zeta\boldsymbol{\sigma}_n-\omega\boldsymbol{\tau}_n})=\frac{\gamma}{\gamma+\zeta+\lambda(1-\beta(\omega))}, \quad Re\,\zeta \geq 0, Re\,\omega \geq 0, \tag{2.10}$$

yielding the covariance of $\boldsymbol{\sigma}_n$ and $\boldsymbol{\tau}_n$

$$Cov(\boldsymbol{\sigma}_n, \boldsymbol{\tau}_n) = \frac{\lambda\beta_1}{\gamma^2}, \tag{2.11}$$

and the coefficient of correlation

$$correl(\boldsymbol{\sigma}_n, \boldsymbol{\tau}_n) = \left[1 + \frac{\gamma\beta_2}{\lambda\beta_1^2}\right]^{-1/2} \geq 0. \tag{2.12}$$

Note that $correl(\boldsymbol{\sigma}_n, \boldsymbol{\tau}_n) \uparrow 1$ for $\gamma \downarrow 0$ and for $\lambda \to \infty$, whereas $correl(\boldsymbol{\sigma}_n, \boldsymbol{\tau}_n) \downarrow 0$ for $\gamma \to \infty$ and for $\lambda \downarrow 0$. For $\gamma \to \infty$ the queue approaches an ordinary $M/G/1$ queue, as individual customers are collected instantaneously.

Related literature

A few studies have appeared that analyse a queueing system with correlation between the interarrival and service times. Cidon et al.[31] consider a correlation between the service time of a customer and the *subsequent* interarrival time; the waiting time for such a correlated queue can be analysed by studying the recurrence relation for the waiting time,

$$\mathbf{W}_{n+1} = \max\{0, \mathbf{W}_n + \boldsymbol{\tau}_n - \boldsymbol{\sigma}_{n+1}\},$$

in an ordinary $GI/GI/1$ queue with similarly distributed $\boldsymbol{\tau}_n - \boldsymbol{\sigma}_{n+1}$ and with $\mathbf{W}_n$ and $\boldsymbol{\tau}_n - \boldsymbol{\sigma}_{n+1}$ independent.

A number of papers has been devoted to the single-server queue with correlation between the service time of a customer and the *preceding* interarrival time. Conolly[40] and Conolly and Hadidi[42] consider an $M/M/1$ queue in which the service time and the preceding interarrival time are linearly related: $\boldsymbol{\tau}_n = \alpha\boldsymbol{\sigma}_n$. The same linear dependence structure is studied in Cidon et al.[32].

Conolly and Choo[41] study an $M/M/1$ queue in which $\boldsymbol{\sigma}_n$ and $\boldsymbol{\tau}_n$ have a bivariate exponential distribution with density

$$g(s,t) = \lambda\mu(1-r)e^{-\lambda s-\mu t}I_0[2\{\lambda\mu rst\}^{1/2}], \tag{2.13}$$

where $I_0[z]$ is a zero order modified Bessel function of the first kind, and where $r \in [0,1)$ is the correlation between $\boldsymbol{\sigma}_n$ and $\boldsymbol{\tau}_n$. For $r = 0$ the queue reduces to an ordinary $M/M/1$ queue. The marginal distributions of $\boldsymbol{\sigma}_n$ and $\boldsymbol{\tau}_n$ are negative exponential. Conolly and Choo analyse the waiting time distribution for this correlated $M/M/1$ queue, showing that its density can be expanded in a series of partial fraction terms. Their numerical calculations reveal that the positive correlation leads to a considerable reduction in mean waiting time. For the same model, (i) Hadidi[54] shows that the waiting times are hyper-exponentially distributed; (ii) Hadidi[55] examines the sensitivity of the waiting

time distribution to the value of the correlation coefficient; (iii) Langaris[68] studies the busy period distribution. However, the starting point of the latter study is wrong: it is assumed that a customer who starts a general busy period has an ordinary service time, whereas in this correlated queue a customer who starts a busy period is likely to have a relatively long interarrival time and hence a relatively long service time. This flaw is discussed in more detail in Borst & Combé[21], cf. Section 2.5.

Linear dependence as well as the bivariate exponential distribution are both examples of the generalized model that will be discussed in Chapter 3. A correlation structure that does not belong to the framework of the latter model is studied in Jacobs[62], in which heavy traffic results are obtained for the waiting time in queues with sequences of ARMA correlated negative exponentially distributed interarrival and service times. In [62] both interarrival and service times of the n–th customer are a weighted sum of an independent exponentially distributed component and a function of a one-step autoregressive random variable $\mathbf{A}_n$. The autoregressive structure is given by $\mathbf{A}_n = \theta \mathbf{A}_{n-1} + \mathbf{L}_n \mathbf{C}_n$, in which $\mathbf{C}_n$ is an exponentially distributed random variable, $\mathbf{L}_n$ is a $0-1$ random variable, and θ a given constant in $[0, 1)$.

The paper that is closest related to the present study is Takahashi[95]. He uses the terminology of 'gates' instead of 'busses'. Arriving customers first join a queue at the gate of the system; at exponentially distributed intervals the gate opens, and the customers at the first queue move to the second queue, where s servers are available. The gate closes immediately after all customers in the first queue have moved to the second queue. Takahashi's study is more restrictive than ours in the sense that he only allows exponentially distributed service times, and that he only studies sojourn times and queue lengths of *individual* customers. His study is more general than ours in the sense that he allows multiple servers. Furthermore, he also briefly considers the case in which the gate opens at fixed intervals. With the generalization presented in Chapter 4, together with the analysis of a special multi-server queue that is presented in Chapter 7 we are also able to analyse multi-server queues with dependence between the interarrival and service time. In addition the distributions of the interarrival times of customers and the gate opening intervals are more general than the exponential distribution.

Ishizaki et al.[60], inspired by [19], study a discrete-time, slotted version of the collector model. In [60] the number of time-slots passing between consecutive collecting epochs is geometrically distributed. At the end of each time-slot customers arrive in batches at the pick-up point, the batch sizes are independent and identically distributed. The main result presented in [60] is the joint distribution of the number of customers at the pick-up point and at the server.

Several other studies have been devoted to queueing systems with two stages of waiting, with a gate at the first queue and service provided only at the second queue; see e.g. Coleman[35] and Ali & Neuts[6]. However, in these studies the opening and closing of the gate is determined by the queue lengths, rather than by a Poisson point process.

In connection with reservation protocols in communication networks a few papers have analysed queueing models with collection procedures for customers for the case of *deterministic* collecting intervals. In a performance evaluation by Boxma, Levy and Yechiali [24] of the Cyclic-Reservation Multiple-Access (CRMA) protocol a customer collection procedure occurs for customers arriving according to a Poisson process. The deterministic collecting interval gives rise to a $D/G/1$ queue - in which obviously the interarrival and service times are *not* dependent. Bisdikian et al.[12] study the $D^X/D/1$ queue with generally distributed batch sizes. They also consider the case of batch size distributions that are the result of a collecting procedure for Poisson customers.

Remark 2.1.2
A more detailed model of the CRMA protocol, taking its back-pressure mechanism into account (Nassehi [78]) would lead to a model that is similar to ours, but in which the arrival process is closer to a deterministic process than to a Poisson process. In Chapter 4 we discuss a generalization of the current collector model that allows more general distributions for the collecting interval, including distributions that lead to more deterministic interval lengths.
A second approach for a model of the CRMA protocol would be to combine the known $D/G/1$ results and the results from the present chapter, to study the performance of CRMA with back-pressure. Performance measures like the sojourn time distributions might be approximated by a weighted sum of these distributions for the present $M/G/1$ case and the $D/G/1$ case with weight factors the squared coefficient of variation of the interarrival times of batch customers and one minus this coefficient. □

Organization of the chapter
In Section 2.2 we derive the LST of the joint distribution of the waiting time and the service time of a supercustomer. The main part of the analysis is devoted to the LST of their sum, the sojourn time of a supercustomer. In Section 2.3 we expand the results of Section 2.2, considering the sojourn time $\mathbf{R}$ and the waiting time $\mathbf{W}$ for supercustomers. We also examine these performance measures for *individual* customers. We relate the results for these performance measures to those in the $M/G/1$ queue *without* the collection procedure, i.e. the $M/G/1$ queue in which the individual customers do not wait to be picked up but immediately join the queue for service. We also compare mean waiting times in the model with the dependence structure to mean waiting times in the model with the same service time characteristics (cf. equation (2.4)), but in which the interarrival and service times are *independent*. Section 2.4 considers the joint distribution of the number of individual customers at the pick-up point and at the service facility. We also present the generating function of the queue length distribution for supercustomers. Section 2.5 is devoted to the busy period distribution. The mean busy period is easily obtained; the busy period *distribution* gives rise to mathematical difficulties which are discussed but not solved (cf. also Section 4.3).

Section 2.6 contains some numerical results, exposing the influence of the dependence on mean waiting times and variance of the busy period.

2.2 The joint distribution of the waiting and service time

In this section we derive the LST for the joint stationary distribution of the waiting and service time of an arbitrary supercustomer. The analysis also leads to the LST's of the steady state waiting and sojourn time distributions.
First some notation. Define the vector $(\mathbf{W}_n, \boldsymbol{\tau}_n)$, $n = 1, 2 \ldots$. $\mathbf{W}_n$ denotes the waiting time of a supercustomer, i.e. the time from the arrival of the supercustomer until the start of the service of the first individual customer belonging to the supercustomer. $\mathbf{R}_n$, the sojourn time of the $n-th$ customer, is defined as $\mathbf{R}_n := \mathbf{W}_n + \boldsymbol{\tau}_n$, so $\mathbf{R}_n$ denotes the time from the arrival of the $n-th$ supercustomer until the completion of service of the last individual customer belonging to the supercustomer. We note that if the service time of a supercustomer is zero, i.e. when there are no customers collected by this supercustomer, the waiting time is equal to the sojourn time, which is the time from the arrival of a supercustomer until the departure of that supercustomer.
Our analysis starts with the following recurrence relation for the vector $(\mathbf{W}_n, \tau_n)$, $n = 1, 2, \ldots$:

$$(\mathbf{W}_{n+1}, \boldsymbol{\tau}_{n+1}) = (\max\{0, \mathbf{W}_n + \boldsymbol{\tau}_n - \boldsymbol{\sigma}_{n+1}\}, \boldsymbol{\tau}_{n+1}). \tag{2.14}$$

Next define $F_n(x, y) := \Pr\{\mathbf{W}_n \leq x, \boldsymbol{\tau}_n \leq y\}$ for $x \geq 0, y \geq 0$. Also let $F_n^*(\omega_1, \omega_2)$, $Re\,\omega_1, Re\,\omega_2 \geq 0$ denote the LST of $(\mathbf{W}_n, \boldsymbol{\tau}_n)$. Denote the LST of $\mathbf{R}_n$ by $r_n(\omega)$, $Re\,\omega \geq 0$.
From expression (2.14) follows

$$F_{n+1}(x, y) = \Pr\{\mathbf{W}_n + \boldsymbol{\tau}_n - \boldsymbol{\sigma}_{n+1} \leq x, \boldsymbol{\tau}_{n+1} \leq y\}, \quad x, y \geq 0. \tag{2.15}$$

In equation (2.15) we observe that, conditional on $\boldsymbol{\sigma}_{n+1}$, $(\mathbf{W}_n + \boldsymbol{\tau}_n - \boldsymbol{\sigma}_{n+1})$ and $\boldsymbol{\tau}_{n+1}$ are independent. Then, with $\mathbf{R}_n = \mathbf{W}_n + \boldsymbol{\tau}_n$, we obtain

$$F_{n+1}(x, y) = \int_{u=0}^{\infty} \Pr\{\mathbf{R}_n \leq x + u\} \Pr\{\boldsymbol{\tau}_{n+1} \leq y \,|\, \boldsymbol{\sigma}_{n+1} = u\} \gamma e^{-\gamma u} du. \tag{2.16}$$

Next we derive $F_{n+1}^*(\omega_1, \omega_2)$. First define $R_n(x) := \Pr\{R_n \leq x\}$ for $x \geq 0$. Then from equation (2.16) and (2.4) we find for $x \geq 0, Re\,\omega_2 \geq 0$,

$$\begin{aligned} \hat{F}_{n+1}(x, \omega_2) &:= \int_{y=0}^{\infty} e^{-\omega_2 y} dF_{n+1}(x, y) = \omega_2 \int_{y=0}^{\infty} e^{-\omega_2 y} F_{n+1}(x, y) dy \\ &= \int_{u=0}^{\infty} R_n(x + u) e^{-\lambda(1-\beta(\omega_2))u} \gamma e^{-\gamma u} du, \end{aligned}$$

which leads to

$$\begin{aligned} F^*_{n+1}(\omega_1,\omega_2) &:= \int_{x=0}^{\infty} e^{-\omega_1 x} d\hat{F}_{n+1}(x,\omega_2) = \omega_1 \int_{x=0}^{\infty} e^{-\omega_1 x} \hat{F}_{n+1}(x,\omega_2) dx \\ &= \omega_1 \int_{u=0}^{\infty} \gamma e^{-(\gamma+\lambda(1-\beta(\omega_2)))u} \int_{x=0}^{\infty} e^{-\omega_1 x} R_n(x+u) dx\, du \\ &= \omega_1 \int_{z=0}^{\infty} e^{-\omega_1 z} R_n(z) \int_{u=0}^{z} \gamma e^{-(\gamma+\lambda(1-\beta(\omega_2))-\omega_1)u} du\, dz \end{aligned}$$

$$= \frac{\gamma\omega_1}{\gamma+\lambda(1-\beta(\omega_2))-\omega_1}\left[\frac{r_n(\omega_1)}{\omega_1} - \frac{r_n(\gamma+\lambda(1-\beta(\omega_2)))}{\gamma+\lambda(1-\beta(\omega_2))}\right], \qquad (2.17)$$

with $Re\omega_1, Re\omega_2 \geq 0$. The third equality in (2.17) follows from the substitution $z = x + u$ and changing the order of integration.
In (2.17) we see that $F^*_{n+1}(\cdot,\cdot)$ is expressed in terms of $r_n(\cdot)$. Moreover, $\mathbf{W}_{n+1} + \boldsymbol{\tau}_{n+1} = \mathbf{R}_{n+1}$ implies $F^*_{n+1}(\omega,\omega) = r_{n+1}(\omega)$, which leads for $Re\,\omega \geq 0$ to

$$r_{n+1}(\omega) = \frac{\gamma\omega}{\gamma+\lambda(1-\beta(\omega))-\omega}\left[\frac{r_n(\omega)}{\omega} - \frac{r_n(\gamma+\lambda(1-\beta(\omega)))}{\gamma+\lambda(1-\beta(\omega))}\right]. \qquad (2.18)$$

As observed in Section 2.1, for $\rho < 1$ the vectors $(\mathbf{W}_n, \boldsymbol{\tau}_n)$ and also the sojourn times $\mathbf{R}_n$ have a proper limiting distribution for $n \to \infty$. Let $\mathbf{W}$, $\boldsymbol{\tau}$ and $\mathbf{R}$ denote the random variables with the simultaneous limiting distributions of $\mathbf{W}_n, \boldsymbol{\tau}_n$ and $\mathbf{R}_n$ respectively, hence $\mathbf{R} \stackrel{d}{=} \mathbf{W} + \boldsymbol{\tau}$, with $\stackrel{d}{=}$ denoting equality in distribution. Denote the LST's of $(\mathbf{W}, \boldsymbol{\tau})$ and $\mathbf{R}$ by $F^*(\omega_1,\omega_2)$ and $r(\omega)$ respectively, $Re\,\omega_1, Re\,\omega_2, Re\,\omega \geq 0$.
By letting $n \to \infty$ in (2.17) we find for $Re\,\omega_1, Re\,\omega_2 \geq 0$,

$$F^*(\omega_1,\omega_2) = \frac{\gamma\omega_1}{\gamma+\lambda(1-\beta(\omega_2))-\omega_1}\left[\frac{r(\omega_1)}{\omega_1} - \frac{r(\gamma+\lambda(1-\beta(\omega_2)))}{\gamma+\lambda(1-\beta(\omega_2))}\right]. \qquad (2.19)$$

Letting $n \to \infty$ in (2.18) leads to

$$r(\omega) = \frac{\gamma\omega}{\omega-\lambda(1-\beta(\omega))}\,\frac{r(\gamma+\lambda(1-\beta(\omega)))}{\gamma+\lambda(1-\beta(\omega))}, \quad Re\,\omega \geq 0. \qquad (2.20)$$

In (2.19) we again observe that $F^*(\cdot,\cdot)$ is expressed in terms of $r(\cdot)$. Hence for the analysis of $F^*(\omega_1,\omega_2)$ it remains to determine $r(\omega)$. We proceed by solving (2.20) for $r(\omega)$.

Define

$$f(\omega) := \frac{\gamma\omega}{\omega-\lambda(1-\beta(\omega))}, \qquad Re\,\omega \geq 0, \qquad (2.21)$$

$$g(\omega) := \gamma + \lambda(1 - \beta(\omega)), \qquad Re\,\omega \geq 0. \tag{2.22}$$

Then we can write

$$r(\omega) = \frac{f(\omega)}{g(\omega)} r(g(\omega)). \tag{2.23}$$

Let

$$\begin{aligned} g^{(0)}(\omega) &:= \omega, && Re\,\omega \geq 0, \\ g^{(h)}(\omega) &:= g(g^{(h-1)}(\omega)), && Re\,\omega \geq 0,\ h = 1, 2, \ldots . \end{aligned}$$

Iterating (2.23) we find

$$r(\omega) = r(g^{(M+1)}(\omega)) \prod_{h=0}^{M} \frac{f(g^{(h)}(\omega))}{g^{(h+1)}(\omega)}, \qquad Re\,\omega \geq 0,$$

for any non-negative integer M.

Lemma 2.2.1

(i). The equation $\omega = g(\omega)$, $Re\,\omega \geq 0$ has a unique solution ω^*. ω^* is real.

(ii). $\lim_{M \to \infty} g^{(M)}(\omega) = \omega^*$ for all ω with $Re\,\omega \geq 0$.

(iii). $\prod_{h=0}^{\infty} \frac{f(g^{(h)}(\omega))}{g^{(h+1)}(\omega)}$ converges for all ω with $Re\,\omega \geq 0$.

Proof

The proofs of (i), (ii) and (iii) are based on the fact that $g(\cdot)$ is a contraction on $\{\omega \in \mathbb{C} | Re\,\omega \geq 0\}$. Moreover ω^* is real because $g(\cdot)$ is also a contraction on $\{\omega \in \mathbb{R} | \omega \geq 0\}$. Details of the proof can be found in Appendix 2.B. □

Lemma 2.2.1 implies that

$$r(\omega) = r(\omega^*) \prod_{h=0}^{\infty} \frac{f(g^{(h)}(\omega))}{g^{(h+1)}(\omega)}, \qquad Re\,\omega \geq 0. \tag{2.24}$$

Putting $\omega = 0$ in (2.24) we find

$$r(\omega^*) = 1 \Big/ \prod_{h=0}^{\infty} \frac{f(g^{(h)}(0))}{g^{(h+1)}(0)},$$

which leads to:

Theorem 2.2.1
The LST of the sojourn time distribution of a supercustomer is

$$r(\omega) = \prod_{h=0}^{\infty} \left[\frac{f(g^{(h)}(\omega))}{g^{(h+1)}(\omega)} \Big/ \frac{f(g^{(h)}(0))}{g^{(h+1)}(0)}, \right] \qquad Re\,\omega \geq 0, \tag{2.25}$$

with $f(\cdot)$ and $g(\cdot)$ given by (2.21) and (2.22). □

By substituting $r(\omega)$ in (2.19) we obtain an explicit expression for $F^*(\cdot,\cdot)$, the LST of the joint distribution of the waiting and service time of a supercustomer. Rewriting (2.19), using (2.22), we find:

Theorem 2.2.2
The LST of the joint distribution of the waiting and service time of a supercustomer is

$$F^*(\omega_1,\omega_2) = \frac{\omega_1}{\omega_1 - g(\omega_2)}[\omega_2 + \gamma - g(\omega_2)]\frac{r(\omega_2)}{\omega_2} - \frac{\gamma}{\omega_1 - g(\omega_2)} r(\omega_1), \tag{2.26}$$

with $Re\,\omega_1, Re\,\omega_2 \geq 0$, and with $r(\cdot)$ given by (2.25). □

Finally, the LST of the marginal distributions of the waiting time $\mathbf{W}$ and the service time τ follow from (2.26), or more easily from (2.19). In particular,

Theorem 2.2.3
The LST of the waiting time distribution of a supercustomer is

$$F^*(\omega,0) = \mathrm{E}\left(e^{-\omega \mathbf{W}} \right) = \frac{\omega r(\gamma) - \gamma r(\omega)}{\omega - \gamma}, \quad Re\,\omega \geq 0. \tag{2.27}$$

□

Remark 2.2.1
The LST's in (2.25) and (2.26) can be numerically inverted using a procedure outlined in Abate and Whitt[1]. Their procedure is based on the Fourier inversion formula for Laplace transforms, the resulting integral expression is numerically evaluated in a relatively elementary way. An interesting application of this method is the numerical analysis of the transient behaviour of queueing systems (cf. Choudhury et al.[29, 30]).

We conclude this section with a number of results which follow directly from the so far presented analysis. In the next section we interpret our results by relating the $M/G/1$ queue with the dependence structure between interarrival and service time to the ordinary $M/G/1$ queue.

Corollary 2.2.1
Moments of $\mathbf{W}$ and $\mathbf{R}$ can be obtained using (2.25). In particular we find by differentation:

$$\mathrm{E}\mathbf{R} = \sum_{h=0}^{\infty} \frac{g^{(h)'}(0)[f(g^{(h)}(0))g'(g^{(h)}(0)) - f'(g^{(h)}(0))g(g^{(h)}(0))]}{f(g^{(h)}(0))g(g^{(h)}(0))}. \tag{2.28}$$

From the relation $\mathbf{R} = \mathbf{W} + \boldsymbol{\tau}$ we also obtain (cf. (2.9))

$$\mathrm{E}\mathbf{W} = \sum_{h=0}^{\infty} \frac{g^{(h)'}(0)[f(g^{(h)}(0))g'(g^{(h)}(0)) - f'(g^{(h)}(0))g(g^{(h)}(0))]}{f(g^{(h)}(0))g(g^{(h)}(0))} - \frac{\lambda\beta_1}{\gamma} \quad (2.29)$$

Expressions (2.28) and (2.29) are used for numerical calculations in Section 2.6. □

Corollary 2.2.2
Letting $\omega \to \infty$ in (2.20) we obtain

$$\Pr\{\mathbf{R} = 0\} = \frac{\gamma}{\gamma + \lambda} r(\gamma + \lambda). \quad (2.30)$$

This formula may also be obtained from the recurrence relation
$\mathbf{R}_{n+1} = \max\{0, \mathbf{R}_n - \boldsymbol{\sigma}_{n+1}\} + \boldsymbol{\tau}_{n+1}$, $n = 1, 2, \ldots$, and (2.8):

$$\Pr\{\mathbf{R}_{n+1} = 0\} = \int_{u=0}^{\infty} \gamma e^{-\gamma u} \Pr\{\boldsymbol{\tau}_{n+1} = 0, \mathbf{R}_n < \boldsymbol{\sigma}_{n+1} \mid \boldsymbol{\sigma}_{n+1} = u\} du$$

$$= \int_{u=0}^{\infty} \gamma e^{-\gamma u} e^{-\lambda u} \Pr\{\mathbf{R}_n < u\} du = \frac{\gamma}{\gamma + \lambda} r_n(\gamma + \lambda).$$

□

Corollary 2.2.3
Letting $\omega \downarrow 0$ in (2.20) we obtain

$$r(\gamma) = 1 - \lambda\beta_1. \quad (2.31)$$

Again, this formula may also be obtained directly:

$$r_n(\gamma) = \int_{t=0}^{\infty} e^{-\gamma t} d\Pr\{\mathbf{R}_n < t\} = \Pr\{\mathbf{R}_n < \boldsymbol{\sigma}_{n+1}\}.$$

The latter term equals the probability that an arriving supercustomer sees the server idle, while obviously the probability that the server is idle equals $1 - \lambda\beta_1$. Because of the PASTA property both probabilities are equal. □

Corollary 2.2.4
Letting $\omega \to \infty$ in (2.27) we derive

$$\Pr\{\mathbf{W} = 0\} = 1 - \lambda\beta_1. \quad (2.32)$$

This result might also be obtained using the recurrence relation
$\mathbf{W}_{n+1} = \max\{0, \mathbf{R}_n - \boldsymbol{\sigma}_{n+1}\}$, $n = 1, 2, \ldots$, from which follows

$$\Pr\{\mathbf{W}_{n+1}=0\} = \int_{u=0}^{\infty} \gamma e^{-\gamma u} \Pr\{\mathbf{R}_n < u\} du = r_n(\gamma). \tag{2.33}$$

Letting $n \to \infty$ in (2.33) and using (2.31) leads to (2.32). □

Corollary 2.2.5
Expression (2.26) also provides the covariance of $\mathbf{W}$ and $\boldsymbol{\tau}$, i.e., the covariance of the waiting and service time of a customer:

$$\mathrm{Cov}(\mathbf{W}, \boldsymbol{\tau}) = -\frac{\phi'(0)}{\gamma}\left[\frac{1-r(\gamma)}{\gamma} + r'(\gamma)\right] < 0, \tag{2.34}$$

the inequality following from the Taylor expansion of $r(0)$ around γ. This result is intuitively clear; a customer having a relatively long (short) interarrival time is likely to have a relatively short (long) waiting time, but also a relatively long (short) service time, the latter being due to the dependence. □

Finally we study the sojourn and waiting times of a non-empty batch customer. Denote by $\mathbf{K}$ the number of individual customers constituting a batch customer. First, letting $\omega_2 \to \infty$ in (2.19) we obtain for $Re\,\omega \geq 0$

$$\mathrm{E}\left(e^{-\omega \mathbf{W}} I_{\{\mathbf{K}=0\}}\right) = \frac{\gamma(\gamma+\lambda)r(\omega) - \gamma\omega r(\gamma+\lambda)}{(\gamma+\lambda)(\gamma+\lambda-\omega)}. \tag{2.35}$$

Then, using (2.8),

$$\mathrm{E}\left(e^{-\omega \mathbf{W}} \,|\, \mathbf{K}>0\right) = \frac{\mathrm{E}\left(e^{-\omega \mathbf{W}}\right) - \mathrm{E}\left(e^{-\omega \mathbf{W}} I_{\{\mathbf{K}=0\}}\right)}{\Pr\{\mathbf{K}>0\}}$$

$$= \frac{-\gamma\lambda(\gamma+\lambda)r(\omega) - \gamma\omega(\gamma-\omega)r(\gamma+\lambda) + \omega(\gamma+\lambda)(\gamma+\lambda-\omega)r(\gamma)}{\lambda(\gamma-\omega)(\gamma+\lambda-\omega)}. \tag{2.36}$$

From (2.36) we obtain, using (2.9) and (2.31),

$$\mathrm{E}(\mathbf{W}\,|\,\mathbf{K}>0) = \mathrm{E}\mathbf{R} + \frac{\gamma}{\lambda(\gamma+\lambda)}[1-r(\gamma+\lambda)] - \frac{(\gamma+\lambda)\beta_1}{\gamma}$$

$$= \mathrm{E}\mathbf{W} + \frac{1}{\lambda}\left[r(\gamma) - \frac{\lambda}{\gamma+\lambda}r(0) - \frac{\gamma}{\gamma+\lambda}r(\gamma+\lambda)\right]. \tag{2.37}$$

As $r(\cdot)$ is a convex function, (2.37) confirms that a non-empty batch customer is likely to have a relatively short waiting time.

For the sojourn time of a non-empty batch customer we use the equality

$$\mathrm{E}\left(e^{-\omega \mathbf{W}} I_{\{\mathbf{K}=0\}}\right) = \mathrm{E}\left(e^{-\omega \mathbf{R}} I_{\{\mathbf{K}=0\}}\right). \tag{2.38}$$

So, similarly to (2.36), using (2.35)

$$\mathrm{E}\left(e^{-\omega \mathbf{R}} | \mathbf{K} > 0\right) = \frac{(\gamma+\lambda)(\lambda-\omega)r(\omega) + \gamma\omega r(\gamma+\lambda)}{\lambda(\gamma+\lambda-\omega)}. \tag{2.39}$$

From (2.39) we derive

$$\mathrm{E}(\mathbf{R} | \mathbf{K} > 0) = \mathrm{E}\mathbf{R} + \frac{\gamma}{\lambda(\gamma+\lambda)}(1 - r(\gamma+\lambda)). \tag{2.40}$$

Recall that a non-empty batch customer is likely to have a relatively long service time, but also a relatively short waiting time. From (2.40) we see that the on average longer service time outweighs the on average shorter waiting time.

Remark 2.2.2
One can also derive the LST for the joint distribution of the waiting time, the service time *and* the event that there are exactly k customers collected. The expression for this random variable contains the $j-th$ derivative of $r(\cdot)$ for j from 0 up to k. The result might be obtained by working out a recurrence relation similar to (2.14). In connection with Remark 2.1.1 we note that in the case of multiple arrival streams of single customers, one can obtain waiting times for each individual class by conditioning on the number of customers that are collected and using the LST of the three-dimensional random variable just introduced. □

2.3 Work decomposition

In this section we examine the waiting and sojourn times of supercustomers in more detail. We interpret the results of the previous section and explain the behaviour of these performance measures for our model. The main contribution of this section is the derivation of a decomposition property for the amount of work in the system. Finally, we also consider the effect the dependence structure has on the waiting and sojourn times of the individual customers.

Let us start by clarifying the meaning of the terms composing (2.20), thus relating $\mathbf{R}$ to the waiting time in an ordinary $M/G/1$ queue. Denote by $\mathbf{W}_{M/G/1}$ a stochastic variable with distribution the stationary distribution of the waiting time in an ordinary $M/G/1$ queue with arrival rate λ and service time distribution function $B(\cdot)$. In the sequel we refer to this queue as the corresponding $M/G/1$ queue without collection. From Cohen[34] p.255, we have:

$$\mathrm{E}\left(e^{-\omega \mathbf{W}_{M/G/1}}\right) = \frac{(1-\lambda\beta_1)\omega}{\omega - \lambda(1-\beta(\omega))}, \qquad Re\,\omega \geq 0.$$

Denote by $\mathbf{H}$ the sojourn time of a supercustomer leaving no supercustomers behind. Such a sojourn time has distribution function $H(\cdot)$ with

$$dH(t) = \frac{e^{-\gamma t} dR(t)}{\int_{u=0}^{\infty} e^{-\gamma u} dR(u)}, \qquad t > 0,$$

and, cf. (2.30),

$$H(0+) = \frac{\gamma}{\gamma+\lambda}\frac{r(\gamma+\lambda)}{r(\gamma)}.$$

Denote by $\mathbf{U}$ the amount of work arriving during such a sojourn time. Then for $Re\,\omega \geq 0$,

$$\mathrm{E}\left(e^{-\omega\mathbf{U}}\right) = \frac{\int_{t=0}^{\infty} e^{-\lambda(1-\beta(\omega))t}e^{-\gamma t}dR(t)}{\int_{u=0}^{\infty} e^{-\gamma u}dR(u)} = \frac{r(\gamma+\lambda(1-\beta(\omega)))}{r(\gamma)}. \tag{2.41}$$

Using expression (2.31), we have arrived at the following

Theorem 2.3.1
The LST of the sojourn time distribution of a batch customer can be decomposed as

$$r(\omega) = \mathrm{E}(e^{-\omega\mathbf{W}_{M/G/1}})\mathrm{E}(e^{-\omega\boldsymbol{\tau}})\mathrm{E}(e^{-\omega\mathbf{U}}), \qquad Re\,\omega \geq 0. \tag{2.42}$$

□

To provide additional insight, we now give a more intuitive derivation of (2.42). The sojourn time of a supercustomer consists of two phases, viz.:
(i). its waiting time, i.e. the time needed to do the work associated with the individual customers present at the server upon the supercustomer's arrival;
(ii). its service time, i.e. the time needed to do the work associated with the individual customers present at the bus-stop upon the supercustomer's arrival.
(Recall that the arrival of the supercustomer at the server coincides with the arrival of the collector at the bus-stop.)
So the sojourn time of a supercustomer equals the amount of work associated with the individual customers present upon its arrival (at the bus-stop as well as at the server). Denote by $\mathbf{V}$ the steady state amount of work associated with the individual customers (at the bus-stop as well as at the server). Because of the PASTA property

$$\mathbf{R} \stackrel{d}{=} \mathbf{V}. \tag{2.43}$$

Denote by $\mathbf{V}_{M/G/1}$ a stochastic variable with distribution the stationary distribution of the amount of work in the corresponding $M/G/1$ queue without collection. Denote by $\mathbf{Y}$ a stochastic variable, independent of $\mathbf{V}_{M/G/1}$, with distribution the stationary distribution of the amount of work associated with the individual customers at an arbitrary epoch in a non-serving interval, i.e. the amount of work associated with the individual customers present at the bus-stop when the server is idle.

Now the following work decomposition property holds, cf. Boxma[22]:

$$\mathbf{V} \stackrel{d}{=} \mathbf{V}_{M/G/1} + \mathbf{Y}, \tag{2.44}$$

with $\mathbf{V}_{M/G/1}$ and $\mathbf{Y}$ independent.
Because of the PASTA property

$$\mathbf{V}_{M/G/1} \stackrel{d}{=} \mathbf{W}_{M/G/1}. \tag{2.45}$$

The amount of work associated with individual customers at an arbitrary epoch in a non-serving interval, $\mathbf{Y}$, consists of two components, viz.:
(i). the amount of work associated with individual customers that have arrived during the sojourn time of the last supercustomer (possibly empty), $\mathbf{Y}^{(i)}$. This sojourn time has distribution function $H(\cdot)$. So

$$\mathbf{Y}^{(i)} \stackrel{d}{=} \mathbf{U}. \tag{2.46}$$

(ii). the amount of work associated with individual customers that have arrived during the past non-serving period since the departure of the last supercustomer (possibly empty), $\mathbf{Y}^{(ii)}$. This past non-serving period is negative exponentially distributed with parameter γ, since the non-serving period is a (residual) collecting interval. So

$$\mathbf{Y}^{(ii)} \stackrel{d}{=} \boldsymbol{\tau}. \tag{2.47}$$

Moreover, $\mathbf{Y}^{(i)}$ and $\mathbf{Y}^{(ii)}$ are independent, since the individual customers arrive according to a Poisson process.
Combining (2.43) - (2.47) yields (2.42).

From (2.42), using (2.31) and (2.41),

$$\begin{aligned} \mathrm{E}\mathbf{R} &= \mathrm{E}\mathbf{W}_{M/G/1} + \mathrm{E}\boldsymbol{\tau} + \mathrm{E}\mathbf{U} \\ &= \mathrm{E}\mathbf{W}_{M/G/1} + \mathrm{E}\boldsymbol{\tau} + \frac{\lambda\beta_1}{1-\lambda\beta_1} \int_{t=0}^{\infty} te^{-\gamma t} dR(t), \end{aligned} \tag{2.48}$$

and

$$\begin{aligned} Var(\mathbf{R}) &= Var(\mathbf{W}_{M/G/1}) + Var(\boldsymbol{\tau}) + \frac{(\lambda\beta_1)^2}{1-\lambda\beta_1} \int_{t=0}^{\infty} t^2 e^{-\gamma t} dR(t) \\ &\quad + \frac{\lambda\beta_2}{1-\lambda\beta_1} \int_{t=0}^{\infty} te^{-\gamma t} dR(t) - \left(\frac{\lambda\beta_1}{1-\lambda\beta_1} \int_{t=0}^{\infty} te^{-\gamma t} dR(t) \right)^2. \end{aligned}$$

Remark 2.3.1
Using (2.21) and (2.22) we find that the factor for $h = 0$ of the infinite product (2.25) equals $\mathrm{E}(e^{-\omega \mathbf{W}_{M/G/1}})\mathrm{E}(e^{-\omega \boldsymbol{\tau}})$. So the remainder of the infinite product equals $\dfrac{r(\gamma + \lambda(1 - \beta(\omega)))}{r(\gamma)}$. Similarly we find that the term for $h = 0$ of the infinite sum (2.28) equals $\mathrm{E}\mathbf{W}_{M/G/1} + \mathrm{E}\boldsymbol{\tau}$. So, using (2.31), the remainder of the infinite sum equals $\dfrac{\lambda\beta_1}{1 - \lambda\beta_1} \int_{t=0}^{\infty} te^{-\gamma t} dR(t)$. □

The sojourn times of individual customers
Next we turn to the sojourn time distribution of an *individual* customer. Using the decomposition argument we derive the LST for the sojourn time distribution as well as for the total number of individual customers in the system.
Denote by $\tilde{\mathbf{R}}$ the random variable with distribution the stationary distribution of the sojourn time of an individual customer and let $\tilde{r}(\omega) := \mathrm{E}(e^{-\omega \tilde{\mathbf{R}}})$ for $Re\,\omega \geq 0$. First we study $\tilde{\mathbf{N}}$, the random variable with distribution the stationary distribution of the number of individual customers. We will find $\tilde{r}(\omega)$ from $\tilde{r}(\lambda(1 - z)) = \mathrm{E}(z^{\tilde{\mathbf{N}}}), |z| \leq 1$. Denote by $\mathbf{N}_{M/G/1}$ a stochastic variable with distribution the stationary distribution of the number of customers in the corresponding $M/G/1$ queue without collection. From Cohen[34] p.247, we have:

$$\mathrm{E}(z^{\mathbf{N}_{M/G/1}}) = \frac{(1 - \lambda\beta_1)(1 - z)\beta(\lambda(1 - z))}{\beta(\lambda(1 - z)) - z}, \qquad |z| \leq 1. \tag{2.49}$$

Denote by $\mathbf{X}^{(i)}$ the number of individual customers present at an arbitrary epoch in a non-serving interval that have arrived during the sojourn time of the last supercustomer (possibly empty). This sojourn time has distribution function $H(\cdot)$. So

$$\mathrm{E}(z^{\mathbf{X}^{(i)}}) = \frac{r(\gamma + \lambda(1 - z))}{r(\gamma)}, \qquad |z| \leq 1. \tag{2.50}$$

Denote by $\mathbf{X}^{(ii)}$ the number of individual customers present at an arbitrary epoch in a non-serving interval that have arrived during the past non-serving period since the departure of the last supercustomer (possibly empty). This past non-serving period is negative exponentially distributed with parameter γ, since the non-serving period is a (residual) collect interval. So

$$\mathrm{E}(z^{\mathbf{X}^{(ii)}}) = \frac{\gamma}{\gamma + \lambda(1 - z)}, \qquad |z| \leq 1. \tag{2.51}$$

Observe that $\mathbf{X}^{(i)}$ and $\mathbf{X}^{(ii)}$ are independent, since the individual customers arrive according to a Poisson process and the non-serving period is a (residual) collect interval, not depending on the sojourn time of the last supercustomer.

Now the following queue length decomposition holds, cf. Fuhrmann & Cooper[49]:

$$\mathrm{E}(z^{\tilde{\mathbf{N}}}) = \mathrm{E}(z^{\mathbf{N}_{M/G/1}})\mathrm{E}(z^{\mathbf{X}^{(\mathrm{i})}})\mathrm{E}(z^{\mathbf{X}^{(\mathrm{ii})}}), \qquad |z| \leq 1. \tag{2.52}$$

Denote by $\mathbf{R}_{M/G/1}$ a stochastic variable with distribution the stationary distribution of the sojourn time in the corresponding $M/G/1$ queue without collection. Substituting $\omega = \lambda(1-z)$ in (2.52) leads to:

Theorem 2.3.2
The LST of the sojourn time distribution of an individual customer is

$$\begin{aligned} \tilde{r}(\omega) &= \mathrm{E}(e^{-\omega\mathbf{R}_{M/G/1}})\mathrm{E}(e^{-\omega\boldsymbol{\sigma}})\mathrm{E}(e^{-\omega\mathbf{H}}) \\ &= \mathrm{E}(e^{-\omega\mathbf{W}_{M/G/1}})\beta(\omega)\frac{\gamma}{\gamma+\omega}\frac{r(\gamma+\omega)}{r(\gamma)}, \quad Re\,\omega \geq 0, \end{aligned} \tag{2.53}$$

with $r(\cdot)$ given by Theorem 2.2.1. □

Remark 2.3.2
As mentioned in the introduction, Takahashi[95] studies sojourn times and queue lengths of *individual* customers in the same model as the present chapter, for the special case of negative exponentially distributed service times. For that case, formula (2.53) reduces to formula (2.29) of Takahashi[95]. To verify this, note that $\mathrm{E}(e^{-\omega\mathbf{W}_{M/G/1}})\beta(\omega)\frac{\gamma}{\gamma+\omega}$ equals $g_0(1-\omega/\lambda)/g_0(1)$ in [95], and that the n-th term in the infinite product (2.25) matches the term $g_{n+1}(1-\omega/\lambda)/g_{n+1}(1)$ in the infinite product (2.19) of [95]. □

From (2.53), using (2.31),

$$\begin{aligned} \mathrm{E}\tilde{\mathbf{R}} &= \mathrm{E}\mathbf{W}_{M/G/1} + \beta_1 + \mathrm{E}\boldsymbol{\sigma} + \mathrm{E}\mathbf{H} \\ &= \mathrm{E}\mathbf{W}_{M/G/1} + \beta_1 + \frac{1}{\gamma} + \frac{1}{1-\lambda\beta_1}\int_{t=0}^{\infty} te^{-\gamma t}dR(t), \end{aligned} \tag{2.54}$$

and

$$\begin{aligned} Var(\tilde{\mathbf{R}}) &= Var(\mathbf{W}_{M/G/1}) + \beta_2 - \beta_1^2 + \frac{1}{\gamma^2} + \\ &\frac{1}{1-\lambda\beta_1}\int_{t=0}^{\infty} t^2e^{-\gamma t}dR(t) - \left(\frac{1}{1-\lambda\beta_1}\int_{t=0}^{\infty} te^{-\gamma t}dR(t)\right)^2. \end{aligned}$$

The waiting time of supercustomers
Next we turn to the waiting times for supercustomers. From (2.27), using (2.9) and (2.31),

$$\mathrm{E}\mathbf{W} = \mathrm{E}\mathbf{R} - \mathrm{E}\boldsymbol{\tau}, \tag{2.55}$$

as should be the case since $\mathbf{R}_n = \mathbf{W}_n + \boldsymbol{\tau}_n$ for $n = 1, 2, \ldots$. Combining (2.48) and (2.55),

$$\mathrm{E}\mathbf{W} = \mathrm{E}\mathbf{W}_{M/G/1} + \frac{\lambda\beta_1}{1-\lambda\beta_1} \int_{t=0}^{\infty} te^{-\gamma t} dR(t). \tag{2.56}$$

The following Lemma gives some bounds for the integral occurring in (2.56).

Lemma 2.3.1

(i). $\displaystyle\int_{t=0}^{\infty} te^{-\gamma t} dR(t) \leq \min\left\{ \frac{e^{-1}}{\gamma}(1 - \frac{\gamma}{\gamma+\lambda} r(\gamma+\lambda)), \frac{\lambda\beta_1}{\gamma} \right\}.$

(ii). $\displaystyle\int_{t=0}^{\infty} te^{-\gamma t} dR(t) \geq \frac{1-\lambda\beta_1}{\gamma} \frac{\lambda(1-\beta(\gamma))}{\gamma - \lambda(1-\beta(\gamma))} \frac{\gamma + \lambda(1-\beta(\gamma)) + 2\lambda\gamma\beta'(\gamma)}{\gamma + \lambda(1-\beta(\gamma))}.$

Proof
See Appendix 2.C
□

Lemma 2.3.1 (i) implies that $\lim_{\gamma\to\infty} \mathrm{E}\mathbf{W} = \mathrm{E}\mathbf{W}_{M/G/1}$. Because of the PASTA property the mean *amount of work* at the server also approaches the mean amount of work in the corresponding $M/G/1$ queue without collection, which is obvious as for $\gamma \to \infty$ the individual customers are collected instantaneously.

We now compare the mean waiting time of a supercustomer with the mean waiting time in an ordinary $M/G/1$ queue with identical traffic characteristics, but without dependence. Denote by $\mathbf{W}_I$ the waiting time in an ordinary $M/G/1$ queue with arrival rate γ and service time distribution having LST $\frac{\gamma}{\gamma+\lambda(1-\beta(\omega))}$, cf. (2.6). In the sequel we refer to this queue as the corresponding $M/G/1$ queue without dependence. Using (2.9),

$$\mathrm{E}\mathbf{W}_I = \frac{\gamma\left(\frac{\lambda\beta_2}{\gamma} + 2\left(\frac{\lambda\beta_1}{\gamma}\right)^2\right)}{2(1-\lambda\beta_1)} = \mathrm{E}\mathbf{W}_{M/G/1} + \frac{\lambda^2\beta_1^2}{\gamma(1-\lambda\beta_1)}. \tag{2.57}$$

Combining (2.56) and (2.57),

$$\mathrm{E}\mathbf{W} = \mathrm{E}\mathbf{W}_I + \frac{\lambda\beta_1}{1-\lambda\beta_1}\left[\int_{t=0}^{\infty} te^{-\gamma t} dR(t) - \frac{\lambda\beta_1}{\gamma}\right]. \tag{2.58}$$

The integral in (2.58) equals $-r'(\gamma)$, and $\frac{\lambda\beta_1}{\gamma} = \frac{1-r(\gamma)}{\gamma}$ (cf. Corollary 2.2.3). From Corollary 2.2.5 we find

$$\mathrm{E}\mathbf{W} = \mathrm{E}\mathbf{W}_I + \frac{\gamma}{1-\lambda\beta_1} Cov(\mathbf{W}, \boldsymbol{\tau}). \tag{2.59}$$

From $Cov(\mathbf{W}, \boldsymbol{\tau}) < 0$ (cf. (2.34)) follows $\mathrm{E}\mathbf{W} \leq \mathrm{E}\mathbf{W}_I$.
Equation (2.59) can also be obtained from an application of the generalized form of Little's law; from Wolff[101] p.279 we have the following formula for the mean total amount of work in the system:

$$\mathrm{E}\mathbf{V} = \gamma \mathrm{E}[\mathbf{W}\boldsymbol{\tau}] + \gamma \mathrm{E}[\boldsymbol{\tau}^2]/2. \tag{2.60}$$

Working out (2.60) results in (2.59).
The following argument explains why $\mathrm{E}\mathbf{W} \leq \mathrm{E}\mathbf{W}_I$. From Corollary 2.2.5 we know that $Cov(\mathbf{W}, \boldsymbol{\tau}) < 0$; a supercustomer having a relatively short/long interarrival time is likely to have a relatively long/short waiting time, but also a relatively short/long service time, due to the dependence. Imposing a positive correlation structure on the interarrival and service times thus has a reducing effect on mean waiting times. From (2.12), (2.34) and (2.59) we see that the difference between $\mathrm{E}\mathbf{W}$ and $\mathrm{E}\mathbf{W}_I$ increases for $\gamma \downarrow 0$, i.e. as the correlation between the interarrival and service times approaches 1. However, Lemma 2.3.1 (ii) implies that both $\mathrm{E}\mathbf{W}$ and $\mathrm{E}\mathbf{W}_I$ tend to infinity for $\gamma \downarrow 0$. Not surprisingly, for $\gamma \downarrow 0$ the mean amount of work at the server tends to infinity. Because of the PASTA property the mean waiting time tends to infinity too. Observe that the Poisson character of the collection process is essential here. In case of generally distributed collection intervals having first moment $\frac{1}{\gamma}$ the mean amount of work tends to infinity as well for $\gamma \downarrow 0$. However, the mean waiting time does not need to tend to infinity too. In case of e.g. deterministic collection intervals of length $\frac{1}{\gamma}$

$$\begin{aligned} E\boldsymbol{\tau} &= \frac{\lambda\beta_1}{\gamma}; \\ E\boldsymbol{\tau}^2 &= \frac{\lambda\beta_2}{\gamma} + \left(\frac{\lambda\beta_1}{\gamma}\right)^2. \end{aligned}$$

Using Chebyshev's inequality,

$$\begin{aligned} Pr\{\boldsymbol{\tau}_{n-1} > \boldsymbol{\sigma}_n + \epsilon\} &\leq Pr\left\{ \mid \boldsymbol{\tau} - \mathrm{E}\boldsymbol{\tau} \mid > \frac{1-\lambda\beta_1}{\gamma} + \epsilon \right\} \\ &\leq \frac{Var(\boldsymbol{\tau})}{\left(\frac{1-\lambda\beta_1}{\gamma} + \epsilon\right)^2} \\ &= \frac{\gamma\lambda\beta_2}{(1-\lambda\beta_1+\gamma\epsilon)^2}, \quad \text{for } \epsilon > 0. \end{aligned}$$

Hence $Pr\{\mathbf{W} = 0\} \to 1$ for $\gamma \downarrow 0$.
For deterministic collecting intervals we observe that increasing the mean collecting interval also increases the correlation between interarrival and service times, and hence reduces the influence of second and higher moments of the service time distribution, finally resulting in a decreasing mean waiting time of a supercustomer. Elaborating on this argument, one might view the collecting procedure as an instrument to control the variance of the arrival process of work at the server station.
In Chapter 4 we discuss a technique for modelling dependence between interarrival times and service times that also allows more general intercollecting times.

The waiting times of individual customers
We conclude this section by studying the waiting time of an *individual* customer. Denote by $\tilde{\mathbf{W}}$ the waiting time of an individual customer. Let $\tilde{w}(\omega) := \mathrm{E}(e^{-\omega \tilde{\mathbf{W}}})$ for $Re\,\omega \geq 0$. From (2.53),

$$\tilde{w}(\omega) = \mathrm{E}(e^{-\omega \mathbf{W}_{M/G/1}}) \frac{\gamma}{\gamma + \omega} \frac{r(\gamma + \omega)}{r(\gamma)}, \qquad Re\,\omega \geq 0, \tag{2.61}$$

since the waiting time and the subsequent service time of an *individual* customer are independent. From (2.61) using (2.31), or immediately from (2.54),

$$\mathrm{E}\tilde{\mathbf{W}} = \mathrm{E}\mathbf{W}_{M/G/1} + \frac{1}{\gamma} + \frac{1}{1 - \lambda\beta_1} \int_{t=0}^{\infty} te^{-\gamma t} dR(t), \tag{2.62}$$

and

$$Var(\tilde{\mathbf{W}}) = Var(\mathbf{W}_{M/G/1}) + \frac{1}{\gamma^2} + \tag{2.63}$$

$$\frac{1}{1 - \lambda\beta_1} \int_{t=0}^{\infty} t^2 e^{-\gamma t} dR(t) - \left(\frac{1}{1 - \lambda\beta_1} \int_{t=0}^{\infty} te^{-\gamma t} dR(t) \right)^2. \tag{2.64}$$

From (2.56) and (2.62), $\mathrm{E}\mathbf{W}_{M/G/1} \leq \mathrm{E}\mathbf{W} \leq \mathrm{E}\tilde{\mathbf{W}}$, as expected.
An alternative derivation of (2.62) proceeds as follows. The waiting time of an individual customer is the sum of three terms (different from the terms occurring in (2.61)), viz.:
(i). the time from its arrival until its collection, $\tilde{\mathbf{W}}^{(i)}$;
(ii). the waiting time of the supercustomer to which it belongs, $\tilde{\mathbf{W}}^{(ii)}$;
(iii). the time from its admission until its service, $\tilde{\mathbf{W}}^{(iii)}$.
These three terms are dependent, however

$$\mathrm{E}\tilde{\mathbf{W}} = \mathrm{E}\tilde{\mathbf{W}}^{(i)} + \mathrm{E}\tilde{\mathbf{W}}^{(ii)} + \mathrm{E}\tilde{\mathbf{W}}^{(iii)}. \tag{2.65}$$

$\tilde{\mathbf{W}}^{(i)}$ is the length of a residual, negative exponentially distributed, collecting interval.

So

$$\mathrm{E}\tilde{\mathbf{W}}^{(i)} = \frac{1}{\gamma}. \tag{2.66}$$

$\mathrm{E}\tilde{\mathbf{W}}^{(ii)}$ does not equal $\mathrm{E}\mathbf{W}$ because the supercustomer containing a tagged individual customer is not typical but is likely to have a long interarrival time and hence a short waiting time. However, applying Little's formula,

$$\mathrm{E}\tilde{\mathbf{W}}^{(ii)} = \frac{\mathrm{E}\tilde{\mathbf{N}}^{(ii)}}{\lambda}, \tag{2.67}$$

where $\tilde{\mathbf{N}}^{(ii)}$ denotes the number of individual customers belonging to waiting supercustomers. Furthermore

$$\mathrm{E}\tilde{\mathbf{N}}^{(ii)} = \frac{\mathrm{E}\mathbf{V}^{(ii)}}{\beta_1}, \tag{2.68}$$

where $\mathbf{V}^{(ii)}$ denotes the amount of work associated with individual customers belonging to waiting supercustomers. From Wolff[101], p.279 we have:

$$\mathrm{E}\mathbf{V}^{(ii)} = \gamma \mathrm{E}[\mathbf{W}\boldsymbol{\tau}]. \tag{2.69}$$

Combining (2.48), (2.43), (2.60), (2.67), (2.68) and (2.69),

$$\mathrm{E}\tilde{\mathbf{W}}^{(ii)} = \mathrm{E}\mathbf{W}_{M/G/1} - \frac{\lambda\beta_1}{\gamma} + \frac{1}{1-\lambda\beta_1} \int_{t=0}^{\infty} te^{-\gamma t} dR(t). \tag{2.70}$$

$\tilde{\mathbf{W}}^{(iii)}$ is the amount of work that arrived during a past negative exponentially distributed collecting interval. So

$$\mathrm{E}\tilde{\mathbf{W}}^{(iii)} = \frac{\lambda\beta_1}{\gamma}. \tag{2.71}$$

Substituting (2.66), (2.70) and (2.71) in (2.65) yields (2.62).

2.4 The number of customers

In this section we focus our attention on the number of batch and individual customers in the system. Using the distributional form of Little's law, cf. Keilson & Servi[64], it is easily seen that the generating function of $\mathbf{N}$, the number of batch customers at an arbitrary time, is given by $\mathrm{E}\{z^{\mathbf{N}}\} = \mathrm{E}\{e^{-\gamma(1-z)\mathbf{R}}\}$. From the (PASTA) property it follows that $\mathbf{N}_A$, the number of batch customers at arrival epochs, and $\mathbf{N}$ are identically distributed.

The remainder of this section is devoted to the number of individual customers in the system. At time t this quantity is $\mathbf{A}(t) + \mathbf{Z}(t)$, with $\mathbf{A}(t)$ the number of individual customers which have arrived but have not yet been collected at time t and $\mathbf{Z}(t)$ the number of individual customers which have been collected

but have not yet departed at time t. Also define $\mathbf{Z}_n$ and $\mathbf{A}_n$, the values of $\mathbf{Z}(t)$ and $\mathbf{A}(t)$ immediately after the departure of the n-th individual customer. Notice that $\{\mathbf{Z}_n, \mathbf{A}_n\}_{n\geq 1}$ is a two-dimensional Markov chain, whereas $\{\mathbf{Z}_n\}_{n\geq 1}$ and $\{\mathbf{A}_n\}_{n\geq 1}$ are not Markov chains. Let $\{\mathbf{Z}, \mathbf{A}\}$ be a vector with distribution the steady state distribution of this Markov chain. By letting $n \to \infty$ in the generating functions of $\{\mathbf{Z}_n, \mathbf{A}_n\}$ we will derive the generating function of $\{\mathbf{Z}, \mathbf{A}\}$.

For our analysis we need to define:

$\hat{\tau}_n :=$ Service time of individual customer n.

$\hat{\sigma}_n :=$ Interarrival time of the first batch customer after and counted from the start of service of individual customer n. If $\hat{\tau}_n < \hat{\sigma}_n$ no batch customers arrive during this service, otherwise at least one batch customer arrives.

$\mu_n :=$ Number of individual customers which arrive during the service of individual customer n and are collected.

$\nu_n :=$ Number of individual customers which arrive after the last batch arrival during the service of individual customer n. If no batch arrival occurs, then ν_n is the total number of individual customers which arrive during the service of individual customer n.

$\zeta_n :=$ Number of individual customers which arrive between the end of service of individual customer $n-1$ and the start of service of individual customer n. If $\mathbf{Z}_{n-1} > 0$, then $\zeta_n = 0$.

To illustrate these definitions, an example of the arrival and departure processes of customers is presented in Figure 2.1.

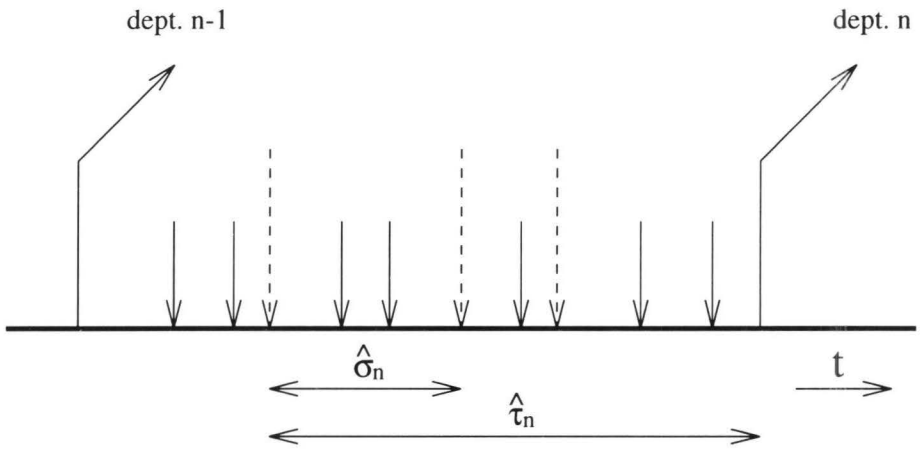

FIGURE 2.1: ARRIVALS AND DEPARTURES OF INDIVIDUAL CUSTOMERS, AND ARRIVALS OF BATCH CUSTOMERS (DASHED ARROWS). IN THIS EXAMPLE, $\zeta_n = 2$, $\mu_n = 3$, $\nu_n = 2$.

With the above-introduced notation, it is readily seen that the transition equations of $\{\mathbf{Z}_n, \mathbf{A}_n\}_{n\geq 1}$ are given by

$$\{\mathbf{Z}_{n+1}, \mathbf{A}_{n+1}\} = \begin{cases} \{\mathbf{Z}_n - 1, \mathbf{A}_n + \boldsymbol{\nu}_{n+1}\} & \text{if } \mathbf{Z}_n \geq 1, \hat{\boldsymbol{\tau}}_{n+1} < \hat{\boldsymbol{\sigma}}_{n+1}, \\ \{\mathbf{Z}_n - 1 + \mathbf{A}_n + \boldsymbol{\mu}_{n+1}, \boldsymbol{\nu}_{n+1}\} & \text{if } \mathbf{Z}_n \geq 1, \hat{\boldsymbol{\tau}}_{n+1} \geq \hat{\boldsymbol{\sigma}}_{n+1}, \\ \{\mathbf{A}_n + \boldsymbol{\zeta}_{n+1} - 1 + \boldsymbol{\mu}_{n+1}, \boldsymbol{\nu}_{n+1}\} & \text{if } \mathbf{Z}_n = 0. \end{cases} \tag{2.72}$$

Now define for $|z| \leq 1, |q| \leq 1, n = 1, 2, \ldots$, the generating functions $\Phi_n(z, q) := \mathrm{E}\{z^{\mathbf{Z}_n} q^{\mathbf{A}_n}\}$. Then

$$\begin{aligned} \mathrm{E}\{z^{\mathbf{Z}_{n+1}} q^{\mathbf{A}_{n+1}}\} &= \mathrm{E}\{z^{\mathbf{Z}_{n+1}} q^{\mathbf{A}_{n+1}} I_{\{\mathbf{Z}_n \geq 1\}} I_{\{\hat{\boldsymbol{\tau}}_{n+1} < \hat{\boldsymbol{\sigma}}_{n+1}\}}\} + \\ &\quad \mathrm{E}\{z^{\mathbf{Z}_{n+1}} q^{\mathbf{A}_{n+1}} I_{\{\mathbf{Z}_n \geq 1\}} I_{\{\hat{\boldsymbol{\tau}}_{n+1} \geq \hat{\boldsymbol{\sigma}}_{n+1}\}}\} + \\ &\quad \mathrm{E}\{z^{\mathbf{Z}_{n+1}} q^{\mathbf{A}_{n+1}} I_{\{\mathbf{Z}_n = 0\}}\}. \end{aligned} \tag{2.73}$$

From (2.72) and (2.73) it follows

$$\begin{aligned} \mathrm{E}\{z^{\mathbf{Z}_{n+1}} q^{\mathbf{A}_{n+1}}\} &= \mathrm{E}\{z^{\mathbf{Z}_n - 1} q^{\mathbf{A}_n + \boldsymbol{\nu}_{n+1}} I_{\{\mathbf{Z}_n \geq 1\}} I_{\{\hat{\boldsymbol{\tau}}_{n+1} < \hat{\boldsymbol{\sigma}}_{n+1}\}}\} + \\ &\quad \mathrm{E}\{z^{\mathbf{Z}_n - 1 + \mathbf{A}_n + \boldsymbol{\mu}_{n+1}} q^{\boldsymbol{\nu}_{n+1}} I_{\{\mathbf{Z}_n \geq 1\}} I_{\{\hat{\boldsymbol{\tau}}_{n+1} \geq \hat{\boldsymbol{\sigma}}_{n+1}\}}\} + \\ &\quad \mathrm{E}\{z^{\mathbf{A}_n + \boldsymbol{\zeta}_{n+1} + \boldsymbol{\mu}_{n+1} - 1} q^{\boldsymbol{\nu}_{n+1}} I_{\{\mathbf{Z}_n = 0\}}\}. \end{aligned} \tag{2.74}$$

Using that $\mathbf{Z}_n, \mathbf{A}_n, \boldsymbol{\zeta}_{n+1}$, $I_{\{\mathbf{Z}_n \geq 1\}}$ and $I_{\{\mathbf{Z}_n = 0\}}$ are independent of $\boldsymbol{\nu}_{n+1}, \boldsymbol{\mu}_{n+1}, I_{\{\hat{\boldsymbol{\tau}}_{n+1} \geq \hat{\boldsymbol{\sigma}}_{n+1}\}}$ and $I_{\{\hat{\boldsymbol{\tau}}_{n+1} < \hat{\boldsymbol{\sigma}}_{n+1}\}}$, it follows for $|z| \leq 1, |q| \leq 1$ that

$$\begin{aligned} \Phi_{n+1}(z, q) &= \frac{1}{z}\mathrm{E}\{z^{\mathbf{Z}_n} q^{\mathbf{A}_n} I_{\{\mathbf{Z}_n \geq 1\}}\}\mathrm{E}\{q^{\boldsymbol{\nu}_{n+1}} I_{\{\hat{\boldsymbol{\tau}}_{n+1} < \hat{\boldsymbol{\sigma}}_{n+1}\}}\} + \\ &\quad \frac{1}{z}\mathrm{E}\{z^{\mathbf{Z}_n + \mathbf{A}_n} I_{\{\mathbf{Z}_n \geq 1\}}\}\mathrm{E}\{z^{\boldsymbol{\mu}_{n+1}} q^{\boldsymbol{\nu}_{n+1}} I_{\{\hat{\boldsymbol{\tau}}_{n+1} \geq \hat{\boldsymbol{\sigma}}_{n+1}\}}\} + \\ &\quad \frac{1}{z}\mathrm{E}\{z^{\mathbf{A}_n + \boldsymbol{\zeta}_{n+1}} I_{\{\mathbf{Z}_n = 0\}}\}\mathrm{E}\{z^{\boldsymbol{\mu}_{n+1}} q^{\boldsymbol{\nu}_{n+1}}\}. \end{aligned} \tag{2.75}$$

We consider each of the three terms in the right-hand side of (2.75) in turn.

$$\begin{aligned} &\mathrm{E}\{q^{\boldsymbol{\nu}_{n+1}} I_{\{\hat{\boldsymbol{\tau}}_{n+1} < \hat{\boldsymbol{\sigma}}_{n+1}\}}\} = \\ &\quad \int_{\tau=0}^{\infty} dB(\tau)\, e^{-\gamma\tau} \sum_{k=0}^{\infty} e^{-\lambda\tau} \frac{(\lambda\tau)^k}{k!} q^k = \beta(\gamma + \lambda(1-q)), \end{aligned} \tag{2.76}$$

$$\mathrm{E}\{z^{\boldsymbol{\mu}_{n+1}} q^{\boldsymbol{\nu}_{n+1}} I_{\{\hat{\boldsymbol{\tau}}_{n+1} \geq \hat{\boldsymbol{\sigma}}_{n+1}\}}\} =$$

$$\int_{\tau=0}^{\infty} dB(\tau) \int_{u=0+}^{\tau} \mathrm{E}\{z^{\boldsymbol{\mu}_{n+1}} q^{\boldsymbol{\nu}_{n+1}} \mid \textit{last batch arr. at } u\}$$

$$dP\{\textit{last batch arr. before } u\} =$$

$$\int_{\tau=0}^{\infty} dB(\tau) \int_{u=0+}^{\tau} \sum_{k=0}^{\infty} e^{-\lambda u} \frac{(\lambda u)^k}{k!} z^k \sum_{j=0}^{\infty} e^{-\lambda(\tau-u)} \frac{(\lambda(\tau-u))^j}{j!} q^j$$

$$d_u\{(1-e^{-\gamma u})\, e^{-\gamma(\tau-u)}\} =$$

$$\int_{\tau=0}^{\infty} dB(\tau) \int_{u=0+}^{\tau} e^{-\lambda(1-z)u} e^{-\lambda(1-q)(\tau-u)} e^{-\gamma(\tau-u)} \gamma du =$$

$$\frac{\gamma}{\gamma+\lambda(z-q)}[\beta(\lambda(1-z)) - \beta(\gamma+\lambda(1-q))], \tag{2.77}$$

$$\mathrm{E}\{z^{\boldsymbol{\mu}_{n+1}} q^{\boldsymbol{\nu}_{n+1}}\} =$$

$$\mathrm{E}\{z^{\boldsymbol{\mu}_{n+1}} q^{\boldsymbol{\nu}_{n+1}} I_{\{\hat{\boldsymbol{\tau}}_{n+1} \geq \hat{\boldsymbol{\sigma}}_{n+1}\}}\} + \mathrm{E}\{q^{\boldsymbol{\nu}_{n+1}} I_{\{\hat{\boldsymbol{\tau}}_{n+1} < \hat{\boldsymbol{\sigma}}_{n+1}\}}\} =$$

$$\frac{\gamma}{\gamma+\lambda(z-q)}[\beta(\lambda(1-z)) - \beta(\gamma+\lambda(1-q))] + \beta(\gamma+\lambda(1-q)) =$$

$$\frac{\gamma}{\gamma+\lambda(z-q)}\beta(\lambda(1-z)) + \frac{\lambda(z-q)}{\gamma+\lambda(z-q)}\beta(\gamma+\lambda(1-q)). \tag{2.78}$$

If $\mathbf{Z}_n = 0$ and $\mathbf{A}_n > 0$ then the service of individual customer $n+1$ starts immediately after the arrival of the first batch customer after the completion of service of individual customer n. Then, $\boldsymbol{\zeta}_{n+1}$ is the number of individual customers which arrive during a residual arrival interval of a batch customer. If $\mathbf{Z}_n = 0$ and $\mathbf{A}_n = 0$ then the service of individual customer $n+1$ starts immediately after the arrival of the first non-empty batch customer after the completion of service of individual customer n. In this case, $\boldsymbol{\zeta}_{n+1} = 1+$ {the number of individual customers which arrive in the residual arrival interval of the first batch customer after the arrival of individual customer $n+1$}.

Since the batch arrival intervals are negative exponentially distributed it follows that

$$\mathrm{E}\{z^{\mathbf{A}_n+\zeta_{n+1}} I_{\{\mathbf{Z}_n=0\}}\} =$$

$$\mathrm{E}\{z^{\mathbf{A}_n+\zeta_{n+1}} I_{\{\mathbf{Z}_n=0\}} I_{\{\mathbf{A}_n>0\}}\} + \mathrm{E}\{z^{\mathbf{A}_n+\zeta_{n+1}} I_{\{\mathbf{Z}_n=0\}} I_{\{\mathbf{A}_n=0\}}\} =$$

$$\frac{\gamma}{\gamma+\lambda(1-z)}\mathrm{E}\{z^{\mathbf{A}_n} I_{\{\mathbf{Z}_n=0\}} I_{\{\mathbf{A}_n>0\}}\}+$$

$$\frac{\gamma z}{\gamma+\lambda(1-z)}\mathrm{E}\{I_{\{\mathbf{Z}_n=0\}} I_{\{\mathbf{A}_n=0\}}\}. \tag{2.79}$$

We next let $n \to \infty$; note that $\{\mathbf{Z}_n, \mathbf{A}_n\}_{n\geq 1}$ is an irreducible, aperiodic Markov chain, so that $\Phi(z,q) := \lim_{n\to\infty} \Phi_n(z,q)$ and related limits exist. Using (2.75)-(2.79) together with

$$\begin{aligned}
\mathrm{E}\{z^{\mathbf{Z}_n+\mathbf{A}_n} I_{\{\mathbf{Z}_n\geq 1\}}\} &= \Phi_n(z,z) - \Phi_n(0,z),\\
\mathrm{E}\{z^{\mathbf{Z}_n} q^{\mathbf{A}_n} I_{\{\mathbf{Z}_n\geq 1\}}\} &= \Phi_n(z,q) - \Phi_n(0,q),\\
\mathrm{E}\{I_{\{\mathbf{Z}_n=0\}} I_{\{\mathbf{A}_n=0\}}\} &= \Phi_n(0,0),\\
\mathrm{E}\{z^{\mathbf{A}_n} I_{\{\mathbf{Z}_n=0\}} I_{\{\mathbf{A}_n>0\}}\} &= \Phi_n(0,z) - \Phi_n(0,0),
\end{aligned}$$

we obtain the functional equation

$$\begin{aligned}
&\Phi(z,q)[z-\beta(\gamma+\lambda(1-q))] =\\
&\quad -\Phi(0,q)\beta(\gamma+\lambda(1-q))\\
&\quad +\Phi(z,z)\frac{\gamma}{\gamma+\lambda(z-q)}[\beta(\lambda(1-z))-\beta(\gamma+\lambda(1-q))]\\
&\quad -\Phi(0,0)\frac{\gamma}{\gamma+\lambda(z-q)}\frac{1-z}{\gamma+\lambda(1-z)}[\gamma\beta(\lambda(1-z))+\lambda(z-q)\beta(\gamma+\lambda(1-q))]+\\
&\quad +\Phi(0,z)\frac{\gamma}{\gamma+\lambda(z-q)}\frac{1}{\gamma+\lambda(1-z)}\times\\
&\qquad\qquad [(\gamma+\lambda(1-q))\beta(\gamma+\lambda(1-q))-\lambda(1-z)\beta(\lambda(1-z))].
\end{aligned} \tag{2.80}$$

We now solve this functional equation, by relating components in the generating function expression (2.80) to the sojourn time distribution $r(\cdot)$ of a supercustomer. A direct analysis of (2.80) can also be performed by first expressing $\Phi(z,z)$ into $\Phi(0,z)$ and $\Phi(0,0)$, and subsequently determining $\Phi(0,z)$. The latter proceeds in an iterative way similar to the analysis of $r(\cdot)$ in Section 2.3. Without going into details of the exact formulas, the functional equation for $\Phi(0,z)$ is of the form

$$\Phi(0,z) = f_1(z)\Phi(0,\delta(z)) - f_2(z)\Phi(0,0),$$

with $f_1(z)$, $f_2(z)$, and $\delta(z)$ functions of z. Details of the derivation are presented in Borst et al.[18].

A first step in solving (2.80) is to consider $\Phi(z,z) = \mathrm{E}\{z^{\mathbf{Z}+\mathbf{A}}\}$. Since individual customers arrive and depart one by one, we can use a level crossing argument to show that the number of individual customers immediately before the arrival of an individual customer in the system and the number of individual customers immediately after the departure of an individual customer are equally distributed. Using PASTA it is seen that $\Phi(z,z)$ is the generating function of the number of individual customers at an arbitrary time. This function we obtained in Section 2.3 using a decomposition argument. Hence, (cf. (2.49)-(2.52)),

$$\Phi(z,z) = \frac{(1-z)\beta(\lambda(1-z))}{\beta(\lambda(1-z))-z}\frac{r(\gamma+\lambda(1-z))\gamma}{\gamma+\lambda(1-z)}, \tag{2.81}$$

for $|z| \leq 1$.

Substituting $z = 0$ in (2.81) leads to

$$\Phi(0,0) = \frac{\gamma}{\gamma+\lambda}r(\gamma+\lambda), \tag{2.82}$$

or $\Phi(0,0) = \Pr\{\mathbf{R} = 0\}$, (cf. (2.30)). Indeed $\Phi(0,0) = \mathrm{E}\{I_{\{\mathbf{Z}=0\}}I_{\{\mathbf{A}=0\}}\} = \Pr\{\mathbf{R} = 0\}$, the last equation following from a level-crossing argument and the PASTA property.

Taking $z = q$ in (2.80) and using (2.81) and (2.82) we find

$$\Phi(0,q) = \frac{\gamma}{\lambda(\gamma+\lambda)}\left[(\gamma+\lambda)r(\gamma+\lambda(1-q)) - \gamma r(\gamma+\lambda)\right]. \tag{2.83}$$

Finally, substituting the expressions for $\Phi(z,z)$, $\Phi(0,0)$ and $\Phi(0,q)$ in (2.80), together with $r(\cdot)$ as given by (2.25), we obtain an explicit expression for $\Phi(z,q)$.

Theorem 2.4.1

The generating function $\Phi(z,q)$ of the joint stationary distribution of the number of single customers at the bus-stop and at the service facility is given by (2.80), with $\Phi(z,z)$ and $\Phi(0,0)$ given by expressions (2.81) and (2.82) respectively, and with $\Phi(0,q)$ and $\Phi(0,z)$ given by (2.83). □

We conclude this section by considering $\Phi(z,z) = \mathrm{E}\{z^{\mathbf{Z}+\mathbf{A}}\}$ in more detail. Taking $z = q$ directly in (2.80) and rewriting the resulting functional equation, we derive

$$\Phi(z,z) = \frac{(1-\lambda\beta_1)(1-z)\beta(\lambda(1-z))}{\beta(\lambda(1-z))-z}\frac{\gamma}{\gamma+\lambda(1-z)}\frac{\frac{\lambda}{\gamma}\Phi(0,z)+\Phi(0,0)}{1-\lambda\beta_1}. \tag{2.84}$$

In (2.84) we again recognize a decomposition (cf. (2.44) and (2.52) of the previous section).
Comparing (2.84) with expressions (2.49)-(2.52) gives

$$\frac{\lambda}{\gamma}\Phi(0,z) + \Phi(0,0) = r(\gamma + \lambda(1-z)). \tag{2.85}$$

One can verify (2.85) algebraically, but a more intuitive proof is given below.
Equation (2.85) shows a relation between the number of individual customers in the system at times when an individual customer leaves the server idle, and at times when a batch customer leaves no batch customers behind (see Section 2.3).
Denote by $\mathbf{X}$ the number of individual customers immediately after the departure of a batch customer. Denote by $\mathbf{R} < \boldsymbol{\sigma}$ the event that the sojourn time of a batch customer is smaller than the interarrival time of the next batch customer. Consequently such a batch customer leaves no batch customers behind. We have (cf. (2.50))

$$\begin{aligned} r(\gamma + \lambda(1-z)) &= E\{z^{\mathbf{X}} I_{\{\mathbf{R}<\boldsymbol{\sigma}\}}\} \\ &= \mathrm{E}\{z^{\mathbf{X}} I_{\{\mathbf{R}<\boldsymbol{\sigma}\}} I_{\{\mathbf{R}=0\}}\} \\ &+ \mathrm{E}\{z^{\mathbf{X}} I_{\{\mathbf{R}<\boldsymbol{\sigma}\}} I_{\{\mathbf{R}>0\}}\}. \end{aligned} \tag{2.86}$$

A batch customer having $\mathbf{R} < \boldsymbol{\sigma}$ and $\mathbf{R} > 0$ corresponds to a unique individual customer having $\mathbf{Z} = 0$, namely the individual customer leaving at the same moment as the batch customer, and vice versa. Using this one-to-one correspondence we have

$$\begin{aligned} \mathrm{E}\{z^{\mathbf{X}} I_{\{\mathbf{R}<\boldsymbol{\sigma}\}} I_{\{\mathbf{R}>0\}}\} &= \mathrm{E}\{z^{\mathbf{X}} \mid \mathbf{R} < \boldsymbol{\sigma}\,,\, \mathbf{R} > 0\} \Pr\{\mathbf{R} < \boldsymbol{\sigma}\,,\, \mathbf{R} > 0\} \\ &= \mathrm{E}\{z^{\mathbf{A}+\mathbf{Z}} \mid \mathbf{Z} = 0\} \Pr\{\mathbf{R} < \boldsymbol{\sigma}\,,\, \mathbf{R} > 0\} \\ &= \Phi(0,z) \frac{\Pr\{\mathbf{R} < \boldsymbol{\sigma}\,,\, \mathbf{R} > 0\}}{\Pr\{\mathbf{Z} = 0\}}. \end{aligned} \tag{2.87}$$

Since at an arbitrary time t the number of batch customers until t having $\mathbf{R} < \boldsymbol{\sigma}$ and $\mathbf{R} > 0$ equals the number of *individual* customers until t having $\mathbf{Z} = 0$, we obtain as a limiting result

$$\frac{\Pr\{\mathbf{R} < \boldsymbol{\sigma}\,,\, \mathbf{R} > 0\}}{\Pr\{\mathbf{Z} = 0\}} = \frac{\lambda}{\gamma}. \tag{2.88}$$

Combining (2.86) - (2.88) and using (2.82) yields (2.85).

2.5 The busy period

For the sojourn time of a supercustomer we were able to derive the LST, starting from recurrence relation (2.1). Such a relation is a typical starting point in $M/G/1$ analysis of waiting time and sojourn time processes. A similar

starting point in the ordinary $M/G/1$ queue for busy period analysis is the branching argument, cf. Cohen[34] p.249 or Takács[94] p.32.
During the service time of the customer that started the busy period new customers (so-called descendants) may arrive. As these first generation descendants are served, again new customers may arrive. We can define the busy period as the service time of a customer plus the time it takes to serve the work associated with the descendants of that customer (this work also including second generation descendants and so on). When the server starts working at the first generation descendants, the time it takes to finish a customers service plus the work associated with the descendants of such a customer again can be viewed as a busy period. So, a busy period is distributed as the convolution of the service time of the customer starting the busy period and the busy periods generated by the customers that arrived during this first service. The branching argument for $M/G/1$ queues states that all busy periods are identically distributed. Moreover, the busy periods associated with first generation customers are independent. This argument leads to the determination of the busy period distribution in the $M/G/1$ queue.
Unfortunately, this argument does not apply to our model. This is due to:

- The distribution of the service time of a supercustomer initiating a busy period is hard to determine because the interarrival time of this customer is atypical. The fact that the previous busy period has ended during its interarrival interval suggests that this interarrival interval is relatively large.
- The length of a (sub)busy period initiated by a supercustomer arriving during the service of the first supercustomer in a busy period depends on the number of supercustomers arriving during that service.

Closely connected to the branching argument in the $M/G/1$ queue is the semi-group property of the busy period length; define for $x \geq 0$ the random variable $\mathbf{Bp}(x)$, which denotes the length of a busy period that started with an amount of work x at the server, then $\mathbf{Bp}(x_1 + x_2) \stackrel{d}{=} \mathbf{Bp}(x_1) + \mathbf{Bp}(x_2)$, $x_1, x_2 \geq 0$, where $\mathbf{Bp}(x_1)$ and $\mathbf{Bp}(x_2)$ are independent. Again, this property does not hold for our model; at the end of the busy period that started with x_1, there might be customers at the pick-up point waiting to be collected. The workload associated with these customers has an influence on $\mathbf{Bp}(x_2)$.

To be able to apply the branching argument, or to use the semi-group property, the state of the process describing the number of customers at the pick-up point has to be taken into account. The latter is done in the framework of the generalization discussed in Chapter 4. There we can apply the above mentioned $M/G/1$ type arguments for the busy period analysis. However, the results are expressed in terms of a set of functional equations of infinite dimension.

A more direct approach to the busy period in a queue with dependence between interarrival and service time is presented in Langaris[68]. The dependence structure in [68] is described by the bivariate density function (2.13).

A considerable part of the analysis in [68] is devoted to $f_n(t,s)$, the probability density function of the following event: the busy period length is t and the number of customers served during this busy period is n, given that the first service of that busy period has length $s < t$. After obtaining the Laplace Transform of $f_n(t,s)$, Langaris derives the Laplace Transform of $b_n'(t)$, the probability density function of the following event: the busy period has length t and exactly n customers are served during this busy period. For this derivation the next relation is used

$$b_n'(t) = \int_{s=0}^{t} f_{n-1}(t,s)dA(s), \tag{2.89}$$

in which $A(\cdot)$ is the marginal distribution function of the service time of an arbitrary customer.

Langaris conditions the length of a new busy period on the length of the service time of the first customer in that busy period, assuming that this customer's service time has an ordinary distribution. This is where a problem arises: that assumption holds for the first busy period, but not for a busy period in steady state. The fact that the previous busy period has ended during the interarrival time of the customer starting the new busy period, makes the interarrival time and, due to the dependence, the service time of this customer atypical.

To obtain $b_n'(t)$, $A(\cdot)$ in (2.89) should be replaced by $A^1(\cdot)$, the service time distribution of the first customer in a busy period. The latter distribution follows from the decomposition property for queues with server vacations; this property was already applied in Section 2.3 (cf. (2.44)).

To determine $A^1(\cdot)$, one can condition on the interarrival time of this customer. The interarrival time of the first customer in a busy period consists of two independent parts, viz.: 1. The sojourn time of his predecessor, which has LST $\frac{\alpha(\lambda+\omega)}{\alpha(\lambda)}$, with λ the parameter of the negative exponentially distributed interarrival times, and $\alpha(\cdot)$ the LST of the sojourn time, as given in Conolly and Choo [41]; 2. The remaining interarrival time, which is again negative exponentially distributed, and has LST $\frac{\lambda}{\lambda+\omega}$. Unfortunately, using these results, (2.89) becomes numerically intractable.

While deriving the LST of the busy period seems a difficult problem, the average length of a busy period is easily obtained using a balancing argument.

With $\mathbf{EB}$ the mean busy period length, $\mathbf{EI} = \frac{1}{\gamma}$ the mean idle period length and using

$$\frac{\mathbf{EB}}{\mathbf{EB} + \mathbf{EI}} = \lambda\beta_1 = \rho\,,$$

the mean busy period length is given by

$$\mathbf{EB} = \frac{\lambda\beta_1/\gamma}{1-\rho}\,,$$

which is the same as $\mathrm{E}\mathbf{B}_I$, the mean busy period length in the corresponding $M/G/1$ queue without dependence. Note that a busy period can have length zero.
Of more interest are 'real' busy periods, viz. busy periods initiated by a supercustomer with positive service time. This conditioned mean busy period length is obviously larger than $\mathrm{E}\mathbf{B}$ and $\mathrm{E}\mathbf{B}_I$.
Define:

$\boldsymbol{\tau}^1$:= service time of a customer initiating a new busy period.
$\tilde{\mathbf{B}}$:= the length of a non-zero busy period.

Then,

$$\mathrm{E}\mathbf{B} = \mathrm{E}\tilde{\mathbf{B}} \cdot \Pr\{\boldsymbol{\tau}^1 > 0\}.$$

Using

$$\begin{aligned}\Pr\{\boldsymbol{\tau}^1 > 0\} &= \Pr\{\boldsymbol{\tau} > 0 \mid \mathbf{W} = 0\} = 1 - \Pr\{\boldsymbol{\tau} = 0 \mid \mathbf{W} = 0\} \\ &= 1 - \frac{\Pr\{\boldsymbol{\tau} = 0, \mathbf{W} = 0\}}{\Pr\{\mathbf{W} = 0\}} = 1 - \frac{\Pr\{\mathbf{R} = 0\}}{1-\rho}\end{aligned}$$

and

$$\Pr\{\mathbf{R} = 0\} = \frac{\gamma}{\gamma + \lambda} r(\gamma + \lambda) \quad \text{(see Corollary 2.2.1)},$$

we obtain

$$\mathrm{E}\tilde{\mathbf{B}} = \frac{\lambda\beta_1}{\gamma} \frac{1}{1 - \rho - \frac{\gamma}{\gamma+\lambda} r(\gamma+\lambda)}. \tag{2.90}$$

Remark 2.5.1
Letting $\gamma \to \infty$ and using Remark 2.2.2 gives as expected $\mathrm{E}\tilde{\mathbf{B}} \to \frac{\beta_1}{1-\rho}$, the mean of $\mathbf{B}_{M/G/1}$, the busy period length in an ordinary $M/G/1$ queue. □

Remark 2.5.2
Letting $\omega = \gamma$ in (2.20) and using the convexity of $r(\cdot)$ gives $\mathrm{E}\tilde{\mathbf{B}} > \mathrm{E}\mathbf{B}_{M/G/1}$.
□

Remark 2.5.3
Although $\mathrm{E}\tilde{\mathbf{B}}$ is larger than $\mathrm{E}\mathbf{B}_{M/G/1}$ and $\mathrm{E}\mathbf{B}_I$, we expect $cv(\tilde{\mathbf{B}})$, the coefficient of variation of the length of a non-zero busy period, to be smaller than $cv(\mathbf{B}_{M/G/1})$ and $cv(\mathbf{B}_I)$. Due to the positive correlation between service and interarrival time of a supercustomer the injection of workload is more regulated than in an ordinary $M/G/1$ queue. This regulation has a stabilizing effect on the busy period length. Moreover, the stronger the correlation, the smaller we expect $cv(\tilde{\mathbf{B}})$ to be. Our conjectures, also stated by Hadidi[54] for a similar model, are supported by simulation results which we present in the next section. □

2.6 Numerical results

In this section we present some numerical results to see the quantitative effects of arrival and traffic intensities in our model. We consider the influence of different service time distributions and we also consider which effect higher moments of the service time distribution have. Finally, we support our claims about the coefficient of variation of the busy period length.

The results have mostly been obtained by numerical evaluation of the infinite product (2.24), its derivative (2.28) and well known formulas for the ordinary $M/G/1$ queue. The infinite product (2.24) and the infinite sum (2.28) converge very fast unless ρ is close to one; it is easily verified that the difference between the k-th term in (2.24) and 1 is of order $O(\rho^k)$ and the k-th term in (2.28) is of order $O(\rho^k)$ for $k = 1, 2, \ldots$. We have only taken recourse to simulation for determining the coefficient of variation of the busy period length.

Waiting times of supercustomers and individual customers

In Tables 2.1a, 2.1b and 2.1c we compare the mean waiting times of supercustomers, $E\mathbf{W}$, of customers in the corresponding $M/G/1$ queue without dependence, $E\mathbf{W}_I$, and of individual customers, $E\tilde{\mathbf{W}}$. We also compare these mean waiting times with the mean waiting times of customers in the corresponding $M/G/1$ queue without collection, $E\mathbf{W}_{M/G/1}$ (in the tables represented by $\gamma = \infty$). Here and in the rest of this section λ is fixed with value 1.

	$\rho = 0.5$				$\rho = 0.9$			
γ	$cor.(\sigma,\tau)$	$E\mathbf{W}$	$E\mathbf{W}_I$	$E\tilde{\mathbf{W}}$	$cor.(\sigma,\tau)$	$E\mathbf{W}$	$E\mathbf{W}_I$	$E\tilde{\mathbf{W}}$
0.1	0.9129	3.392	5.500	16.292	0.9129	23.197	89.100	34.875
0.5	0.7071	0.987	1.500	3.474	0.7071	9.951	24.300	12.157
1	0.5774	0.701	1.000	1.902	0.5774	8.729	16.200	9.798
2	0.4472	0.574	0.750	1.148	0.4472	8.290	12.150	8.811
4	0.3333	0.524	0.625	0.797	0.3333	8.152	10.125	8.408
10	0.2182	0.504	0.550	0.609	0.2182	8.109	8.9100	8.210
∞	0	0.500	0.500	0.500	0	8.100	8.100	8.100

Table 2.1a: Waiting times in the collector model and the associated queue without collecting. Exponential service times.

	$\rho = 0.5$				$\rho = 0.9$			
γ	$cor.(\sigma,\tau)$	$E\mathbf{W}$	$E\mathbf{W}_I$	$E\tilde{\mathbf{W}}$	$cor.(\sigma,\tau)$	$E\mathbf{W}$	$E\mathbf{W}_I$	$E\tilde{\mathbf{W}}$
0.1	0.9535	3.213	5.250	16.176	0.9535	20.494	85.050	32.321
0.5	0.8165	0.790	1.250	3.330	0.8165	6.452	20.250	8.718
1	0.7071	0.491	0.750	1.731	0.7071	4.937	12.150	6.036
2	0.5774	0.346	0.500	0.943	0.5774	4.327	8.100	4.859
4	0.4472	0.282	0.375	0.565	0.4472	4.121	6.075	4.378
10	0.3015	0.255	0.300	0.361	0.3015	4.060	4.860	4.161
∞	0	0.250	0.250	0.250	0	4.050	4.050	4.050

Table 2.1b: Waiting times in the collector model and the associated queue without collecting. Deterministic service times.

	$\rho = 0.5$				$\rho = 0.9$			
γ	$cor.(\sigma,\tau)$	EW	$\mathbf{EW}_I$	$\mathrm{E}\tilde{\mathbf{W}}$	$cor.(\sigma,\tau)$	EW	$\mathbf{EW}_I$	$\mathrm{E}\tilde{\mathbf{W}}$
0.1	0.9393	3.521	5.666	16.374	0.9393	25.160	91.800	36.756
0.5	0.7746	1.128	1.666	3.589	0.7746	12.463	27.000	14.647
1	0.6547	0.853	1.166	2.040	0.6547	11.364	18.900	12.426
2	0.5222	0.734	0.916	1.302	0.5222	10.974	14.850	11.493
4	0.3974	0.689	0.791	0.961	0.3974	10.850	12.825	11.105
10	0.2641	0.671	0.671	0.775	0.2641	10.809	11.610	10.910
∞	0	0.666	0.666	0.666	0	10.800	10.800	10.800

TABLE 2.1C: WAITING TIMES IN THE COLLECTOR MODEL AND THE ASSOCIATED QUEUE WITHOUT COLLECTING. HYPER-EXPONENTIAL SERVICE TIMES.

In Table 2.1c the service time of a message is with probability $\frac{1}{4}$ exponentially distributed with parameter μ_1 and with probability $\frac{3}{4}$ exponentially distributed with parameter μ_2. $\mu_1 = 1$ and $\mu_2 = 3$ for $\rho = 0.5$; $\mu_1 = \frac{5}{9}$ and $\mu_2 = \frac{5}{3}$ for $\rho = 0.9$.

Comparing EW and $\mathbf{EW}_I$ in Tables 2.1a-c we conclude that the positive correlation between interarrival and service times leads to a reduction of mean waiting times. The reduction is particularly strong in heavy traffic. A similar observation has been made by Hadidi [54] for the dependence structure displayed in formula (2.13). Tables 2.1a, 2.1b and 2.1c also show that the influence of the service time distribution of individual customers decreases when γ approaches 0. In fact, it can be seen from (2.6) that for any service time distribution $B(\cdot)$, $\lim_{\gamma \downarrow 0} \mathrm{E}\left(e^{-\gamma\omega\boldsymbol{\tau}_i}\right) = (1 + \lambda\beta_1\omega)^{-1}$, i.e. the distribution of the scaled service time of a batch customer converges to the negative exponential distribution with mean $\lambda\beta_1$.
Finally we see that $\mathrm{E}\tilde{\mathbf{W}}$ converges slower towards $\mathbf{EW}_{M/G/1}$ than EW for $\gamma \to \infty$, but most of the difference is due to the remaining collecting interval.

The influence of higher moments
In an ordinary $M/G/1$ queue, the influence of the service time distribution on mean waiting time is limited to its first and second moment. The question arises whether that is the case in our model. Formula (2.48) shows that $r'(\gamma)$ contributes to the mean waiting time and therefore we would suspect that the whole service time distribution plays a role. Tables 2.2a and 2.2b indicate that this conjecture is correct but also that the influence of higher moments is almost negligible. In these tables we consider mixtures of exponential distributions for the service time of individual customers.

γ	EW1	EW2
0.1	5.5479	5.5960
0.5	3.4816	3.5304
1	3.3154	3.3451
2	3.2559	3.2697
4	3.2344	3.2393
10	3.2252	3.2260

	β_1	β_2	β_3
Mix.1	0.5	3.222	38.407
Mix.2	0.5	3.222	52.345

γ	EW1	EW2
0.1	71.597	72.293
0.5	64.657	64.943
1	64.291	64.426
2	64.173	64.226
4	64.113	64.149
10	64.116	64.119

	β_1	β_2	β_3
Mix.1	0.9	12.822	307.207
Mix.2	0.9	12.822	423.554

TABLES 2.2A (LEFT) AND 2.2B (RIGHT): THE INFLUENCE OF HIGHER MOMENTS OF THE SERVICE TIME DISTRIBUTION ON THE MEAN WAITING TIMES. IN TABLE 2.2A $\rho = 0.5$, IN TABLE 2.2B $\rho = 0.9$.

In Table 2.2 **EW1** and **EW2** are the mean waiting times for a supercustomer composed of individual customers with service time distribution mixtures 1 and 2, respectively.

The busy period

In Section 5 we suggested that in our model the coefficient of variation for real busy periods, $cv(\tilde{\mathbf{B}})$, would be smaller than $cv(\mathbf{B}_{M/G/1})$ and $cv(\mathbf{B}_I)$. As explained there, analytical and numerical results are not available, so to obtain column $cv(\tilde{\mathbf{B}})$ in Tables 3a, 3b and 3c below we have used a simulation. The simulation was performed with the queueing simulation software package Q+, running the process for 10^6 time units. $\mathrm{E}\tilde{\mathbf{B}}$ has been obtained using formula (2.90).

	$\rho = 0.5$				$\rho = 0.9$			
γ	$\mathrm{E}\tilde{\mathbf{B}}$	$cv(\tilde{\mathbf{B}})$	$\mathrm{E}\mathbf{B}_I$	$cv(\mathbf{B}_I)$	$\mathrm{E}\tilde{\mathbf{B}}$	$cv(\tilde{\mathbf{B}})$	$\mathrm{E}\mathbf{B}_I$	$cv(\mathbf{B}_I)$
0.1	10.076	0.954	10.000	1.844	90.057	1.473	90.000	4.583
0.5	2.439	1.230	2.000	2.236	20.423	2.763	18.000	5.385
1	1.603	1.407	1.000	2.646	13.356	3.546	9.000	6.245
2	1.233	1.557	0.500	3.317	10.517	3.991	4.500	7.681
4	1.080	1.652	0.250	4.359	9.465	4.123	2.250	9.950
10	1.016	1.711	0.100	6.557	9.089	4.266	0.900	14.799
∞	1.000	1.732	0.000	∞	9.000	4.359	0.000	∞

TABLE 2.3A: THE BUSY PERIOD LENGTH, MEAN AND COEFFICIENT OF VARIATION. EXPONENTIAL SERVICE TIME.

	$\rho = 0.5$				$\rho = 0.9$			
γ	$\mathrm{E}\tilde{\mathbf{B}}$	$cv(\tilde{\mathbf{B}})$	$\mathrm{E}\mathbf{B}_I$	$cv(\mathbf{B}_I)$	$\mathrm{E}\tilde{\mathbf{B}}$	$cv(\tilde{\mathbf{B}})$	$\mathrm{E}\mathbf{B}_I$	$cv(\mathbf{B}_I)$
0.1	10.053	0.894	10.000	1.789	90.012	1.168	90.000	4.472
0.5	2.374	0.983	2.000	2.000	19.525	2.043	18.000	4.899
1	1.534	0.995	1.000	2.236	12.324	2.540	9.000	5.385
2	1.167	0.994	0.500	2.646	9.702	2.856	4.500	6.245
4	1.032	1.002	0.250	3.317	9.061	3.002	2.250	7.811
10	1.001	0.999	0.100	4.796	9.000	2.971	0.900	10.909
∞	1.000	1.000	0.000	∞	9.000	3.000	0.000	∞

TABLE 2.3B: THE BUSY PERIOD LENGTH, MEAN AND COEFFICIENT OF VARIATION. DETERMINISTIC SERVICE TIME.

	$\rho = 0.5$				$\rho = 0.9$			
γ	$\mathrm{E}\tilde{\mathbf{B}}$	$cv(\tilde{\mathbf{B}})$	$\mathrm{E}\mathbf{B}_I$	$cv(\mathbf{B}_I)$	$\mathrm{E}\tilde{\mathbf{B}}$	$cv(\tilde{\mathbf{B}})$	$\mathrm{E}\mathbf{B}_I$	$cv(\mathbf{B}_I)$
0.1	10.086	1.010	10.000	1.880	90.090	1.591	90.000	4.655
0.5	2.463	1.390	2.000	2.380	20.720	3.284	18.000	5.686
1	1.625	1.645	1.000	2.887	13.614	4.068	9.000	6.758
2	1.249	1.863	0.500	3.697	10.675	4.492	4.500	8.505
4	1.089	1.973	0.250	4.933	9.548	4.831	2.250	11.210
10	1.019	2.052	0.100	7.506	9.107	4.852	0.900	16.902
∞	1.000	2.082	0.000	∞	9.000	5.066	0.000	∞

TABLE 2.3C: THE BUSY PERIOD LENGTH, MEAN AND COEFFICIENT OF VARIATION. HYPER-EXPONENTIAL SERVICE TIME (SAME AS IN TABLE 2.1C).

In Tables 3a, 3b, 3c $cv(\tilde{\mathbf{B}})$ is smaller than $cv(\mathbf{B}_{M/G/1})$ and $cv(\mathbf{B}_I)$, supporting the conjectures made in Remark 2.5.3.

APPENDICES

2.A ERGODICITY OF THE WAITING AND SERVICE TIME PROCESS

In this appendix we prove that the Markov process $\{(\mathbf{W}_n, \boldsymbol{\tau}_n), n = 0, 1, \ldots\}$, as defined in Section 2.2 is ergodic if $\lambda\beta_1 < 1$. At first sight this appears to be a natural condition; the amount of work entering the system per unit of time is less than 1. However, the server can be idle while there might be customers waiting at the bus-stop, hence the system is not work-conserving. This inefficiency might have an effect on the ergodicity criterion. That this is not the case here is due to the fact that the system *is* work-conserving with respect to the work present at the service facility and that on the average the amount of work arriving per unit of time at this service facility is $\lambda\beta_1 < 1$. Retaining customers at the bus-stop only affects their waiting time if the server becomes idle before they are collected.

In this appendix we show the ergodicity of a slightly different process, namely the embedded process $\{(\mathbf{W}_n, \mathbf{K}_n), n = 1, 2, \ldots\}$, where the embedded time

points are the moments of arrival, either of a single customer or of a bus, and with $\mathbf{K}_n$ denoting the number of customers at the bus-stop just prior to the $n-th$ epoch. So the time between two events defining embedded time points is exponentially distributed with parameter $\lambda + \gamma$.
The ergodicity of the joint waiting and service time process of *batch customers* follows from the observation that the service time of a batch customer is unequivocally defined by the number of customers in the batch and applying the PASTA property.
The technique we use is adapted from Laslett et al.[69] and is summarized in Appendix B. The basic idea is to identify a bounded subset of the state space which is positive recurrent. Then, under certain conditions which are described in Laslett et al.[69], the Markov process is ergodic. We next apply the method of Laslett et al.[69], by performing the three-step scheme as described in Appendix B.

1. ϕ-irreducibility.
With $[0,\infty) \times \{0,1,\ldots\}$ being the state-space of $\{(\mathbf{W}_n, \mathbf{K}_n)\}$ we choose for ϕ an arbitrary non-zero measure with an atom at $\{(0,0)\}$. It is obvious that $\{(0,0)\}$ can be reached from any subset $A \subset [0,\infty) \times \{0,1,\ldots\}$ in at most two steps. Hence, $\{(\mathbf{W}_n, \mathbf{K}_n)\}$ is ϕ-irreducible.

2. Identifying possible test sets.
It is readily verified that according to Theorem 3.2 of [69] every set A of the form $[0,x_1) \times \{0,1,\ldots,N\}$, $x_1 > 0, N < \infty$ is a test set.

3. Applying Theorem B.1 of Appendix B.
We apply Theorem B.1 of Appendix B, using the function $g((x,i)) := x + i\beta_1$. First, let $\mathbf{X}_1 := (x_1, i_1) = (\mathbf{W}_1, \mathbf{K}_1)$ and $\mathbf{X}_2 := (x_2, i_2) = (\mathbf{W}_2, \mathbf{K}_2)$. Then by conditioning on the time between epochs we obtain

$$\begin{aligned} \mathrm{E}[g(\mathbf{X}_2)|\mathbf{X}_1] &= i_1\beta_1 + \frac{\lambda\beta_1}{\lambda+\gamma} + \int_{u=0}^{\infty} \mathrm{E}(\max\{0, x_1 - u\})(\lambda+\gamma)e^{-(\lambda+\gamma)u}du \\ &= i_1\beta_1 + x_1 - \frac{1 - \lambda\beta_1 - e^{-(\lambda+\gamma)x_1}}{\lambda+\gamma}. \end{aligned} \tag{2.91}$$

With $g(\mathbf{X}_0) = i_1\beta_1 + x_1$ it follows from (2.91) that $\mathrm{E}[g(\mathbf{X}_1)|\mathbf{X}_0] - g(\mathbf{X}_0) < 0$ if $e^{-(\lambda+\gamma)x_1} < 1 - \lambda\beta_1$, which is the case for x_1 large enough. If we choose $A = [0,x_1) \times \{0,\ldots,i_1\}$ such that the latter is the case then the proof of the ergodicity of $\{(\mathbf{W}_n, \mathbf{K}_n), n = 1,2,\ldots\}$ follows from [69]. □

Remark 2.A.1
A first observation is that for the test set A the value x_1 can not be arbitrary. This seems logical because after a transition the mean amount of work in the system should be less than before the transition. A second observation is that i_0 can be any number. This reflects the property that the total amount of work in the *system* does not change at the moment a bus collects the customers.

2.B Proof of Lemma 2.2.1

(i) & (ii). We prove that $\mid g(\omega_1) - g(\omega_2) \mid \leq \lambda\beta_1 \mid \omega_1 - \omega_2 \mid$ for all ω_1, ω_2 with $Re\,\omega_1, Re\,\omega_2 \geq 0$. Distinguish two cases.
a. $Re\,\omega_1 \geq Re\,\omega_2$. Using (2.22),

$$\begin{aligned} & \mid g(\omega_1) - g(\omega_2) \mid \\ = \; & \lambda \mid \beta(\omega_1) - \beta(\omega_2) \mid \\ = \; & \lambda \mid \int_{t=0}^{\infty} e^{-\omega_2 t}[1 - e^{-(\omega_1 - \omega_2)t}]dB(t) \mid \\ \leq \; & \lambda\beta_1 \mid \omega_1 - \omega_2 \mid, \end{aligned}$$

since $\mid e^{-\omega} \mid \leq 1$ and $\mid 1 - e^{-\omega} \mid \leq \mid \omega \mid$ for all ω with $Re\,\omega \geq 0$.
b. $Re\,\omega_1 \leq Re\,\omega_2$.
This case proceeds similarly.

Moreover $Re\,g(\omega) \geq 0$ for all ω with $Re\,\omega \geq 0$. As we assumed $\rho = \lambda\beta_1 < 1$, we conclude from the fixed-point theorem for contractions, cf. Apostol [7], that the equation $g(\omega) = \omega, Re\,\omega \geq 0$ has a unique solution ω^* and that $\lim_{M\to\infty} g^{(M)}(\omega) = \omega^*$ for all ω with $Re\,\omega \geq 0$. Since $g(\gamma) = \gamma + \lambda(1 - \beta(\gamma)) \geq \gamma$ and $g(\gamma + \lambda) = \gamma + \lambda(1 - \beta(\gamma + \lambda)) \leq \gamma + \lambda$, ω^* is real and $\gamma \leq \omega^* \leq \gamma + \lambda$.

(iii). From the theory of infinite products, cf. Titchmarsh [96] p.18,

$$\prod_{h=0}^{\infty} \frac{f(g^{(h)}(\omega))}{g^{(h+1)}(\omega)}$$

converges iff

$$\sum_{h=0}^{\infty} [1 - \frac{f(g^{(h)}(\omega))}{g^{(h+1)}(\omega)}]$$

converges. Using (2.21) and (2.22),

$$1 - \frac{f(g^{(h)}(\omega))}{g^{(h+1)}(\omega)} = \frac{\gamma(1 - \frac{g^{(h)}(\omega)}{g^{(h+1)}(\omega)}) + g^{(h)}(\omega) - g^{(h+1)}(\omega)}{\gamma + g^{(h)}(\omega) - g^{(h+1)}(\omega)}.$$

Since $\mid g^{(h+1)}(\omega) - g^{(h)}(\omega) \mid \leq \lambda\beta_1 \mid g^{(h)}(\omega) - g^{(h-1)}(\omega) \mid$, cf. the proof of (i) & (ii),

$$\sum_{h=0}^{\infty} \frac{\gamma(1 - \frac{g^{(h)}(\omega)}{g^{(h+1)}(\omega)}) + g^{(h)}(\omega) - g^{(h+1)}(\omega)}{\gamma + g^{(h)}(\omega) - g^{(h+1)}(\omega)}$$

converges and so

$$\prod_{h=0}^{\infty} \frac{f(g^{(h)}(\omega))}{g^{(h+1)}(\omega)}$$

converges.

□

2.C Proof of Lemma 2.3.1

(i). We successively prove both parts of the inequality.

$$\frac{d}{dt}[te^{-\gamma t}] = (1-\gamma t)e^{-\gamma t}. \tag{2.92}$$

Using (2.30) and (2.92),

$$\int_{t=0}^{\infty} te^{-\gamma t} dR(t) = \int_{t=0+}^{\infty} te^{-\gamma t} dR(t)$$

$$\leq \sup_{t\in(0,\infty)} [te^{-\gamma t}] \int_{t=0+}^{\infty} dR(t) = \frac{e^{-1}}{\gamma}(1 - \frac{\gamma}{\gamma+\lambda} r(\gamma+\lambda)).$$

Using (2.31),

$$\int_{t=0}^{\infty} te^{-\gamma t} dR(t) - \frac{\lambda\beta_1}{\gamma} = \int_{t=0}^{\infty} \left[te^{-\gamma t} + \frac{1}{\gamma}e^{-\gamma t} - \frac{1}{\gamma}\right] dR(t).$$

$$\frac{d}{dt}\left[te^{-\gamma t} + \frac{1}{\gamma}e^{-\gamma t} - \frac{1}{\gamma}\right] = -\gamma te^{-\gamma t} \leq 0, \qquad t \geq 0. \tag{2.93}$$

Using (2.93),

$$te^{-\gamma t} + \frac{1}{\gamma}e^{-\gamma t} - \frac{1}{\gamma} \leq \left[te^{-\gamma t} + \frac{1}{\gamma}e^{-\gamma t} - \frac{1}{\gamma}\right]|_{t=0} = 0, \qquad t \geq 0.$$

(ii). Recalling Remark 2.3.1,

$$\frac{\lambda\beta_1}{1-\lambda\beta_1} \int_{t=0}^{\infty} te^{-\gamma t} dR(t)$$

$$= \sum_{h=1}^{\infty} \frac{g^{(h)'}(0)(f(g^{(h)}(0))g'(g^{(h)}(0)) - f'(g^{(h)}(0))g(g^{(h)}(0)))}{f(g^{(h)}(0))g(g^{(h)}(0))}.$$

Using (2.21) and (2.22), we find that the first term in the sum equals

$$\frac{\lambda\beta_1}{\gamma} \frac{\lambda(1-\beta(\gamma))}{\gamma - \lambda(1-\beta(\gamma))} \frac{\gamma + \lambda(1-\beta(\gamma)) + 2\lambda\gamma\beta'(\gamma)}{\gamma + \lambda(1-\beta(\gamma))}.$$

Moreover it is straightforward to verify by induction that all terms in the sum are strictly positive, the latter following from the contraction property of $g(\cdot)$ and $\rho < 1$.

□

Chapter 3

The single-server queue with a correlated input process

3.1 Introduction

In this chapter we discuss a broader framework of $M/G/1$ queues with a positive correlation between interarrival and service times of customers. This framework also includes the dependence structure we studied in the previous chapter. We consider the dependence structure in the present chapter from the perspective of queues with gated admission. Work arrives at a single-server queue according to a process with stationary non-negative independent increments. This work, however, does not immediately enter the queue of the service facility; instead it accumulates behind a gate. At exponential intervals the gate is opened and - after the addition of an independent component - the work is collected and delivered as a single customer at the queue of the service facility. The additional component might be viewed as a set-up time. Like in the previous chapter, we can view the service facility as an M/G/1 queue in which the interarrival and service time for each customer are positively correlated. Again, if the interval between two consecutive openings of the gate is relatively long (short), it is likely that a relatively large (small) amount of work has accumulated in that interval.

Our motivation for studying the above-sketched dependence structure - and the main contribution of this chapter - is that the present model is a unification and generalization of the so far studied M/G/1 queues with a positive correlation between interarrival and subsequent service times [19, 20, 32, 41].

The collecting model of the previous chapter actually describes the following situation: given that the interarrival time $\boldsymbol{\sigma}_n$ between the $(n-1)st$ and the $n{-}th$ batch equals u, the amount of work $\boldsymbol{\tau}_n$ in that $n{-}th$ batch is distributed as the state of a Compound Poisson Process (CPP) $\mathbf{Y}(\cdot)$ at time u, where the

intensity of the jumps is λ, the jump-sizes have distribution $B(\cdot)$ and $\mathbf{Y}(0) = 0$. The Laplace-Stieltjes Transform (LST) of $\mathbf{Y}(u)$ is given by

$$\mathrm{E}(e^{-\omega \boldsymbol{\tau}_n} \mid \boldsymbol{\sigma}_n = u) = e^{-u\lambda(1-\beta(\omega))}, \qquad Re\,\omega \geq 0, \tag{3.1}$$

in which $\beta(\cdot)$ is the LST of the service time (jump-size) distribution $B(\cdot)$.
Cidon et al.[32] analyse the M/G/1 queue in which the interarrival and service time are related through $\boldsymbol{\tau}_n = \alpha \boldsymbol{\sigma}_n + \tilde{\boldsymbol{\tau}}_n$, $(0 < \alpha < 1)$, with $\tilde{\boldsymbol{\tau}}_n$ an amount of work that is independent of the interarrival time. For this model

$$\mathrm{E}(e^{-\omega \boldsymbol{\tau}_n} \mid \boldsymbol{\sigma}_n = u) = E(e^{-\omega \tilde{\boldsymbol{\tau}}_n}) e^{-u\alpha\omega}, \qquad Re\,\omega \geq 0. \tag{3.2}$$

As Compound Poisson Processes and the process $\mathbf{Y}(u) = \alpha u$ are examples of processes with stationary non-negative independent increments, these models are special cases of the model of the present chapter.

In the remainder of this introduction we present a more detailed model description and an overview of the chapter. The literature that is most relevant for this chapter will be discussed in Section 3.2, where we relate existing studies on dependent interarrival and service times to the framework of the present chapter. A more general overview of related literature on the subject has already been presented in the introduction of Chapter 2.

Model description
In the present study we extend the analysis of [19, 20, 32, 41] to M/G/1 queues with arrival rate γ in which the LST of the service time $\boldsymbol{\tau}_n$, given that the interarrival time $\boldsymbol{\sigma}_n$ equals u, is of the following form

$$\mathrm{E}(e^{-\omega \boldsymbol{\tau}_n} \mid \boldsymbol{\sigma}_n = u) = v(\omega) e^{-\phi(\omega)u}, \qquad Re\,\omega \geq 0, \tag{3.3}$$

with $\phi(0) = 0$, and $\phi(\omega)$ having a completely monotone derivative; $v(\cdot)$ is the LST of the probability distribution of a non-negative random variable.
The service time $\boldsymbol{\tau}_n$ consists of two parts: a component which depends on the interarrival time, represented by $e^{-\phi(\omega)u}$, and an 'ordinary' M/G/1 service time with LST $v(\omega)$, which does not depend on the interarrival time. The dependent part of the service time will be shown to represent increments during an exponential gate opening interval of an arbitrary process with stationary non-negative independent increments. In the next section this process is described in detail; here we only mention that $\phi'(0)$ and $(-\phi''(0) + (\phi'(0))^2)$ are respectively the first and second moment of the amount of work arriving at the gate per unit of time. For the independent component of the service time, the first two moments are $-v'(0)$ and $v''(0)$.
The bivariate LST of $\boldsymbol{\sigma}_n$ and $\boldsymbol{\tau}_n$ follows from (3.3):

$$\mathrm{E}(e^{-\omega_1 \boldsymbol{\sigma}_n - \omega_2 \boldsymbol{\tau}_n}) = \frac{\gamma v(\omega_2)}{\gamma + \omega_1 + \phi(\omega_2)}, \qquad Re\,\omega_1 \geq 0, Re\,\omega_2 \geq 0. \tag{3.4}$$

Expression (3.4) leads to

$$\begin{aligned} \mathrm{E}\boldsymbol{\tau}_n &= -v'(0) + \frac{\phi'(0)}{\gamma}, \\ \mathrm{E}\boldsymbol{\tau}_n^2 &= v''(0) - 2\frac{v'(0)\phi'(0)}{\gamma} - \frac{\gamma\phi''(0) - 2(\phi'(0))^2}{\gamma^2}, \end{aligned} \tag{3.5}$$

$$Cov(\boldsymbol{\sigma}_n, \boldsymbol{\tau}_n) = \frac{\phi'(0)}{\gamma^2} \geq 0, \tag{3.6}$$

$$correl(\boldsymbol{\sigma}_n, \boldsymbol{\tau}_n) = \left[1 + \frac{-\gamma\phi''(0)}{(\phi'(0))^2} + \left(\frac{\gamma}{\phi'(0)}\right)^2 (v''(0) - v'(0)^2)\right]^{-1/2} \in [0,1]. \tag{3.7}$$

Note that $correl(\boldsymbol{\sigma}_n, \boldsymbol{\tau}_n) \to 1$ if $\gamma \to 0$ and $correl(\boldsymbol{\sigma}_n, \boldsymbol{\tau}_n)$ decreases for increasing γ and increasing variability of the independent service time component.

Overview of the chapter
Section 3.2 describes the dependence structure in more detail. We give a few examples and show how previously analysed models for M/G/1 queues with dependence fit into this structure. Section 3.3 is devoted to the analysis of the joint distribution of the waiting and service time of an arbitrary customer in steady state. There we also present a vacation-type workload decomposition for the M/G/1 queue with the exponential gating mechanism that was described in the beginning of this section.

3.2 The dependence structure

In Section 3.1 the dependent part of the service time distribution was characterized by the LST $e^{-\phi(\omega)u}$; here $e^{-\phi(\omega)}$ is the LST of a non-negative random variable with an infinitely divisible distribution (cf. Theorem 1 on p.450 of Feller[46]):

Theorem 3.2.1
The function $\psi(\omega)$ is the LST of an infinitely divisible probability distribution if and only if $\psi(\omega) = e^{-\phi(\omega)}$, where $\phi(\omega)$ has a completely monotone derivative and $\phi(0) = 0$. □

Remark 3.2.1
An equivalent definition of an infinitely divisible distribution is that it has an LST of the form $\psi(\omega) = e^{-\phi(\omega)}$ where

$$\phi(\omega) = \int_0^\infty \frac{1 - e^{-\omega x}}{x} dP(x), \quad \omega \geq 0, \tag{3.8}$$

and P is a measure such that

$$\int_1^\infty x^{-1} dP(x) < \infty.$$

□

Remark 3.2.2
In Feller ([46],p.303) it is shown that the following classes of probability distributions are identical:

(i) Infinitely divisible distributions.

(ii) Distributions of increments in processes with stationary independent increments.

(iii) Limits of sequences of compound Poisson distributions. □

A process $\{\mathbf{Y}(u),\ u \geq 0\}$ has stationary independent increments when the distribution of $\mathbf{Y}(t+s) - \mathbf{Y}(s)$ is independent of s, for all $t, s \geq 0$. In terms of a collecting procedure, this means that the rate of increments of work at the pick-up point (gate) does not depend on the length of the elapsed collecting (gate opening) interval. An important characterization of a process $\mathbf{Y}(u)$ with stationary independent increments with $\mathbf{Y}(0) = 0$ is: for $s, t > 0$, $\mathbf{Y}(t+s) \stackrel{d}{=} \mathbf{Y}(t) + \mathbf{Y}(s)$, where $\mathbf{Y}(t)$ and $\mathbf{Y}(s)$ are independent, here $\stackrel{d}{=}$ denotes equality in distribution (cf. Feller[46], p.180). Obviously

$$\mathrm{E}\left(e^{-\omega \mathbf{Y}(u)}\right) = \left[\mathrm{E}\left(e^{-\omega \mathbf{Y}(1)}\right)\right]^u = e^{-u\phi(\omega)}, \qquad u, \omega \geq 0. \tag{3.9}$$

We next present a few examples from the class of processes with stationary independent increments, in which the character of the dependence structure comes more to light. We also show how some studied dependence structures fit into this class. In example i, the ϕ function studied will be denoted by $\phi_i(\cdot)$, $i = 1, \ldots, 4$.

Example 1: The Compound Poisson Process.
This is the dependence structure as studied in Chapter 2. From (3.1) and (3.3) we see that $\phi_1(\omega)$ can be expressed as $\phi_1(\omega) = \lambda(1 - \beta(\omega))$, with λ the intensity of jump occurrences, and $\beta(\cdot)$ the LST of the jump-size distribution $B(\cdot)$. For this example $dP(x) = \lambda x dB(x)$.

Example 2: Linear Dependency.
This is the dependence structure as studied by Cidon et al.[32]. Here $\phi_2(\omega) = \alpha\omega$ (cf. (3.2)). It is readily verified that the underlying infinitely divisible distribution is the limit of a sequence of Compound Poisson Processes, i.e., $\phi_2(\omega) = \lim_{k \to \infty} \lambda_k(1 - \beta_k(\omega))$, for all $\omega \geq 0$, when $\lambda_k \to \infty$, $\lambda_k \beta_k \to \alpha$ and

$\frac{\beta_k^{(2)}}{\beta_k} \to 0$ as $k \to \infty$. Here β_k and $\beta_k^{(2)}$ denote the first and second moment of jump-size distribution $B_k(\cdot)$ respectively. For linear dependency we find $dP(x) = \alpha\delta(x)$, where $\delta(x)$ is the Dirac delta function.

Example 3: Gamma Distributions.
In this example $\mathbf{Y}(u)$ has a Gamma(ζ^{-1}, u) distribution ($\zeta > 0$), i.e. $\psi(\omega) = \frac{1}{1+\zeta\omega}$. Rewriting to the standard exponential form (cf. (3.8)) we obtain: $\phi_3(\omega) = \int_0^\infty \frac{1-e^{-\zeta\omega y}}{y} e^{-y} dy$, leading to $dP(x) = e^{-x/\zeta} dx$. This process can be obtained as the limit of a sequence of CPP's with $\lambda_k = k$, and B_k being a Gamma(ζ^{-1}, k^{-1}) distribution, $k = 1, 2, \ldots$.

Example 4: Subordination of Processes with Stationary Independent Increments.
Let $\mathbf{Z}(t)$ and $\mathbf{T}(t)$, $t \geq 0$, be processes with stationary non-negative independent increments with $\mathbf{Z}(0) = \mathbf{T}(0) = 0$. Define the process $\mathbf{Y}(t), t \geq 0$, as follows. Let $U_t(\cdot)$ and $Q_t(\cdot)$, $t \geq 0$, denote the distributions of the states of $\mathbf{Z}$ and $\mathbf{T}$ at time t respectively. Then $U_t^0(\cdot)$, the distribution of the state of $\mathbf{Y}(t)$ at time $t \geq 0$, is defined by

$$U_t^0(x) := \int_{s=0}^{\infty} U_s(x) dQ_t(s), \qquad x \geq 0. \tag{3.10}$$

We can view $\mathbf{T}$ as a transformer of the time in the process $\mathbf{Z}$. A simple illustration of such a subordination is with $\mathbf{Z}$ a CPP as described in example 1 and $\mathbf{T}$ a Gamma distribution as described in example 3. This Gamma randomization yields another process $\mathbf{Y}$ with stationary non-negative independent increments, with

$$\phi_4(\omega) = \int_0^\infty \frac{1 - e^{-\zeta\lambda(1-\beta(\omega))x}}{x} e^{-x} dx, \quad \omega \geq 0. \tag{3.11}$$

Comparison with example 3 shows that $\phi_4(\omega) = \phi_3(\phi_1(\omega))$.

So far in this section we only considered the dependent component of the service time of a customer. Adding an independent 'ordinary' M/G/1 component further widens the range of the class of service time distributions. For example, this class, as described by (3.3), now also includes the case where the interarrival and service time have a bivariate exponential joint distribution with density function:

$$g(s,t) = \zeta\mu(1-p)e^{-\zeta s - \mu t} I_0[2\{\zeta\mu p s t\}^{1/2}], \tag{3.12}$$

where $I_0[\cdot]$ is a zero-order modified Bessel function of the first kind, and $p \in [0,1)$ is the correlation between $\boldsymbol{\sigma}_n$ and $\boldsymbol{\tau}_n$. The marginal distributions of the interarrival and service time are exponential.

To verify that the dependence structure as described by (3.12) also fits in the framework of (3.3) we characterize the bivariate distribution in a reliability context. The following description for the bivariate exponential distribution is adapted from Al-Saadi et al.[5]. Suppose that we have two components which can survive a number of shocks before a failure occurs. The number of shocks to failure is the same for each component but is geometrically distributed with parameter p. Furthermore, the time between two shocks for component 1 (component 2) is exponentially distributed with parameter ζ (μ). After conditioning on the number of shocks to failure we find that the joint distribution of the failure times of both components is the same as the joint distribution of the interarrival and service time as given by (3.12). We can interpret our collector model in the same way by viewing the arrival of a customer at the bus-stop as a 'shock', and the collecting of the customers as a 'failure'. Then the number of shocks to failure has a geometric distribution with parameter $\frac{\lambda}{\gamma+\lambda}$, and with the time between two shocks being exponentially distributed with parameter $\gamma+\lambda$ this leads to an exponentially distributed failure time (customer interarrival time) with parameter λ.
Moreover, the number of shocks minus one equals the size of the batch collected (cf. expression (2.2) of Chapter 2). Let the service time of a single customer be exponentially distributed with parameter μ_1, and add an independent $\exp(\mu_1)$ distributed service time to each batch (the independent component of (3.3)). After interpreting the interarrival time as the time to failure of the first component, now we view the total service time of a batch customer as the time to failure of the second component (or equivalently the total number of customers in a batch as the number of shocks to failure). Then with $\zeta = \gamma+\lambda$, $\mu = \mu_1$ and $p = \frac{\lambda}{\gamma+\lambda}$ it is readily verified that (3.4) and (3.12) describe the same bivariate exponential distribution. We remark that this equality could also have been obtained by writing down the Laplace-Stieltjes Transform of (3.12).

Remark 3.2.3
In the context of reliability models we can interpret the correlation factor p in (3.12). For $p \downarrow 0$, or equivalently $\gamma \to \infty$, a component always fails after the first 'shock', and the interarrival and service times are independent. For $p \uparrow 1$, or equivalently $\gamma \downarrow 0$, the number of 'shocks' to failure tends to infinity. □

Cidon et al.[32] also discuss an extension of the linear dependence structure as described by (3.2) (cf. also example 2).
Their model can be characterized as a gated model with alternating On/Off periods for the accumulation process of work at the gate; during an On period work accumulates at a constant rate α, during an Off period no work arrives at the gate (rate 0). Moreover, at each opening of the gate according to a Bernoulli distribution it is decided whether the next period is an On or an Off period.
For this model Cidon et al.[32] analyse the stationary sojourn time distribution by means of an iteration technique that is similar to the iteration procedure we apply for our model.

It appears that this On/Off model can be extended to the general dependence structure - including the independent component - as described by (3.3), and that this extension still can be analysed by means of an iterative procedure.
In Remark 3.3.2 we discuss a model that is again somewhat more general than the above-mentioned On/Off models. First we analyse the model with dependence structure (3.3).

3.3 The waiting and sojourn time

In this section we derive the LST for the joint steady-state distribution of the waiting and service time of an arbitrary customer. The analysis of the joint waiting and service time distribution proceeds along the same way as the analysis in Section 2.2. With $\mathbf{W}_n$ denoting the waiting time of the $n-th$ customer, we start from the recurrence relation for the vector $(\mathbf{W}_n, \boldsymbol{\tau}_n)$, $n = 1, 2, \ldots$:

$$(\mathbf{W}_{n+1}, \boldsymbol{\tau}_{n+1}) = (\max\{0, \mathbf{W}_n + \boldsymbol{\tau}_n - \boldsymbol{\sigma}_{n+1}\}, \boldsymbol{\tau}_{n+1}). \tag{3.13}$$

From (3.4) it follows that the workload of the server $\rho = \frac{\mathrm{E}\boldsymbol{\tau}}{\mathrm{E}\boldsymbol{\sigma}} = -\gamma v'(0) + \phi'(0)$. We assume that $\rho < 1$, in Appendix 3.A we show that this is a sufficient condition for the existence of a proper limiting distribution of $(\mathbf{W}_n, \boldsymbol{\tau}_n)$ and the sojourn time $\mathbf{R}_n$ for $n \to \infty$.
Let $\mathbf{W}, \boldsymbol{\tau}$ and $\mathbf{R}$ denote the random variables with the simultaneous limiting distributions of $\mathbf{W}_n, \boldsymbol{\tau}_n$ and $\mathbf{R}_n$ respectively, hence $\mathbf{R} \stackrel{d}{=} \mathbf{W} + \boldsymbol{\tau}$.
Define $F(x, y) := \Pr\{\mathbf{W} \leq x, \boldsymbol{\tau} \leq y\}$, and let $F^*(\omega_1, \omega_2)$, $Re\,\omega_1, Re\,\omega_2 \geq 0$ denote the LST of this joint distribution. Denote the LST of $\mathbf{R}$ by $r(\omega)$, $Re\,\omega \geq 0$.
Similarly to Section 2.2 we derive $F^*(\omega_1, \omega_2)$. We remark that the analysis is possible due to the exponential form of the LST of the dependent component of the service time (cf. (3.3)). We obtain for $Re\,\omega_1, Re\,\omega_2 \geq 0$

$$F^*(\omega_1, \omega_2) = \frac{\gamma v(\omega_2)}{\omega_1 - \gamma - \phi(\omega_2)} \left[\frac{\omega_1}{\gamma + \phi(\omega_2)} r(\gamma + \phi(\omega_2)) - r(\omega_1) \right], \tag{3.14}$$

and for $Re\,\omega \geq 0$,

$$r(\omega) = \frac{\gamma\ \omega\ v(\omega)}{(\omega + \gamma v(\omega) - \gamma - \phi(\omega))(\gamma + \phi(\omega))} r(\gamma + \phi(\omega)). \tag{3.15}$$

Next define for $Re\,\omega \geq 0$,

$$\begin{aligned} g(\omega) &:= \gamma + \phi(\omega), \quad g^{(0)}(\omega) := \omega, \\ g^{(k)}(\omega) &:= g(g^{(k-1)}(\omega)), \quad k = 1, 2, \ldots. \end{aligned} \tag{3.16}$$

With (3.15) and (3.16) we find after M iterations for $Re\,\omega \geq 0$, $M = 0, 1, \ldots$

$$r(\omega) = r(g^{(M+1)}(\omega)) \times \prod_{k=0}^{M} \frac{\gamma\ g^{(k)}(\omega)\ v(g^{(k)}(\omega))}{\left(g^{(k)}(\omega) + \gamma v(g^{(k)}(\omega)) - g^{(k+1)}(\omega)\right) g^{(k+1)}(\omega)}. \tag{3.17}$$

Lemma 3.3.1

(i). The equation $\omega = g(\omega)$, $Re\,\omega \geq 0$, has a unique real solution ω^*.

(ii). $\lim_{M\to\infty} g^{(M)}(\omega) = \omega^*$ for all $Re\,\omega \geq 0$.

(iii). The right-hand side of (3.17) converges for $M \to \infty$, for all $Re\,\omega \geq 0$.

Proof

The proof is analogous to the proof of Lemma 2.2.1 in Chapter 2 (cf. Appendix 2.B). The function $\phi(\omega)$ is completely monotone (cf. Theorem 3.2.1) and $\rho < 1$. This implies that $g(\cdot)$ is a contraction on $\{\omega \in \mathbb{C} | Re\,\omega \geq 0\}$ and also a contraction on $\{\omega \in \mathbb{R} | \omega \geq 0\}$. □

Lemma 3.3.1 leads to

Theorem 3.3.1

The LST of the sojourn time of a customer is

$$r(\omega) = r(\omega^*) \times \prod_{k=0}^{\infty} \frac{\gamma\ g^{(k)}(\omega)\ v(g^{(k)}(\omega))}{\left(g^{(k)}(\omega)+\gamma v(g^{(k)}(\omega))-g^{(k+1)}(\omega)\right) g^{(k+1)}(\omega)}, \quad (3.18)$$

with $Re\,\omega \geq 0$. □

Here $r(\omega^*)$ follows from $r(0) = 1$.

Substituting $r(\omega)$ in (3.14) and using (3.15), we obtain the LST of the joint distribution of the waiting and service time.

Theorem 3.3.2

$$\begin{aligned} F^*(\omega_1,\omega_2) &= \frac{\gamma\omega_1 v(\omega_2)}{\omega_1 - g(\omega_2)} \left[\frac{r(g(\omega_2))}{g(\omega_2)} - \frac{r(\omega_1)}{\omega_1}\right] \\ &= \frac{\omega_1}{\omega_1 - g(\omega_2)} \left[\omega_2 + \gamma v(\omega_2) - g(\omega_2)\right] \frac{r(\omega_2)}{\omega_2} \\ &\quad - \frac{\gamma v(\omega_2)}{\omega_1 - g(\omega_2)} r(\omega_1), \end{aligned} \quad (3.19)$$

$Re\,\omega_1, Re\,\omega_2 \geq 0$. □

The LST of the marginal distributions of the waiting time $\mathbf{W}$ and the service time $\boldsymbol{\tau}$ follow from (3.19). For the waiting time we obtain

Theorem 3.3.3

$$F^*(\omega,0) = \mathrm{E}\left(e^{-\omega\mathbf{W}}\right) = \frac{\omega r(\gamma) - \gamma r(\omega)}{\omega - \gamma}, \quad Re\,\omega \geq 0. \quad \square \quad (3.20)$$

Obviously (3.20) is of the same form as the waiting time distribution in the model of Chapter 2 (cf. Theorem 2.2.3), since the waiting time of a customer in both models only depends on the sojourn times of previous customers.

Remark 3.3.1
For this model it is easy to derive a number of results like those presented in Corollaries (2.2.1)-(2.2.5) of Chapter 2. We confine ourself to the following results:

$$r(\gamma) = \Pr\{\mathbf{W} = 0\} = 1 - \rho,$$

and

$$Cov(\mathbf{W}, \boldsymbol{\tau}) = -\frac{\phi'(0)}{\gamma}\left[\frac{1 - r(\gamma)}{\gamma} + r'(\gamma)\right] < 0.$$

□

In the following remark we reflect on the limits to further extending the dependence model. We also discuss the range of the iteration technique we applied in this and the previous chapter.

Remark 3.3.2
At the end of Section 3.2 we introduced an On/Off model with dependence between interarrival and service times of customers. In Cidon et al.[32] the analysis of a special case of this model is translated to a linear non-homogeneous functional equation, which allows a solution by means of iteration.
An appealing generalization of the On/Off model is the following. Let the arrivals of customers be generated at the moments of transitions in a finite state Markov process $\{\mathbf{J}(t), t \geq 0\}$; the time between transitions in the Markov process determines the interarrival time of customers. Next, let the duration of the service request of a customer depend on the time between transitions in the Markov process in the way described by dependence structure (3.3). We remark that both the sojourn time (i.e. the interarrival time) and the service time may depend on the state of the Markov process. To summarize, the Markov process $\{\mathbf{J}(t), t \geq 0\}$ remains in a state i for an exponentially(λ_i) distributed period of time, during this time work accumulates according to a process with independent increments characterized by a function $\phi_i(\cdot)$, and at the moment of a transition an independent service component with LST $v_i()$ is added to the accumulated work.
It is readily verified that this dependence structure contains the model analysed in this chapter as well as the On/Off model of [32].
Unfortunately, the analysis of the stationary distribution of the sojourn times of customers in this general model leads to a set of non-homogeneous linear functional equations for functions $r_i(\omega)$. Assuming ergodicity of the queueing process, the set of functional equations characterizes an ergodic process, hence theoretically this set must in some sense converge under iteration. Practically however, iteration does not appear to be a feasible method. For this there are two main reasons:

(i) the number of unknown terms in the set of functional equations grows exponentially with the number of iterations. To compare, in the right-hand side of expression (3.17) there is just the one unknown term $r(g^{M-1}(\omega))$.
(ii) the iteration of the multi-dimension analog of $r(g^{M-1}(\omega))$ leads to complex terms such as $r_i(\lambda_i + \phi_i(\lambda_j + \phi_j(\omega)))$; in the one-dimensional case in each step the same function is applied to the argument, in the multi-dimensional case this need not be the case. It is unclear how such terms converge to fixed points in the way $r(g^{M-1}(\omega))$ converges to ω^*.

It appears that to analyse a model by iteration, the model must not feature one of the above problems. But this severely restrains the generality of the dependence structure.
In Chapter 4 we present a method of modelling dependence between interarrival and service times which to some extent approximates the generalized On/Off model. □

Work decomposition
We conclude this chapter by making a comparison between single server queues with and without gates. In Chapter 2 the M/G/1 queue with dependence structure (3.1) is related to an ordinary M/G/1 queue *without* collection. Here, we look at the model from the more general perspective of single-server queues, or dam models (cf. Prabhu[83]), with arrival process a process with stationary non-negative independent increments, and we compare the model with an exponential gate to a single-server queue in which work directly arrives at the server queue, according to a process $\mathbf{Y}(u), u \geq 0$ with stationary non-negative independent increments. For dependence structure (3.3) the corresponding process would be the superposition of the dependent arrival process and a Compound Poisson Process with arrival rate γ and jump-size LST $v(\omega)$, the latter representing the independent part of the service time of a customer. This superposition again is a process with stationary non-negative independent increments and is characterized by

$$\mathrm{E}\left(e^{-\omega \mathbf{Y}(u)}\right) = e^{-\chi(\omega)u}, \quad u \geq 0, Re\,\omega \geq 0, \tag{3.21}$$

with $\chi(\omega) = \phi(\omega) + \gamma(1 - v(\omega))$.
For the single-server queue with the arrival process characterized by (3.21) the steady-state distribution of the amount of work $\mathbf{V}_{nogate}$ in the system exists if $\chi'(0) < 1$ and in that case its LST $\tilde{V}_{nogate}(\omega)$ is given by (cf. Prabhu[83] p.249)

$$\tilde{V}_{nogate}(\omega) = \frac{\omega(1 - \chi'(0))}{\omega - \chi(\omega)}, \quad Re\,\omega \geq 0. \tag{3.22}$$

From $r(\gamma) = 1 - \rho$ and expressions (3.4), (3.14) and (3.22) it follows that

$$r(\omega) = \mathrm{E}(e^{-\omega \mathbf{V}_{nogate}})\mathrm{E}(e^{-\omega \boldsymbol{\tau}})\frac{r(\gamma + \phi(\omega))}{r(\gamma)}, \quad Re\,\omega \geq 0. \tag{3.23}$$

The interpretation of the term $\dfrac{r(\gamma+\phi(\omega))}{r(\gamma)}$ is as follows.

In the model with the gating mechanism, denote by $\mathbf{H}$ the sojourn time of a customer leaving no customers behind at the server. Also denote by $\mathbf{U}$ the amount of work arriving at the gate during such a sojourn time. Then, for $Re\,\omega \geq 0$, similar to equation (2.41) in Chapter 2, we obtain

$$\mathrm{E}\left(e^{-\omega \mathbf{U}}\right) = \frac{\int_{t=0}^{\infty} e^{-\phi(\omega)t} e^{-\gamma t} dR(t)}{\int_{u=0}^{\infty} e^{-\gamma u} dR(u)} = \frac{r(\gamma+\phi(\omega))}{r(\gamma)}. \tag{3.24}$$

In fact

$$\mathbf{R} \stackrel{d}{=} \mathbf{V}_{nogate} + \boldsymbol{\tau} + \mathbf{U} \stackrel{d}{=} \mathbf{V}_{nogate} + \boldsymbol{\tau}_{dep} + \boldsymbol{\tau}_{indep} + \mathbf{U}, \tag{3.25}$$

$\stackrel{d}{=}$ denoting equality in distribution, $\boldsymbol{\tau}_{dep}$ having LST $\frac{\gamma}{\gamma+\phi(\omega)}$ and $\boldsymbol{\tau}_{indep}$ having LST $v(\omega)$. Moreover, the terms in (3.25) are independent.

This decomposition result and in particular the last property follow from the following probabilistic reasoning. Because of the PASTA property, $\mathbf{R} \stackrel{d}{=} \mathbf{V} + \boldsymbol{\tau}_{indep}$, with $\mathbf{V}$ the steady-state amount of work in the system (at the gate and at the server). Denote by $\boldsymbol{\eta}$ a stochastic variable with distribution the stationary distribution of the amount of work at the gate at times when the server is idle. According to the work decomposition property, cf. (2.44) in Chapter 2,

$$\mathbf{V} \stackrel{d}{=} \mathbf{V}_{nogate} + \boldsymbol{\eta}, \tag{3.26}$$

$\mathbf{V}_{nogate}$ and $\boldsymbol{\eta}$ being independent. The interpretation of the decomposition (3.26) is completed by observing that $\boldsymbol{\eta}$ consists of two independent components: (i) the amount of work $\mathbf{U}$ at the gate that has arrived during the sojourn time of the last customer before the server became idle, and (ii) the amount of work at the gate that has arrived during the past part of the idle period since the departure of the last customer. The latter term has the same distribution as $\boldsymbol{\tau}_{dep}$.

APPENDIX

3.A ERGODICITY OF THE WAITING AND SERVICE TIME PROCESS.

In this appendix we prove that the Markov process $\{(\mathbf{W}_n, \boldsymbol{\tau}_n), n = 0, 1, \ldots\}$ as defined in Section 3.3 is ergodic if $1 - \phi'(0) + \gamma v'(0) > 0$. We remark that $\mathbf{W}_n$ ($\boldsymbol{\tau}_n$) is the amount of work present at the server (gate) just prior to the $n-th$ opening of the gate. Moreover, $\boldsymbol{\tau}_n$ only contains the component of the service time that is dependent on the interarrival time, i.e. $\boldsymbol{\tau}_n$ is the amount of work that accumulated behind the gate since the previous opening of the gate.
Similar to Appendix 2.A of Chapter 2 the ergodicity is obtained by performing a three-step scheme of Laslett et al.[69] (cf. Appendix B). The aim is to identify a bounded positive recurrent subset of the state-space.

1. ϕ-irreducibility.
The state-space of $\{(\mathbf{W}_n, \boldsymbol{\tau}_n), n = 0, 1, \ldots\}$ is $[0, \infty) \times [0, \infty)$. Due to the fact that the work that accumulates behind the gate may arrive according to a continuous process (viz. linear growth) rather than to a jump-process (viz. arrivals of single customers) the state $(0, 0)$ in general is not an atom for any ϕ for which the Markov process is ϕ-irreducible. However, in queueing applications naturally arising test sets are those which are connected to the empty state of the system. It is clear that every subset of the form $A_{x_0,y_0} := [0, x_0) \times [0, y_0)$, $x_0, y_0 > 0$ can be entered from any state $(x, y) \in [0, \infty) \times [0, \infty)$ in at most two transitions. Hence, by choosing for ϕ an arbitrary measure on the $\sigma-$algebra of Borel subsets of the state-space with $\phi(A_{x_0,y_0}) > 0$, $x_0, y_0 > 0$, it follows that $\{(\mathbf{W}_n, \boldsymbol{\tau}_n), n = 0, 1, \ldots\}$ is ϕ-irreducible. To avoid confusion, we remark that we have adopted the terminology from Laslett et al.[69], so that the ϕ used in this step of the scheme is in no way connected to the function ϕ that characterizes the accumulation process of work at the gate (cf. (3.3)).

2. Identifying possible test sets.
In case ϕ contains an atom, the identification of a test set is trivial (cf. Appendix 2.A of Chapter 2). When atoms are absent, in queueing applications possible test sets still naturally arise; being those subsets of the state space that are connected to an empty system. In such cases, the verification that such a subset is indeed a test set (i.e. positive recurrence of this set indeed is sufficient for ergodicity of the Markov chain) is based on a continuity condition of the transitions (cf. theorem 4.1 of [69]). At first sight this continuity condition appears to be difficult to verify since the transition structure of the Markov chain leads to a multi-dimensional recurrence relation (cf. expression (3.13)). However, in section 4 of Laslett et al.[69] it is shown how this verification can be decomposed into a number of relatively elementary steps. The idea is that a complex looking random-walk relation can be decomposed as a sequence of simple transitions.
To conclude, analogously to section 11 of [69] it is readily verified that sets

$A_{x_0,y_0} = [0, x_0) \times [0, y_0)$, $x_0, y_0 > 0$ are test sets.

3. Applying Theorem B.1 of Appendix B.
We apply Theorem B.1 of Appendix B, using the function $g((x,y)) := x + y$. Define $\mathbf{X}_0 := (x_0, y_0) = (\mathbf{W}_1, \boldsymbol{\tau}_1)$ and $\mathbf{X}_1 := (x_1, y_1) = (\mathbf{W}_2, \boldsymbol{\tau}_2)$, and let $\boldsymbol{\sigma}_n$ be the time between the $n-th$ and $(n-1)th$ opening of the gate. Finally, let $\boldsymbol{\tau}_{ind}$ be a random variable with distribution the distribution of the independent component of a customer's service request. Hence the LST of $\boldsymbol{\tau}_{ind}$ is given by $v(\cdot)$ of (3.3). Under the condition $\boldsymbol{\sigma}_1 = u$ we find from recurrence relation (3.13)

$$g(\mathbf{X}_1) = \max\{x_0 + y_0 + \boldsymbol{\tau}_{ind} - u, 0\} + y_1. \tag{3.27}$$

In equation (3.13) the component y_1 depends on u, and has LST $e^{-\phi(\omega)u}$ (in the remainder of this appendix $\phi(\omega)$ is connected to the work accumulation process).
By conditioning on $\boldsymbol{\tau}_{ind}$ we derive

$$\mathrm{E}[g(\mathbf{X}_1)|\mathbf{X}_0] = (x_0 + y_0) - \frac{1}{\gamma}(1 - \phi'(0) + \gamma v'(0) - e^{-(x_0+y_0)\gamma} v(\gamma)). \tag{3.28}$$

From (3.28) if follows that $\mathrm{E}[g(\mathbf{X}_1)|\mathbf{X}_0] - g(\mathbf{X}_0) < 0$ if $1 - \phi'(0) + \gamma v'(0) > 0$ for $x_0 + y_0$ large enough. This completes the scheme of [69]. □

Chapter 4

Modelling dependence with Markov modulated arrival processes

4.1 Introduction

In the previous chapter we have extended the dependence structure of Chapter 2 to a model where the service time distribution of a customer (the collected work) consists of two components: a component that depends on the time between two collectings, the amount of work collected being characterized by a process with non-negative independent increments, and a component that is independent of the collecting interval, an 'ordinary' $M/G/1$ service request. However, the interarrival times of batches still are exponentially distributed. As pointed out in Remark 2.1.1 in Chapter 2, in some applications the collecting procedure might be closer to a deterministic process than to a Poisson process. Moreover, also the dependent component of the work accumulation process at the bus-stop, or gate, might not possess the property of independent increments.

In this chapter we discuss a second generalization of the collector model from Chapter 2. In Chapter 3 we approached the collector model from the viewpoint of queues with a gate, here we consider the arrivals of batches from the perspective of Markov processes. The approaches in Chapter 2 and 3 are of a theoretical nature; exact analysis leading to explicit expressions which give structural insights into the effect of correlated interarrival and service times on the behaviour of the queueing system. In this chapter we present a more pragmatic approach to queueing models with a dependence between interarrival and service times, but in which the (batch) arrival process and work accumulation process deviate from respectively the Poisson process or the process with non-negative independent increments.

In the generalization we use Markov Modulated Queueing Systems (MMQS) (cf. Chapter 7) for modelling more general customer arrival and collecting processes. In particular we apply the framework of the Batch Markovian Arrival Process (BMAP, cf. Appendix A). The main reasons for using the BMAP are that the $BMAP/G/1$ queue, the single-server queue with the BMAP as arrival process, allows a detailed exact analysis and that there exist convenient numerical procedures for most performance measures of interest. Moreover, the BMAP framework allows a constructional approach to modelling dependence between interarrival and service times. In a general MMQS both the interarrival and service time distribution may depend on the state of the directing Markov process. However, it appears that an observed or assumed correlation structure between interarrival and service times is more convenient to model with the BMAP than via other MMQS.

Overview of the chapter

In Section 4.2 we describe how an MMQS can be used to model queueing systems with correlation between the interarrival and service time of a customer. Section 4.3 illustrates the modelling technique with a number of examples. Section 4.4 numerically shows the potential of the modelling technique to obtain performance measures and insights for queueing systems with dependence between interarrival and service times. In Section 4.5 we consider queueing models in which various dependence structures occur simultaneously.

4.2 Modelling dependence with the BMAP

In this section we concentrate on modelling correlation between interarrival and service times of customers with the use of the Batch Markovian Arrival Process (BMAP). First we describe the characteristics of the BMAP. A detailed overview of the BMAP and results on the $BMAP/G/1$ queue can be found in Appendix A.

The Batch Markovian Arrival Process

In the BMAP, the interarrival and service times of customers are directed by a continuous time Markov process $\{\mathbf{J}(t), t \geq 0\}$ on a finite state space E. The essential mechanism in the BMAP is that, conditional on a transition from a state i to a state j, a batch arrival is generated, the size of the batch being k with probability q_{ij}^k, $k = 0, 1, \ldots, \sum_k q_{ij}^k = 1$ where q_{ij}^0 may be interpreted as the probability of having no arrival or as the probability of the arrival of an empty batch. With λ_{ij} being the transition rate from state i to j, $i \neq j$, and $\lambda_{ii} := -\lambda_i$ the parameter of sojourn time in state i the BMAP is completely characterized by the sequence of matrices $D_k = (\lambda_{ij} q_{ij}^k)$, $k = 0, 1, \ldots, \sum_k D_k = D$.

The $BMAP/G/1$ queue is defined as the single-server queue with the BMAP describing the arrival process of batch customers, the service times of single customers being independent and identically distributed with a general distribution function $H(\cdot)$. The $BMAP/G/1$ has been studied in Lucantoni[73], in which one can find results for most of the performance measures of interest,

such as the number of customers, waiting times, and the busy period.

In important observation is that a batch arrival can be viewed as the arrival of a group of individual customers, or as the arrival of a (super) customer whose service time is distributed as the sum of the service requests of the customers contained in the batch. In this paper we will consider the BMAP from the latter viewpoint.

Modelling correlation between interarrival and service times
A key observation of the chapter is that dependence between interarrival and service times can be modelled by viewing $\{\mathbf{J}(t), t \geq 0\}$ as a two-dimensional Markov process, $\{(\mathbf{J}_1(t), \mathbf{J}_2(t)), t \geq 0\}$. $\mathbf{J}_1$ generates the arrivals of batches, and the state of $\mathbf{J}_1$ contains information about the remaining interarrival time; if the state of $\mathbf{J}_1$ is j_1, this might for example imply that the remaining interarrival time consists of j_1 exponential phases. The component $\mathbf{J}_2$ contains information about the batch size distribution; for example, the state of $\mathbf{J}_2$ might stand for the number of customers in a batch (here and in the remainder of the chapter we use the convenient notation $\mathbf{J}_i$ for $\{\mathbf{J}_i(t), t \geq 0\}$, i=1,2).
The key idea is to let the interarrival time, the time in $\mathbf{J}_1$ between two batch generating epochs, be the time parameter in $\mathbf{J}_2$. So, as the time between two arrivals goes by, the batch size distribution is described by the evolution of $\mathbf{J}_2$. The state of $\mathbf{J}_2$ just before a batch arrival contains information about the batch size.
A typical example of this mechanism is the collector model of Chapter 2. As described in Section 3.2, in particular in example 1, the dependent component is characterized by the Compound Poisson Process (CPP); the interarrival time of a batch is exponentially distributed, the size of a batch is the number of customers generated by a second Poisson process during that interval. In this example, $\mathbf{J}_1$ is a one state Markov process, describing the exponentially distributed interarrival times of batch customers, $\mathbf{J}_2$ is a Markov chain with state space $\{0, 1, \ldots\}$, describing the current size of the batch, i.e., the number of customers that have arrived in the second Poisson process since the last batch arrival. Again we remark that we treat the dependence structure from the angle of batch customers whose service times are the sum of the service requests associated to the number of single customers constituting the batch customer.

4.3 Exploring the model

In this section we use the BMAP for modelling a number of variants (and generalizations) of the collector model.

• *Variant 1: Finite bus capacity*
In the original model the number of customers in the bus can be arbitrarily large. However, the method used in Section 2.2 to analyse this case does not seem to be applicable for the natural variant where the number of customers in

the bus can not exceed M. With the BMAP, there are two ways of modelling this restriction.

-First approach. If the number of customers at the bus-stop has reached M, all future arrivals are rejected until the bus has collected the M customers at the stop. This is an example of the most pure form of the construction; the Markov process $\mathbf{J}_1$ only describes the phase of the interarrival process, the Markov process $\mathbf{J}_2$ only describes the number of customers in the bus, and the interaction of the two chains is limited to the resetting of $\mathbf{J}_2$ to the state corresponding to an empty bus-stop, at the moment at which $\mathbf{J}_1$ generates a batch arrival. The Markov chain of $\mathbf{J}_1$ has a single state because the interarrival time is exponentially distributed, the Markov chain of $\mathbf{J}_2$ has state space $E = \{0, 1, \ldots, M\}$, each state representing the number of customers at the bus-stop. In this second Markov chain M is an absorbing state.
Note that the arrival process of the busses is still a Poisson process.
In Figure 4.1 the transition rate diagram for $(\mathbf{J}_1, \mathbf{J}_2)$ is presented for the case $M = 3$. Figure 4.2 shows the generator matrix D and the matrices $D_0, \ldots, D_3$ it decomposes into. In Figure 4.1, the nodes are numbered $0, \ldots, 3$, representing the states of $\mathbf{J}_2$.

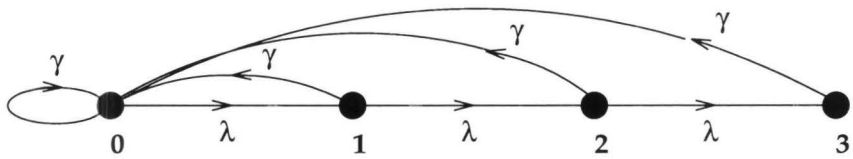

FIGURE 4.1: TRANSITION RATE DIAGRAM FOR THE COLLECTOR MODEL WITH FINITE BUS CAPACITY AND BLOCKING OF CUSTOMERS.

$$D = \begin{bmatrix} -\lambda & \lambda & 0 & 0 \\ \gamma & -(\lambda+\gamma) & \lambda & 0 \\ \gamma & 0 & -(\lambda+\gamma) & \lambda \\ \gamma & 0 & 0 & -\gamma \end{bmatrix}, D_0 = \begin{bmatrix} -\lambda & \lambda & 0 & 0 \\ 0 & -(\lambda+\gamma) & \lambda & 0 \\ 0 & 0 & -(\lambda+\gamma) & \lambda \\ 0 & 0 & 0 & -\gamma \end{bmatrix},$$

$$D_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ \gamma & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, D_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \gamma & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, D_3 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \gamma & 0 & 0 & 0 \end{bmatrix}.$$

FIGURE 4.2: GENERATOR MATRICES FOR THE COLLECTOR MODEL WITH FINITE BUS CAPACITY AND BLOCKING OF CUSTOMERS.

For the $BMAP/G/1$ queue with such generator matrices and a general distribution function $H(\cdot)$ for the service time of a single customer, analytical results and numerical evaluation procedures are presented in Lucantoni[73] for many performance measures of interest, such as the waiting time, the queue length and the busy period. For $M \to \infty$ the $BMAP/G/1$ queue reduces to the model

studied in Chapter 2. As a particular result of this, we are able to approximate higher moments of the busy period in the latter model as accurately as we wish. In Section 2.5 we were only able to obtain the first moment of the busy period. This limited result was due to the fact that basic characteristics for the busy period in the $M/G/1$ queue such as the branching argument and the semi-group property do not immediately apply for the collector model. There we argued that also some information about the state of the process describing the number of customers at the bus-stop should be taken into account. With the BMAP formulation the latter has been done and indeed it is possible (cf. Appendix A) to analyse the busy period in greater detail.

-Second approach. Here we let $\mathbf{J}_1$ generate batch arrivals, but we also have the bus visit the bus-stop when the number of customers at the bus-stop reaches M. Hence every time $\mathbf{J}_2$ is in state $M-1$, the next transition, bus or customer, generates a batch arrival. For this model, for the case $M=3$, Figures 4.3 and 4.4 show the transition rate diagram for $(\mathbf{J}_1, \mathbf{J}_2)$ and the matrices $D_0, D_1, \ldots, D_M$ respectively.

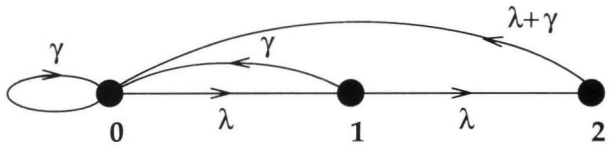

FIGURE 4.3: TRANSITION RATE DIAGRAM FOR THE COLLECTOR MODEL WITH FINITE BUS CAPACITY AND NO BLOCKING OF CUSTOMERS.

$$D = \begin{bmatrix} -\lambda & \lambda & 0 \\ \gamma & -(\lambda+\gamma) & \lambda \\ \lambda+\gamma & 0 & -(\lambda+\gamma) \end{bmatrix}, D_0 = \begin{bmatrix} -\lambda & \lambda & 0 \\ 0 & -(\lambda+\gamma) & \lambda \\ 0 & 0 & -(\lambda+\gamma) \end{bmatrix},$$

$$D_1 = \begin{bmatrix} 0 & 0 & 0 \\ \gamma & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, D_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \gamma & 0 & 0 \end{bmatrix}, D_3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \lambda & 0 & 0 \end{bmatrix}.$$

FIGURE 4.4: GENERATOR MATRICES FOR THE COLLECTOR MODEL WITH FINITE BUS CAPACITY AND NO BLOCKING OF CUSTOMERS.

Remark 4.3.1
In the first example, the batch size of an arrival is equal to the state of $\mathbf{J}_2$ at the moment of an arrival. In the second example, a batch arrival may be caused by the arrival of a bus or by the arrival of the $M-th$ customer at the bus-stop. Hence, when a batch arrival occurs while $\mathbf{J}_2$ is in state $M-1$, the batch size is with probability $\frac{\gamma}{\lambda+\gamma}$ equal to $M-1$, and with probability $\frac{\lambda}{\lambda+\gamma}$ equal to M. In general, the state of $\mathbf{J}_2$ may describe a batch size *distribution*, rather than just being the number of customers in a batch. □

• *Variant 2: General arrival processes*
In the original collector model the arrival process of customers at the bus-stop and the arrival process of busses are both Poisson. For more general arrival processes it seems hard to extend the approach of Chapter 2 to obtain analytical results. However, when the interarrival times of busses and customers are of semi-Markov type, the BMAP modelling can provide approximations for the case of infinite bus capacity, and exact results for the case of finite bus capacity. Here we remark that most results on the $BMAP/G/1$ theoretically seem to extend for the case of a countable underlying Markov chain. However, numerical methods available are mainly developed for the case of a finite underlying Markov chain.
As an example of more general arrival processes we present the case in which the interarrival times of busses are Erlang-2 distributed, the customers arrive according to a Poisson process at the bus-stop, and customers are rejected once the bus is full. Figure 4.5 shows the corresponding transition rate diagram for the case $M = 3$. The nodes $(j_1, j_2) \in \{0,1\} \times \{0,1,2,3\}$ in Figure 4.5 represent the state of $(\mathbf{J}_1, \mathbf{J}_2)$, $\mathbf{J}_1$ being the state of the bus interarrival process, $\mathbf{J}_2$ representing the number of customers at the bus-stop.

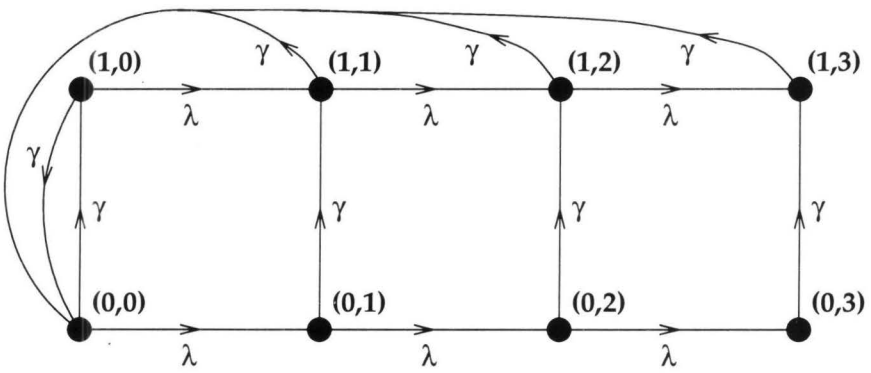

FIGURE 4.5: TRANSITION RATE DIAGRAM FOR THE COLLECTOR MODEL WITH ERLANG-2 DISTRIBUTED INTERARRIVAL TIMES OF BATCHES.

• *Variant 3: Other types of correlation*
The above mentioned variants of the model with customer collection all consider a positive correlation between the interarrival and service time of a batch customer. More specific, the batch size distribution is stochastically non-decreasing as a function of the interarrival time. In particular, the Markov chain in the first example (cf. Figure 4.1) is related to a birth process. An obvious extension is to construct a negative correlation between interarrival and service times, for example by letting the batch size distribution be stochastically non-increasing as a function of the interarrival time.

A second extension is to model customer behaviour at the bus-stop; customers may for example leave the bus-stop after a while when the bus is 'late'. This queueing model with impatient customers can be adequately modeled with Markov processes $\mathbf{J}_1$ and $\mathbf{J}_2$.

Remark 4.3.2
In the generalized dependence structure that is considered in Chapter 3 the service time of a customer consists of two components: a component that depends on the interarrival time and a component that is independent of the interarrival time. We remark that the analysis of the virtual waiting time process of the $BMAP/G/1$ queue (cf. (A.6) of Appendix A) can be extended to the case where a batch arrival consists of a batch of service requests plus an additional service component. The batch size is defined in the standard way by the matrices D_i and a distribution function $H(\cdot)$ (with LST $h(\cdot)$) for a single service request. The additional service component that is added to each batch has distribution function $\hat{H}(\cdot)$ with LST $\hat{h}(\cdot)$. This arrival process leads to an expression for the LST of the virtual waiting time that is similar to expression (A.6). Without going into details we remark that the boundary state probabilities that have to be determined follow from exploiting the analytical properties of LST's of proper distribution functions (for similar derivations of boundary state probabilities we refer to Chapter 7).
However, this extended BMAP process allows us to generalize the dependence structure of Chapter 3 for the case where the dependent component is a compound Poisson process (cf. Section 3.2). Moreover, it follows from Remark 3.2.2 (iii) that also general processes with non-negative independent increments can be approximated by the above-introduced extended $BMAP$ arrival process. □

4.4 Numerical results

In this section we numerically illustrate the possibilities of our model. We show how collector models with unbounded batch capacity might adequately be approximated by a $BMAP/G/1$ queue. We pay attention to the effect of the maximum batch size and we also investigate the influence of the distribution of the intercollection interval on performance measures of the queueing model. The performance measures considered are the waiting time and the busy period length.
We have restricted ourself to a few examples, because numerical exploration of $BMAP/G/1$ procedures is not the main purpose of this chapter. Lucantoni[73] presents for the $BMAP/G/1$ queue many results and numerical procedures.

Approximating infinite bus capacity
While discussing variant 2 we remarked that most numerical methods for the $BMAP/G/1$ queue are mainly for the case of a finite underlying Markov chain. In Tables 4.1 and 4.2 we examine various performance measures as a function of the dimension of this Markov chain for the case of Poisson arrivals of customers and exponentially distributed intercollection intervals (variant 1). For $M \to \infty$

exact results are adopted from Table 2.1 of Section 2.6. Unforced collect and forced collect respectively are the first and second approach of variant 1. In the example, the arrival rate λ of customers at the bus stop is 1, the mean service time of a customer is 0.5, and the intercollecting distribution has parameter 1. For this collector model we distinguish between two types of busy periods: busy periods **B** that also include busy periods of length 0, i.e. busy periods started by empty busses, and busy periods **B'**, for which only strictly positive busy periods are taken into account.

	unforced collect		forced collect	
M	**EW**	**EB'**	**EW**	**EB'**
5	0.594280	1.507388	0.630328	1.527809
6	0.639372	1.554330	0.657011	1.562283
7	0.666197	1.578507	0.674647	1.581514
8	0.681677	1.590791	0.685685	1.591904
9	0.690428	1.596987	0.692327	1.597393
10	0.695301	1.600099	0.696207	1.600245
20	0.701151	1.603123	0.701182	1.603130
∞	0.701155	1.603122	0.701155	1.603122

TABLE 4.1: MEAN WAITING TIMES AND MEAN BUSY PERIOD LENGTHS AS A FUNCTION OF THE BUS CAPACITY M. EXPONENTIAL SERVICE TIMES.

	unforced collect		forced collect	
M	**EW**	**EB'**	**EW**	**EB'**
5	0.398027	1.454060	0.417376	1.441434
6	0.436232	1.490878	0.445759	1.486589
7	0.459451	1.511221	0.464061	1.509814
8	0.473076	1.522049	0.475293	1.521601
9	0.480881	1.527681	0.481952	1.527540
10	0.485275	1.530565	0.485796	1.530522
20	0.490637	1.533424	0.701182	1.533416
∞	0.490648	1.533418	0.490648	1.533418

TABLE 4.2: MEAN WAITING TIMES AND MEAN BUSY PERIOD LENGTHS AS A FUNCTION OF THE BUS CAPACITY M. DETERMINISTIC SERVICE TIMES.

We notice that for moderate M the values of the mean waiting times **EW** and mean busy period length **EB'** are already reasonably close to these values for $M = \infty$. The difference for M small is explained by the fact that for unforced collect not all customers arriving at the bus-stop will receive service; when $\mathbf{J}_2 = M$, newly arriving customers are rejected. In connection with Remark 4.3.1 we note that this might be avoided by having a batch size *distribution* rather than fixing the batch size to M for $\mathbf{J}_2 = M$.

The distribution of the collecting interval

Figure 4.6 illustrates variant 2. Mean busy period lengths and mean waiting times are presented as functions of the coefficient of variation of Erlang distributed intercollection times. The mean collecting interval is 1, the number of phases in the Erlang intercollection distribution ranges from 1 to 7, and we also examined $\mathbf{EW}$, $\mathbf{EB}$, and $\mathbf{EB}'$ for a deterministically distributed collecting interval. For the last two instances we performed a simulation experiment. Customers arrive according to a Poisson process with rate 1, the mean service time of a customer is 0.5.

Figures 4.6a and 4.6 illustrate the possibility of the BMAP framework to obtain insight in queueing systems with dependence between interarrival and service times, in particular when this dependence is the result of a collection/reservation mechanism. The first observation in Figures 4.6a and 4.6b is that both mean waiting times and busy period lengths are decreasing as the intercollection interval becomes 'more deterministic'. Secondly, the performance measures are almost linear functions of the coefficient of variation of the intercollection interval distribution. Finally, not shown here, we observed that the coefficient of variation of the busy period remains almost constant.

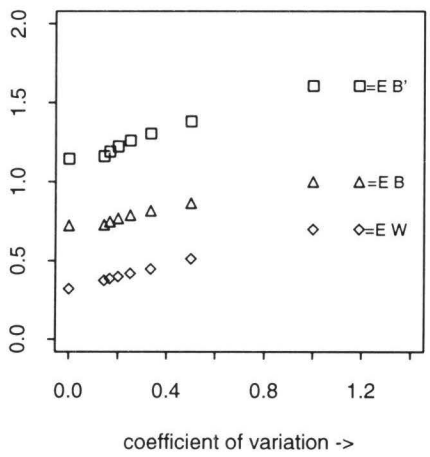

FIGURE 4.6A: THE EFFECT OF THE INTERCOLLECTION INTERVAL DISTRIBUTION ON MEAN WAITING TIMES AND MEAN BUSY PERIOD LENGTHS. EXPONENTIAL SERVICE TIMES.

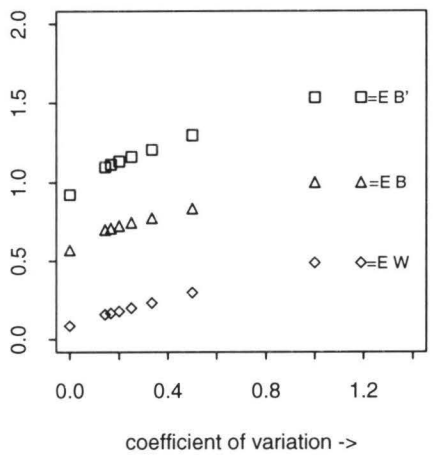

FIGURE 4.6B: THE EFFECT OF THE INTERCOLLECTION INTERVAL DISTRIBUTION ON MEAN WAITING TIMES AND MEAN BUSY PERIOD LENGTHS. DETERMINISTIC SERVICE TIMES.

4.5 INTEGRATING VARIOUS DEPENDENCE STRUCTURES

In general one considers three types of dependence in queueing systems: between consecutive interarrival times, between consecutive service times, and between the interarrival and service times of customers. Fendick et al.[47] state that in packet queues with multiple classes of traffic and variable packet lengths all three types of dependence occur *simultaneously*. The BMAP allows us to model this phenomenon. To illustrate this, we present an example in which the arrival process of busses is a two-state Markov Modulated Poisson Process (MMPP), the batch size corresponds to a collecting procedure, and consecutive batch sizes are positively correlated. The example features all three types of dependence.

The dependence between two consecutive batch sizes can be modeled in several ways; for example by letting the arrival process of customers at the bus-stop be a Markov Modulated Arrival Process, or by assuming that when the number of customers at the bus-stop equals the maximum capacity M of the bus, newly arriving customers will wait for the next bus. The latter can be modeled as follows; at the moments $\mathbf{J}_1$ generates a bus arrival, also allow transitions in $\mathbf{J}_2$ to states other than 0 (this state representing the empty batch).

For our example we choose to model the dependency of consecutive service times by a two-state MMPP. The two-state MMPP's for the arrivals of busses and customers are as follows: while the underlying Markov process is in state

i, the arrival rate for the arrival process of busses (of customers at the bus-stop) is γ_i (λ_i), and the sojourn time in state i is exponentially distributed with parameter η_i (ν_i), $i = 1, 2$. For this model it is convenient to describe the evolution of the batch size distribution by a two-dimensional Markov chain $\mathbf{J}_2 = (\mathbf{J}_2^1, \mathbf{J}_2^2)$, $\mathbf{J}_2^1$ representing the state of the MMPP for the individual customers, and the state of $\mathbf{J}_2^2$ representing the number of customers at the bus-stop.
In Figure 4.7 the transition rate diagram is presented for the case $M = 3$. We remark that higher values of M only affect component $\mathbf{J}_2$, so that Figure 4.7 will only be extended along the axis of $\mathbf{J}_2$.

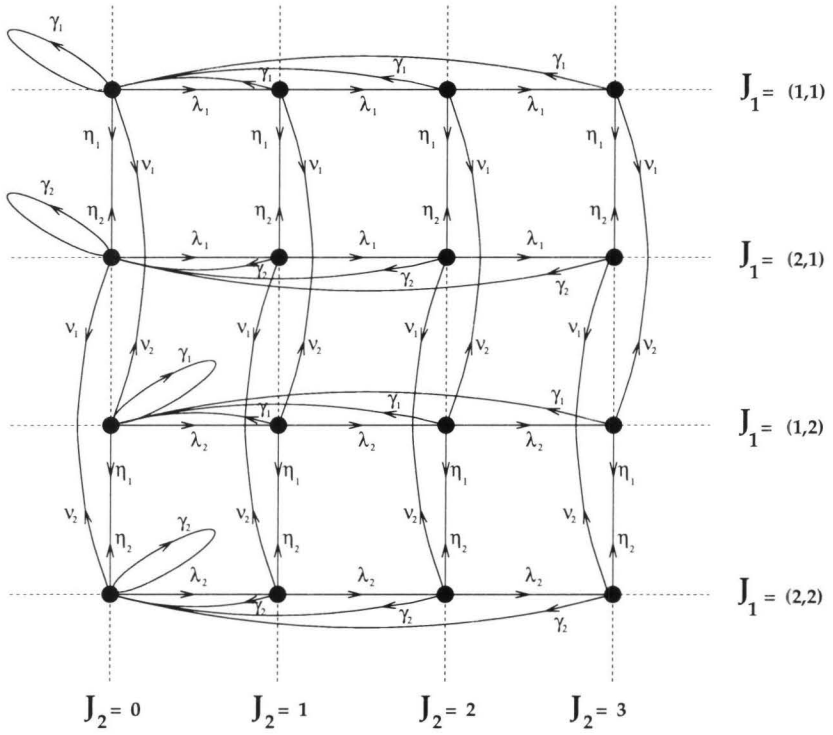

FIGURE 4.7: TRANSITION RATE DIAGRAM FOR THE COLLECTOR MODEL. BUSSES AND CUSTOMERS ARRIVE ACCORDING TO TWO-STATE MARKOV MODULATED POISSON PROCESSES.

To clarify the example we discuss the transition structure in Figure 4.7 for the state $(\mathbf{J}_1, \mathbf{J}_2) = ((1,2),2)$. For simplicity assume that the total transition rate in $((1,2),2)$ is equal to 1. Then we can write down the following list of possible events with their respective probabilities and the states they lead to.

$\lambda_2 \rightarrow ((1,2),3)$ probability of the arrival of a customer at the bus-stop.

$\gamma_1 \rightarrow ((1,2),0)$ probability of the arrival of a bus.

$\nu_2 \rightarrow ((1,1),2)$ probability of a change in the arrival rate of customers at the bus-stop (from λ_2 to λ_1).

$\eta_1 \rightarrow ((2,2),2)$ probability of a change in the bus arrival rate (from γ_1 to γ_2).

Chapter 5

Impatient customers in the $MAP/G/1$ queue

This chapter discusses a single-server queueing model with customer impatience. We extend results on the $M/G/1$ queue with impatient customers to the case of customers arriving according to a Markovian Arrival Process. We derive the Laplace transform for the virtual waiting time, from which we obtain expressions for customer waiting time and the probability of a premature departure of a customer.

5.1 Introduction

In queueing models of communication systems various forms of interaction between the arrival process of customers and the workload of the system may occur. An example of such interaction occurs in a queueing network with limited buffer space and admission control; a customer might be rejected when his service time would cause the overflow of a buffer.

In this chapter we consider a second form of interaction, so-called customer impatience. In a queueing model with customer impatience each customer is prepared to wait a limited period of time in the queue, and when the actual waiting time is larger, the customer's patience runs out and he leaves the system. Such a situation might for example occur in a telephone system, where customers disconnect if the waiting time is too large.

Boxma & de Waal[26] give two examples of customer impatience that are directly connected to communication networks:

(i) Real-time communication (voice, video) in which the content of a message loses its importance after a certain amount of time.

(ii) Data communication networks with a time-out protocol.

A key reference to single-server queues with customer impatience is Baccelli et al.[10], which considers $GI/G/1$ queueing systems where customers at the moment of arrival decide whether to join the queue or not, their decision depending on the amount of work present at that time. There are a number of variants of such an impatience structure; a customer might immediately leave the system without joining the queue, or the decision to leave system without receiving service might be based on the sojourn time instead of the waiting time.

In this chapter we study a single-server system with customer impatience for the case of a Markovian Arrival Process (MAP) and exponentially(α) distributed patience. The latter means that when a customer enters the system and the total remaining work in the system is equal to x, the customer will receive service with probability $e^{-\alpha x}$, and will leave the queue without receiving service with probability $1 - e^{-\alpha x}$. The (FCFS) queueing model under consideration will be denoted by $MAP/G/1 + M$.

The motivation for the analysis of this queueing system is twofold. Firstly, it extends the results of Baccelli et al.[10] on the $M/G/1 + M$ queue to the case of the more general MAP as arrival process. The MAP allows a more accurate description of the arrival processes in modern queueing systems than the Poisson process. For example the Markov Modulated Poisson Process (MMPP), frequently used to model On/Off sources in a communication network, fits into the framework of the MAP. In particular for examples (i) and (ii) a MAP might enable a better description of the actual arrival process than the Poisson process.

Secondly, any G(I) arrival process can be approximated arbitrarily close by a MAP, hence the analysis of the $MAP/G/1 + M$ queue can be used to provide (approximate) results for the more general $G(I)/G/1 + M$ queue. Note that no explicit expressions are known for the latter model, cf. Baccelli et al.[10], as their efforts on the $G(I)/G/1 + M$ did not result in explicit expressions for the waiting times of customers.

Overview of the chapter

In Section 5.2, we first describe the $MAP/G/1 + M$ queue in more detail, then present an analysis of the workload process (also referred to as the virtual waiting time process), resulting in the Laplace-Stieltjes Transform (LST) of its stationary distribution. In Section 5.3 we consider the moments of the workload process, and we also present the LST of the waiting time and the probability that a customer prematurely leaves the queue. In Section 5.4 we discuss numerical aspects of the $MAP/G/1 + M$ queue and we present some numerical results for the $MMPP/M/1 + M$ queue.

5.2 THE WORKLOAD PROCESS

In this section we present an analysis of the workload process in the $MAP/G/1+M$ queue. First we describe the $MAP/G/1+M$ queue in more detail.

Mathematical description of the $MAP/G/1+M$ queue

We consider a single-server queue in which the arrival epochs of customers are directed by a continuous time Markov process, and in which for a customer at the moment of his arrival it is decided whether this customer will eventually receive service or not, this decision being based on the amount of work present in the system at that moment, i.e. on the actual waiting time of the customer. We remark that because the server will not serve customers that leave the system prematurely, their service times do not add to the waiting times of other customers.

Let us first consider the arrival process. The MAP is a particular case of the BMAP that was used in Chapter 4. Below we define the characteristics of the MAP that is used in this chapter; details of the BMAP and the $BMAP/G/1$ queue can be found in Appendix A.

The interarrival times of customers are directed by a continuous time Markov process $\{\mathbf{J}(t), t \geq 0\}$ on a finite state space E. A transition from a state i to a state j may induce the arrival of a customer. Let the sojourn time in state $i \in E$ be exponentially distributed with parameter $\lambda_i > 0$, and denote by λ_{ij} the rate of transition from state i to j, $i \neq j$, with $\lambda_{ii} := -\lambda_i$ (the probability of a transition from i to j is $p_{ij} = \frac{\lambda_{ij}}{\lambda_i}$). The generator of the Markov process $\{\mathbf{J}(t), t \geq 0\}$ is given by the matrix $D = (\lambda_{ij})$. Next, conditional on a transition from a state i to a state j, the probability of an arrival is denoted by q_{ij}. With this notation, $q_{ij}\lambda_{ij}$ is the transition rate from state i to state j inducing an arrival. Defining $q_{ii} := 0$, then the MAP is completely characterized by the matrices $D_0 := (\lambda_i p_{ij}(1 - q_{ij}))$ and $D_1 := (\lambda_i p_{ij} q_{ij})$. The stationary probability (row) vector of $\{\mathbf{J}(t), t \geq 0\}$, is denoted by π and satisfies $\pi D = 0$ and $\pi e = 1$, with e the $|E|$ dimensional unit (column) vector.

Next we consider the impatience structure. If at the moment of an arrival of a customer the amount of work in the system is x, $(x \geq 0)$, then this customer will eventually receive service with probability $e^{-\alpha x}$, and will eventually leave the queue without receiving service with probability $1 - e^{-\alpha x}$. Hence the time before a customer runs out of patience and leaves the queue without receiving service has an exponential distribution with parameter α.

The description of the $MAP/G/1+M$ queue is completed with a distribution function $B_s(\cdot)$ for the service time of a customer. The LST of $B_s(\cdot)$ is denoted by $\beta(\cdot)$.

We remark that for $\alpha \to \infty$ the model reduces to a $MAP/G/1/1$ loss system, and with $\alpha \to 0$ we return to the standard $MAP/G/1$ queue.

The workload process

The workload (virtual waiting time) of the system at time t is defined as the

total remaining work of the customer presently in service plus the sum of the service times of the customers in the queue who will receive service. So, the service times of customers who will leave the system prematurely do not add to the amount of work in the system. Denote by $\{\mathbf{V}(t), t \geq 0\}$ the process describing the workload of the system, and let $\{\mathbf{J}(t), t \geq 0\}$ describe the state of the underlying Markov chain of the MAP with state space E. The workload process is defined as the Markov process $\{(\mathbf{V}(t), \mathbf{J}(t)), t \geq 0\}$, which has state space $[0, \infty) \times E$.
Let $\{\mathbf{V}, \mathbf{J}\}$ be the random variable with distribution the stationary distribution of $\{(\mathbf{V}(t), \mathbf{J}(t)), t \geq 0\}$. In Appendix 5.A we show that a unique stationary distribution exists if $\alpha > 0$.
We start our analysis of the process $\{(\mathbf{V}(t), \mathbf{J}(t)), t \geq 0\}$ by writing down a vector equivalent of the Takács integro-differential equation.
First we describe the evolution of the workload process during a small period of time Δ.
Defining for $t \geq 0, x \geq 0, j \in E$: $V_j(t,x) := \Pr\{\mathbf{V}(t) \leq x, \mathbf{J}(t) = j\}$, then for $t > 0$, $x > 0$, and $\Delta > 0$ small,

$$V_j(t+\Delta, x) = (1 - \lambda_j \Delta) V_j(t, x+\Delta) + \sum_{i \in E, i \neq j} [\lambda_{ij}(1 - q_{ij})\Delta] V_i(t, x+\Delta) + \sum_{i \in E} \lambda_{ij} q_{ij} \Delta \int_{y=0}^{x} (e^{-\alpha y} B_s(x-y) + 1 - e^{-\alpha y}) dV_i(t,y) + o(\Delta). \quad (5.1)$$

To avoid the technical details that concern the continuity and differentiability of $V_j(t,x)$ we directly turn to the LST of the amount of work. This LST is defined for $t > 0$ and $Re\,\omega \geq 0$ as

$$\Phi_j(t,\omega) := \mathrm{E}\{e^{-\omega \mathbf{V}(t)} I_{\{\mathbf{J}(t)=j\}}\} = \int_{x=0}^{\infty} e^{-\omega x} dV_j(t,x).$$

The technical concerns are mainly caused by the distribution function of the service time, which might not be absolutely continuous, e.g. when the service time is deterministic. For details we refer to Hasofer[57]. However, in an analogous way to the derivation of similar equations by Loynes[72] and Takács[94, p.52-53], we obtain from (5.1)

$$\frac{\partial}{\partial t}\Phi_j(t,\omega) = \omega \Phi_j(t,\omega) - \omega V_j(t,0) - \lambda_j \Phi_j(t,\omega) + \sum_{i \in E, i \neq j} \lambda_{ij} \Phi_i(t,\omega) - \sum_{i \in E} \lambda_{ij} q_{ij} (1 - \beta(\omega)) \Phi_i(t, \omega + \alpha), \quad t > 0, Re\,\omega \geq 0. \quad (5.2)$$

Next we define the LST related to the process $\{\mathbf{V}, \mathbf{J}\}$:
$\Phi_j(\omega) := \int_{x=0}^{\infty} e^{-\omega x} dV_j(x)$, $Re\,\omega \geq 0$,
with $V_j(x) := \lim t \to \infty V_j(t,x)$, $x \geq 0, j \in E$.
Letting $t \to \infty$, and using the (row)vector notation $\Phi(\omega) = (\Phi_j(\omega))$ together with the (row)vector $V(0)$ of probabilities $\lim_{t\to\infty} V_j(t,0)$, we derive from (5.2)

$$\Phi(\omega)[\omega I + D] = \omega V(0) + (1 - \beta(\omega))\Phi(\omega + \alpha)D_1, \quad Re\,\omega \geq 0. \tag{5.3}$$

In (5.3), I is the identity-matrix of dimension $|E|$, $D = D_0 + D_1$ is the generator of $\{\mathbf{J}(t), t \geq 0\}$.

To assure that (5.3) has a unique solution and also to avoid some technical difficulties in the analysis we make the assumption that the Markov chain connected to $\{\mathbf{J}(t), t \geq 0\}$ is irreducible. This assumption is not restrictive since we can analyse the queueing processes on the irreducible sub-chains of $\{\mathbf{J}(t), t \geq 0\}$ separately and merge the results later on.
We remark that for $\alpha = 0$ equation (5.3) gives the LST equation for the stationary workload process in an ordinary $MAP/G/1$ queue, provided that this workload process is ergodic. For $\alpha \to \infty$ the model can be interpreted as the $MAP/G/1/1$ loss model; in particular for $E = \{1\}$, equation (5.3) reduces to the LST equation of the stationary workload distribution in an $M/G/1/1$ loss model.

We proceed with solving equation (5.3) by iteration.
First, D is a generator of a Markov process on a finite state space E, hence the matrix $[\omega I + D]$ is non-singular for $Re\,\omega \geq 0$, except for at most $|E|$ values of ω. Defining

$$\zeta := \{\omega | Re\,\omega \geq 0, \mathrm{Det}[\omega I + D] \neq 0\}, \quad (\text{note that } 0 \notin \zeta),$$

we rewrite (5.3) for $\omega \in \zeta$ into

$$\Phi(\omega) = V(0)\omega[\omega I + D]^{-1} + \Phi(\omega + \alpha)D_1[\omega I + D]^{-1}(1 - \beta(\omega)). \tag{5.4}$$

We remark that $\omega[\omega I + D]^{-1}$ is a matrix of rational functions and that the set ζ contains the poles of these functions.
Defining $A(\omega) := \omega[\omega I + D]^{-1}$, and $B(\omega) := D_1[\omega I + D]^{-1}(1 - \beta(\omega))$, equation (5.4) can be rewritten into

$$\Phi(\omega) = V(0)A(\omega) + \Phi(\omega + \alpha)B(\omega), \quad \omega \in \zeta. \tag{5.5}$$

Next define $\omega_k := \omega + k\alpha$, $k = 0, 1, \ldots$, and $\prod_{l=0}^{k} B(\omega_l) := B(\omega_k) \ldots B(\omega_1)B(\omega_0)$, with the convention that an empty product ($k < 0$) is equal to I.

Iterating (5.5) K times we derive

$$\Phi(\omega) = V(0)\sum_{k=0}^{K} A(\omega_k)\prod_{l=0}^{k-1} B(\omega_l) + \Phi(\omega_{K+1})\prod_{l=0}^{K} B(\omega_l), \quad \omega \in \hat{\zeta}, \tag{5.6}$$

with $\hat{\zeta} = \{\omega \in \mathbb{C} | Re\ \omega \geq 0, \omega_l \in \zeta, l = 0, 1, \ldots\}$. Since $\alpha > 0$, the complement of the set $\hat{\zeta}$ is finite.

Assumption 5.2.1
For the sake of simplicity we assume in our analysis of $\{\mathbf{V}, \mathbf{J}\}$ that the eigenvalues of D are distinct. The analysis can be extended to the case of non-simple eigenvalues.

In proving the convergence of (5.6) for $K \to \infty$ we use the following lemma.

Lemma 5.2.1
(*i*) As $k \to \infty$, $A(\omega_k) \to I$, for $Re\ \omega \geq 0$.
(*ii*) For $K = 0, 1, \ldots$, $||\prod_{l=0}^{K} B(\omega_l)|| \leq c_\omega \tau^K$, for $\omega \in \hat{\zeta}$, with $\tau < 1$, $c_\omega < \infty$, and where $||C|| = |E| \max_{i,j} |C_{ij}|$ is a matrix norm of $C \in \mathbb{C}^{|E| \times |E|}$.

Proof
(*i*) Let η_j, $j = 1, \ldots, |E|$, be the eigenvalues of D. Since D is a generator matrix of a non-degenerate Markov process, there exists a constant $\xi < \infty$ such that $|\eta_j| < \xi$. According to the spectral theorem for matrices (cf. Lancaster & Tismenetsky[67] p.314), for $\omega \in \zeta$ there exist matrices $Z_j, j = 1, \ldots, |E|$ such that $[\omega I + D]^{-1} = \sum_{j=1}^{|E|} \frac{1}{\omega + \eta_j} Z_j$. Since $\alpha > 0$ and $\xi < \infty$, there exists a $K_0 > 0$ such that $\omega_k \in \zeta$ for $k > K_0$. Moreover, $\frac{\eta_j}{\omega_k} \to 0$, as $k \to \infty$. It readily follows that $||A(\omega_k) - I|| \to 0$. Applying proposition 1 on page 361 of Lancaster & Tismenetsky[67] finishes the proof of (i).
(*ii*) Next, define $\theta := \max_{i,j \in E} D_1(i,j)$. Then $||D_1|| = |E|\theta \leq |E| \max_{i \in E} \lambda_i$. From (i) it also follows that $[\omega_K I + D]^{-1} \to 0$ when $K \to \infty$, hence there exists a K_1 such that for all $k \geq K_1$, $||\sum_{j=1}^{|E|} \frac{1}{\omega_k + \eta_j} Z_j|| \leq \frac{\tau}{2\theta|E|}$ with $\tau < 1$. Since $||\cdot||$ is a matrix norm, $||C_1 C_2|| \leq ||C_1||.||C_2||$, for $C_1, C_2 \in \mathbb{C}^{|E| \times |E|}$. It follows that $||B(\omega_k)|| \leq \tau$, for $k \geq K_1$. Hence for each $\omega \in \hat{\zeta}$ there exists a $c_\omega < \infty$, such that $||\prod_{l=0}^{K} B(\omega_l)|| \leq c_\omega \tau^K$ for all $k = 0, 1, \ldots$. □

Application of Lemma 5.2.1 to (5.6) yields

Theorem 5.2.1

$$\Phi(\omega) = V(0)\sum_{k=0}^{\infty}\left(A(\omega_k)\prod_{l=0}^{k-1}B(\omega_l)\right), \quad \omega \in z\hat{e}ta, \quad \Phi(0)=\pi. \quad (5.7)$$

□

Next we derive $V(0) = (V_j(0))$, with $V_j(0)$ the steady state joint probability that the server is idle and **J** is in state j.
The $MAP/G/1+M$ is ergodic for $\alpha > 0$ (cf. Appendix 5.A), hence equation (5.3) has a unique solution, as given by (5.7). Moreover, $V(0)$ is the only vector that fits (5.7) for all $\omega \in \mathbb{C}, Re\ \omega \geq 0$. The latter remains valid if we post-multiply both sides of (5.3) with the non-singular matrix $R = (r_j)$, r_j being the right (column) eigenvector of D, associated with eigenvalue η_j, $j = 1, \ldots, |E|$. In particular, $r_1 = e$ is the eigenvector for the eigenvalue $\eta_1 = 0$. The matrix R is non-singular because all eigenvalues of D are distinct (cf. Lancaster & Tismenetsky[67] p.153). After the multiplication of (5.3) with R, dividing by ω and observing that the limit $\omega \downarrow 0$ exists, results in the following set of equations

$$\Phi(\omega)\left(1+\frac{\eta_j}{\omega}\right)r_j = V(0)r_j + \frac{1-\beta(\omega)}{\omega}\Phi(\omega+\alpha)D_1 r_j, \quad (5.8)$$

$j = 1, \ldots, |E|$, $Re\ \omega \geq 0$.
Subsequently, we fixate $\omega = -\eta_j$ in the $j-th$ equation of (5.8). Rewriting the right-hand side of (5.7) as $V(0)C(\omega)$, and defining constants $\beta_j := \lim_{\omega \downarrow -\eta_j} \frac{1-\beta(\omega)}{-\omega}$, then (5.8) becomes

$$V(0)r_j = I_{\{1=j\}} + [V(0)C(\alpha-\eta_j)\beta_j D_1]r_j, \quad j = 1, \ldots, |E|, \quad (5.9)$$

with $I_{\{\cdot\}}$ the indicator function.

The set of equations (5.9) has a unique positive solution. This follows from a probabilistic argument. Since the workload process is ergodic, $\Phi(\omega)$ is analytic for $\omega \in \{Re\ \omega \geq 0\}$. Moreover, $\Phi(\omega)$ is unique. Since the solution of (5.9) establishes an analytic solution of the functional equation (5.3), the uniqueness of $\Phi(\omega)$ forces $V(0)$ to be the only solution of (5.9).
The vector $V(0)$ is obtained from (5.9) by post-multiplying the $j-th$ equation of (5.9) with l_j, the left (row) eigenvector of D, associated with r_j ($l_i r_j = I_{\{i=j\}}, i, j = 1, \ldots, |E|$), and applying $\sum_{j=1}^{|E|} r_j l_j = I$. Note that $l_1 = \pi$, since π is the unique solution of $xD = 0, xe = 1$.

Finally, we have obtained

Theorem 5.2.2

$$V(0) = \pi \left[I - \sum_{j=1}^{|E|} C(\alpha - \eta_j)\beta_j D_1 r_j l_j \right]^{-1}. \tag{5.10}$$

□

Remark 5.2.1
For $\omega = -\eta_j$, $j = 2, \ldots, |E|$, $\Phi(\omega)$ can be derived by taking limits in (5.7); subsequently the remaining $\Phi(\omega)$ for $\omega \notin \hat{\zeta}$ can be obtained using relation (5.3). Since $\Phi(\omega)$ is analytic for $\omega \in \{Re\,\omega \geq 0\}$, the limits $\lim_{\omega \to -\eta_j} \Phi(\omega)$ exist. □

Remark 5.2.2
From expression (5.9) it follows that for the analysis it is necessary that $\alpha \in \hat{\zeta}$. Without losing generality we assume that this is the case; by slightly altering the parameters we can approximate any $MAP/G/1 + M$ queue arbitrarily closely by a $MAP/G/1 + M$ queue for which this assumption does hold. □

The results of our analysis are summarized in the following theorem.

Theorem 5.2.3
For $\omega \in \hat{\zeta} = \{\omega \in \mathbb{C} | Re\ \omega \geq 0, \omega_l \in \zeta, l = 0, 1, \ldots\}$ the vector LST $\Phi(\omega)$ of the workload process in the $MAP/G/1 + M$ queue is given by (5.7). The vector $V(0)$, denoting the stationary probability that the server is idle, is given by (5.10). For $\omega \notin \hat{\zeta}$ the values of $\Phi(\omega)$ can be obtained in the way described in Remark 5.2.1. □

Remark 5.2.3
We end this section by reflecting on other MAP generalizations of queueing models with impatient customers. In Baccelli et al.[10] the only other model that could be solved completely was the $M/M/1 + D$ queue, i.e. the $M/M/1$ queue with deterministically distributed patience. Unfortunately its MAP equivalent results in a differential equation which seems hard to solve. The general $M/G/1 + G$ queue was not explicitly solvable, so one could not expect more for the general $MAP/G/1 + G$ queue. □

5.3 Related results

Moments of $\{\mathbf{V}, \mathbf{J}\}$
The moments of $\{\mathbf{V}, \mathbf{J}\}$ can be obtained from (5.3). These are not obtainable by differentiating (5.7), because $0 \notin \hat{\zeta}$. Below we present the first moment, higher moments follow analogously.
Multiplying both sides of (5.3) with the unit vector e, using $De = 0$, dividing by ω, and taking derivatives with respect to ω, we obtain

$$\Phi'(\omega)e = \left(\frac{1-\beta(\omega)}{\omega}\right)' \Phi(\omega+\alpha)D_1 e + \left(\frac{1-\beta(\omega)}{\omega}\right) \Phi'(\omega+\alpha)D_1 e. \tag{5.11}$$

Letting $\omega \downarrow 0$ we find the mean total workload of the system

$$\Phi'(0)e = -\frac{\beta^{(2)}}{2}\Phi(\alpha)D_1e + \beta\Phi'(\alpha)D_1e, \tag{5.12}$$

in which β and $\beta^{(2)}$ respectively are the first and second moment of the service time distribution. $\Phi(\alpha)$ and $\Phi'(\alpha)$ can be derived from (5.7). By taking derivatives directly in equation (5.3), we also find an expression for $\Phi'(0)D$. Post-multiplying in (5.12) with π, and noting that $[e\pi + D]$ is non-singular, we finally obtain the mean workload vector

$$\begin{aligned} \mathbf{EV} &= -\Phi'(0) \\ &= [\pi - V(0) - \beta\Phi(\alpha)D_1][e\pi + D]^{-1} + \left[\frac{\beta^{(2)}}{2}\Phi(\alpha) - \beta\Phi'(\alpha)\right] D_1e\pi. \end{aligned} \tag{5.13}$$

Remark 5.3.1
As stated in the beginning of this section, for $\alpha \downarrow 0$, the $MAP/G/1+M$ queue reduces to the ordinary $MAP/G/1$ queue, although the $MAP/G/1+M$ queue is always stable for $\alpha > 0$ while the corresponding $MAP/G/1$ might not be. The stability condition for the ordinary $MAP/G/1$ queue is $\pi D_1 e\beta < 1$ (cf. Ramaswami[85]). Provided that this condition is satisfied, then, for $\alpha \downarrow 0$, the equation (5.3) for the LST of the workload process holds for the $MAP/G/1$ queue. Moreover, the expressions (5.12) and (5.13) are also consistent with the ordinary $MAP/G/1$ queue when $\alpha \downarrow 0$. However, they do not yield the moments for the ordinary $MAP/G/1$ queue because of the appearance of the term $\Phi'(0)D_1$ in the right-hand sides of (5.12) and (5.13) when $\alpha \downarrow 0$. The moments of the workload vector can be obtained as follows: rewrite (5.3) for $\alpha \downarrow 0$ as $\Phi(\omega)[\omega I + D_0 + \beta(\omega)D_1] = \omega V(0)$, postmultiply with the eigenvector $r_1(\omega)$ which belongs to the smallest eigenvalue $\eta(\omega)$ of $[\omega I + D_0 + \beta(\omega)D_1]$. Taking derivatives with respect to ω and letting $\omega \downarrow 0$ gives an expression for $E\mathbf{V}$, containing the unknown vector $V(0)$. This vector can be obtained in the same way as for the $MAP/G/1+M$ queue. A second method to derive this vector is presented in Lucantoni [73]; it involves the iteration of the matrix equivalent of the $M/G/1$ busy period equation (cf. Appendix A). □

The waiting time
$\Phi(\omega)$ is the LST vector of the virtual waiting time and the probability that $\mathbf{J}$ is in state j. Let us now show how this leads to the LST vector of the waiting time and to the probability that a customer prematurely leaves the system. A main observation is that the time between transitions in the Markov process is exponentially distributed, hence we can apply the PASTA property to link the virtual waiting time process to the actual waiting times of customers. Denote by $q_j = \sum_{i \in E} p_{ji}q_{ji}$ the probability that an arrival occurs given a transition out of state j, $j \in E$. Then, with $q = (q_j)$, we find for $x \geq 0$

$$\begin{aligned}\Pr\{\mathbf{W} \leq x\} &= \sum_{j \in E} \Pr\{\mathbf{V} \leq x, \mathbf{J} = j | \text{arrival of a customer}\} \\ &= \sum_{j \in E} \frac{\Pr\{\mathbf{V} \leq x, \mathbf{J} = j, \text{arrival of a customer}\}}{\Pr\{\text{arrival of a customer}\}} \\ &= \frac{V(x)q}{\pi q}. \end{aligned} \tag{5.14}$$

In the first equality of (5.14) we have applied the PASTA property.
From this we find $W(\omega)$, the LST of the waiting time of an arbitrary customer in steady state

$$W(\omega) = \frac{\Phi(\omega)q}{\pi q}, \quad Re\,\omega \geq 0. \tag{5.15}$$

By a derivation similar to (5.14) we find $\tilde{W}(\omega)$, the LST of the waiting time of a customer who will receive service.

$$\tilde{W}(\omega) = \frac{\Phi(\omega + \alpha)q}{\Phi(\alpha)q}, \quad Re\,\omega \geq 0. \tag{5.16}$$

The probability that a customer leaves the system prematurely
Analogous to the derivation of the LST for the waiting time we find P_α, the probability of a premature departure

$$P_\alpha = 1 - \int_{x=0}^{\infty} e^{-\alpha x} d\frac{V(x)q}{\pi q} = 1 - \frac{\Phi(\alpha)q}{\pi q} = 1 - W(\alpha). \tag{5.17}$$

5.4 Numerical results

In this section we numerically investigate the applicability of the results we obtained in Sections 5.2 and 5.3. We study the quantitative effects of customer impatience for the single-server queue with exponentially distributed service times and a two-state Markov Modulated Poisson Process (MMPP) as the customers arrival process. For this $MMPP/M/1+M$ queue we evaluate the mean amount of work in the system and the probability that a customer prematurely leaves the system.

Evaluation of performance measures for the $MAP/G/1 + M$ queue
Expression (5.7) of Theorem 5.2.1 shows that $\Phi(\omega)$ contains an infinite sum of terms, each of which consists of a product of complex valued matrices. In a numerical evaluation procedure the infinite sum has to be truncated, which

might introduce numerical inaccuracies. Moreover, all eigenvalues and eigenvectors of the generator D have to be computed in order to determine $V(0)$ in a numerical way. These difficulties form the main motivation to perform the experiments: we wish to answer the question how complicated it is to efficiently evaluate various performances measures such as the mean amount of work in the system **EVe** (cf. (5.13)) and the probability of a premature departure P_α (cf. (5.17)).
The results are obtained by implementing our expressions in the mathematical software environment MAPLE. This tool provides the algebraical structures and operations (e.g., complex matrices and matrix computation), as well as the numerical procedures (e.g., determining eigenvalues and eigenvectors) to make the implementation quite straightforward. Moreover, the problem of truncating the infinite sum in (5.7) can be treated fairly easily.
We conclude that expressions such as (5.7), (5.13) and (5.17) can successfully be evaluated.

Numerical results for a two-state $MMPP/M/1+M$ queue
In the remainder of this section we present and discuss some of our results on the $MMPP/M/1+M$ queue. First we describe the MMPP in some detail.
In the MMPP the arrival rate of customers is specified by the state of a continuous-time Markov process on a finite state space E; when the Markov process is in state $i \in E$ the arrival rate is λ_i. When a transition is made to state $i \in E$, the sojourn time in i is exponentially distributed with parameter σ_i, at the moment of transitions the next state is determined via some matrix of routing probabilities. It is readily seen that the MMPP is a special variant of the MAP; transitions from i to itself (with rate λ_i) induce the arrival of a customer, transitions to $j \neq i$ do not induce an arrival. (In the formal definition of the MAP presented in Section 5.2 transitions from i to itself are not allowed. However, self-transitions do not influence the analysis of the $MAP/G/1$ and $MAP/G/1+M$ queues.)
In our numerical experiments the state space of the MMPP has dimension two. From the definition of the MMPP it follows that

$$D_0 := \begin{bmatrix} -\lambda_1 - \sigma_1 & \sigma_1 \\ \sigma_2 & -\lambda_2 - \sigma_2 \end{bmatrix}, \qquad D_1 := \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}.$$

In our numerical examples $\sigma_1 = \sigma_2 = 1$, i.e. the Markov process is symmetric, hence the stationary probabilities for the states of the directing Markov process $\{\mathbf{J}(t), t \geq 0\}$ of the MMPP are $\pi_1 = \pi_2 = 1/2$.
Figures 5.1, 5.2 and 5.3 for various loads of the system show the effect that the impatience parameter α has on the probability of a premature departure and the mean amount of work present in the system. In each figure these performance measures are shown for varying degree of "burstiness" in the arrival process, i.e. for varying ratios $\lambda_1 : \lambda_2$. The total arrival rate remains a constant $\lambda' := \pi D_1 e = \pi_1\lambda_1 + \pi_2\lambda_2 = 1\frac{1}{2}$. In the figures α ranges between 0.19 and 1.39. In Figures 5.1,5.2 and 5.3 the numbers attached to the curves show the

values of λ_1 (note that $\lambda_2 = 3 - \lambda_1$). We remark that for $\lambda_1 = 1\frac{1}{2}$ the MMPP actually is a Poisson process with rate $1\frac{1}{2}$, and for $\lambda_1 = 0$ the MMPP is an interrupted Poisson process. Finally, the load $\rho := \lambda'/\mu$ is varied by changing the service rate μ. Note that in Figures 5.2 and 5.3 the load $\rho \geq 1$; without the impatience structure these queues would be saturated.

Discussion of Figures 5.1, 5.2 and 5.3

The curves in the figures show a behaviour that is logical. Firstly, as a function of α, P_α is increasing and $\mathbf{EV}e$ is decreasing. Secondly, as a function of ρ, both P_α and $\mathbf{EV}e$ are increasing. Thirdly, P_α is higher when the arrival process is "more bursty", i.e. P_α is smallest for $\lambda_1 = 1\frac{1}{2}$ (when the MMPP actually is an ordinary Poisson process), P_α is highest when $\lambda_1 = 0$ (when the arrival process is an interrupted Poisson process).

A fourth effect that we observed in the numerical results, but which is difficult to observe in the figures since the curves are so close together, is the fact that $\mathbf{EV}e$ is not monotone with respect to the degree of "burstiness", i.e. $\lambda_1 = 0$ does not for all values of α result in a lower value of $\mathbf{EV}e$ than $\lambda_1 = 1\frac{1}{2}$. To be more specific, for α small "not bursty" is better than "bursty", for α high the opposite holds. This property appears to be the result of two counteracting effects: the increasing probability of rejection as a function of α, versus the varying characteristics of the arrival process during alternating periods of high and low load. Apparently for α large, the higher probability of a premature departure has a more reducing effect on the amount of work for "bursty" arrival processes than it has for "not bursty" arrival processes. For α small, the relatively low probability of a premature departure results in a higher mean amount of work for "bursty" arrival processes.

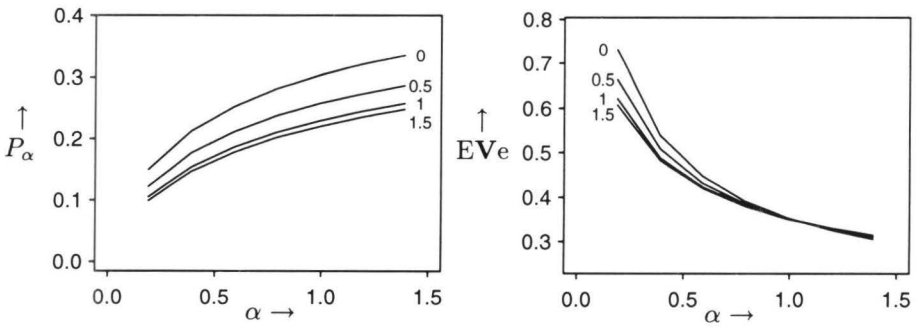

FIGURE 5.1: PROBABILITY OF A PREMATURE DEPARTURE AND MEAN AMOUNT OF WORK IN THE SYSTEM IN AN $MMPP/M/1{+}M$ QUEUE WITH LOAD $\rho = 0.7$.

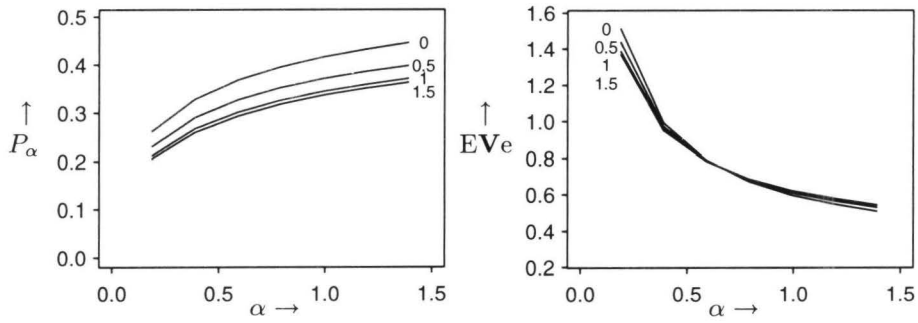

FIGURE 5.2: PROBABILITY OF A PREMATURE DEPARTURE AND MEAN AMOUNT OF WORK IN THE SYSTEM IN AN $MMPP/M/1+M$ QUEUE WITH LOAD $\rho = 1$.

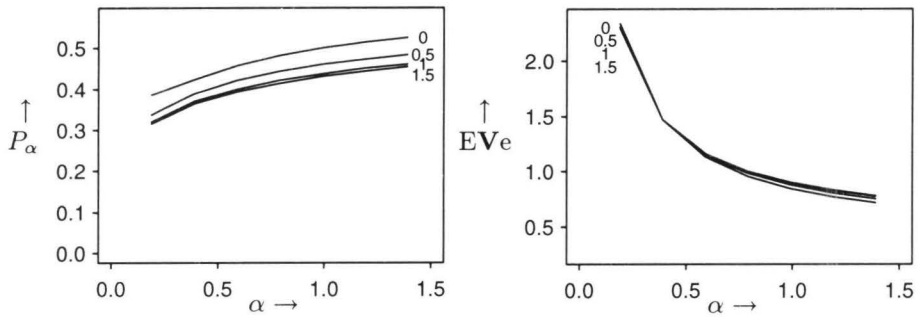

FIGURE 5.3: PROBABILITY OF A PREMATURE DEPARTURE AND MEAN AMOUNT OF WORK IN THE SYSTEM IN AN $MMPP/M/1+M$ QUEUE WITH LOAD $\rho = 1.3$.

APPENDIX

5.A ERGODICITY OF THE $MAP/G/1+M$ QUEUE

In this appendix we prove that the workload process $\{(\mathbf{V}(t), \mathbf{J}(t)), t \geq 0\}$ of the $MAP/G/1+M$ queue is ergodic for $\alpha > 0$.

Intuitively this is clear, due to the impatience of the customers: as the amount of work in the system increases, the number of customers joining the queue gets smaller and smaller. Eventually, the work that joins the queue per unit of time is less than 1, hence there exists an $x_0 \geq 0$ such that the drift of the Markov process is towards the origin for all states $x \in [x_0, \infty) \times E$, (we recall that $[0, \infty) \times E$ is the state space of $\{(\mathbf{V}(t), \mathbf{J}(t)), t \geq 0\}$). The general idea of the proof follows this intuition. If one can show the existence of a positive recurrent subset $A \subset [0, \infty) \times E$ then, under certain conditions, this is sufficient for the existence of a unique stationary distribution of $\{(\mathbf{V}(t), \mathbf{J}(t)), t \geq 0\}$. The general idea is worked out in Laslett et al.[69]; in Appendix B we present a

3-step scheme from [69] for proving the ergodicity of multi-dimensional Markov processes. In this appendix we apply that scheme.
For the workload process $\{(\mathbf{V}(t), \mathbf{J}(t)), t \geq 0\}$ of the $MAP/G/1 + M$ queue, we prove ergodicity of the embedded Markov chain $\{(\mathbf{V}_n, \mathbf{J}_n), n = 0, 1, \ldots\}$ $= \{(\mathbf{V}(t_n), \mathbf{J}(t_n)), n = 0, 1, \ldots\}$, where t_n is the time just before the $n-th$ transition in $\{\mathbf{J}(t), t \geq 0\}$. Applying the conditional PASTA property, we find that the stationary distribution of the embedded Markov chain equals the stationary distribution of $\{(\mathbf{V}(t), \mathbf{J}(t)), t \geq 0\}$.

1. We choose for ϕ an arbitrary nonzero measure on F (F is the σ-algebra induced by the Borel subsets of the state space), such that ϕ has an atom at $(0, i) \in [0, \infty) \times E$, where i is an arbitrary element of E. Since the Markov chain $\{\mathbf{J}(t), t \geq 0\}$ is irreducible it follows that $(0, i)$ can be reached from any subset $A \subset [0, \infty) \times E$ in at most $|E|$ steps. Hence, $\{(\mathbf{V}_n, \mathbf{J}_n), n = 0, 1, \ldots\}$ is ϕ-irreducible.

2. According to Theorem 3.2 of Laslett et al.[69], any set A containing $\{(0, i)\}$ is a test set if for some integer N and some $\delta > 0$, $\max_{n \leq N} P^n(y, (0, i)) \geq \delta$, for all $y \in A$. With $N = |E|$ this condition holds when A is of the form $[0, x_0) \times E$, with $x_0 > 0$.

3. We apply Theorem B.1, using the function $g((x, i)) = x$, $(x, i) \in [0, \infty) \times E$. t_0 is defined as the time just before the first transition in $\{\mathbf{J}(t), t \geq 0\}$. Denote by (x_0, i) the state of $(\mathbf{V}_0, \mathbf{J}_0)$ and $\tau_{(x_0,i)}$ the amount of work joining the queue at time t_0. Define by $\sigma_{(x_0,i)} := t_1 - t_0$ the time until the first jump after t_0 in $\{\mathbf{J}(t), t \geq 0\}$, conditioned on the state of $(\mathbf{V}_0, \mathbf{J}_0)$. Then

$$\begin{aligned}
&\mathrm{E}[g(\mathbf{V}_1, \mathbf{J}_1) - g(\mathbf{V}_0, \mathbf{J}_0) | (\mathbf{V}_0, \mathbf{J}_0) = (x_0, i)] \\
&\qquad = \mathrm{E}[\max[x_0 + \tau_{(x_0,i)} - \sigma_{(x_0,i)}, 0] - x_0] \\
&\qquad = \mathrm{E}[\max[\tau_{(x_0,i)} - \sigma_{(x_0,i)}, -x_0]] \\
&\qquad = \int_{u=-\infty}^{-x_0} (-x_0) dH_{(x_0,i)}(u) + \int_{u=-x_0}^{\infty} u dH_{(x_0,i)}(u),
\end{aligned}$$

with $H_{(x_0,i)} = \tau_{(x_0,i)} - \sigma_{(x_0,i)}$, and $H_{(x_0,i)}(u) = \Pr\{H_{(x_0,i)} \leq u\}$, $u \in \mathbb{R}$.
It is readily verified that $\mathrm{E}H_{(x_0,i)}$ is a continuous non-increasing function of x_0 when $\alpha > 0$. Finally, $\lim_{x_0 \to \infty} \mathrm{E}H_{(x_0,i)} \leq -\frac{1}{\hat{\theta}} < 0$, with $\hat{\theta} = \max_{j \in E} \lambda_j$.
It follows that for $\alpha > 0$, $\exists x_0 \geq 0$ such that for all $i \in E$, $\mathrm{E}[g(\mathbf{V}_1, \mathbf{J}_1) - g(\mathbf{V}_0, \mathbf{J}_0) | (\mathbf{V}_0, \mathbf{J}_0) = (x_0, i)] < 0$, hence $\sup_{(x,i) \in A} \mathrm{E}[T_A | (\mathbf{V}_0, \mathbf{J}_0) = (x, i)] < \infty$, with $A = [0, x_0) \times E$ (we remark that in general x_0 depends on α). This completes the last step of our scheme. □

Chapter 6

Optimization of static traffic allocation policies

We consider the traffic allocation problem: customers arriving at a service facility have to be assigned to one of a group of servers. The aim is to optimize system performance measures, such as mean waiting time of a customer or total number of customers in the system, under a given static allocation policy. Two static policies are considered: probabilistic assignment and allocation according to a fixed pattern. For these two policies general properties as well as optimization aspects are discussed.

6.1 Introduction

In a distributed computer system, tasks generated by a group of users can be distributed over a number of available processors. This contrasts with systems in which a single processor provides (global) computer capacity for all users, or systems in which each user is provided with its own local processor, usually with very limited capacity.
An operational aspect of such a distributed system is the availability of a *load balancing protocol*. Such a protocol balances the workload over the servers, aiming to optimize performance measures for the system, such as mean amount of workload, throughput, or mean waiting times of jobs.
Load balancing is required in many situations where a workload is offered to a number of servers with limited capacity. Apart from distributed systems, one may e.g. think of the transmission of messages along one of several available paths of a communication network.
An important element of a load balancing protocol is the information it requires to operate. This information can range from total knowledge about the system

at any point in time, to only information about some basic characteristics, like arrival rate and service times. In general, the term *dynamic* is used for policies which operate under time dependent information, whereas protocols operating under time independent characteristics of the system are called *static*.

It is clear that the more information is available for making decisions, the better the allocation of workload can be. Dynamic policies in general perform better than static policies. However, static load balancing protocols are also of considerable interest. First of all, the situation of total knowledge at all times is unrealistic. From a viewpoint of costs, overhead grows as the amount of information to be exchanged, stored and processed increases. Moreover, dynamic policies are not always that effective: there will always be some kind of time delay between updates of the system's current state, and this time delay can have a considerable effect on the quality of the protocol.

A second reason for studying static allocation policies is that they can be useful tools in the design phase of a computer- or communication network. Static policies can provide performance bounds for dynamically controlled systems; the performance measures under static policies are in general evaluated reasonably quickly, whereas dynamic policies are harder to analyse and their performance can only be evaluated with time consuming methods.

In this chapter we consider the *traffic allocation problem* for two *static* allocation protocols for the model of a single Poisson stream of jobs offered to a fixed number of server stations. The allocation protocols we study are static in the sense described above; only the traffic intensity and the server characteristics are used. We give an overview of the results for these policies and also extend optimization procedures for some models.

In the remainder of this section we present a brief survey of related literature and an outline of the chapter.

Related literature

Several papers have addressed the load balancing problem. Below we refer to two overview papers for the general load balancing problem, before giving a more extensive overview of the traffic allocation problem. Wang & Morris[100] give a taxonomy for the current load balancing protocols. They formulate the load balancing problem in its most general form, also discriminating between *server initiative* protocols, i.e. the servers determine from which input sources they draw their customers, and *source initiative* protocols, i.e. at the moment of arrival in the system jobs are (irrevocably) routed to one of the servers. Wang & Morris[100] provide numerical comparisons, based on analysis and simulation, of various allocation protocols. An overview of load balancing policies and their performances is also given by Boel & Van Schuppen[15]. They consider the problem from a control point of view, and discuss the question what amount of information is required at the routing points to achieve good system performance. Their paper concentrates on analytically and numerically tractable models. Numerical comparisons (cf. Wang & Morris[100]) reveal that indeed dynamic allocation policies lead to considerably better results than probabilistic

allocation (better with respect to performance measures such as mean waiting times and mean amount of work in the system).

Two static allocation policies have been proposed for the traffic allocation problem: viz. probabilistic assignment and pattern allocation. With the probabilistic policy each arriving customer is routed to one of the servers with fixed probabilities. Under pattern allocation each arriving customer is routed to a server according to an allocation table.

For probabilistic allocation, Buzen & Chen[28] present an algorithm for determining the allocation which minimizes the mean sojourn time of a customer. Their mathematical programming formulation can easily be extended for various other performance measures and fits into the framework of Ibaraki & Katoh[59] for Resource Allocation Problems (RAP). Probabilistic load balancing has been studied by Jean-Marie[63] for the case of two parallel exponential servers and resequencing.

Yum[102] proposed the pattern allocation policy ('semi-dynamic deterministic routing'), which performs notably better than the probabilistic allocation policy (cf. Agrawala & Tripathi[3] and Yum[102]). The reason for this is that the arrival processes at the servers under the pattern allocation policy are less irregular than under probabilistic allocation. However, constructing *the* optimal allocation pattern is yet an unsolved problem. For the case of two identical exponential servers, Ephremides et al.[44] proved that alternately assigning customers to each queue is optimal with respect to minimizing the expected total completion time for all service requests that arrive before time T, a result which was extended by Ramakrishnan[84] for the model with more than two identical exponential servers. Ramakrishnan[84] also proposed a useful approximation procedure for determining the optimal pattern allocation (optimal with respect to mean sojourn time) for the case of non-identical exponential servers.

The problem of determining good pattern allocations for the case of non-identical exponential servers has also been studied by Hordijk et al.[58]. In [58] the problem is analyzed in the context of Markov decision theory.

The present chapter extends the approximation procedure proposed by Ramakrishnan[84] in several directions, in particular allowing *general* service time distributions. We also give an overview of the results for the two above-mentioned static allocation policies. By comparing both policies from a more theoretical viewpoint than in most previous studies, we develop insights into general allocation problems and clarify some reported, but hitherto unexplained, properties.

Outline of the chapter

In Section 6.2 a mathematical description of the allocation problem is presented, and the probabilistic allocation policy is discussed. In Section 6.3 we argue that allocation policies which result in more regular arrival processes than

the Poisson arrival process are to be preferred to probabilistic allocation. There also the pattern allocation policy is introduced. In Section 6.4 an optimization procedure for pattern allocation is presented. In Section 6.5 the performance measures under both allocation policies are numerically compared for various models. We also compare both policies with a dynamic policy that is expected to outperform most policies for the objective functions we consider. Section 6.6 discusses three extensions of the basic traffic allocation problem. The first extension deals with the case of a general arrival process. The second model describes the case in which all server stations receive a ('dedicated') Poisson arrival stream, on top of which an extra arrival stream has to be allocated. The third extension considers allocation to multiple server stations.

6.2 Probabilistic allocation

Before studying the probabilistic allocation policy we first present a mathematical description of the traffic allocation problem.

Model description

Customers arrive at a routing point according to a Poisson process with rate Λ. At the instance of arrival, a customer has to be assigned to one of N single servers in parallel. This assignment is irrevocable.

The service time $\mathbf{B}_i$ of a customer that is assigned to server i has general distribution $B_i(\cdot)$, with first and second moment β_i and $\beta_i^{(2)}$ respectively. All service times are independent.

Let P denote an allocation policy and p_i, $i = 1, \ldots, N$, be the fraction of the customers that is routed to server i under policy P.

In our traffic allocation problem, the aim is to minimize

$$\sum_{i=1}^{N} f_i(P) C_i \mathrm{E}\mathbf{W}_i(P). \tag{6.1}$$

In (6.1) $\mathrm{E}\mathbf{W}_i(P)$ denotes the mean waiting time of a customer assigned to server i under allocation policy P. C_i is the cost associated with waiting one time unit at queue i. The factors $f_i(P)$ are additional, load dependent, weight factors. The objective function can have various interpretations by varying $f_i(\cdot)$ and C_i. For example, if $f_i(P) = p_i$ and $C_i = 1$, $i = 1, \ldots, N$, then the objective function represents the mean waiting time of an arbitrary customer. Or, with $f_i(P) = \Lambda p_i$ and $C_i = \beta_i$, Little's law shows that the objective is to minimize the mean total amount of work in the queues. Instead of $\mathrm{E}\mathbf{W}_i(P)$, also $\mathrm{E}\mathbf{R}_i(P)$, the mean sojourn time of a customer assigned to queue i, could have been used in (6.1).

Probabilistic allocation

As described in the introduction, the assignment of an arriving customer to a queue can depend on all kinds of information contained in the history and the present state of the system. In this section we discuss the probabilistic

allocation policy, also known as random splitting. Under this policy, a fraction p_i of the arrivals is routed to queue i by assigning a customer, arriving at the routing point, to server station i with probability p_i, $i = 1, \ldots, N$ ($\sum_i p_i = 1$). These probabilities p_i are the same for all customers, and do not change in time. Let P_{pr} denote the class of probabilistic allocation policies. This class can be completely described by $P_{pr} = \{p \mid p \in [0,1]^N, \sum_{i=1}^N p_i = 1\}$.
The probabilistic allocation policy is static in the sense that when a customer has to be routed to one of the queues, no information about the history and the present state of the system is used. Under $P \in P_{pr}$, the arrival process at queue i is Poisson with intensity $\lambda_i = p_i\Lambda$, $i = 1, \ldots, N$, and the objective function becomes

$$\sum_{i=1}^{N} f_i(P) C_i \mathrm{E}\mathbf{W}_i(P) \quad = \quad \sum_{i=1}^{N} f_i(P) C_i \frac{p_i \Lambda \beta_i^{(2)}}{2(1 - p_i \Lambda \beta_i)}. \tag{6.2}$$

Among the first to study the probabilistic allocation policy were Buzen & Chen[28]. Their aim was to minimize the mean sojourn time of a customer for a model with generally distributed service times at the server stations. They solved the problem using standard mathematical programming techniques.
As an example, we take $f_i(P) = \frac{\lambda_i}{\Lambda}$ in (6.2) and solve the allocation problem. In this case, the objective is to minimize the mean weighted waiting time of a customer or, using Little's law, to minimize a weighted sum of the mean numbers of waiting customers in the system. To obtain the assignment probabilities $p_i^* = \frac{\lambda_i^*}{\Lambda}$ which minimize this function, the following Mathematical Programming Problem has to be solved
PA1:

$$\min \quad \sum_{i=1}^{N} \lambda_i C_i \frac{\lambda_i \beta_i^{(2)}}{2(1 - \lambda_i \beta_i)} \tag{6.3}$$

$$s.t. \quad \sum_{i=1}^{N} \lambda_i = \Lambda, \tag{6.4}$$

$$0 \leq \lambda_i < \frac{1}{\beta_i}, \quad i = 1, \ldots, N. \tag{6.5}$$

Note that the objective function in PA1 can be separated in terms $T_i = \lambda_i C_i \frac{\lambda_i \beta_i^{(2)}}{2(1-\lambda_i\beta_i)}$, which are strictly convex functions in λ_i. It can also be verified that PA1 has a feasible solution provided that $\sum_i \frac{1}{\beta_i} > \Lambda$, i.e. the arrival rate does not exceed the total service capacity. Here and in the remainder of the chapter it is assumed that such is the case.
Problem PA1 allows an analytical solution. To find this solution, we first relax PA1 by dropping constraint (6.5). Using standard Lagrange-multiplier techniques we obtain, with δ denoting the Lagrange-multiplier, the following first order Kuhn-Tucker constraints:

$$\frac{\partial}{\partial \lambda_i}\left\{\lambda_i C_i \frac{\lambda_i \beta_i^{(2)}}{2(1-\lambda_i\beta_i)}\right\} = \delta, \qquad i = 1,\ldots,N, \tag{6.6}$$

$$\sum_{i=1}^{N} \lambda_i - \Lambda = 0. \tag{6.7}$$

From (6.6) and remarking that (6.3) concerns a sum of terms T_i which are convex functions of λ_i we find the unique optimal values λ_i^*

$$\lambda_i^* = \frac{1}{\beta_i} - \frac{1}{\beta_i}\left(\sqrt{1+\frac{2\beta_i\delta}{C_i\beta_i^{(2)}}}\right)^{-1}, \qquad i = 1,\ldots,N, \tag{6.8}$$

in which the value of the Lagrange-multiplier δ is determined by the constraint (6.7). The optimal splitting probabilities are given by $p_i^* = \frac{\lambda_i^*}{\Lambda}$, $i = 1,\ldots,N$. In (6.8) we see that $0 \leq \lambda_i^* < \frac{1}{\beta_i}$, $i = 1,\ldots,N$; so the vector λ^* is also the solution of PA1.

This example shows us the structure of our traffic allocation problem. The objective function is separable in (strictly convex) terms T_i, each term T_i being a function of λ_i. Hence the solution of PA1 is determined by the derivatives of the terms (cf. (6.6)) rather than their values.
A second observation follows from (6.7) and (6.8): if $\Lambda > 0$, then $\lambda_i^* > 0$ for all i. This is a direct consequence of the above-mentioned properties; in the example $\frac{\partial}{\partial \lambda_i} T_i \mid_{\lambda_i=0} = 0$ and $\frac{\partial}{\partial \lambda_i} T_i$ is an increasing function in λ_i, hence $\lambda_i^* > 0$, provided that $\Lambda > 0$.
However, the latter property not necessarily holds in all situations; for other naturally arising objective functions, such as $\sum_i C_i \mathrm{E}\mathbf{W}_i$ or $\sum_i f_i C_i \mathrm{E}\mathbf{R}_i$, with $\mathbf{R}_i$ denoting the sojourn time of a customer routed to queue i, the optimal values of some λ_i's (and p_i's) may be equal to zero. For these objective functions $\frac{\partial}{\partial \lambda_i} T_i \mid_{\lambda_i=0}$ can be so large, relative to the other queues, that it is advantageous not to assign customers to queue i, but to allocate all arrivals to the other queues.
For the latter objective functions the Mathematical Programming Problems have the same structure as PA1. Usually these MPP's do not allow an explicit analytical solution. However, PA1 can be solved quite easily numerically, due to its special structure: the control variables only interact through the linear restriction (6.4). This characteristic is typical for the class of Resource Allocation Problems (RAP), as studied in Ibaraki & Katoh[59]. In their book they also consider a RAP which has almost the same form as PA1, the only difference being that the control variables are allowed to equal the upper bounds.
In Appendix 6.A we present an algorithm to solve the traffic allocation problem that has a separable objective function, consisting of strictly convex terms, and that has strict upper bounds on the control variables. The algorithm is a variant of the procedure RANK in Katoh & Ibaraki([59], p.19). The algorithm first

determines the set of queues for which $\lambda_i^* > 0$; for that set it subsequently solves a set of equations of the form of (6.6) and (6.7).
The algorithm strongly depends on the strict convexity of the terms. Due to this property, there is only one local minimum, which consequently has to be the optimal solution for the allocation problem. If the objective function is not separable into strictly convex terms then in general there may exist several local minima for the allocation problem. Moreover, in most situations only approximately optimal allocation probabilities can be obtained. One of the cases in which the property of strictly convex terms may not hold is the traffic allocation problem with a general arrival process, as studied in Tang & van Vliet[99]. Their method involves an algorithm for quadratic programming and provides one of the local minima. They also argue that this local minimum should be close to the global minimum.

6.3 Reducing variance in arrival processes

Intuitively one expects that when traffic allocation leads to a more regular arrival process, then the mean waiting times are reduced and consequently also the value of the objective function. However, it is very difficult to prove such statements, except for special cases. A detailed investigation of these issues would not fit into the framework of this chapter, and hence we restrict ourself to presenting some basic results on comparison between queueing systems, along with a special case to support the above-mentioned intuition.
In this section we discuss the single-server queue with an arrival process that is more regular than the Poisson process, and we argue that for an important class of such arrival processes, the behaviour of the mean waiting time as a function of the load is better than for Poisson arrivals.

General comparisons of GI/G/1 systems are presented in Stoyan[93]. Particularly useful for our purposes is his theorem 5.2.1, which states the following monotonicity property for the waiting times:

Lemma 6.3.1
Consider two GI/G/1 queueing systems with identically distributed service times. If for the interarrival times $\mathbf{A}_1$ and $\mathbf{A}_2$, $\mathbf{A}_1 \leq_c \mathbf{A}_2$, then also for the steady state waiting times $\mathbf{W}_1 \leq_c \mathbf{W}_2$.

In Lemma 6.3.1 $\leq_c$ denotes the convex stochastic ordering for random variables, and indices 1,2 refer to the two queuing systems. Since $\mathbf{W}_1$ and $\mathbf{W}_2$ are positive random variables, $\mathbf{W}_1 \leq_c \mathbf{W}_2$ implies $\mathrm{E}\,\mathbf{W}_1^r \leq \mathrm{E}\,\mathbf{W}_2^r$, $r = 1, 2, \ldots$.

In particular, if $\mathrm{E}\,\mathbf{A}_1 = \mathrm{E}\,\mathbf{A}_2$, $\mathbf{A}_1 \leq_c \mathbf{A}_2$ holds in the following two cases:
(i) $\mathbf{A}_1$ is constant (cf. Stoyan[93], example 1.9(a)).
(ii) $\mathbf{A}_1$ is NBUE and $\mathbf{A}_2$ has an exponential distribution.
A stochastic variable $\mathbf{X}$ with distribution function F is New Better than Used in Expectation (NBUE) if $\int_t^\infty (1-F(x))dx/(1-F(t)) \leq \mathrm{E}\mathbf{X}$ for all $t \geq 0$. Note that

if $\mathbf{X}$ has an increasing failure rate, $\mathbf{X}$ is NBUE. As examples, Gamma(Λ, y) with $y \geq 1$, Weibull(Λ, y) with $y \geq 1$ and Uniformly distributed random variables are all NBUE (cf. Stoyan[93] Chapter 1). The Gamma(Λ, y) case is now discussed in more detail, as it plays an important role in the remainder of the chapter.

Case 6.3.1:
Consider a Gamma(Λ, y)/M/1 queueing model with $y > 1$, so that the arrival process has a coefficient of variation which is smaller than that of a Poisson process. Let μ be the service rate and $\frac{\Lambda}{\mu y} < 1$, i.e. the queue is stable. In this queue the mean waiting time of a customer $\mathrm{E}\mathbf{W}^G$ is given by $\frac{\omega}{\mu(1-\omega)}$, with $\omega = Pr\{\mathbf{W}^G > 0\}$ the smallest positive real solution of $x = \alpha(\mu(1-x))$, with $\alpha(\cdot)$ being the Laplace-Stieltjes Transform of the arrival process (cf. Cohen[34]). For Gamma(Λ, y) we have $\alpha(x) = \left(\frac{\Lambda}{\Lambda + x}\right)^y$.

Firstly, for this queueing model $\mathrm{E}\mathbf{W}^G < \mathrm{E}\mathbf{W}^M$, which follows from Lemma 6.3.1 Here $\mathrm{E}\mathbf{W}^M$ denotes the mean waiting time in the M/M/1 queue with arrival rate $\frac{\Lambda}{y}$ and service rate μ.

Secondly, it is readily verified that $\frac{\partial}{\partial \Lambda}\mathrm{E}\mathbf{W}^G \downarrow 0$ as $\Lambda \downarrow 0$.

These two properties have the following consequence for the traffic allocation problem with N Gamma(λ_i, y_i)/M/1 queues ($y_i > 1$, $i = 1, \ldots, N$). If one has to determine intensities λ_i such that the overall arrival rate $\sum_i \frac{\lambda_i}{y_i} = \Lambda$ while the objective is to minimize $\sum_i \mathrm{E}\mathbf{W}_i^G$, then $\lambda_i > 0$ for all i. Moreover, the value of the objective function will be lower than if a Poisson Λ arrival stream had been allocated probabilistically to the N stations.

For the Gamma(Λ, y)/M/1 queueing model we also find that $\frac{\partial}{\partial y}\mathrm{E}\mathbf{W}^G \downarrow 0$ as $y \to \infty$. As a consequence, in the traffic allocation problem for the queueing system with N parallel Gamma(Λ, y_i)/M/1 queues where one has to assign y_i's under the condition that $\sum_i \frac{1}{y_i} = 1$ (i.e., the sum of the arrival rates at the queues is Λ) we find that $\frac{1}{y_i} > 0$ for all i. This special queueing model is used in the next section as an approximation for a queue with a special non-renewal arrival process.

Pattern allocation - the MAP/G/1 queue
Next we introduce a traffic allocation policy which allocates the Poisson arrival stream such that the arrival processes at the queues are less variable than under probabilistic assignment, but which still is static in the sense that no state information of the queues is used and that the allocations are time independent. This policy is pattern allocation. Pattern allocation uses an infinite string of integers $\{a_1, a_2, \ldots, a_{n-1}, a_n, a_{n+1}, \ldots\}$, where a_n denotes the number of the queue to which the n-th customer in the arrival process is routed. For practical reasons it is assumed that this string contains a sub-pattern S of finite length M which is repeated over and over. Thus $a_i = a_{i+kM}$ for all $i = 1, \ldots, M$ and $k = 1, 2, \ldots$. Like for the probabilistic allocation policy we can completely describe P_{pa}, the class of pattern allocations: $P_{pa} = \{a \mid a \in [1..N]^k, k = 1, 2, \ldots\}$.

Let A_{i_n} be the time between the n-th and $n+1$-st arrival at queue i. Under pattern allocation, the distributions of A_{i_n} form a repeated sequence of Erlang distributions. For example, if $S = \{1,2,1,3,4,1,2\}$, then the sequence of the interarrival distributions at queue 1 is a repetition of {Erlang$(\Lambda,2)$, Erlang$(\Lambda,3)$, Erlang$(\Lambda,2)$}.
The pattern allocation policy was first introduced by Yum[102] as semi-dynamic deterministic routing. For the cases of two and infinitely many identical exponential server stations, Yum[102] shows a considerable reduction in mean waiting time if the pattern allocation policy is used instead of probabilistic allocation.

Again, as in Chapters 4,5 and 7, we encounter the class of Markovian arrival processes; the arrival process under pattern allocation can be seen as a special case of the Batch Markovian Arrival Process(BMAP), more specifically, since customers arrive separately, it is a special case of the MAP. Details on the BMAP and the $BMAP/G/1$ queue are presented in Appendix A, below we characterize the MAP arising under pattern allocation.
Arrivals are generated at transition epochs of a continuous time Markov process with finite state space $\{1,\ldots,M\}$, with M being the length of the allocation pattern. The sojourn time in each state is Λ, the rate of the Poisson customers arriving at the routing point. In the MAP arising under pattern allocation, the Markov chain has the special property that only transitions from state i to state (i mod M) $+1$ can occur. Finally, for queue k, the transition from i generates an arrival if the *i–th* element of S equals k. We remark that Agrawala & Tripathi[4] analyse the waiting times in the MAP/M/1 queue for the typical MAP that we consider. Their analysis is based on classical complex analysis techniques, the method of Lucantoni[73] for the more general $BMAP/G/1$ queue is more of a probabilistic nature.
Observe that the Markovian Arrival Process in general is not a renewal process. The earlier mentioned comparisons from Stoyan[93] are for GI/G/1 queues and do not apply to MAP/G/1 queues. Besides, a useful characterization of the irregularity of a MAP is much more complicated than for GI arrival processes. However, in order to compare the MAP/G/1 queue with an M/G/1 queue with the same service time distribution we state the following conjecture, which is based on the observations made earlier in this section and supported by numerical experience.

Conjecture 6.3.1
Consider a stable M/G/1 queue in which the arrival rate is $p\Lambda$, with $p < 1$, and the service time has distribution $B(\cdot)$. Then there exists a MAP with transition rate Λ in all states of the underlying Markov chain, and overall arrival rate closely approximating $p\Lambda$ from above, such that in the MAP/G/1 queue with the same service time distribution $B(\cdot)$, $\mathrm{E}\mathbf{W}^{MAP} < \mathrm{E}\mathbf{W}^{M}$, where $\mathbf{W}^{MAP}$ and $\mathbf{W}^{M}$ denote the steady state waiting times of customers in the MAP/G/1 queue and M/G/1 queue respectively. □

Conjecture 6.3.1 is clarified by viewing the Poisson($p\Lambda$) arrival process as the result of a probabilistic allocation and the MAP as the result of a pattern allocation. Let M be the number of phases in the MAP. Then in the pattern allocation out of every M arriving customers an exact fraction p is routed to the queue, whereas under probabilistic allocation this fraction is only in expectation equal to p. Moreover, in the MAP the arrivals can be better regulated, e.g., for $p = \frac{2}{5}$ every second and fifth customer can be routed to the queue.
Note that not for *all* MAP with phase intensity Λ and overall arrival rate $p\Lambda$, $\mathrm{E}\mathbf{W}^{MAP} < \mathrm{E}\mathbf{W}^{M}$; for example, again viewing the MAP as the result of pattern allocation, when of every $2M$ customers the first M are routed to the queue, then for $\frac{1}{2} < \Lambda\beta < 1$ the mean waiting time of a customer at the queue tends to infinity as $M \to \infty$, while $p = \frac{1}{2}$ and $\mathrm{E}\mathbf{W}^{M} < \infty$.
Also note that the refinement "closely approximating $p\Lambda$ from above" has to be made, because for p irrational there does not exist a MAP with finite state space of the underlying Markov process such that the overall arrival rate is exactly equal to $p\Lambda$.
A benefit of the pattern allocation policy, besides lowering the value of the objective function, is that it is more robust than probabilistic allocation. For example, from the explicit expression for the mean waiting times in an exponential server queue (cf. case 6.3.1), it follows that a change of a few percent of the arrival intensity in an Erlang$(\Lambda, 2)$/M/1 queue has less influence on the wating times than a similar change of the arrival intensity in an M/M/1 queue.

In this section we have argued that in the traffic allocation problem the pattern allocation policy is to be preferred to the probabilistic allocation policy, because of the reduction of variability in the arrival processes. In the next section we turn our attention to an optimization procedure for the pattern allocation policy.

6.4 Optimal pattern allocation

The mean waiting time of a customer in the MAP/G/1 queue can be evaluated from expressions (A.7)-(A.9) in Appendix A. However, these expressions are not very suitable for a direct optimization procedure; their matrix structure makes an exact analytical optimization actually impossible. This contrasts with probabilistic allocation where only the N optimal assignment probabilities p_i^* have to be determined and where the simple structure of the objective function (6.3) allows an analytical solution of the Mathematical Programming Problem PA1.
Moreover, for pattern allocation it is impossible to determine the optimal allocation pattern by comparing patterns; there are too many patterns with length smaller than some practical bound, and the matrix operations involved in the evaluation of expression (A.8) are too time consuming.
We therefore have to resort to an approximate optimization procedure. Our procedure consists of two steps:

(1) Approximate p_i^*, $i = 1, \ldots, N$, the queue assignment frequencies in the optimal allocation pattern.
(2) Use these frequencies for the construction of the allocation pattern.

In this section, our attention is mainly devoted to step 1. The problems related to step 2 are more of a combinatorial nature, and in fact a quite difficult cyclic scheduling problem has to be solved. In Remark 6.4.4 we mention some of the difficulties occurring here, and in Appendix 6.B we present a heuristic for building an allocation pattern from a set of allocation frequencies.

Due to the matrix operations involved, comparing assignment frequencies directly using expressions (A.7)-(A.9) is too time consuming, so further approximations have to be made. To avoid the matrix operations we approximate the MAP with a GI arrival process.
An obvious option for this GI arrival process is the Poisson arrival process. In step 1, the fractions p_i^*, $i = 1, \ldots, N$ are then approximated by the optimal probabilistic allocation. However, in general this does not lead to the optimal allocation pattern, as illustrated in Agrawala & Tripathi[3] for the traffic allocation problem with the mean sojourn time of a customer as objective function.
A good choice for a GI approximation of the MAP is the Gamma arrival process, an arrival process with Gamma distributed interarrival times. The first step then is to determine the optimal allocation fractions for a model in which we are to assign customers from an infinite reservoir of customers to N parallel Gamma/G/1 queues maintaining an overall arrival rate Λ. This we call the Gamma approximation procedure.
The idea of approximating the arrival process with a Gamma arrival process was first used by Ramakrishnan[84], who studied various allocation policies for the case of exponentially distributed service times. Using the exact expression for the mean waiting times in the Gamma/M/1 queue (cf. case 3.1 in Section 6.3) Ramakrishnan numerically solved the Gamma/M/1 allocation problem for the case of two queues.
The Gamma(Λ, y) arrival process looks to be a reasonable approximation for the MAP with overall arrival intensity $\frac{\Lambda}{y}$ that arises under pattern allocation. It possesses the same phase character as the MAP, and if y is an integer and the MAP is as regular as possible, both arrival processes have the same Erlang interarrival times.
The Gamma arrival process can be viewed as the ideal MAP; if from an infinite reservoir of customers a_i customers out of every M have to be routed to queue i such that the interarrival times of the customers are i.i.d. and the sum of a_i interarrival times has an Erlang(Λ, M) distribution (the length of the arrival pattern), then the interarrival time of a customer has a Gamma$(\Lambda, \frac{M}{a_i})$ distribution. This implies that a Gamma$(\Lambda, \frac{M}{a_i})$ arrival process is more regular than the MAP with the same arrival intensity. Hence we expect the mean waiting times in the MAP/G/1 queue to be bounded from below by the mean waiting times in the corresponding Gamma/G/1 queue. Again, such a statement is

hard to prove, except for the case of exponential servers, for which the proof readily follows from the results in Hajek[56].

Unfortunately, the expression for the mean waiting times of customers in a Gamma/G/1 queue (cf. Cohen[34]) is too complicated to be useful in an optimization procedure, and hence we have to resort to more simple approximate expressions for this mean waiting time.

The next part of this section is devoted to the actual determination of the allocation fractions. For the mean waiting times in a Gamma/G/1 queue we apply the two-moment approximation proposed by Krämer and Langenbach-Belz (KLB; cf. Krämer & Langenbach-Belz[66]) for GI/G/1 queues:

$$\mathrm{EW} = \frac{\rho\beta}{2(1-\rho)}\left[c_a^2 + c_s^2\right] exp\left\{-\frac{2(1-\rho)}{3\rho}\frac{(1-c_a^2)^2}{c_a^2+c_s^2}\right\} \tag{6.9}$$

in which β is the mean service time, ρ is the load of the queue, and c_a^2 and c_s^2 denote the squared coefficient of variation (variance divided by squared mean) of the arrival time and service time distributions respectively. As can be readily verified, (6.9) is exact if the arrival process is Poisson.
A number of approximations for the mean waiting time in the GI/G/1 queue are compared in Shanthikumar and Buzacott[89]. From [89] it appears that the Marshall approximation can be a good alternative for the KLB approximation. For a Gamma(Λ,y) process the arrival rate λ is given by $\frac{\Lambda}{y}$ and $c_a^2 = \frac{1}{y}$, and for the Gamma/G/1 queue (6.9) thus becomes

$$\mathrm{EW} = \frac{\Lambda\beta^2}{2(y-\Lambda\beta)}\left[\frac{1}{y} + \frac{\beta^{(2)}-\beta^2}{\beta^2}\right] exp\left\{-\frac{2(y-\Lambda\beta)}{3\Lambda\beta}\frac{(1-\frac{1}{y})^2}{\frac{1}{y}-1+\frac{\beta^{(2)}}{\beta^2}}\right\} \tag{6.10}$$

With (6.10) we can formulate the Mathematical Programming Problem for the Gamma approximation procedure. For objective function (6.2), substituting $\alpha_i := f_i = \frac{\lambda_i}{\Lambda} = \frac{1}{y_i}$ we find
GA1:

$$\begin{aligned}
\min \quad & \sum_{i=1}^{N} \frac{\Lambda\alpha_i^2\beta_i^2 C_i}{2(1-\alpha_i\Lambda\beta_i)}\left[\alpha_i + \frac{\beta_i^{(2)}-\beta_i^2}{\beta_i^2}\right] \times \\
& \qquad exp\left\{-\frac{2(1-\alpha_i\Lambda\beta_i)}{3\alpha_i\Lambda\beta_i}\frac{(1-\alpha_i)^2}{\alpha_i - 1 + (\beta_i^{(2)}/\beta_i^2)}\right\} \qquad (6.11) \\
s.t. \quad & \sum_{i=1}^{N}\alpha_i = 1, \\
& 0 \le \alpha_i < \frac{1}{\Lambda\beta_i}, \quad i = 1,\ldots,N.
\end{aligned}$$

Problem GA1 has the same structure as PA1 in Section 6.2, and hence it can easily be solved numerically with the algorithm presented in Appendix 6.A.

Note that $\lim_{\{\epsilon \downarrow 0\}} \frac{\partial \mathrm{E}\mathbf{W}_i}{\partial \alpha_i} |_{\alpha_i=\epsilon} = 0$. Hence not only the optimal assignment frequencies resulting from GA1 are all greater than 0, but this would also be the case for objective function $\sum_i C_i \mathrm{E}\mathbf{W}_i$. The latter was not always the case for the optimal probabilistic allocation.
Earlier in this section we stated that the mean waiting times in the MAP/G/1 queue are bounded from below by the mean waiting times in the corresponding Gamma/G/1 queue. Consequently, the solution of GA1 provides an approximate lower bound for the mean waiting costs under the optimal allocation pattern.

Remark 6.4.1
An important observation is that for the optimal pattern allocation more load than under probabilistic allocation is assigned to the queues with relatively high first moment of the service time distribution. This property was first reported by Agrawala & Tripathi[3].
The explanation of this property is that the effect of regularizing is stronger for the queues with relatively small assignment probabilities. For example: consider a traffic allocation problem with two queues for which the optimal probabilistic assignment fractions are $p_1^* = \frac{8}{9}$ and $p_2^* = \frac{1}{9}$. Then the MAP for the first queue would approximately be equal to a Poisson arrival process with arrival intensity $\frac{8}{9}\Lambda$, hence the switch from probabilistic to pattern allocation would not cause great changes in the arrival process at queue 1. However, for queue 2, switching from probabilistic to pattern allocation also changes the arrival process at queue 2 from a Poisson($\frac{1}{9}\Lambda$) into an Erlang($\Lambda, 9$) arrival process. The switch from probabilistic to pattern allocation has a more regularizing effect on queue 2 than on queue 1, and hence the relative decrement of the mean waiting times is larger for queue 2 than for queue 1.
This example also shows why the Gamma approximation procedure has a better performance than the approximations obtained from probabilistic allocation: *the Gamma arrival process better captures the influence of assignment fractions on the degree of regularization.* □

Remark 6.4.2
Elaborating on Remark 6.4.1, we expect that the effect of a transition from probabilistic allocation to pattern allocation will be stronger when the assignment fractions are closer to each other. In that situation all servers will profit from regularization. An interesting conclusion is that for the case of non-identical service rates, comparing the Gamma approximation procedure with probabilistic patterns, the difference in patterns is in particular pronounced for low system loads. When the load increases both methods will lead to allocation fractions close to the capacities of the queues, but for low load probabilistic allocation tends to assign many more customers to the faster queue than the Gamma approximation. Another interesting conclusion is that when the number of servers increases, the effect of regularizing becomes stronger. For example, consider the case of k identical servers with service rate 1 and

$\Lambda = \rho k$, $\rho < 1$ and all servers receiving the same fraction $\frac{1}{k}$ of the arrivals. The allocation pattern based on these fractions leads to k Erlang$(\rho k, k)$ arrival processes. Stoyan[93] example 1.5.1(e) shows that Erlang$(\rho(k+1), k+1) \leq_c$ Erlang$(\rho k, k)$. Hence the value of the objective function decreases when k increases. Note that for $k \to \infty$ the arrival processes at the queues all converge towards a deterministic arrival process. □

Remark 6.4.3
In Hordijk et al.[58] the traffic allocation problem is studied for the case of non-identical exponential servers. The method in [58] is based on Markov decision theory and results in a value-iteration algorithm. The algorithm treats both optimization steps (finding good fractions and building a pattern) simultaneously. In [58] the performance of their algorithm is compared with our results, which showed that the objective function, as well as the allocation fractions and final patterns are quite similar in both methods. Regarding the quality of the allocation patterns, it does not appear that one the methods should receive clear preference. □

We conclude this section with a remark concerning the validity of our optimization procedure. In this remark we also reveal some problems which occur in the second step of the procedure, where allocation frequencies are to be translated into patterns.

Remark 6.4.4
Assignment fractions p_i do not determine a unique allocation pattern. Firstly, as explained in the previous section, the p_i's can be irrational, so in general a finite pattern with corresponding assignment fractions p_i for $i = 1, \ldots, N$ does not exist. And secondly, even if there exist integer numbers a_i such that $p_i = \frac{a_i}{\sum_j a_j}$ for $i = 1, \ldots, N$, the orders in which the queue numbers can be placed in a pattern are numerous.
However, the natural requirement that the arrival processes should be as regular as possible causes a set of allocation fractions to lead to a more or less uniquely determined allocation pattern. Let us now consider the translation of assignment fractions into patterns.
First of all $(a_1, \ldots, a_N)$ are defined in the following way. For all $\epsilon > 0$, there exists an integer $\bar{m}$ such that $\bar{m} := \min\{m > N | \parallel (p_i m - [p_i m])/[p_i m] \parallel < \epsilon, \frac{[p_i m]}{m}\Lambda < \beta_i, \text{ for all } p_i > 0\}$. Let $a_i := [p_i \bar{m}]$, $i = 1, \ldots, N$. Hence, the a_i's are uniquely defined by a chosen $\epsilon > 0$. Note that the value of ϵ has a strong influence on the length of the pattern.
Secondly, in Section 6.3 we saw that the mean waiting time decreases with increasing regularity of the arrival process; so given numbers a_i, $i = 1, \ldots, N$, we try to construct an allocation pattern in which the occurrences of the queue numbers are as uniformly distributed as possible. In this way, given $\epsilon > 0$, assignment fractions p_i correspond to a more or less uniquely determined allocation pattern.

This does not imply monotonicity of the waiting times as a function of the assignment fractions. For certain values p_i, $i = 1, \ldots, N$, placing the queue numbers into a pattern in a uniformly distributed way can be rather difficult, whereas after slightly altering the frequencies, a much more regular pattern would arise. This property also has consequences for the value of the objective function. This contrasts with probabilistic allocation where, given the assignment fractions, the model is equivalent to N independent M/G/1 queues.
The actual construction of an allocation pattern is an interesting combinatorial problem, for which we present a heuristic in Appendix 6.B. Here the main problem is that the interests of the queues interfere, i.e. we try to make the arrival process as regular as possible for all queues simultaneously. An example of such interference is with $N = 3$, $a_1 = 1$, $a_2 = 2$, $a_3 = 3$. The reader can easily check that there exists no pattern of length 6 in which the arrival process at all three queues is a renewal process.
In general *the* optimal allocation pattern can not be determined, hence it is not to be expected that the optimal assignment frequencies are determined by applying the Gamma approximation procedure. However, our numerical experience indicates that this procedure results in a pattern under which the objective function is close to the approximate lower bound for the optimal arrival pattern, this lower bound being the value of the solution of GA1.

6.5 Numerical results

In this section we present some numerical results. We compare various allocation policies and also discuss the quality of the Gamma approximation procedure. We show five instances for the case of two servers and three instances for the case of three servers in parallel. For each instance the objective is to minimize the mean waiting time of a customer. As a function of the load of the total system, we present absolute and relative values of the objective function for various optimized allocation policies and for the solutions of the mathematical programs.

Description of numerical instances and presented results
In Figures 6.1-6.8 we show numerical results for 8 instances. We have considered the problem of minimizing the mean waiting time of an arbitrary customer (taking $f_i(P) = p_i$ and $C_i = 1$ in (6.2), $i = 1, \ldots, N$). For each instance we have optimized various allocation policies for system loads ρ that we increased from 0.05 to 0.95 with steps of 0.05. The system load we defined as $\rho := \Lambda(\sum_i \frac{1}{\beta_i})^{-1}$, i.e., the offered traffic to the system divided by the total service capacity of the system.
Figures 6.1-6.5 concern the case of two servers, 6.6-6.8 the case of three servers. We have considered three types of service time distributions: Exponential, Erlang 2 and Hyper-Exponential. For the Hyper-Exponential distribution the coefficient of variation is 2. In cases 6.1-6.3 and 6.6-6.8 the servers are of the same type, but differ in service rates. For cases 6.4 and 6.5 the servers are not

of the same type; in case 6.4 both servers have identical service rate, in case 6.5 the rates are different.
For each instance two figures are presented, one displaying absolute value of the objective function, the other showing this value relative to the value for optimal probabilistic allocation. The abbreviations in the figures stand for:

prob := mean waiting times, under the optimal probabilistic allocation.

prop := pattern that is based on the optimal probabilistic pattern.

klbp := pattern obtained via the gamma approximation procedure, using the KLB approximation for the Gamma/G/1 queue.

klbb := approximate lower bound for the mean waiting times under the optimal allocation pattern (see Section 6.4).

lb := strict lower bound for the mean waiting times under the optimal allocation pattern. This is for the case of exponential servers (see Section 6.4).

jlw := mean waiting times under the dynamic policy that allocates a customer to the queue with the least waiting time. These are simulation results.

Comparing probabilistic and pattern allocation
In Section 6.3 we argued that regularizing arrival processes leads to lower mean waiting times. Also, in Conjecture 6.3.1 we stated that for each M/G/1 queue there exists a MAP/G/1 queue with lower mean waiting times, where the Poisson arrival process has intensity $p\Lambda$, $p < 1$, and the MAP has the same arrival rate and phase intensity Λ. We concluded that pattern allocation leads to lower mean waiting times than probabilistic allocation. This conclusion is supported by our numerical results. In all cases considered, the allocation pattern based on the optimal probabilistic allocation fractions (curves **prop** in fig. 6.1-6.8) performs better than the probabilistic allocation itself (**prob**). The relative differences vary from 7 to 40% for low loads up to about 40% for high load, except for the case of non-identical servers with identical rate, where the difference for low loads is even 50%. Also, the effect of a transition from probabilistic to pattern allocation is stronger for the case of three servers. All observations are illuminated by Remarks 6.4.1 and 6.4.2.
Figures 6.1-6.3 and 6.6-6.8 suggest that for identical servers the effect of a transition from probabilistic to pattern allocation is stronger when the service time distribution has a smaller coefficient of variation.
The non-smoothness of the curves (**prop**) in Figures 6.6-6.8 is caused by the way the patterns were constructed in our numerical experiments. Due to pattern length limitations, imposed by computer capacity, some inaccuracies occur. The assigned fractions in the pattern are for some values of ρ closer to the optimal probabilistic allocation than for others. It is interesting to see that when the deviation results in - relatively speaking - more (less) load at a slower server, this decreases (increases) the value of the objective function.

In Remark 6.4.1 this observation is explained. The effect is most pronounced for ρ=0.1, where the constructed allocation pattern actually assigns no customers to the slowest server.

Comparing the Gamma approximation procedure with pattern allocation based on optimal probabilistic allocation

In Section 6.4 we concluded that the Gamma approximation procedure would lead to better allocation patterns than probabilistic allocation because the Gamma arrival process better captures the behaviour of the MAP than the Poisson arrival process. This conclusion is supported by the numerical results. The difference between objective functions for Gamma approximation based patterns(**klbp**) and probabilistically based patterns(**prop**) is larger for lower loads than for higher loads. The difference ranges from 0% to 45%.

Optimal pattern allocation

Finally we turn to the questions (i) how good is pattern allocation compared to the best policy, and (ii) how close is the pattern obtained with the Gamma approximation procedure to the optimal allocation pattern?

Concerning (i), it is very hard to determine the optimal - probably a dynamic - allocation policy. Hence we have considered a dynamic policy which is expected to perform better than most policies, and considerably better than the static policies. This dynamic policy operates under complete knowledge of the system at the moment of arrival and sends each customer to the queue with smallest waiting time. Our claim that this is a nearly optimal allocation policy, is based on the fact that this policy uses all information available at moments of arrival and seems to use this information in a very sensible way. In the figures one can see that this dynamic policy(**jlw**) performs from 40% up to about 95% better than probabilistic allocation. The difference with the optimal Gamma approximated pattern ranges from 20% up to 45%.

Concerning question (ii), we know that it is very hard to determine *the* optimal allocation pattern. However, in an indirect way we are able to make a statement about the quality of the Gamma approximation procedure. In Section 6.4 we stated that waiting times in a MAP/G/1 queue are bounded from below by the waiting times in the corresponding Gamma/G/1 queue. Numerical experience shows that the approximation of the mean waiting times in the MAP/G/1 queue, using the KLB formula for waiting times in the corresponding Gamma/G/1 queue, is fairly accurate. Hence the value of the objective function for the solution of mathematical program GA1(**klbb**) is an approximate lower bound for the optimal allocation pattern.

We also notice that, in particular for high loads, this value reasonably accurately approximates the mean waiting times under the allocation pattern that is constructed from the solution of GA1.

So GA1 provides an approximate lower bound for the mean waiting times under the optimal allocation pattern, as well as an accurate approximation of the mean waiting times under the allocation pattern that is based on the solution of GA1. We conclude that the Gamma approximation procedure provides us with a nearly optimal allocation pattern.

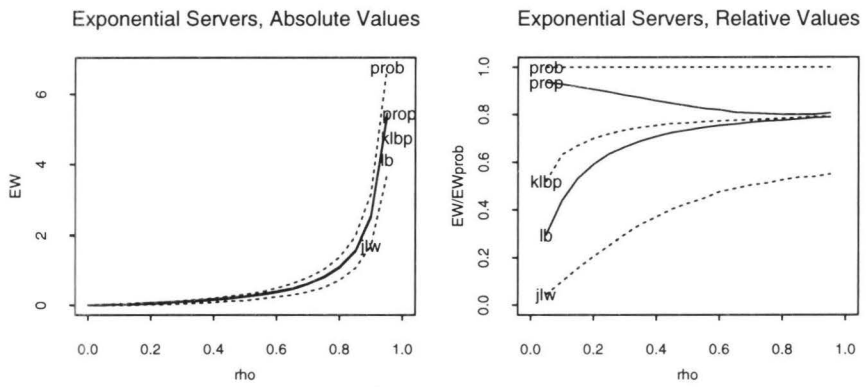

FIGURE 6.1: TWO EXPONENTIAL SERVERS, WITH SERVICE RATE 1 AND 4 RESPECTIVELY.

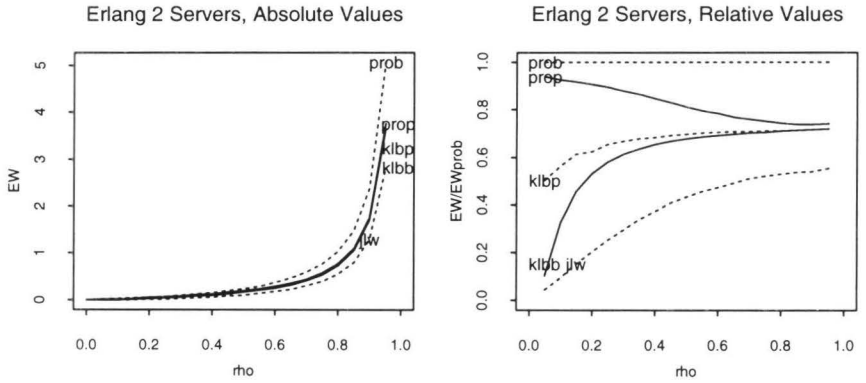

FIGURE 6.2: TWO ERLANG 2 SERVERS, WITH SERVICE RATE 1 AND 4 RESPECTIVELY.

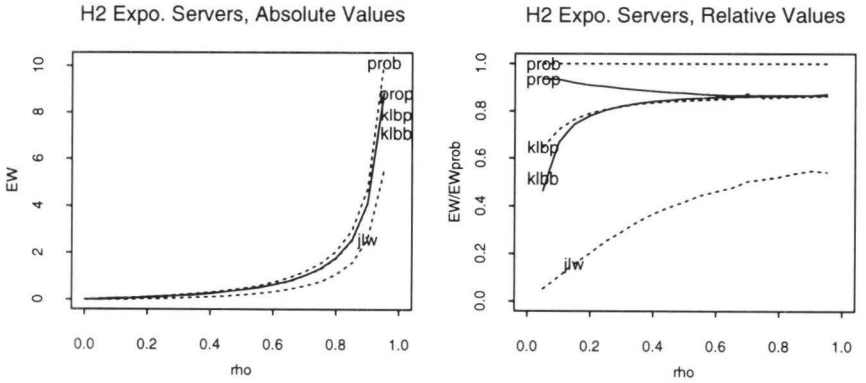

FIGURE 6.3: TWO HYPER-EXPONENTIAL SERVERS, WITH SERVICE RATE 1 AND 4 RESPECTIVELY. THE SERVICE TIME OF A CUSTOMER IS WITH PROBABILITY $q = \frac{1}{3}$ EXPONENTIALLY DISTRIBUTED WITH PARAMETER $\frac{1}{2}\mu$, AND HAS WITH PROBABILITY $1 - q = \frac{2}{3}$ AN EXPONENTIAL DISTRIBUTION WITH PARAMETER 2μ. IN THIS WAY THE SERVICE RATE IS μ AND THE COEFFICIENT OF VARIATION IS 2.

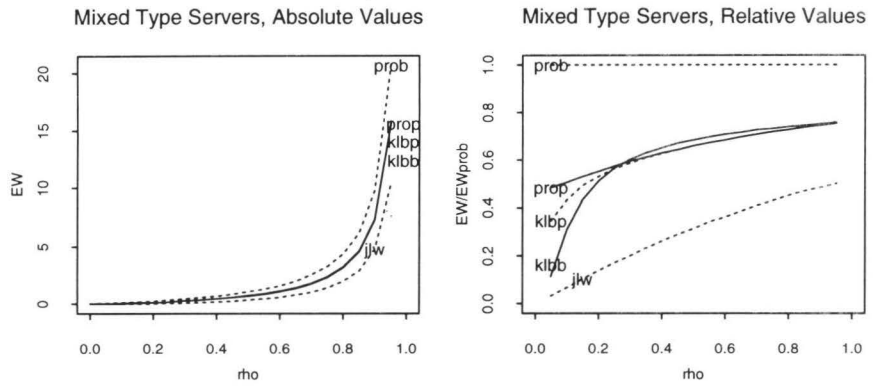

FIGURE 6.4: TWO SERVERS WITH SERVICE RATE 1. THE FIRST SERVER HAS ERLANG 2 DISTRIBUTED SERVICE TIMES, THE SECOND SERVER HAS HYPER-EXPONENTIALLY DISTRIBUTED SERVICE TIMES AS DESCRIBED BY FIGURE 6.3.

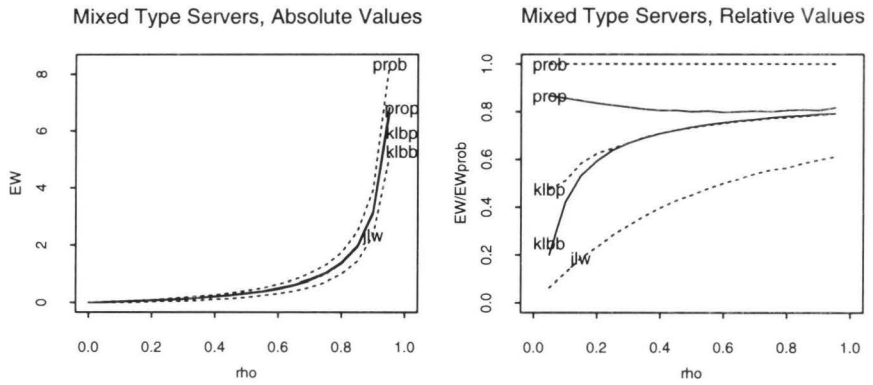

FIGURE 6.5: TWO SERVERS. THE FIRST SERVER HAS ERLANG 2 DISTRIBUTED SERVICE TIMES WITH RATE 1. THE SECOND SERVER HAS HYPER-EXPONENTIALLY DISTRIBUTED SERVICE TIME AS DESCRIBED BY FIGURE 6.3.

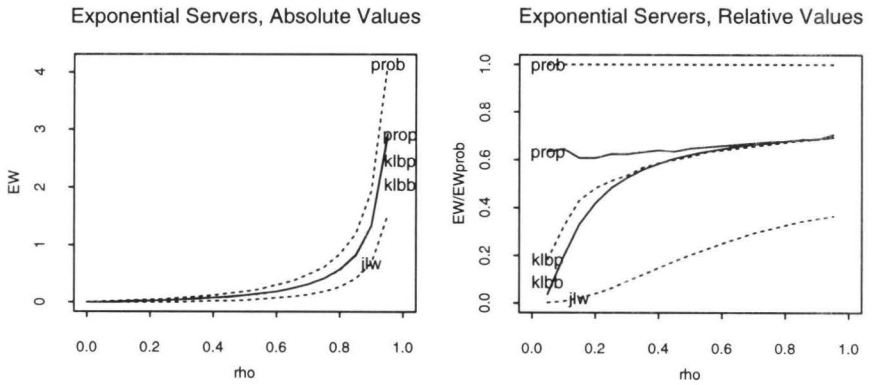

FIGURE 6.6: THREE EXPONENTIAL SERVERS, WITH SERVICE RATES 1,4 AND 7 RESPECTIVELY.

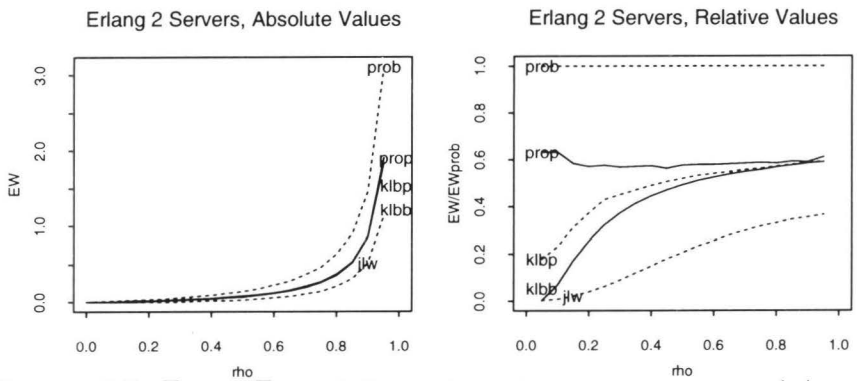

FIGURE 6.7: THREE ERLANG 2 SERVERS, WITH SERVICE RATES 1,4 AND 7 RESPECTIVELY.

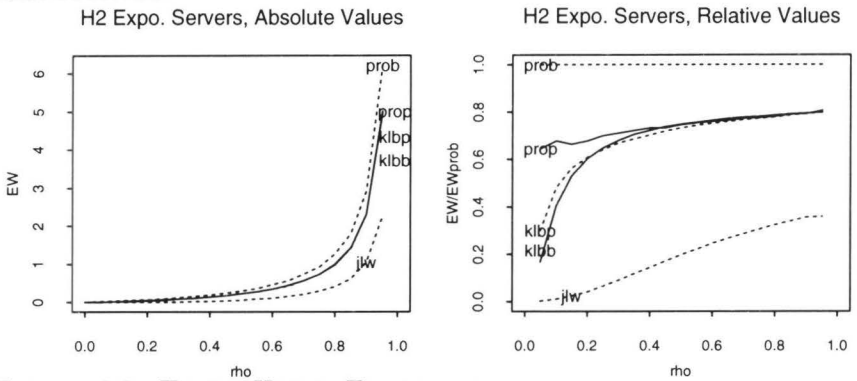

FIGURE 6.8: THREE HYPER-EXPONENTIAL SERVERS, WITH SERVICE RATES 1,4 AND 7 RESPECTIVELY. SERVICE TIME DISTRIBUTIONS AS DESCRIBED BY FIGURE 6.3.

6.6 EXTENSIONS OF THE TRAFFIC ALLOCATION PROBLEM

In this section we briefly discuss the traffic allocation problem for three extensions of the model that was discussed in the previous sections. For all extensions, the traffic allocation problem can be approached in a similar way as the original problem.

First we discuss the case of general arrival processes. The second model considers the situation in which one has to allocate a Poisson arrival stream to N queues, where each queue already receives a Poisson arrival stream. This problem is known as the traffic allocation problem with dedicated arrival streams. The third model under consideration is the allocation problem with multiple server stations.

In Section 6.2 we argued that regularizing arrival streams tends to reduce the mean waiting times. Based on similar intuitive arguments, we now make the same conjecture for the extended allocation problems, realizing that proving the same statements for the extended models can only be harder than for the original allocation problem.

According to this assumption, the customers are allocated using an allocation pattern rather than assigning them to the queues probabilistically.

Allocation for the case of general arrival processes

In many traffic allocation situations, the arrival process will not be a Poisson process. We believe that our approach can be extended for such situations. The first step would be to approximate the arrival process by a Gamma process, fitting the parameters $\hat{\Lambda}$ and c_a^2, the arrival rate and squared coefficient of variation respectively. Hence the interarrival time has the LST $\left(\frac{\Lambda}{\Lambda+\omega}\right)^{\frac{1}{c_a^2}}$, with $\Lambda = \frac{\hat{\Lambda}}{c_a^2}$. Subsequently, applying pattern allocation would result in sending one out of each y_i customers to queue i. Of course y_i need not be integral. However, in an ideal pattern allocation, the interarrival time at queue i will have an LST $\left(\frac{\Lambda}{\Lambda+\omega}\right)^{\frac{y_i}{c_a^2}}$, and hence will be Gamma distributed. We thus can again approximate the optimal assignment frequencies by solving a mathematical programming problem of the form of optimization problem GA1, with constraint $\sum_{i=1}^{N} \alpha_i = 1$ replaced by $\sum_{i=1}^{N} \alpha_i = c_a^2$.

The allocation problem with dedicated arrival streams

A Poisson arrival stream with intensity Λ has to be allocated to N queues Q_i, each queue already receiving a Poisson arrival stream with intensity $\lambda_i^d \geq 0$, $i = 1, \ldots, N$. Note that the original problem returns if $\lambda_i^d = 0$ for $i = 1, \ldots, N$. For the allocation problem with dedicated arrival streams, for the case of exponential servers, Ni & Hwang[82] optimize the probabilistic allocation policy with the mean sojourn time of a customer as objective function.

The benefit of using pattern allocation is less substantial than in the original allocation problem without dedicated arrivals, because allocated arrivals from the additional Poisson arrival stream (forming a MAP) join in with the arrivals from the dedicated Poisson (λ_i^d) arrival stream. So the arrival processes at the queues are not as regular as the MAP in the original problem, they are the sum of such a MAP and a Poisson arrival process. Although the sum of two MAP's is also a MAP, and the Poisson process is just a special MAP, the arrival processes at the queues are hard to approximate by any GI arrival process, in particular the Gamma arrival process.

As a result, we try to approximate the optimal assignment fractions for the allocation pattern with probabilistic allocation or with the Gamma optimization procedure, depending on the ratio between the sum of the dedicated arrival rates and the extra arrival rate. For example: if the sum of the dedicated arrival rates is large compared to the rate of the extra arrival stream, then the resulting arrival processes will resemble a Poisson process more than a MAP, hence it makes more sense to use the assignment fractions from probabilistic allocation for the allocation pattern.

Below the Mathematical Programming Problem is formulated for the probabilistic allocation policy; for the Gamma optimization procedure the formulation is quite similar.
DA1:

$$\min \quad \sum_{i=1}^{N} f_i C_i \frac{(\lambda_i + \lambda_i^d)\beta_i^{(2)}}{2(1-(\lambda_i+\lambda_i^d)\beta_i)} \tag{6.12}$$

$$s.t. \quad \sum_{i=1}^{N} \lambda_i = \Lambda,$$

$$\lambda_i + \lambda_i^d < \frac{1}{\beta_i}, \ i = 1, \ldots, N,$$

$$\lambda_i \geq 0, \ i = 1, \ldots, N.$$

The structure of DA1 is similar to the earlier presented Mathematical Programming Problems PA1 and GA1. Again, the solution of DA1 can easily be determined using the allocation algorithm in Appendix 6.A.
Note that under probabilistic allocation, the original probabilistic allocation problem reappears with the additional constraints that the arrival rate at Q_i should be at least λ_i^d, $i = 1, \ldots, N$.
We finish this discussion of load balancing with dedicated arrival streams by mentioning two references on this topic.
Bonomi & Kumar[16] discuss an adaptive probabilistic allocation policy for the case of exponential and the case of identical servers with objective function the mean sojourn time of a customer. They consider the situation where not all system parameters are known, or where some of the parameters may change from time to time. Their main concern is the speed of convergence of the allocation policy towards the optimal assignment probabilities.
Ross & Yao[88] consider the following N server model. At server i a set S_i of customer types arrives according to Poisson processes with arrival intensities λ_{ij}, $j \in S_i$, $i = 1, \ldots, N$. The $j-th$ arrival stream at server i has service time distribution $B_{ij}(\cdot)$, $j \in S_i$, $i = 1, \ldots, N$. Furthermore, each server generates additional customers, according to a Poisson process, which may be routed to one of the other servers. The service time of such a customer at server i has distribution $B_i(\cdot)$. The aim in Ross & Yao[88] is to find the probabilistic allocation policy that minimizes the sum of the mean sojourn time and some rerouting delay of a customer from the additional arrival process, under the constraints that the mean sojourn time of the $j-th$ dedicated customer stream at server i is less than or equal to α_{ij}, $j \in S_i$, $i = 1, \ldots, N$. Ross & Yao[88] allow local priority scheduling of customer types, which also involves the additional customers. The essential problem is to derive an expression for $\mathrm{E}\mathbf{R}_i$, the mean sojourn time at server i of an additional customer when local priority scheduling of customers is allowed. Using matroid theory Ross & Yao[88] prove that $x_i \mathrm{E}\mathbf{R}_i$ is a convex function in x_i, where x_i denotes the additional load assigned to

server i. The remaining problem, determining the optimal assignment vector $x^* = (x_1^*, \ldots, x_N^*)$, proceeds in a way that is similar to solving a common RAP. It might be interesting to use the resulting assignment vector for determining a good pattern allocation.

Allocation for the case of Multiple Server Queues

In this model, a Poisson arrival stream with intensity Λ has to be allocated over N multiple server queues, where the number of servers at $Q_1, \ldots, Q_N$ are $s_1, \ldots, s_N$.

Except for a few special cases, no explicit expressions for the mean waiting times in GI/G/s or MAP/G/s queueing systems are available. Hence in this model, for both probabilistic assignment and pattern allocation an optimal allocation can not be determined analytically.

We again assume that regularizing the arrival streams decreases mean waiting times, and again we expect to obtain better allocation fractions using the Gamma/G/s_i approximation for the MAP/G/s_i queue than when using the M/G/s_i approximation.

Using this assumption, the path towards an allocation pattern is reasonably straightforward; by choosing a suitable strictly convex approximation for the mean waiting times in a Gamma/G/s queue one can formulate a Mathematical Programming Problem which possesses all the required properties for applying the algorithm in Appendix 6.A. Bitran & Dasu[13] and Buzacott & Shanthikumar[27] mention several approximations for the mean waiting time in the GI/G/s queue.

Remark 6.6.1

In multi-processor computer networks a job (customer) can often be decomposed into a number of sub-tasks. The sequential and parallel precedence relations for the order of service of these sub-tasks are described by means of a a *task graph*. This special structure for the service requests of customers gives rise to interesting single-server and multiple-server queueing models.

However, these models are hard to analyse. One reason being that in general it is difficult to explicitly describe the distribution of the total service time of a job (the total service time associated with a task graph). In the multiple-server case the analysis looks even more complicated since two or more jobs might be in service at the same time, this can affect the scheduling of sub-tasks on the processors. Finally, the task graphs associated to the service request of a customer might not be identical for all customers but instead be randomly drawn from a set of task graphs.

One model that can be analysed is the single-queue multiple-server system with Poisson arrivals, in which all sub-tasks are identically exponentially distributed and where all tasks can be processed independently from one another; this leads to the $M/M/s$ queue with batch arrivals.

In the traffic allocation problem that is connected to this model, customers arriving according to a Poisson process have to be assigned to one of a group of N multi-server queues; under pattern allocation this gives rise to N parallel

$MAP/M/s$ queues with batch arrivals, in the common framework of notation these queues are denoted as $BMAP/M/s$ queues. The $BMAP/M/s$ queueing model will be examined in the next chapter. Again, due to the matrix structure of the analytical results, solving the allocation problem directly using these results seems inefficient, hence suitable approximations for performance measures have to be found. □

APPENDICES

6.A TRAFFIC ALLOCATION ALGORITHM

In this appendix we present an algorithm to solve the Mathematical Programming Problems that where defined in Sections 6.2 and 6.4.

The Mathematical Programming Problems for the traffic allocation problem fit into the class of Resource Allocation Problems (RAP) with a separable, strictly convex continuous objective function. In chapter 2 of Ibaraki & Katoh[59] this class is studied. They also present several algorithms from which we have derived the algorithm for the traffic allocation problem.
The basic formulation of a RAP as studied in chapter 2 of Ibaraki & Katoh[59] is:

RAP:

$$\min \quad \sum_{i=1}^{N} T_i(x_i) \tag{6.13}$$

$$s.t. \quad \sum_{i=1}^{N} x_i = C, \tag{6.14}$$

$$0 \leq x_i, \quad i = 1, \ldots, N. \tag{6.15}$$

Here $T_i(\cdot)$ is a strictly convex function, $i = 1, \ldots, N$. For the traffic allocation problem, the x_i's have upper bounds u_i, which follow from the stability conditions of the queues. Hence, we obtain the following Mathematical Programming Problem:

MP:

$$\min \quad \sum_{i=1}^{N} T_i(x_i)$$

$$s.t. \quad \sum_{i=1}^{N} x_i = C,$$

$$0 \leq x_i < u_i, \quad i = 1, \ldots, N.$$

For the traffic allocation problem with dedicated arrival streams (cf. Section 4), the x_i's have lower bounds $l_i \geq 0$. However, using a translation of the variables,

we can always formulate a Mathematical Programming Problem which has the form of MP.

Using the separability and convexity of the objective function, the first order Kuhn-Tucker constraints lead to the following lemma:

Lemma 6.A.1
Provided that a feasible solution exists, the unique optimum $x^* = (x_1^*, \ldots, x_N^*)$ of MP fulfills the following set of equations:

$$\sum_{i=1}^{N} x_i = C, \tag{6.16}$$

$$\text{if } x_i > 0 \text{ then } T_i'(x_i) = \delta, \tag{6.17}$$

$$\text{if } x_i = 0 \text{ then } T_i'(x_i) \geq \delta, \tag{6.18}$$

where δ is the Lagrange multiplier connected to constraint (6.14).

Proof
Knowing that $T_i(\cdot)$ is strictly convex and that for the optimum $x_i^* < u_i$, $i = 1, \ldots, N$, these equations follow directly from the first order Kuhn-Tucker constraints for the Lagrange relaxation of MP. □

The following lemma is an extension of lemma 2.2.1 in Ibaraki & Katoh[59].

Lemma 6.A.2
Let T_i be strictly convex and continuously differentiable over the interval $[0, u_i)$ and $T_i'(x_i) \to \infty$ as $x_i \to u_i$, $i = 1, \ldots, N$, and let the indices of T_i be arranged such that

$$T_1'(0) \leq T_2'(0) \leq \ldots \leq T_N'(0).$$

If $x^* = (x_1^*, \ldots, x_N^*)$ is an optimal solution of MP, then there exists an index $i^* \in \{1, \ldots, N\}$ such that

$$\begin{aligned} x_i^* &> 0, \quad i = 1, \ldots, i^*, \\ x_i^* &= 0, \quad i = i^* + 1, \ldots, N. \end{aligned}$$

Proof
Since x^* is optimal, there exists a δ for which (6.17) and (6.18) hold. Let i^* be the smallest index such that $x_{i^*+1}^* = 0$. Then, from Lemma 6.A.1, $T_{i^*+1}'(0) \geq \delta$. Suppose $x_i^* > 0$ for a certain $i > i^*$. Then, from the strict convexity of the T_i's and (6.18), it follows that $T_i'(x_i^*) > T_i'(0) \geq T_{i^*}'(0) \geq \delta$, which is in contradiction with (6.17). Thus $x_i^* = 0$ for $i = i^* + 1, \ldots, N$. □

Below we present an algorithm for solving MP, based on Lemma 6.A.1 and 6.A.2. The algorithm consists of two phases; first i^* is computed, subsequently MP is solved with a reduced set of variables $x_1, \ldots, x_{i^*}$, putting $x_{i^*+1}, \ldots, x_N$ equal to zero.

Algorithm 6.A.1:

Phase 1:

0. Order the terms T_i such that $T_1'(0) \leq T_2'(0) \leq \ldots \leq T_N'(0)$.
1. Compute $i_0 = \min\{i \mid \sum_{j=1}^{i} u_j > C\}$. (from Lemma 6.A.1 and (6.14) it follows that i_0 is the minimal number of positive x_i's that is required for a feasible solution.)
2. $k := i_0$.
3. Compute x_i for $i = 1, \ldots, k$, such that $T_i'(x_i) = T_{k+1}'(0)$ (binary search, or Newton's method).
4. if $\sum_{i=1}^{k} x_i > C$ then GOTO 6.
5. $k := k + 1$, if k =N then GOTO 6 else GOTO 3.
6. $i^* = k$.

Phase 2:

7. Compute δ and $x_i(\delta)$ such that $T_i'(x_i(\delta)) = \delta$, $i = 1, \ldots, i^*$, and $\sum_{i=1}^{i^*} x_i = C$. This can be done by a binary search for δ in $[T_{i^*}'(0), T_{i^*+1}'(0)]$. (if $i^* = N$ then set $T_{i^*+1}'(0)$ equal to $\max_i T_i(y_i)$ for an arbitrary feasible allocation y.)
8. $x_i^* = x_i(\delta)$, $i = 1, \ldots, i^*$, $\quad x_i^* = 0$, $i = i^* + 1, \ldots, N$.
9. STOP.

(end algorithm 6.A.1) □

6.B Constructing the allocation pattern

In this appendix we discuss the problem of constructing an allocation pattern from a given set $(p_1, \ldots, p_N)$ of allocation fractions. First of all, these frequencies are translated into the vector $(a_1, \ldots, a_N)$, in which a_i is the number of occurrence of index i in the pattern. The integers a_i are computed by defining $\bar{m} = \min\{m > N \mid \parallel (p_i m - [p_i m])/[p_i m] \parallel < \epsilon, \ \frac{[p_i m]}{m}\Lambda < \beta_i, \ \text{for all } p_i > 0\}$ and taking $a_i = [p_i \bar{m}]$, $i = 1, \ldots, N$. Note that the choice of ϵ has a strong influence on the length of the pattern. After this translation there remains the problem of determining an allocation pattern, such that the number of arriving customers at the routing point between two consecutive allocations to queue i is as constant as possible. Moreover, one has to achieve this for all queues simultaneously. In various optimization problems this combinatorial cyclic scheduling problem has been encountered. Itai & Rosberg[61] suggest the so-called Golden Ratio method for a cyclic scheduling problem that arises in the access control for a multi-access channel. Boxma et al.[23] study a polling model in which a server visits the queues according to a polling table. For this more or less dual

problem of the traffic allocation problem they follow an optimization procedure which is similar to our approach for the traffic allocation problem. First good visit frequencies are computed for a polling model in which the server chooses his next queue probabilistically, subsequently a polling table based on these frequencies is constructed, using the Golden Ratio method.
The combinatorial complexity of the cyclic scheduling problem is yet undetermined. However, it seems to be a hard problem; it can be translated to known NP-hard problems, although with special structures, but those special structures do not seem to reduce the problem to a polynomially solvable one.
In this appendix a heuristic based on the paper by Hajek[56] on extremal splittings of point processes is presented. This heuristic is an alternative for the Golden Ratio policy as described in Itai & Rosberg[61].
First some notation and a mathematical criterion for optimality are introduced. A pattern S is defined by $S := \{s_1, \ldots, s_k\}$ in which $k := \mid S \mid$ is the length of S. Let S_0 be the class of patterns of length $M = \sum_i [f_i \bar{m}]$ in which index i occurs exactly a_i times. In the rest of this appendix we assume that $a_1 \geq a_2 \geq \ldots \geq a_N$ and we set N equal to the number of queues with $a_i > 0$. Under the allocation pattern $S \in S_0$ the interarrival times at queue i form a repeated sequence of a_i Erlang$(\Lambda, d_{i_j}(S))$ distributed variables, $j = 1, \ldots, a_i$. The 'distances' $d_{i_j}(S)$ are the numbers of arriving customers at the routing point between two consecutive allocations to queue i.
Next, the problem is to determine $S^* = \text{argmax}_{S \in S_0} V(S)$, in which $V(S) = \{\sum_i a_i \sum_{j=1}^{a_i} d_{i_j}^2(S)\}$. The objective function $V(\cdot)$ tries to capture the notion of even spreading in the pattern by a kind of second moment function. The weights a_i have been chosen such that in the optimal allocation, i.e. $d_{i_j} = \frac{M}{a_i}$, $j = 1, \ldots, a_i$, $i = 1, \ldots, N$, the contributions of the queues to the objective function are all equal. The optimality criterion is quite arbitrary, for example the weight factors could also have favored the queues with high or those with low frequencies. The same holds for the order in which the indices are included. At the moment it is unclear which objective function is best. However, our numerical experience suggests that slightly altering these factors does not have a substantial influence on the value of the objective function.

Algorithm 6.B.1
The algorithm consists of two phases. In phase 1 a basic pattern is created with a method derived from Hajek[56]. In phase 2, this pattern is improved with the use of a local search method.

Phase 1:
A basic pattern is constructed in an iterative way, starting with an empty pattern and consecutively inserting the indices of the queues into the pattern. After step i, the algorithm has produced a sub-pattern S_i, which contains the indices of queues $1, \ldots, i$. The method operates as follows: If in step i in sub-pattern S_{i-1} of length k, the index of queue i has to be inserted a_i times, then first the distances d_{i_j} for the next sub-pattern S_i are computed, following Hajek[56], by $d_{i_j} = [j \frac{k+a_i}{a_i}]$, $j = 1, \ldots, a_i$. In this way, the distances for queue i

are regularly placed around their mean $\frac{k+a_i}{a_i}$. After computing these distances d_{i_j}, the indices still can be inserted in various ways into S_{i-1}. To illustrate, if $S_2 = \{1,1,2\}$ and $a_3 = 1$, then there are three different patterns to choose S_3 from: {3,1,1,2}, {1,3,1,2} and {1,1,3,2}. In this example, the insertion can start from three different points in S_2. In general, there can be k different ways of inserting, creating possible new sub-patterns $S_i^1, \ldots, S_i^k$. From these patterns, S_i is chosen such that $V(S_i) = \min_{1 \leq j \leq k} V(S_i^j)$.

We see that in the ith step of phase 1, index i is optimally placed in the sub-pattern. However, this optimality could be ruffled in subsequent iteration steps. In phase 2 therefore a local search method is applied, trying to restore some of the regularity.

Phase 2:
In the local search S_N is replaced by $S'_N(k,l)$ if $V(S'_N(k,l)) < V(S_N)$, where $S'_N(k,l) = S_N$ except for entries k and l, which in $S'_N(k,l)$ are interchanged compared to S_N. This local search is repeated until no further improvements can be made.

(end algorithm 6.B.1) □

Remark B.1:
For the first phase also the Golden Ratio method, as described by Itai & Rosberg[61], could have been applied. The local search method does improve the Golden Ratio pattern, but in general the above-described heuristic based on Hajek[56] performs better. In the cases that we ran, the latter method provides a pattern S for which the objective function $V(S)$ lies between 0 and 5 percent of the theoretical minimum, whereas Golden Ratio's relative error is in most cases between 2 and 4 times as high. □

Chapter 7

The $BMAP/M/s$ queue

In this chapter we analyse the $BMAP/M/s$ queue and present an overview of current methods for analysing Markov modulated queueing models.

7.1 Introduction

In previous chapters we frequently used Markov processes in the modelling of queueing systems. In Chapter 4 we modelled dependence between the inter-arrival and service time of a customer with a batch Markovian arrival process (BMAP), in Chapter 5 we analysed the single-server queue with impatient customers with customers arriving according to a Markovian arrival process (MAP), and in the study on traffic allocation that we presented in Chapter 6 the pattern allocation policy resulted in a special MAP. These examples illustrate the increasing interest for the application of Markov modulated processes in queueing theory, for which two main reasons can be given. Firstly, the performance measures of Markov modulated queueing systems can often efficiently and rapidly be evaluated, which makes these queueing models of practical use. Secondly, modern queueing characteristics such as dependence structures in the input processes (cf. Chapter 1) can fairly adequately be modelled by relatively simple Markov Processes, whereas the classical $GI/G/1$ queueing models often just lack the degree of freedom for modelling such features. An illustrative example is traffic in a communication network; the arrival process of data packets is typically not a Poisson or renewal process, but the characteristics of such an arrival stream might already be captured by a two- or three-state Markov process, where the state of such a Markov process describes the actual characteristics of the arrival process in more detail than the (single-parameter) Poisson process (cf. Grünenfelder & Robert[53] and Bonomi et al.[17]).

Currently there exist a number of methods to analyse Markov modulated queueing models; most of them can be viewed as matrix generalizations of the analysis methods that exist for classical queueing models and Markov processes.

In this chapter we analyse the queue length process of the $BMAP/M/s$ queue, i.e. the multi-server single-queue system in which customers arrive in batches according to a BMAP and in which the service times of individual customers are exponentially distributed.

One of our motivations for studying the $BMAP/M/s$ queue is that it has many interesting applications. For instance we are able to evaluate the multi-server generalization of the collector model as presented in Chapters 2,3 and 4. A second application is the evaluation of pattern allocation of batch customers to multiple-server stations, which is an extension of the single-server queueing model of Chapter 6.

A second motivation for studying the $BMAP/M/s$ queue is that is allows us to give an overview and classification of current methods for analysing Markov modulated queueing models. In this thesis we already have encountered two different analysis techniques: the standard framework of the $BMAP/G/1$ queue (cf. Chapters 4 and 6), and a spectral analysis method (cf. Chapter 5). In this chapter we connect these two methods and a third by showing how closely they are related.

In the remainder of this section we present a description of the $BMAP/M/s$ queue, a survey of literature related to this queueing model, and an outline of the chapter.

Model description

The $BMAP/M/s$ queue is the multi-server single-queue system where customers arrive in batches according to a BMAP (cf. Appendix A) and in which the service times of customers are independent and identically exponentially distributed with parameter μ. Customers of different batches are served in order of the arrival of the batches they belong to, and customers arriving in a same batch are served in random order. The BMAP is described in detail in Appendix A. In this chapter we adopt the notation of that appendix, for instance the matrices D_k, $k = 0, 1, \ldots$ describe the arrival process of batches, with $D := \sum_{k=0}^{\infty} D_k$ being the generator of the underlying Markov process $\{\mathbf{J}(t), t \geq 0\}$, $(\mathbf{J}(t) \in E = \{1, \ldots, M\})$. Furthermore, the stationary probability vector of $\{\mathbf{J}(t), t \geq 0\}$ is π, with π satisfying $\pi D = 0$, and $\pi e = 1$.

In this chapter we concentrate on the analysis of the stationary distribution of the process $\{(\mathbf{X}(t), \mathbf{J}(t)), t \geq 0\}$, where $\mathbf{X}(t)$ denotes the number of customers in the system at time t, $(\mathbf{X}(t) \in \{0, 1, \ldots\})$. We remark that in general $\mathbf{X}(t)$ is not a Markov process, however the process $\{(\mathbf{X}(t), \mathbf{J}(t)), t \geq 0\}$ is Markovian.

We observe that when $\mathbf{X}(t) \geq s$ the transition dynamics of the process $\{(\mathbf{X}(t), \mathbf{J}(t)), t \geq 0\}$ are identical to the transition dynamics of an associated $BMAP/M/1$ queue, for which the same BMAP describes the arrival process of

batches but in which the service times of individual customers are exponentially distributed with parameter $s\mu$. Obviously, as the set $\{(i,j)|0 \leq i < s, j \in E\}$ is finite, the ergodicity condition of $\{(\mathbf{X}(t), \mathbf{J}(t)), t \geq 0\}$ of the $BMAP/M/s$ queue is $\pi \sum_{k=1}^{\infty} kD_k e < s\mu$, which is the stability condition of the associated $BMAP/M/1$ queue (cf. Appendix A).

Related literature for this model

Over the years many studies on multi-server single-queue models have appeared. The first model we mention here is the $GI/M/s$ queue; this extension of the $GI/M/1$ model is well understood (cf. Wolff[101, p.398-399]). Considering the embedded Markov chain of the number of customers at arrival moments, it is seen that the transition structures in the $GI/M/1$ and $GI/M/s$ are quite similar; they only differ for the subset of boundary states. As a consequence, the tails of the stationary distributions in both embedded Markov processes have a geometrical form (cf. Neuts[80]). When the arrival process is of MAP type (cf. Appendix A) the queueing process is generally referred to as a Quasi Birth-and-Death process (QBD). This type of Markov processes again has the geometric property for the tail of the stationary distribution (cf. Neuts[80]).

The $GI^X/M/s$ queue, the multiple exponential server queue with batches arriving according to a renewal process has been studied in Baily & Neuts[11] and recently in Zhao[103]. In [11] the embedded Markov process describing the queue length at moments of arrivals is analysed with matrix-analytic techniques. In Zhao[103] it is shown that the stationary probabilities for the number of customers in the system can be written as a linear combination of s geometric terms.

A special case of the $BMAP/M/s$ queue is the one-dimensional variant, being the $M/M/s$ queue with batch arrivals. This model is included in [11], but here the queue length process itself is Markovian, and the stationary distribution of the number of customers in the system can also be obtained by means of generating function techniques directly applied to the stationary queue length process.

Another special case of the current model is the $N/D/s$ queue, in which the service times of customers are deterministically distributed and where the N denotes the N-process, which is equivalent to the BMAP. This queueing model has been investigated by Neuts[81, section 5.5B]. However, as pointed out in Neuts[81, p.345] the results appear to be of theoretic use only and are hard to implement numerically.

A last group of models we mention here is related to the $GI/Ph/s$ queue. Where in the $BMAP/M/s$ queue the individual customers in the batch arrivals are served separately, in the $GI/Ph/s$ queue we have single arrivals but for special cases this customer can be viewed as a batch of customers. For instance, a single customer with an Erlang distributed service time with k phases can be viewed as a batch of k individual customers (each with an exponentially distributed service time) which all have to be served by the same server.

The $GI/Ph/s$ queue and variants, for instance non-homogeneous servers, are studied in the book of Neuts[80]. A particular variant of the $GI/Ph/s$ queue is the $GI/H_m/s$, with H_m indicating a hyper-exponential distribution. This model is analysed by de Smit[90, 91] by means of the Wiener-Hopf factorization technique. This group of models leads to Markov processes with a two-dimensional state space. A common aspect of these Markov processes is that the state-space component describing the state of the servers grows exponentially with the number of servers and number of phases of the service time.

Outline of the chapter
Section 7.2 is devoted to a general overview of Markov modulated models in queueing theory. We sketch the rise of these models and also the background of various methods for analysis. In Section 7.3 we analyse the stationary joint distribution of the number of customers and the underlying Markov process in the $BMAP/M/s$ queue. We present three methods of deriving this joint distribution. A main result is an iterative scheme to obtain stationary probabilities of boundary states.
In Section 7.4 we numerically apply our results to a multi-server extension of the queueing model with dependence between the interarrival and service time of a customer. This model was the subject of Chapters 2,3 and 4.

7.2 Markov modulated queueing models

The classical $GI/G/1$ queue has been thoroughly analysed, cf. Cohen[34] for a rigorous analytical treatment or Wolff[101] for a more probabilistic approach. However, the results on the $GI/G/1$ queue, e.g. the Pollaczek integral equation for the $GI/G/1$ waiting time (cf. Cohen[34]), have been hard to use directly in numerical applications, as they often involve transform expressions with complex arguments from which performance measures are difficult to evaluate. These numerical difficulties, rather than the limited modelling power of the $GI/G/1$ queue, have led to the first interest for Markov modulated queueing models; non-trivial queueing systems which do allow numerical performance evaluation almost invariably involved exponential (phase-type) distributions.
With the increasing attention for general Markov processes and renewal theory one became to consider sequences of random variables in terms of *processes* rather than *distributions*. For instance, instead of defining the arrival process as a sequence of Erlang distributed random variables, one considers the arrivals as events in a stochastic process in which the time between events is Erlang distributed. In line with this change of perspective, in the Kendall notation the indications of distributions are being replaced by indications of processes, e.g. $BMAP/G/1$ instead of $GI/G/1$.
The connection of this interpretation with the technique of embedded Markov processes has led to the theory of queueing models in which the queueing process is described together with a secondary Markov process. The queueing

process in itself is not a Markov process, however, it *is* when conditioned on the state of the secondary Markov process. As an example we consider the $E_k/GI/1$ queue, where E_k denotes Erlang-k distributed interarrival times. At the moments of departure the process describing the number of customers in the system is in itself not Markovian, but this embedded process together with the number of phases of the Erlang *process* that have expired *does* possess the Markov property.
An early paper recognizing the potential of such a two-dimensional model is Loynes[71], in which a number of applications are mentioned, such as: non-renewal arrival processes, batch arrivals and services, bursty arrival processes, and state dependent service rates.
In fact, the classical transform results have recently become more and more accessible for numerical implementation, an example again being the Pollaczek integral equation for the $GI/G/1$ waiting time, which is numerically inverted in Abate et al.[2]. So the motivation to study queueing models with Markovian features for evaluating $GI/G/1$ queues has somewhat disappeared, but nowadays the study of Markov modulated queueing systems is stimulated by their high modelling potential.
Next we discuss some methods for the analysis of Markov modulated queueing systems. Most attention will be paid to methods that handle the typical Markov processes that arise in these systems. In Section 7.3 the analysis of the $BMAP/M/s$ queue will guide a further discussion of these methods.

Analysis of Markov modulated queueing systems
A pioneer in the field of Markov modulated queueing systems is Neuts (cf. Neuts[80, 81]). The main contribution of Neuts to this field is the introduction of a probabilistic and algorithmic approach for the analysis of these queueing models. Before the work of Neuts, the analysis (including Loynes[71]) remained of a classical nature, following the lines of the complex analysis of Takács (cf. Takács[94]) and spectral analysis from linear algebra. Due to the efforts of Neuts and others (Ramaswami, Lucantoni, Latouche), Markov modulated queueing models have become an important tool in the analysis of modern queueing systems.
In recent years a revival of classical analysis methods can be detected, cf. Mitrani & Mitra[75], de Smit[90, 91] and Regterschot and de Smit[86], the reason for this being that the increasing computer capacity and the availability of efficient algorithms has made these methods as practical as the algorithmic methods of Neuts, hereby relieving earlier mentioned drawbacks of complex analysis techniques.

In a Markov modulated queueing system the queueing process is directed by a continuous time Markov process, where the state of this Markov process contains all necessary information about the system characteristics. This might be extra information on random variables, e.g. the remaining interarrival time, but could also describe the current configuration of the system, e.g. the number

of available servers. For such queueing systems one can often construct a two-dimensional embedded *Markov process* $\{(\mathbf{X}_n, \mathbf{J}_n), n = 1, 2, \ldots\}$, where $\mathbf{J}_n$ denotes the state of the directing Markov process and with $\mathbf{X}_n$ usually being one of the familiar queueing measures, such as the number of customers at a departure moment. $\mathbf{X}_n$ attains non-negative integer values. This typical two-dimensional state space is known as the *semi-infinite strip*.

An important class of analysis methods directly exploits the homogeneity properties in the transition structure of $\{(\mathbf{X}_n, \mathbf{J}_n), n = 1, 2, \ldots\}$. In the one-dimensional case, i.e. in the $M/G/1$ or $GI/M/1$ queue, by the homogeneity property we mean that transitions in the interior of the state-space only depend on the size of the transition and are independent of the actual states that are involved in the transition. To illustrate this: in the Markov process describing the number of customers at departure epochs in the $M/G/1$ queue, transitions are related to the number of customers that arrive during a service, and obviously the distribution of the latter does not depend on the number of customers in the queue. In the two-dimensional case the homogeneity property means that, next to depending on the state of the directing process $\mathbf{J}_n$, transition probabilities in the process only depend on the *size of the transition* rather than on the *state* of the $\mathbf{X}_n$ component. We remark that the property of homogeneity only has to exist for the interior states, transitions involving boundary states actually may depend on the state of $\mathbf{X}_n$.

Analysis methods that exploit this homogeneity property are the matrix-analytic approach, the transform method, and the spectral expansion method (these are the common names, usually adopted from papers by their main contributors). Below we discuss them in some detail, also pointing at similarities and key differences.

1. The matrix-analytic technique

The matrix-analytic technique can be viewed as the matrix extension of probabilistic analysis for random walks with homogeneous properties; the derivations of the results are of a probabilistic nature and also many elements of the results have probabilistic interpretations. Basically, due to the homogeneity structure, performance measures can explicitly be described by compact expressions for the generating functions with a few boundary state probabilities to be determined. In the matrix-analytic technique the boundary probabilities are obtained with results from Markov renewal theory.

The probabilistic analysis of the matrix-analytic technique was first developed for the matrix-geometric approach. The latter name refers to the matrix generalization of a special feature of stationary probabilities of the $GI/M/1$ queue; let $\pi_0, \pi_1, \ldots$ denote the stationary probabilities of the number of customers in the system, then $\pi_n = \rho(1-\alpha)\alpha^{n-1}$, $n = 1, 2, \ldots$, i.e. the probabilities have a geometrically distributed tail (cf. [80]). Put differently, $\pi_n = \pi_{n-1}\alpha$. In Markov modulated queues of $GI/M/1$ type, where π_n is a vector, there exists a matrix R such that $\pi_n = \pi_{n-1}R$ for n sufficiently large, i.e. for levels in the homogeneous part of the state space. Not all Markov modulated queueing

models possess the matrix-*geometric* property, one of the examples being the $BMAP/M/s$ queue (cf. Remark 7.3.3). For this reason the method introduced by Neuts is usually referred to as the matrix-analytic method.
A very appealing aspect of matrix generalizations of $GI/M/1$ and $M/G/1$ queues is that not only stationary probabilities have a form that is analogous to the one-dimensional case, but more generally the expressions for most performance measures can be viewed as matrix-generalizations of one-dimensional results (cf. Appendix A).
Key references for the matrix-analytic technique are the two books of Neuts[80, 81] and papers by Ramaswami[85], Latouche & Ramaswami[70] and Lucantoni[73, 74].

2. The transform method
The transform method, as Gail et al.[50] call it, is the extension of classical generating function and transform analysis. Central in this method are Rouché type arguments. The generating function expressions are obtained in the same way as for the matrix-analytic technique. For continuous valued processes transform expressions follow from recurrence relations at embedded time points, e.g. departure epochs.
With the transform method boundary probabilities are obtained by exploiting the analytic properties of generating functions of a proper distribution. These properties lead to a set of equations from which the boundary state probabilities can be obtained. Key references for this method are Gail et al.[50, 51].
Once the general behaviour of the queueing system has been characterized in terms of generating functions or Laplace transforms, the matrix-analytic method and the transform method can be viewed as different ways of determining the boundary states.

3. The spectral-expansion method
This technique directly considers state-space probabilities. First a general form of the stationary probabilities is derived from the balance equations that describe the interior of the state space. Subsequently the boundary states follow from the balance equations. The general form of the stationary probabilities is determined by a multi-dimensional generalization of the spectral expansion techniques from linear algebra that are used for solving linear difference equations. This method is currently advocated in papers by Mitrani and co-authors (cf. [45, 75, 76]).
The spectral-expansion technique is restricted to models for which only a finite subset of the state space is non-homogeneous, this means that from a certain level in the $\mathbf{X}$ component of a state $(\mathbf{X}, \mathbf{J})$, the transitions no longer depend on the value of $\mathbf{X}$. As a consequence, the batch size distributions that can be dealt with by this method must have a finite support. Under this restriction, the spectral-expansion technique and the matrix-*geometric* technique can be viewed as different methods of obtaining the above-mentioned matrix R. This observation is discussed in some more detail in Remark 7.3.4.

Generally, it is hard to tell which of the methods is preferable. In general the matrix-analytic method appears to be the most stable method for numerical performance evaluation since it mainly involves iterations, matrix inversions and integrations. On the other hand, in theory the root determination in the transform method and the spectral-expansion method is of a lower complexity (cf. Mitrani & Mitra[75]).

A method of analysing Markov modulated queueing models that does not directly utilize homogeneity properties is Wiener-Hopf factorization. With this method the queueing process at embedded time points is treated from the perspective of random walks; starting point is a multi-dimensional variant of the well known Lindley equation for $GI/G/1$ queues (cf. p.140 of Takács[94]). Moreover, the Wiener-Hopf convolution integral equation that is connected to this Lindley equation also has a multi-dimensional counterpart. In a series of papers by de Smit[90, 91], Regterschot & de Smit[86], and in the thesis of Regterschot[87] the classical one-dimensional Wiener-Hopf factorization is extended to Markov modulated queues.
From the viewpoint of generality, it appears that the Wiener-Hopf factorization technique has more modelling freedom than the three above-mentioned methods because it does not explicitly use the homogeneity structure of the two-dimensional Markov process. For the methods which do use this structure often all the matrices describing the transition probabilities in the embedded Markov process have to be calculated. These matrices might be quite hard to obtain, for instance they might concern the joint probabilities of the number of service completions during an arrival together with the transitions in the underlying Markov process.

As a broad class of models can be analysed by any of the above-described methods, it might be interesting to determine the connections between various quantities and steps in the analysis. Asmussen[8] attempts to connect the results from the probabilistic analysis of Neuts and others to the Wiener-Hopf factorization method. This is done by deriving a matrix analog of the probabilistic form of the Wiener-Hopf factorization as it exists for one-dimensional random walks.
In the one-dimensional random walk it is known (cf. p.400 of Feller[46]) that the factorization of an LST into two functions, one function being analytic on $\{\omega \in \mathbb{C} | Re\,\omega \leq \delta\}$ for some $\delta > 0$, the other function being analytic in $\{\omega \in \mathbb{C} | Re\,\omega \geq 0\}$, is equivalent to a factorization of the transition probability function of the random walk into two probability functions, one function having its support on the positive real axis, the other function having its support on the non-positive real axis. Asmussen derives a matrix analog of this probabilistic form of the Wiener-Hopf factorization.
This connects the method of Neuts and the Wiener-Hopf factorization of de Smit. However, except for special cases a close connection between the methods is not established.

Remark 7.2.1
There also exist purely numerical techniques for evaluating performance measures of Markovian queueing systems. One such method is the Power Series Algorithm (cf. the tutorial of Blanc [14]), which is based on a power series expansion property for stationary probabilities in queueing systems. And finally, much progress has also been made in developing numerical methods that can deal with general (large) Markov chains. Recent developments for such methods and references are presented in the conference proceedings [92]. □

7.3 Analysis of the $BMAP/M/s$ queue

In this section the analysis of the $BMAP/M/s$ queue guides a further discussion of the matrix-geometric, transform and the spectral-expansion method. We analyse the two-dimensional process of the number of customers in the $BMAP/M/s$ queue as defined in Section 7.1. As indicated in Section 7.1, the $BMAP/M/s$ queue features homogeneity properties in the interior of the state space; if all servers are busy the transition structure is that of an ordinary $BMAP/M/1$ queue. However, the $BMAP/M/s$ queue has not yet been treated directly, probably because of the technical complications that result from the fact that the batch size distribution not necessarily has a finite support (e.g., geometric batch size distributions are allowed).
One of the consequences is that in general the tail of the stationary distribution of the number of customers in the system is not of the geometric structure as described in the previous section. This will be explained in Remark 7.3.3.

Basic properties
With $D_0, D_1, \ldots$ describing the BMAP arrival process (cf. Appendix A), where D_i is an $M \times M$ matrix, and with I being the $M \times M$ identity matrix, the generator Q of the two-dimensional Markov process $\{(\mathbf{X}(t), \mathbf{J}(t)), t \geq 0\}$ can be represented as

$$Q = \begin{bmatrix}
D_0 & D_1 & D_2 & D_3 & D_4 & D_5 & \ldots & & \\
\mu I & D_0{-}\mu I & D_1 & D_2 & D_3 & D_4 & D_5 & \ldots & \\
 & 2\mu I & D_0{-}2\mu I & D_1 & D_2 & D_3 & D_4 & D_5 & \ldots \\
 & & \ldots & \ldots & \ldots & & & & \\
 & & & \ldots & \ldots & \ldots & & & \\
 & & & & (s{-}1)\mu I & D_0{-}(s{-}1)\mu I & D_1 & D_2 & \ldots \\
 & & & & & s\mu I & D_0{-}s\mu I & D_1 & \ldots \\
 & & & & & & s\mu I & D_0{-}s\mu I & \ldots \\
 & & & & & & & s\mu I & \ldots \\
 & & & & & & & & \ldots
\end{bmatrix}$$

FIGURE 7.1: GENERATOR OF $\{(\mathbf{X}(t), \mathbf{J}(t)), t \geq 0\}$.
We see that Q is represented in blocks of $M \times M$ matrices. For a state (i, j) of $\{(\mathbf{X}(t), \mathbf{J}(t)), t \geq 0\}$ the component i denotes the $i-th$ row of blocks in Q, the component j specifies the $j-th$ element within the M states of the $i-th$ block.

Moreover, the Markov process $\{(\mathbf{X}(t), \mathbf{J}(t)), t \geq 0\}$ is skip-free to the left, i.e. the path from a level i_0 to a level $i_1 < i_0$ visits all levels between i_0 and i_1.
Denote by $\{(\mathbf{X}, \mathbf{J})\}$ the stationary process of $\{(\mathbf{X}(t), \mathbf{J}(t)), t \geq 0\}$, and let $x = \{x_0, x_1, \ldots\}$ be the stationary probability vector of $\{(\mathbf{X}(t), \mathbf{J}(t)), t \geq 0\}$, with $x_i = (x_{i1}, \ldots, x_{iM})$. Then x satisfies

$$xQ = 0, \quad \sum_{i=0}^{\infty} \sum_{j=1}^{M} x_{ij} = 1. \tag{7.1}$$

This results in the set of balance equations

$$\sum_{i=0}^{k} x_i D_{k-i} - k\mu x_k + (k+1)\mu x_{k+1} = 0, \quad k = 0, \ldots, s-1. \tag{7.2}$$

$$\sum_{i=0}^{k} x_i D_{k-i} - s\mu x_k + s\mu x_{k+1} = 0, \quad k = s, s+1, \ldots. \tag{7.3}$$

The set of balance equations given by (7.2) and (7.3) can be written more concisely in terms of generating functions. Define the vector generating function $X(z) := (\sum_{i=0}^{\infty} x_{ij} z^i)$. Then, with $D(z) := \sum_{i=0}^{\infty} D_i z^i$, we obtain

$$X(z)\,[(1-z)s\mu I + zD(z)] = (1-z) \sum_{k=0}^{s-1} (s-k)\mu x_k z^k. \tag{7.4}$$

We observe that the multi-dimensional functional equation has the same form as the one-dimensional case; expression (7.4) is a matrix generalization of the expression for the generating function of the number of customers in the $M/M/s$ queue with batch arrivals.
We next determine the stationary probability vector x with the three different methods that we discussed in the previous section.

The matrix-analytic technique: a probabilistic approach
In (7.4) we observe that $X(z)$ is characterized up to the unknown vectors $x_0, \ldots, x_{s-1}$. Moreover, from (7.2) we find that once x_0 has been determined $x_1, \ldots, x_{s-1}$ follow successively.
From Markov renewal theory (cf. Cinlar[33]) we know that the mean return time, say $\gamma < \infty$, to a state (i, j) of the stationary process $\{(\mathbf{X}, \mathbf{J})\}$ and the stationary probability x_{ij} are connected through $x_{ij} = \frac{1}{\gamma}$. Hence we concentrate on the mean return time to level $0 = \{(0, j) | j = 1, \ldots, M\}$.
Next we translate the continuous-time Markov process into a discrete-time Markov process by uniformization and determine the mean return times to x_0 in this discrete-time queue.

Let $\theta > \max_i(-D_{ii}) + s\mu$. The one-step transition probability matrix of the discrete-time Markov chain that is the result of the uniformization is $P := Q/\theta + I$.

Define

$$\begin{aligned} A_i &:= \frac{D_{i-1}}{\theta}, \quad i = 2, 3, \ldots, \\ A_{1,j} &:= \frac{D_0 - j\mu I}{\theta} + I, \quad j = 0, \ldots, s, \\ A_1 &:= A_{1,s}, \\ A_{0,j} &:= \frac{j\mu I}{\theta}, \quad j = 0, \ldots, s, \\ A_0 &:= A_{0,s}. \end{aligned}$$

Then

$$P = \begin{bmatrix} A_{1,0} & A_2 & A_3 & A_4 & \ldots & & & & & \\ A_{0,1} & A_{1,1} & A_2 & A_3 & A_4 & \ldots & & & & \\ & A_{0,2} & A_{1,2} & A_2 & A_3 & A_4 & \ldots & & & \\ & & \ldots & \ldots & \ldots & & & & & \\ & & & \ldots & \ldots & \ldots & & & & \\ & & & & A_{0,s-1} & A_{1,s-1} & A_2 & A_3 & A_4 & \ldots \\ & & & & & A_0 & A_1 & A_2 & A_3 & \ldots \\ & & & & & & A_0 & A_1 & A_2 & \ldots \\ & & & & & & & A_0 & A_1 & \ldots \\ & & & & & & & & \ldots & \ldots \end{bmatrix}$$

FIGURE 7.2: ONE-STEP TRANSITION MATRIX P.

A next step is the observation that the return time from level i to level 0 is the convolution of the successive first passage times from level l to level $l-1$, for $l = i, i-1, \ldots 1$. As observed in the introduction, and as can be seen from Figure 7.1 or equation (7.3), for levels $i \geq s$ the transitions of the Markov process have the characteristics of the $BMAP/M/1$ queue with service times that are exponentially distributed with parameter $s\mu$. Consequently, the first passage times for levels i to $i-1$, $i \geq s$, follow from the results for the $BMAP/M/1$ queue.

Let $G(z)$ define the matrix generating function for the uniformized Markov chain describing the first passage time from level i to level $i-1$, $i \geq s$, together with the transition probabilities in the directing Markov process.

It follows from the $BMAP/G/1$ theory (cf. Appendix A) that $G(z)$ is a busy period type generating function satisfying the functional equation

$$G(z) = z \sum_{k=0}^{\infty} A_k G^k(z). \tag{7.5}$$

Equation (7.5) is a matrix-generalization of the familiar branching expression for the number of customers served during the busy period in the $M/G/1$ queue (cf. p.32 of Takács[94]).
For the mean return times we need to evaluate the matrices $G(1)$ and $G'(1)$. Starting with some specific initial matrices, $G(1)$ can be obtained by iterating (7.5), and subsequently $G'(1)$ is obtained by iterating the derivative of (7.5). For the case of single arrivals (MAP), an elegant probabilistic algorithm is proposed in Latouche & Ramaswami[70] as an alternative evaluation of $G(1)$. This algorithm may well be extended for the case of batch size distributions with finite support.

Next define for $i = 1, \ldots, s-1$, the matrix generating function $G_i(z)$ that describes the first passage time from level i to level $i-1$.
Then (cf. Figure 7.2), for $i = 1, \ldots, s-1$,

$$G_{s-i}(z) = zA_{0,s-i} + z\left(A_{1,s-i}G_{s-i}(z) + \sum_{k=2}^{i} A_k \prod_{n=k-1}^{0} G_{s-i+n}(z) + \sum_{k=i+1}^{\infty} A_k G^{k-i}(z) \prod_{n=1}^{i} G_{s-n}(z)\right). \quad (7.6)$$

In (7.6) we introduce the convention $\prod_{i=a}^{b} G_i(z) := G_a(z) \times \ldots \times G_b(z)$, for $a \geq b$ as well as for $a \leq b$.
Again, for the mean return times we need to evaluate matrices $G_{s-i}(1)$ and $G'_{s-i}(1)$. In (7.6) we observe that $G_{s-i}(z)$ is expressed in terms of $G(z)$ and $G_j(z)$ with $j \geq s-i$. This leads to a recursive scheme for the matrices $G_{s-i}(1)$, $i = 1, \ldots, s-1$. The matrices $G_{s-i}(1)$ can be expressed directly in earlier computed matrices by rewriting (7.6), or they might be computed by iterating (7.6).
After obtaining $G(1), G_1(1), \ldots, G_{s-1}(1)$ and $G'(1), G'_1(1), \ldots, G'_{s-1}(1)$ we subsequently derive the mean return times from the matrix generating function $K(z)$, where $K_{ij}(z)$ is the generating function of the number of transitions until the first return to level 0 and the probability that the underlying Markov chain is in state j at that time, given that at this moment the two-dimensional Markov chain is in state $(0, i)$. By conditioning on the the first transition that is made after a return to level 0 and using (7.6) we obtain

$$K(z) = z\left(A_{1,0} + \sum_{k=2}^{\infty} A_k G^{[0,k-s]^+}(z) \prod_{l=[k-1,s-1]^-}^{1} G_l(z)\right), \quad (7.7)$$

here $[k_0, k_1]^+ := \max\{k_0, k_1\}$ and $[k_0, k_1]^- := \min\{k_0, k_1\}$.
To derive the mean return times to level 0, we first consider $K(1)$ which is a stochastic matrix describing a Markov chain on the state-space of $\mathbf{J}$, this Markov chain being the embedded Markov process at the moments of visits to

level 0. Let $\bar{k}$ be the stationary probability (row)vector of this process, defined by $\bar{k}K(1) = \bar{k}$, $\bar{k}e = 1$, where e denotes the M state unit (column)vector. Then, from Markov renewal theory (cf. Cinlar[33]) we find that x_0 is given by

$$x_0 = \frac{\bar{k}}{\bar{k}K'(1)e}, \tag{7.8}$$

where $K'(1)$ follows from (7.7).
As stated $x_1, \ldots, x_{s-1}$ now successively follow from (7.2). For example

$$x_1 = \frac{-x_0 D_0}{\mu}, \text{ and } x_2 = x_0 \frac{D_0^2/\mu - D_0 - D_1}{2\mu}. \tag{7.9}$$

Finally, we have arrived at:

Theorem 7.3.1
The generating function of the stationary distribution of the number of customers in the $BMAP/M/s$ queue is given by (7.4), where the vector x_0 is given by (7.8) and the $s-1$ vectors $x_1, \ldots, x_{s-1}$ follow from (7.2). □

The transform method: an analytic approach
In equation (7.4) the generating function $X(z)$ is characterized up to s unknown boundary vectors. With the matrix-analytic method x_0 is determined in an iterative fashion, after which $x_1, \ldots, x_{s-1}$ successively follow. With the transform method the vectors of the boundary states are obtained by exploring the analytic properties of generating functions of proper distribution functions. After post-multiplying in (7.4) with the inverse of $[(1-z)s\mu I + zD(z)]$ we have isolated $X(z)$. Since the queueing process is ergodic (cf. Section 7.1), $X(z)$ is analytical in the open unit disk $|z| < 1$ and continuous in $|z| \leq 1$. Consequently singularities in the inverse matrix somehow have to be canceled. The principal idea of the transform method is to use this condition for each singularity and obtain an equation for the vectors $x_0, \ldots, x_{s-1}$. The idea of this method is quite transparent, but there are some technical details connected to the central question: does the set of equations that arises contain enough, namely $s \times M$, independent equations to determine the vectors $x_0, \ldots, x_{s-1}$. The two main problems for Markov modulated queueing models are: (i) the number of points on $|z| = 1$ for which $[(1-z)s\mu I + zD(z)]$ is singular, and (ii) the multiplicity of the singularities for $|z| < 1$.
In Gail et al.[50] these technical aspects are addressed in detail, resulting in a well-rounded theory on the extension of the classical Rouché type analysis to the Markov modulated queueing models of $M/G/1$ type. A main result of Gail et al.[50] is that if the Markov process is ergodic in general there can indeed be derived sufficiently many independent equations for the determination of the stationary distribution. In Gail et al.[51] these results are summarized for a similar Markov process which only slightly deviates from the general model, but the extra assumptions that are made reduce the technical complexity to a great extent.

In the following we apply the results from [50] and [51] to the $BMAP/M/s$ queue.

When we conform the transition structure (cf. Figures 7.1 and 7.2) to the notation of [50] and derive the functional equation of the generating function $X(z)$ from the balance equations, we find an equation that is equal to expression (7.4) but with both sides multiplied with a factor z^{s-1} and the uniformization constant θ:

$$X(z)z^{s-1}\theta\left[(1-z)s\mu I + zD(z)\right] = z^{s-1}(1-z)\theta\sum_{k=0}^{s-1}(s-k)\mu x_k z^k. \quad (7.10)$$

Due to the ergodicity of the queueing process and the polynomial structure of $[(1-z)s\mu I + zD(z)]$ and $\sum_{k=0}^{s-1}(s-k)\mu x_k z^k$ it follows from [50] that the vectors $x_0, \ldots, x_{s-1}$ can be obtained from the set of equations that results from exploring the analytic properties of $X(z)$.

We next concentrate on the determination of the set of equations that provides $x_0, \ldots, x_{s-1}$.

First we make the assumption that $[(1-z)s\mu I + zD(z)]$ has a single singularity for $|z| = 1$, being $z = 1$. This is assumption (A2) from theorem 1 of Gail et al.[51]. While this assumption is not necessary from a theoretical viewpoint, in practice it avoids a number of additional technical difficulties. In [51] it is claimed that this condition is not restrictive. Under this assumption, it follows from theorem 1 of [51] that the determinant of $z^{s-1}\left[(1-z)s\mu I + zD(z)\right]$ has exactly $sM-1$ zeros in the open unit disk and a simple zero at $z = 1$.

In the set of $sM-1$ zeros inside the unit disk $z = 0$ has multiplicity $(s-1)M$, the remaining $M-1$ zeros of the determinant are connected to the eigenvalues of $D(z)$.

From Lemma 3 of [50] it follows that $z = 0$ results in $M \times (s-1)M$ equations of which exactly $(s-1)M$ are independent. Fortunately, as this set of equations is connected to the derivatives of (7.10), it follows that the independent set consists of the first $s-1$ equations of (7.2). We remark that the independence of this set also follows from the fact that x_0 automatically results in $x_1, \ldots, x_{s-1}$.

Returning to the functional equation (7.4), we can derive a set of M additional equations which together with (7.2) determine $x_0, \ldots, x_{s-1}$. This is done in the following manner. For certain values of z, $|z| \leq 1$, the determinant of $[(1-z)s\mu I + zD(z)]$ equals zero. By multiplying both sides of (7.4) for these values of z with the right eigenvector belonging to eigenvalue 0, the left-hand side of (7.4) is canceled. From the analytical properties of $X(z)$ it then follows that also the right-hand side must vanish.

Doing so, let η_i, $i = 1, \ldots, M$ be the values of z, $|z| \leq 1$, for which $[(1-z)s\mu I + zD(z)]$ is singular, with $\eta_1 = 1$. Next we make the assumption that all η_i are simple. Again this assumption is non-restrictive; parameters can always be altered such that the assumption holds while the model is hardly

affected. We remark that Gail et al.[50] show that again from a theoretical viewpoint this assumption is not necessary.
Next, let r_i be the vector such that $[(1-\eta_i)s\mu I + \eta_i D(\eta_i)]\, r_i = 0$, noting that r_i is an eigenvector of $D(\eta_i)$ and r_1 the unit vector e. From our last assumption it follows that $r_1, \ldots, r_M$ are independent (cf. p.153 of Lancaster & Tismenetsky[67]).
As stated, due to the regularity of $X(z)$ we obtain for $i = 2, \ldots, M$

$$(1-\eta_i)\left(\sum_{k=0}^{s-1}(s-k)\mu x_k \eta_i^k\right) r_i = 0, \tag{7.11}$$

and

$$\sum_{k=0}^{s-1}(s-k)\mu x_k e = \omega_1^*. \tag{7.12}$$

In (7.12) the constant ω_1^* equals $\lim_{z\to 1}\frac{\omega_1(z)}{1-z}$, with $\omega_1(z)$ the eigenvalue belonging to the eigenvector $r_1(z)$ of $[(1-z)s\mu I + zD(z)]$ and $\lim_{z\to 1}\omega_1(z) = \omega_1(\eta_1) = 0$.
In the next theorem we summarize the above:

Theorem 7.3.2
The generating function of the stationary distribution of the number of customers in the $BMAP/M/s$ queue is given by (7.4), where the boundary vectors $x_0, \ldots, x_{s-1}$ follow from equations (7.11),(7.12) and the first $s-1$ equations of (7.2). □

Remark 7.3.1
In both the matrix-analytic method and the transform method an important role in the analysis is played by the functional $[(1-z)s\mu I + zD(z)]$. In the probabilistic approach the aim is to determine the generating function $G(z)$ as the solution of the functional equation (7.5), in the transform method the aim is to find zeros of the determinant of $[(1-z)s\mu I + zD(z)]$, but after rewriting this functional the latter problem is equivalent to determining the zeros of the determinant of $[zI - A(z)]$, with $A(z) := \sum_{v=0}^{\infty} A_v z^v$. By writing $G(z)$ in the spectral decomposition form (cf. Lemma 5.2.1 in Chapter 5), it is readily verified that the right eigenvector r_i is also a right eigenvector of $G(\zeta_i)$ for some $\zeta_i \leq 1$, $i = 1, \ldots, M$. Incidentally, the same property not necessarily holds for the left eigenvectors. □

Remark 7.3.2
When comparing the matrix-analytic method with the transform method we see that with the matrix-analytic method the operations mainly involve iterations

of matrices of size $M \times M$, whereas in the transform method a set of $s \times M$ linear equations with as many variables has to be solved. Although the vectors $x_1, \ldots, x_{s-1}$ can be expressed in terms of x_0 via the balance equations (7.2) (cf. (7.9)), still a set of equations has to be solved. □

In Section 7.4 we apply the matrix-analytic method to a multi-server generalization of the collector model of Chapter 2. In our numerical investigations we mainly concentrate on $\mathrm{E}\tilde{\mathbf{N}}$, being the mean number of customers in the system. This performance measure is defined as $\mathrm{E}\tilde{\mathbf{N}} := X'(1)e$, with $X'(1)$ the derivative of $X(z)$ evaluated in $z = 1$. Below we derive an explicit expression for $\mathrm{E}\tilde{\mathbf{N}}$, also using the analytic results of Gail et al.[50].

For convenience we first define for $|z| \leq 1$

$$C(z) := [(1-z)s\mu I + zD(z)], \tag{7.13}$$

and

$$v(z) := \sum_{k=0}^{s-1} (s-k)\mu x_k z^k. \tag{7.14}$$

Then we can rewrite (7.4) into

$$X(z)C(z) = (1-z)v(z), \quad |z| \leq 1. \tag{7.15}$$

Let $l(z)$ and $r(z)$ be the left and right eigenvector of $C(z)$ respectively belonging to the eigenvalue $\omega_1(z)$ that has the property $\omega_1(1) = 0$. Moreover, let $l(z)$ and $r(z)$ be defined such that for $|z| \leq 1$

$$\begin{aligned} l(z)r(z) &= 1, \\ l(1) &= \pi, \ r(1) = e, \\ l(z)e &= 1. \end{aligned}$$

Then by exploiting the analytical properties of $l(\cdot), r(\cdot)$ and $\omega_1(\cdot)$ (cf. [50]), and the identities

$$[C(z) - \omega_1(z)I]r(z) = 0, \quad l(z)[C(z) - \omega_1(z)I] = 0,$$

we find

$$\begin{aligned} \omega_1'(1) &= \pi C'(1)e, \\ r'(1) &= [e\pi + C(1)]^{-1}[\omega_1'(1)I - C'(1)]e, \\ \omega_1''(1) &= \pi C''(1)e + 2\pi C'(1)r'(1). \end{aligned} \tag{7.16}$$

By multiplying both sides of (7.15) to the right with $r(z)$ and taking derivatives we obtain for $z = 1$ after some rewriting:

Theorem 7.3.3
$\mathrm{E}\tilde{\mathbf{N}}$, the mean number of customers in the $BMAP/M/s$ queue, is given by

$$\mathrm{E}\tilde{\mathbf{N}} = X'(1)e = \frac{\omega_1''(1)}{2\left(\omega_1'(1)\right)^2} - \frac{1}{\omega_1(1)'}\left(v'(1)e + v(1)r'(1)\right), \tag{7.17}$$

with $n_1'(1)$, $n_1''(1)$ and $r'(1)$ given by expressions (7.16), and with $v(1)$ and $v'(1)$ following from (7.14). □

From (7.4) and (7.17) we can readily obtain the *vector* $X'(1)$, with $X_i'(1)$ denoting the mean number of customers in the system together with the probability that the underlying Markov chain is in state i. This is done by directly taking derivatives in (7.4) and taking the limit $z \to 1$. This results in an expression for $X'(1)D$. Multiplying on the right with π in (7.17) gives an expression for $X'(1)e\pi$. Observing that $[e\pi + D]$ is non-singular leads to $X'(1)$. We remark that performance measures such as mean waiting times and number of customers at moments of arrival follow from $X(z)$ and $X'(1)$ via applications of Little's law and the PASTA property.

Spectral expansion: an algebraic approach
This method concentrates on determining the stationary probability vector x directly from the set of balance equations given by (7.2) and (7.3). First a general form for the states in the homogeneous part of the state-space is derived from (7.3), subsequently the vector x is obtained from (7.2) and the normalization constant $\sum_{i=0}^{\infty} x_i e = 1$. The method extends an algebraic technique for solving linear difference equations to the multi-dimensional case. A general outline of the method has been presented in Mitrani & Mitra[75]. The method has been applied to queueing models in several papers; Chakka & Mitrani[76] and Ettl & Mitrani[45] study multi-server models with breakdown and repair of the servers; Elwalid et al.[43] study a communication model with multiplexing of sources by means of a Markov modulated queueing system.
To start the investigation of the $BMAP/M/s$ queue, suppose the batch size distribution has finite support. This means that $\exists k_0$ such that $D_k = 0$ for $k > k_0$.
Then for $k \geq \max\{s, k_0\}$ we can rewrite (7.3) as

$$\sum_{j=0}^{k_0} x_{k-j} D_j - s\mu x_k + s\mu x_{k+1} = 0. \tag{7.18}$$

Define matrices $H_i := D_{k_0-i}$ for $i = 0, \ldots, k_0 - 1$, $H_{k_0} := D_0 - s\mu I$, and $H_{k_0+1} := s\mu I$. Then for $k \geq \max\{s - k_0, 0\}$ equation (7.18) can be written as

$$x_k H_0 + x_{k+1} H_1 + \ldots + x_{k+k_0} H_{k_0} + x_{k+k_0+1} H_{k_0+1} = 0. \tag{7.19}$$

Equation (7.19) is a vector difference equation of order $k_0 + 1$. Its general solution can be derived from the associated characteristic matrix polynomial $H(\lambda)$, which is defined as

$$H(\lambda) := \sum_{i=0}^{k_0+1} H_i \lambda^i. \tag{7.20}$$

It is known (cf. Gohberg et al.[52]) that equation (7.19) allows a "spectral decomposition": the vectors x_i, $i = s-1, s, \ldots$ can be written in terms of the zeros λ_k and corresponding left-eigenvectors l_k of (7.20):

$$x_i = \sum_{k=1}^{M} \zeta_k l_k \lambda_k^i, \quad i = s-1, s, \ldots, \tag{7.21}$$

where the constants ζ_k remain to be determined from the boundary conditions (7.2) and the normalization equation $\sum_{i=0}^{\infty} x_i e = 1$.
For this method one has to verify that indeed there exist M zeros of $H(\lambda)$ inside the unit disk, and also that the set of equations is sufficient for determining all ζ_k.
The first condition follows from the ergodicity of the Markov process; if the radius of one of the eigenvalues λ_k would be larger than 1, the expression (7.21) would not converge for $i \to \infty$. The second condition follows from a probabilistic argument: the set of equations that arises has at least one solution, being the stationary solution; would this set provide more than one solution, we would also have obtained more than one stationary vector of the Markov process, which is in contradiction with the uniqueness of the stationary vector.
So we have obtained the stationary probability vector with a third method; however we started with the assumption that the batch size distribution has finite support. When batch sizes can be arbitrarily large, for example in the case of geometric distributions, the tails of the distribution have to be cut off in order to be able to apply the method. However, similar problems arise for the matrix-analytic method as well as for the transform method; in the first case infinite sums occur in expressions (7.5) and (7.6), in the second case one has to determine the zeros of the determinant of a matrix whose elements contain the batch size generating functions, the latter functions need not be finite polynomials since the batch size distributions not necessarily have finite support (e.g. geometric distributions).

Remark 7.3.3
An important observation for the model for which batch size distributions have finite support is that from a certain level not only the rate of transitions going *out* of a state is independent of the number of customers in the system, but also the rate of transitions *into* that state is independent of the level. To illuminate this observation, consider the ordinary $M/M/s$ queue: when the number of customers is $0 \leq i < s$ then the rate out of state i is $i\mu + \lambda$, the rate into state i is $(i+1)\mu + \lambda$. When the number of customers in the system is $i > s$, then the rate out of i as well as the rate into i are independent of i. On the other hand,

in the $M/M/s$ queue with batch arrivals, when batch sizes can be arbitrarily large, state i can always be entered by a batch size arrival of size i from the boundary state 0, but state $i-1$ can only be entered with a batch arrival of size $i-1$. With all other transition rates into and out of the states i and $i-1$ being equal, we see that for all i the transition rate for state i depends on i. In cases where the tail of the stationary distribution has the matrix-geometric property $\pi_n = \pi_{n-1}R$ for large n, one of the necessary conditions for this matrix geometric form is that the distribution of the upward jumps has finite support (cf. chapter 1 of Neuts[80]). As a consequence, the stationary probability vector of the queueing process in the general $BMAP/M/s$ queue, with batch sizes of infinite support, is not of matrix-geometric type. □

Remark 7.3.4

A second observation is that bounding the batch sizes is not merely a technical issue; for bounded batch sizes the structure of the $BMAP/M/s$ queue is also of $GI/M/1$ type. As a consequence the stationary distribution of the number of customers has the geometric properties we discussed in Section 7.2, this is reflected by the form of expression (7.21). We find that the analysis of the spectral expansion technique is closely related to the matrix-analytic method for Quasi Birth-and-Death models with bounded jumps as is discussed by Neuts[80]. As the transform method concentrates on the solution of a functional equation connected to the matrix G, the spectral expansion method and the matrix-analytic method are connected via the vector difference equation (7.18). Moreover, from the structure of (7.21) it follows that the matrix R in the geometric tail of x_n (as mentioned in our discussion of the matrix-analytic technique in Section 7.2), is determined by the eigenvalues ζ_k and their respective eigenvectors l_k.

□

7.4 Numerical results

In this section we numerically examine our results. We concentrate on testing the iterative scheme that we have developed in the previous section for the probabilistic method.

We have refrained from including numerical performances in our comparison of the three methods. The reason for this is that it appears to be very hard to make such a comparison fair. This is due to a number of reasons:

(i) Implementation of the methods. The methods perform different operations, for which a number of alternative routines may be available.

(ii) The specific parameters of the queueing model that is being evaluated; for each method the choice of parameters might have a different impact on its performance. In one-dimensional single-server queueing models the performance of an implementation often is closely connected to the workload of the system. However, in the multi-dimensional multi-server case other aspects play a role as well.

Typical problems for Markov modulated queueing models arise due to the matrix structure; examples are coinciding eigenvalues of matrices or periodicity of the directing Markov chain.

(iii) Demands for the accuracy of the evaluation. The effect of the desired level of accuracy on running time, number of iterations etc. will not be the same for each method.

(iv) Computer specifications. For larger instances even computer specifications can play a role (memory capacity, instruction set of the processor).

Although a fair numerical comparison appears to be a hard problem, even categorizing the weak and strong points each method will be a rather formidable task, involving a vast set of test cases. Mitrani & Chakka[76] compare the spectral expansion method (which is the main subject of [76]) with the probabilistic method for a special multi-server model with breakdown and repair of the servers, However, this comparison does not lead to strong conclusions.

The remainder of this section is devoted to numerical results obtained with the probabilistic method.
The model we consider is a multi-server generalization of the collector model that is presented in Chapter 2. Similarly to Chapter 4 we apply the BMAP technique for modelling dependence between interarrival and service times. As such, this section can be considered as an extension of the study on dependence between interarrival and service times, which was the subject of Chapters 2,3 and 4.

First we briefly describe the characteristics of the multi-server collector model. For a detailed description of the single-server case we refer to Section 2.1.
Customers arrive at a bus-stop according to a Poisson process with rate λ. With exponentially(γ) distributed intervals the customers are collected, and delivered at a single-queue of a multiple-server facility which has a group of s identical servers in parallel. The service time of a customer is exponentially distributed with parameter μ. As described in Chapter 2, in this model the interarrival times of collectors and the numbers of collected customers (or the amount of work associated with the collected customers) are positively correlated.
With the modelling technique described in Chapter 4 we can fit the multiple exponential server queue with this typical dependence structure - the $M/M/s$ queue with exponential customer collection - into the framework of the $BMAP/M/s$ queue.
Actually, the number of customers in the $M/M/s$ queue with an exponential gate has been studied by Takahashi[95] (similar to Chapter 3, Takahashi uses the terminology "gates" instead of collection). As pointed out in Chapter 4, with the BMAP it is possible to model more general arrival processes of customers (e.g., Markov modulated Poisson processes) and also more general collecting distributions (e.g., Erlang distribution).

For a number of values of the parameters λ, μ, γ and s we have evaluated $\mathrm{E}\tilde{\mathbf{N}}$, being the mean number of customers in the system as given by expression (7.17).
We remark that in all cases the number of states of the underlying Markov chain has been chosen such that the model approximates the $M/M/s$ queue with exponential customer collection as closely as possible (cf. the discussion about approximating infinite bus capacity in Sections 4.3 and 4.4).
In (7.17) the only quantities that are not directly provided by the input are the vectors $x_0, \ldots, x_{s-1}$. This set of vectors that have to be evaluated is further reduced to just x_0, since $x_1, \ldots, x_{s-1}$ are obtained from x_0 via the balance equations (7.3).
The vector x_0 is given by equation (7.8). From expression (7.6) and (7.7) it follows that the only part in the evaluation of $\mathrm{E}\tilde{\mathbf{N}}$ that involves a numerical approximation procedure is the determination of the matrices $G(1), G'(1)$, $G_1(1), \ldots, G_{s-1}(1), G_1'(1), \ldots G_{s-1}'(1)$. Each matrix is obtained by iteration, with the iteration of a matrix C ending after $k+1$ steps if $\max_{ij} |C_{ij}^{(k+1)} - C_{ij}^{(k)}| < 1E-13$ (for $\gamma = 0.1$ in Tables 7.1 and 7.2 the bound $1E-15$ is used).

Remark 7.4.1
For most of the parameter settings the bound $1E-13$ is unnecessarily small, but for some instances a very small bound is vital for the stability of the evaluation. We have observed that instability may occur when x_0 is small. The main observation we make here is that instability is not directly connected to load that is offered to the system (being $\frac{\lambda}{\mu}$), but that it is a consequence of the fact that the queueing system is a multi-server model. In such systems $x_0 e$, the total probability mass at level 0, can be relatively small even for low system load. This occurs when $\frac{\lambda}{\mu s} < 1$ but with $\frac{\lambda}{\mu i} > 1$ for $i = 1, \ldots, j$, $j < s$. In this case it is likely that level 0 will be visited very rarely. As the vectors $x_1, \ldots, x_{s-1}$ are computed from x_0 and the probability masses constituted by these vectors often differ some orders of magnitude from $x_0 e$ (viz. geometric tails), a high accuracy in the computation of vector x_0 is demanded. This example illustrates item (ii) of the difficulties in comparing various analysis methods listed in the beginning of this section. □

We next consider three aspects of the $BMAP/M/s$ queue under consideration.

Dividing service capacity between servers
In Tables 7.1 and 7.2 we study the effect of the collecting mechanism on the number of customers in the system under a varying number of servers between which the total service capacity is divided; with λ fixed to 1, the ratio $\frac{\lambda}{s\mu}$ (being the load of the system) is kept constant while varying the number of servers s (i.e. few fast servers vs. many slower servers). In Table 7.1 the offered load to the system is $\frac{\lambda}{s\mu} = 0.5$, in Table 7.2 this load is $\frac{\lambda}{s\mu} = 0.9$. For reference

we have added the mean number of customers in the queueing model with instantaneous collection ($\gamma \to \infty$), i.e. the ordinary $M/M/s$ queue.

	s							
γ	1	2	3	4	5	6	7	8
0.1	6.7587	6.7913	6.8550	6.9483	7.0698	7.2179	7.3911	7.5878
0.5	1.9728	2.1069	2.3380	2.6387	2.9898	3.3776	3.7925	4.2275
1	1.4023	1.6089	1.9262	2.3064	2.7249	3.1677	3.6266	4.0965
2	1.1475	1.4188	1.7902	2.2092	2.6547	3.1164	3.5888	4.0683
4	1.0473	1.3574	1.7512	2.1832	2.6368	3.1037	3.5795	4.0615
10	1.0090	1.3375	1.7393	2.1755	2.6315	3.0999	3.5768	4.0595
20	1.0024	1.3344	1.7375	2.1743	2.6307	3.0994	3.5763	4.0606
∞	1.0000	1.3333	1.7368	2.1743	2.6304	3.0991	3.5762	4.0590

TABLE 7.1: MEAN NUMBER OF CUSTOMERS IN THE MULTI-SERVER COLLECTOR MODEL WITH A CONSTANT LOAD: $\frac{\lambda}{s\mu} = 0.5$, WITH $\lambda = 1$.

	s							
γ	1	2	3	4	5	6	7	8
0.1	25.7112	25.7576	25.8485	25.9821	26.1559	26.3671	26.6131	26.8932
0.5	11.0515	11.2546	11.6020	12.0509	12.5728	13.1489	13.7663	14.4159
1	9.6977	10.0158	10.4923	11.0580	11.6801	12.3408	13.0297	13.7403
2	9.2111	9.6197	10.1661	10.7821	11.4412	12.1301	12.8411	13.5694
4	9.0583	9.5109	10.0817	10.7127	11.3819	12.0782	12.7948	13.5276
10	9.0104	9.4803	10.0587	10.6941	11.3662	12.0645	12.7827	13.5168
20	9.0029	9.4756	10.0551	10.6912	11.3637	12.0623	12.7800	13.5015
∞	9.0000	9.4737	10.0535	10.6898	11.362	12.0611	12.7796	13.5138

TABLE 7.2: MEAN NUMBER OF CUSTOMERS IN THE MULTI-SERVER COLLECTOR MODEL WITH A CONSTANT LOAD: $\frac{\lambda}{s\mu} = 0.9$, WITH $\lambda = 1$.

A first observation in Tables 7.1 and 7.2 is that the number of customers in the system increases with s. This is to be expected since servers might be idling when there are less than s customers present. At such moments the rate of customers leaving the system is higher for queues with a few fast servers than for queues with many slow servers.

A second observation is that the differences in the mean numbers of customers are more pronounced for higher values of γ (i.e. relatively frequent collecting) than for low values of γ. Moreover, the effect of the collecting procedure is stronger for lower values of s. For example, in Table 7.1 for $s = 1$ the ratio of the mean number of customers for $\gamma = 0.1$ and $\gamma = 20$ is $6.4453 : 1.0024$, which is approximately 3.6 times higher than the same ratio for $s = 8$. This can be explained by the fact that for γ small the idling situation just described occurs less frequently than for γ large; customers are more likely to arrive in large batches than as single individuals (for $\gamma = 0.1$ the mean number of customers collected each time is $\frac{\lambda}{\gamma} = 10$, for $\gamma = 20$ the mean number of customers collected is 0.05). For γ small and for low load the servers are less efficiently used than for γ high and high load. This also explains the fact that in Table

7.2, in which case the load is 0.9, the frequency of collection has less influence on the mean number of customers than in Table 7.1, in which case the load is 0.5.
We remark that as a check for $s = 1$ the results of Tables 7.1 and 7.2 can be compared to those to Table 2.2a of Chapter 2 via Little's law and after subtracting the number of customers at the bus-stop: $E\tilde{\mathbf{N}} = \lambda(E\tilde{\mathbf{W}} + \beta) - \frac{\lambda}{\gamma}$.

The value of extra servers
In Tables 7.3 and 7.4 we examine the effect of adding extra (identical) servers on the mean number of customers in the system. In Table 7.3 (7.4) the starting point is $s = 5$, $\lambda = 1$, $\mu = 0.4$ ($\mu = \frac{2}{9}$).

	s				
γ	5	6	7	8	9
0.1	7.0698	5.7106	4.8755	4.3228	3.9349
0.5	2.9906	2.7264	2.6117	2.5578	2.5310
1	2.7249	2.5769	2.5270	2.5096	2.5034
2	2.6547	2.5441	2.5124	2.5034	2.5009
4	2.6368	2.5365	2.5096	2.5024	2.5006
10	2.6314	2.5344	2.5088	2.5021	2.5005
20	2.6307	2.5340	2.5086	2.5021	2.5005
∞	2.6304	2.5339	2.5086	2.5021	2.5004

TABLE 7.3: MEAN NUMBER OF CUSTOMERS IN THE MULTI-SERVER COLLECTOR MODEL WITH A CONSTANT SERVICE RATE: $\mu = 0.4$. $\lambda = 1$.

	s				
γ	5	6	7	8	9
0.1	26.1559	14.4655	10.8524	8.9809	7.8325
0.5	12.5769	6.4635	5.2974	4.8675	4.6782
1	11.6801	5.9511	4.9957	4.6895	4.5742
2	11.4412	5.8126	4.9180	4.6478	4.5530
4	11.3819	5.7771	4.8980	4.6373	4.5479
10	11.3662	5.7670	4.8922	4.6343	4.5463
20	11.3637	5.7655	4.8913	4.6338	4.5471
∞	11.3624	5.7650	4.8910	4.6336	4.5460

TABLE 7.4: MEAN NUMBER OF CUSTOMERS IN THE MULTI-SERVER COLLECTOR MODEL WITH A CONSTANT SERVICE RATE: $\mu = \frac{2}{9}$. $\lambda = 1$.

In Tables 7.3 and 7.4 we see that adding an extra server has slightly more effect for small values of γ than it has for large values of γ, the differences measured both in absolute and relative value. The reason for this is that for small γ the extra service capacity becomes of greater use due to the increasing number of customers which is the result of less frequently collecting the customers. As observed in Chapter 2, the waiting times of customers for the single server case, to which the number of customers can be connected, behave like $\frac{1}{\gamma}$. For the multi-server case the extra service capacity reduces the effect of γ.

The effect of adding extra servers is more substantial in Table 7.4 than it is in Table 7.3. This is readily explained by the fact that in Table 7.4 the initial load for $s = 5$ is 0.9, which is close to the critical load of the queueing system. Hence in this case an extra server brings greater relief than in the fairly lightly loaded situation of Table 7.3.
We remark that from a balance argument for the average output and input of work for the system, it follows that the mean number of customers is always greater than $\frac{\lambda}{\mu}$ which is the average amount of work entering the system. For the case of Table 7.3 (7.4) this lower bound is 2.5 (4.5).

The character of the accumulation process
In Tables 7.5 and 7.6 we investigate for $s = 1, 2, 3$ the influence of changing the ratio of γ and λ on the mean amount of work in the system. With γ fixed to 1, the value of λ is varied from 1 upto 16, while keeping $\frac{\lambda}{s\mu}$ constant (0.5 and 0.9 in Tables 7.5 and 7.6 respectively). Via this scaling operation the character of the accumulation process of customers at the bus-stop ranges from a jump process ($\lambda : \gamma = 1 : 1$) to a more continuous process ($\lambda : \gamma = 16 : 1$). Simultaneously, as follows from equation (2.12) of Chapter 2, the correlation coefficient of the interarrival time of a group of collected customers and the number of customers collected increases with λ. We remark that for exponential service time distributions the mean amount of work is just the mean number of customers times the mean service time.

	λ				
s	1	2	4	8	16
1	0.7011	0.4932	0.3956	0.3415	0.3100
2	1.6089	1.0534	0.8104	0.6881	0.6214
3	2.8893	1.7535	1.2690	1.0469	0.9359

TABLE 7.5: MEAN AMOUNT OF WORK IN A MULTI-SERVER COLLECTOR MODEL WITH A CONSTANT RATIO BETWEEN THE ARRIVAL AND SERVICE RATE: $\frac{\lambda}{\mu} = 0.5s$. $\gamma = 1$.

	λ				
s	1	2	4	8	16
1	8.7281	4.9732	3.2404	2.3679	1.9209
2	18.0288	10.1291	6.5310	4.7493	3.8454
3	28.3293	15.6627	9.9374	7.1631	5.7787

TABLE 7.6: MEAN AMOUNT OF WORK IN A MULTI-SERVER COLLECTOR MODEL WITH A CONSTANT RATIO BETWEEN THE ARRIVAL AND SERVICE RATE: $\frac{\lambda}{\mu} = 0.9s$. $\gamma = 1$.

A first observation in Tables 7.5 and 7.6 is that the mean amount of work decreases when the ratio $\lambda : \gamma$ increases; i.e. as the coefficient of correlation of the interarrival and service times increases. The fact that this property can also be observed for $s = 1$, together with the observation that compared to $s = 1$ the relative reduction for λ increasing from 1 to 16 does not radically change for $s = 2, 3$, implies that the correlation coefficient of the collecting interval and number of customers collected is an important factor.
For $s = 2$ and $s = 3$ an increment of λ also results in a better spreading of the customers over the servers; as the offered load is divided between more customers it will occur less frequently that servers are unnecessarily idling. This is reflected by the fact that for $\lambda = 16$ the mean amount of work in the system divided by the number of servers is almost equal for $s = 1, 2, 3$. In the limiting situation $\lambda \to \infty$ (and hence $\mu \to \infty$) the operative state of all servers (busy or idle) is the same at all time.

Appendix A

The BMAP and the $BMAP/G/1$ queue

In this appendix we present an overview of the theory of the Batch Markovian Arrival Process (BMAP) and the single server queue $BMAP/G/1$. The theory presented in this appendix is mostly adapted from Lucantoni[73] and the tutorial paper by the same author[74], to which we also refer for examples of the BMAP and extensive surveys of related literature. Two sources discussing the mathematical foundations of the results to be presented are the books by Neuts[80, 81].

When compared with the ordinary Poisson process, the BMAP is a much more powerful instrument to model the characteristics of modern queueing systems, while at the same time the generalization from the Poisson process to the BMAP appears only to be a single evolutionary step. The analysis of the $BMAP/G/1$ queue is also close to the analysis of the $M/G/1$ queue; most results for the $BMAP/G/1$ queue are matrix generalizations of $M/G/1$ theory. Finally, numerical algorithms for the evaluation of performance measures of the $BMAP/G/1$ queue are reasonably easy to implement.

In the remainder of this appendix we first describe the BMAP, then present an overview of results for the $BMAP/G/1$ queue.

The Batch Markovian Arrival Process

In the BMAP, the interarrival and service times of customers are directed by a continuous time Markov process $\{\mathbf{J}(t), t \geq 0\}$ on a finite state space E. A transition from a state i to a state j may induce the arrival of a *batch customer*, the size of the batch depending on i and j. Let the sojourn time in state $i \in E$ be exponentially distributed with parameter $\lambda_i > 0$, and given that a transition

takes place, let p_{ij} be the probability of a transition from i to a state $j \in E\backslash\{i\}$, $\sum_{j\in E\backslash\{i\}} p_{ij} = 1$. Defining $p_{ii} := -1$, then the generator of the Markov process $\{\mathbf{J}(t), t \geq 0\}$ is given by the matrix $D = (p_{ij}\lambda_i)$. The stationary probability (row) vector of the Markov process is usually denoted by π and satisfies the conditions $\pi D = 0$ and $\pi\, e = 1$, where e is the $|E|$ dimensional unit (column) vector.
Next, conditional on a transition from a state i to a state j, the probability of a batch arrival of size k is denoted by q_{ij}^k, $k = 0, 1, \ldots$, where q_{ij}^0 may be interpreted as the probability of having no arrival or of the arrival of an empty batch. With this notation, $p_{ij}q_{ij}^k\lambda_i$ is the transition rate from state i to state j inducing a batch arrival of size k. The BMAP is completely characterized by the sequence of matrices $D_k = (p_{ij}q_{ij}^k\lambda_i)$, $k = 0, 1, \ldots$, $\sum_{k=0}^{\infty} D_k = D$. Associated with this sequence of matrices is the matrix generating function $D(z) = \sum_{k=0}^{\infty} D_k z^k$, $|z| \leq 1$. A special case is the case when $D_k = 0$ for $k \geq 2$, this special non-renewal process is referred to as the Markovian Arrival Process (MAP).

With the above-described mechanism we have constructed a counting process $\{\mathbf{N}(t), \mathbf{J}(t), t \geq 0\}$, where $\mathbf{N}(t)$ has state space {0,1,...}. In the counting process $\mathbf{N}(t)$ describes the number of arrivals up to time t and $\mathbf{J}(t)$ the state of the underlying Markov process at time t. $\{\mathbf{N}(t), \mathbf{J}(t), t \geq 0\}$ is Markov process with a generator Q and a transition probability function $P_{ij}(n,t) = \Pr\{\mathbf{N}(t) = n, \mathbf{J}(t) = j | \mathbf{N}(0) = 0, \mathbf{J}(0) = i\}$. It is readily verified that the matrix generating function $P^*(z,t) = (\sum_{n=0}^{\infty} P_{ij}(n,t)z^n)$, $|z| \leq 1$ is given by $P^*(z,t) = e^{D(z)t}$, $|z| \leq 1, t \geq 0$. An example of the one-dimensional case is the ordinary Poisson process, for which $P^*(z,t) = e^{-\lambda(1-z)t}$.

Before turning to the $BMAP/G/1$ queue we first state some properties of the BMAP:
(i) Stationary BMAP's are dense in the class of all stationary point processes. (cf. Asmussen & Koole[9]).
(ii) The superposition of two BMAP's is again a BMAP.
Finally we mention that there exists a discrete time analog of the BMAP, often referred to as the DBMAP.

The $BMAP/G/1$ queue
The $BMAP/G/1$ queue is defined as the single server queue with the BMAP describing the arrivals of batches of customers. The customers are served in order of arrival, their service times are independent and identically distributed with distribution function $H(\cdot)$, with Laplace-Stieltjes Transform (LST) $h(\cdot)$, and finite first and second moment h_1 and h_2 respectively.
The stability condition for this queue is $\rho = \pi \sum_{k=0}^{\infty} kD_k e\, h_1 < 1$, with $\pi \sum_{k=0}^{\infty} kD_k e$ being the stationary arrival rate of customers (cf. Ramaswami[85]).
The analysis of the $BMAP/G/1$ queue starts by studying the Markov process

describing the vector random variable of the number of customers $\mathbf{X}(t)$ and the state of the Markov process $\mathbf{J}(t)$ directly after the moment of a service completion. This embedded Markov chain is denoted by $\{\mathbf{X}_n, \mathbf{J}_n, n = 1, 2, \ldots\}$. There are two elements that characterize the behaviour of the $BMAP/G/1$ queue:

(i) The probability matrices $A_n(x)$, $n = 0, 1, \ldots$, with
$(A_n(x))_{ij}$:= Pr{ Given that at time 0 a new service starts and $\mathbf{J}(0) = i$, the service is completed at a time $x_1 \leq x$ with $\mathbf{J}(x_1) = j$, and during that service there were exactly n arrivals }.

Defining the matrix $A(z,s) := \sum_{n=0}^{\infty} \int_{x=0}^{\infty} e^{-sx} dA_n(x) z^n$, the following holds

$$A(z,s) = \int_{x=0}^{\infty} e^{-sx} e^{D(z)x} dH(x), \quad |z| \leq 1, \ Re\, s \geq 0. \tag{A.1}$$

Expression (A.1) can be interpreted as the matrix equivalent of the functional describing the length of a service time and the number of customers that arrive during that service in the ordinary $M/G/1$ queue, the latter being given by $A(z,s) = h(s + \lambda - \lambda z)$.

(ii) The busy period. Similarly to $(A_n(x))$ we define matrices $G_n(x,y)$
$(G_n(x,y))_{ij}$:= Pr{ Given that at time 0 the workload consisting of the not completed service requests is y and $\mathbf{J}(0) = i$, the busy period ends at a time $x_1 \leq x$ with $\mathbf{J}(x_1) = j$, and during the busy period exactly n new customers have arrived and have been served }.

Defining $G(z,s) := \int_{y=0}^{\infty} \sum_{n=0}^{\infty} \int_{x=0}^{\infty} e^{-sx} dG_n(x,y) z^n dH(y)$, which describes the number of customers served during a busy period and the length of that busy period, it can be shown that this function satisfies the following functional equation

$$G(z,s) = z \int_{x=0}^{\infty} e^{-sx} e^{D[G(z,s)]x} dH(x), \text{ for } |z| \leq 1, \ Re\ s \geq 0, \tag{A.2}$$

with $D[G(z,s)] := \sum_{k=0}^{\infty} D_k \, (G(z,s))^k$.

Again, expression (A.2) can be seen as the matrix equivalent of the busy period equation for the ordinary $M/G/1$ queue: $G(z,s) = zh(s + \lambda - \lambda G(z,s))$ (cf. Takács[94] p.32) .
An important observation is the exponential form of the argument in the integral in expression (A.2) which reflects the convolution semi-group property of

the matrix random variable describing the busy-period; defining $\widehat{G}(y,z,s) := \sum_{n=0}^{\infty} \int_{x=0}^{\infty} e^{-sx} dG_n(x,y) z^n$, the following holds

$$\widehat{G}(y_1+y_2,z,s) = \widehat{G}(y_1,z,s)\,\widehat{G}(y_2,z,s), \tag{A.3}$$

with $y_1, y_2 \geq 0$, $|z| \leq 1$, $Re\ s \geq 0$.
According to Feller[46, theorem 2 p.303], the class of probability distributions associated with continuous convolution semi-groups (which is the class of distributions of increments in processes with stationary independent increments) is identical to the class of infinitely divisible distributions. The latter class is characterized by the exponential form of the LST as observed in expression (A.3) (cf. [46, section X.9 p.353-355 and theorem 1 p.450]).

The matrix $G(z,s)$ and the derived matrices $G(s) := G(1,s)$ and $G := G(0)$ are key elements in the analysis of the stationary performance measures of the $BMAP/G/1$; the busy period is related to first-passage times and recurrence times. In particular for the boundary states of the process $\{\mathbf{X}(t), \mathbf{J}(t), t \geq 0\}$, which are the states where the server is idle ($\mathbf{X}(t) = 0$), we are interested in recurrence times because these are closely connected to the vector g of stationary probabilities $\lim_{t\to\infty}(\Pr\{\mathbf{J}(t) = j | \mathbf{X}(t) = 0\})$. It can be shown that g satisfies $gG = g$ and $ge = 1$.

Stationary results
Below we present expressions for a few performance measures. The derivations of these results and others are spread over a number of papers by Neuts, Ramaswami and Lucantoni; we refer to Lucantoni[73] for an extensive list of references. Again we remark that the expressions for the performance measures in the $BMAP/G/1$ queue are matrix generalizations of their $M/G/1$ counterparts.
The results presented are for the stationary situation, we remark that recently also work has been performed for the transient case[29, 30]. For the latter case transform expressions of the distributions are quite easily obtained, but formulas for moments and numerical algorithms are not as explicit as for the stationary case.

The number of customers at an arbitrary moment
Define $y_{ki} := \lim_{t\to\infty} \Pr\{\mathbf{X}(t) = k, \mathbf{J}(t) = i\}$ and the vector of generating functions $Y(z) = (\sum_{k=0}^{\infty} z^k y_{ki})$. Then

$$Y(z) = (1-\rho) g(z-1) A(z) [zI - A(z)]^{-1}, \quad |z| \leq 1, \tag{A.4}$$

with $A(z) := A(z,0)$, I the identity matrix, and g the vector of stationary probabilities $\lim_{t\to\infty}(\Pr\{\mathbf{J}(t) = j | \mathbf{X}(t) = 0\})$.
From (A.4) the generating function of the number of customers at arrival and departure moments readily follows; the vector of generating functions $X(z)$ of the queue length at arrival instants is

$$X(z) = -(\pi D'(1)e)^{-1}(1-\rho)gD(z)A(z). \tag{A.5}$$

From (A.4) and (A.5) one can obtain the vectors $Y'(1)$ and $X'(1)$ of first moments.

The virtual waiting time
Define the vector random variable $\mathbf{V}$ with distribution function $V(\cdot)$

$V_j(x) :=$ Pr{ At an arbitrary time the amount of work in the system is at most x and the directing Markov process is in state j }.

Let $v(s)$ denote the LST of $V(\cdot)$, then $v(s)$ satisfies

$$\begin{aligned} v(s) &= s(1-\rho)g[sI + D(h(s))]^{-1}, \ Re\, s \geq 0, \\ v(0) &= \pi. \end{aligned} \tag{A.6}$$

The LST $w(s)$ of the waiting time distribution of an arbitrary batch customer can be obtained from (A.6) by conditioning on the state of the directing Markov process given that an arrival takes place. Let $q_i = (\sum_{k=1}^{\infty} D_k e)_i/\lambda_i$ denote the probability that a transition from i generates a batch arrival, then

$$w(s) = \frac{v(s)q}{\pi q}, \ Re\ s \geq 0. \tag{A.7}$$

For the numerical experiments in Sections (4.4) and (6.5) we calculated the mean virtual waiting time from

$$\mathbf{EV} = (\mathbf{EV})e\pi + \pi - ((1-\lambda'\beta)g + \beta\pi D'(1))[e\pi + D]^{-1}, \tag{A.8}$$

in which $(\mathbf{EV})e$ is given by

$$\begin{aligned} (\mathbf{EV})e = \frac{1}{2(1-\rho)} \Big[& 2(\rho - ((1-\rho)g + \pi\beta D'(1))\ [e\pi + D]^{-1}\beta D'(1)e) \\ & + \pi(h_1^2 D''(1) + h_2 D'(1))e \Big]. \end{aligned} \tag{A.9}$$

Numerical implications
The expressions for the moments of performance measures look complicated; they involve inverses of matrices (e.g. $[e\pi + D]^{-1}$) and solutions of functional equations and linear sets of equations (cf. (A.2), or $gG = g$). In practice, however, the numerical evaluations can be done quite efficiently. For example, when the LST of the service time distribution is explicitly known, the solution of (A.2) reduces to an iterative procedure only involving elementary arithmetic operations.

Appendix B

Ergodicity of multi-dimensional Markov chains

In this appendix we describe a technique for proving the ergodicity of multi-dimensional Markov processes. The technique is discussed in Laslett et al.[69], to which we refer for technical details.

The key idea behind the method is that a Markov process can only be ergodic if there exists at least one finite (or bounded) subset of the state space for which the mean recurrence time is finite. In Laslett et al.[69] it is argued that for certain subsets (so-called *test-sets*) this condition is sufficient for ergodicity.

The method has two attractive features. Firstly, the technique can be applied for a large class of multi-dimensional Markov Processes. For instance, in Chapters 2 and 3 we use the method to prove the ergodicity of the two-dimensional process that describes the waiting and the service time of a customer in the $M/G/1$ queue with correlated interarrival and service times of customers. In Chapter 2 we show ergodicity of this two-dimensional random variable by considering a Markov process with a discrete and a continuous component, in Chapter 3 the ergodicity is obtained by considering a continuously-valued Markov process. In Chapter 5 we use the method for the $MAP/G/1$ queue with impatient customers, the Markov process under consideration there being the two-dimensional process describing the amount of work in the system and the state of a directing Markov chain (cf. Appendix A). In this last case the states of the Markov process have both a discrete and a continuous component.

A second feature is that the method might also be applied when the Markov process has state-dependent transitions, the latter being the case in the $MAP/G/1$ queue with impatient customers.

After this brief introduction we next discuss the method of [69] in some detail.

First we mention a number of concepts that are used in [69].

ϕ-irreducibility.
A Markov chain $\{X_n\}$ with state space χ is called *ϕ-irreducible* if there exists a nonzero measure ϕ on F (F is the σ-algebra of Borel subsets of χ), such that for any $x \in \chi$ and $A \in F$ with $\phi(A) > 0$, there exists an n for which $P^n(x, A) > 0$. $P^n(x, A)$ denotes the n-step transition probability of $\{X_n\}$.

Ergodicity.
If $\{X_n\}$ is ϕ-irreducible then there exists *at most one* stationary distribution, $\{X_n\}$ is *ergodic* if it has a *unique* stationary distribution. If $\{X_n\}$ is ergodic the n-step probability transitions converge to this stationary distribution in the strong Cesaro sense.

Hitting times.
Hitting times T_A are defined as $T_A := \inf\{n > 0 | X_n \in A\}$.

Test set.
Assume $\{X_n\}$ is ϕ-irreducible. Then, if $\sup_{x \in A} \mathrm{E}[T_A | X_0 = x] < \infty$ is a sufficient condition for $\{X_n\}$ to be ergodic, $A \in F$ is called a *test set.*

The main theorem for testing this condition for a set $A \in F$ is (cf. Theorem 2.1 of [69])

Theorem B.1
Let g be a nonnegative function on χ. If for some $\epsilon > 0$ and $A \in F$

$$\mathrm{E}[g(X_1) | X_0 = x] \leq g(x) - \epsilon, \text{ for } x \in A^c,$$

then

$$\sup_{x \in A} \mathrm{E}[T_A | X_0 = x] < \infty.$$

□

The scheme in [69] for proving that a Markov chain $\{X_n\}$ is ergodic consists of three steps:
1. Prove that $\{X_n\}$ is ϕ-irreducible.
2. Identify possible test sets.
3. Apply Theorem B.1 to one of these sets.

Step 1 and 3 are both quite transparent, the technical details of the method are covered in step 2 of the scheme; under what conditions does the application of Theorem B.1 prove ergodicity of the Markov process under consideration. This question is addressed in Tweedie[98]. In [69] the general results of [98] are summarized.

In most queueing applications step 1 of the scheme is readily performed. Usually ϕ can be chosen such that the state-space contains an atom, e.g. the state

connected to an empty system. If it is hard to identify an atom off-hand, still in most cases a bounded subset of the state-space, containing the state(s) connected to an empty system suffices for proving ϕ-irreducibility.
Identifying possible test sets is trivial when ϕ has an atom, say α. In that case $\{\alpha\}$ is a test set (cf. theorem 3.1 of [69]). If step 3 of the scheme can not be applied for the test set $\{\alpha\}$, as was the case in Chapter 5, under mild conditions a subset B of the state-space containing α is a test set (cf. theorem 3.2 of [69]).
If ϕ does not contain an atom, then step 2 can become quite hard. However, in queueing applications the effort of identifying possible test sets might only consist of a number of relatively elementary steps. An example is provided in Appendix 3.A, for an extensive description of such steps we refer to sections 4,5 and 6 of [69].
Step 3 of the scheme in queueing applications is usually quite clear, the function g to be used usually is connected to performance measures such as the amount of work or the number of customers present in the system. In Chapters 2,3 and 5 the function g was connected to the total amount of work in the system.

Bibliography

[1] Abate, J., Whitt, W. (1991). The Fourier-series method for inverting transforms of probability distributions. *Queueing Systems* **10**, 5-88.

[2] Abate, J., Choudhury, G.L., Whitt, W. (1993). Calculation of the $GI/G/1$ waiting-time distribution and its cumulants from Pollaczek's formulas. *AEÜ* **47**, 311-321, *Special Issue on Teletraffic Theory and Engineering in Memory of F. Pollaczek.*

[3] Agrawala, A.K., Tripathi, S.K. (1981). On the optimality of semidynamic routing schemes. *Inf. Proc. Letters* **13**, 20-22.

[4] Agrawala, A.K., Tripathi, S.K. (1982). On an exponential server with general cyclic arrivals. *Acta Inf.* **18**, 319-334.

[5] Al-Saadi, S.D., Young, D.H. (1980). Estimators for the correlation coefficient in a bivariate exponential distribution. *J. Statist. Comput. Simul.* **11**, 13-20.

[6] Ali, O.M.E., Neuts, M.F. (1984). A service system with two stages of waiting and feedback of customers. *J. Appl. Prob.* **21**, 404-413.

[7] Apostol, T.M. (1974). *Mathematical Analysis* (Addison-Wesley, Reading, 2nd ed.).

[8] Asmussen, S. (1989). Aspects of matrix Wiener-Hopf factorization in applied probability. *Math. Scientist* **14**, 101-116.

[9] Asmussen, S., Koole, G.M. (1993). Marked point processes as limits of Markovian arrival streams. *J. Appl. Prob.* **30**, 365-372.

[10] Baccelli, F., Boyer, P., Hebuterne, G. (1984). Single-server queues with impatient customers. *Adv. Appl. Prob.* **16**, 887-905.

[11] Baily, D.E., Neuts, M.F. (1981). Algorithmic methods for multi-server queues with group arrivals and exponential services. *Eur. J. Oper. Res.* **8**, 184-196.

[12] Bisdikian, C., Lew, J.S., Tantawi, A.N. (1992). The generalized $D^{[X]}/D/1$ queue and its application in the analysis of bridged high speed token-ring networks. IBM Res. Rep., RC 18387.

[13] Bitran, G.R., Dasu, S. (1992). A review of open queueing networks models of manufacturing systems. *Queueing Systems* **12**, 95-133.

[14] Blanc, J.P.C. (1993). Performance analysis and optimization with the power-series algorithm. In: *Models and Techniques for Performance Evaluation of Computer and Communication Systems,* Eds. L. Donatiello, R.D. Nelson (Springer, Berlin), 53-80.

[15] Boel, R.K., van Schuppen, J.H. (1989). Distributed routing for load balancing. *Proc. IEEE* **77**, 210-221.

[16] Bonomi, F., Kumar, S. (1990). Adaptive optimal load balancing in a nonhomogeneous multiserver system with a central job scheduler. *IEEE Trans. Computers* **39**, 1232-1250.

[17] Bonomi, F., Meyer, J., Montagna, S., Paglino, R. (1994). Minimal on/off source models for ATM traffic. In: *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks*, Eds. J. Labetoulle, J.W. Roberts (Elsevier, Amsterdam), 387-400.

[18] Borst, S.C., Boxma, O.J., Combé, M.B. (1991). An $M/G/1$ queue with dependence between interarrival and service times. CWI Report BS-R9125.

[19] Borst, S.C., Boxma, O.J., Combé, M.B. (1992). Collection of customers: A correlated M/G/1 queue. *Perf. Eval. Review* **20**, 47-59.

[20] Borst, S.C., Boxma, O.J., Combé, M.B. (1993). An $M/G/1$ queue with customer collection. *Stochastic Models* **9**, 341-371.

[21] Borst, S.C., Combé, M.B. (1992). Busy period analysis of a correlated queue. *J. Appl. Prob.* **29**, 482-483.

[22] Boxma, O.J. (1989). Workloads and waiting times in single-server systems with multiple customer classes. *Queueing Systems* **5**, 185-214.

[23] Boxma, O.J., Levy, H., Weststrate, J.A. (1991). Efficient visit frequencies for polling tables: minimization of waiting costs. *Queueing Systems* **9**, 133-162.

[24] Boxma, O.J., Levy, H., Yechiali, U. (1992). Cyclic reservation schemes for efficient operation of multiple-queue single-server systems. *Annals of Operations Research* **35**, 187-208, *Special Issue on Stochastic Modelling of Telecommunication Systems.*

[25] Boxma, O.J., Combé, M.B. (1993). The correlated $M/G/1$ queue. *AEÜ* **47**, 330-335, *Special Issue on Teletraffic Theory and Engineering in memory of F. Pollaczek.*

[26] Boxma, O.J., de Waal, P.R. (1994). Multiserver queues with impatient customers. In: *The Fundamental Role of Teletraffic in the Evolution of*

Telecommunications Networks, Eds. J. Labetoulle, J.W. Roberts (Elsevier, Amsterdam), 743-756.

[27] Buzacott, J.A., Shanthikumar, J.G. (1992). Design of manufacturing systems using queueing models. *Queueing Systems* **12**, 135-213.

[28] Buzen, J.P., Chen, P.P.-S. (1974). Optimal load balancing in memory hierarchies. In: *Proceedings of IFIP* 1974, Ed. J.L. Rosenfeld (North-Holland, Amsterdam), 271-275.

[29] Choudhury, G.L., Lucantoni, D.L., Whitt, W. (1994). Multidimensional transform inversion with applications to the transient $M/G/1$ queue. To appear in: *Ann. Appl. Prob.*.

[30] Choudhury, G.L., Lucantoni, D.L., Whitt, W. (1994). The transient BMAP/G/1 queue. *Stochastic Models* **10**, 145-182.

[31] Cidon, I., Guérin, R., Khamisy, A., Sidi, M. (1991). On queues with inter-arrival times proportional to service times. Report Technion, EE PUB No. 811, December 1991.

[32] Cidon, I., Guérin, R., Khamisy, A., Sidi, M. (1991). Analysis of a correlated queue in communication systems. Report Technion, EE PUB No. 812, December 1991.

[33] Cinlar, E. (1975). Markov renewal theory: A survey. *Management Science* **21**, 727-752.

[34] Cohen, J.W. (1982). *The Single Server Queue* (North-Holland, Amsterdam, 2nd ed.).

[35] Coleman, R.D. (1973). Use of a gate to reduce the variance of delays in queues with random service. *Bell Syst. Techn. J.* **52**, 1403-1422.

[36] Combé, M.B., Boxma, O.J. (1994). Optimization of static traffic allocation policies. *Theoretical Computer Science A* **125**, 17-43.

[37] Combé, M.B. (1994). Modelling dependence between interarrival and service times with Markovian arrival processes. CWI Report BS-R9412. (submitted for publication).

[38] Combé, M.B. (1994). Impatient customers in the $MAP/G/1$ queue. CWI Report BS-R9413. (submitted for publication).

[39] Combé, M.B. (1994). The $BMAP/M/s$ queue. CWI Report BS-R9435.

[40] Conolly, B.W. (1968). The waiting time for a certain correlated queue. *Oper. Res.* **15**, 1006-1015.

[41] Conolly, B.W., Choo, Q.H. (1979). The waiting time process for a generalized correlated queue with exponential demand and service. *SIAM J. Appl. Math.* **37**, 263-275.

[42] Conolly, B.W., Hadidi, N. (1969). A correlated queue. *J. Appl. Prob.* **6**, 122-136.

[43] Elwalid, A.I., Mitra, D., Stern, T.E. (1991). Statistical multiplexing of Markov modulated sources: theory and computational algorithms. In: *Teletraffic and Datatraffic*, Eds. A. Jensen, V.B. Iversen, 495-500.

[44] Ephremides, A., Varaiya, P., Walrand, J. (1980). A simple dynamic routing problem. *IEEE Trans. Automatic Control* **AC-25**, 690-693.

[45] Ettl, M., Mitrani, I. (1994). Applying spectral expansions in evaluating the performance of multiprocessor systems. In: *Performance Evaluation of Parallel and Distributed Systems — Solution Methods*, Eds: O.J. Boxma, G.M. Koole (CWI Tract 105 & 106, Amsterdam), 45-58.

[46] Feller, W. (1971). *An Introduction to Probability Theory and Its Applications, Vol. II* (Wiley, New York, 2nd ed.).

[47] Fendick, K.W., Saksena, V.R., Whitt, W. (1989). Dependence in packet queues. *IEEE Trans. Comm.* **37**, 1173-1183.

[48] Fischer, W., Meier-Hellstern, K. (1993). The Markov-modulated Poisson process (MMPP) cookbook. *Perf. Eval.* **18**, 149-171.

[49] Fuhrmann, S.W., Cooper, R.B. (1985). Stochastic decompositions in the $M/G/1$ queue with generalized vacations. *Oper. Res.* **33**, 1117-1129.

[50] Gail, H.R., Hantler, S.L., Taylor, B.A. (1992). Spectral analysis of M/G/1 type Markov chains. IBM Res. Rep., RC 17765.

[51] Gail, H.R., Hantler, S.L., Konheim, A., Taylor, B.A. (1992). The transform method for $M/G/1$ type Markov chains. IBM Res. Rep., RC 17891.

[52] Gohberg, I., Lancaster, P., Rodman, L. (1982). *Matrix Polynomials* (Academic Press, New York).

[53] Grünenfelder, R., Robert, S. (1994). Which arrival law parameters are decisive for queueing system performance. In: *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks*, Eds. J. Labetoulle, J.W. Roberts (Elsevier, Amsterdam), 377-386.

[54] Hadidi, N. (1981). Queues with partial correlation. *SIAM J. Appl. Math.* **40**, 467-475.

[55] Hadidi, N. (1985). Further results on queues with partial correlation. *Oper. Res.* **33**, 203-209.

[56] Hajek, B. (1985). Extremal splittings of point processes. *Math. of Oper. Res.* **10**, 543-556.

[57] Hasofer, A.M. (1963). On the integrability, continuity and differentiability of a family of functions introduced by L. Takács, *Ann. Math. Statist.* **34**, 1045-1049.

[58] Hordijk, A., Koole, G.M., Loeve, J.A. (1993). Analysis of a customer assignment model with no state information. To appear in: *Prob. Eng. Info. Sciences.*

[59] Ibaraki, T.I., Katoh, N. (1988). *Resource Allocation Problems* (MIT Press, Cambridge).

[60] Ishizaki, F., Takine, T., Hasegawa, T. (1994). Analysis of a discrete-time queue with a gate. In: *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks*, Eds. J. Labetoulle, J.W. Roberts (Elsevier, Amsterdam), 169-178.

[61] Itai, A., Rosberg, Z. (1984). A golden ratio control policy for a multiple-access channel. *IEEE Trans. Automatic Control* **AC-29**, 712-718.

[62] Jacobs, P.A. (1980). Heavy traffic results for single-server queues with dependent (EARMA) service and interarrival times. *Adv. Appl. Prob.* **12**, 517-529.

[63] Jean-Marie, A. (1988). Load balancing in a system of two queues with resequencing. In: *Proceedings of Performance '87*, Eds. P.J. Courtois, G. Latouche (North-Holland, Amsterdam), 75-88.

[64] Keilson, J., Servi, L.D. (1990). The distributional form of Little's law and the Fuhrmann-Cooper decomposition. *Oper. Res. Letters* **9**, 239-247.

[65] Kleinrock, L. (1964). *Communication Nets* (McGraw-Hill, New York).

[66] Krämer, W., Langenbach-Belz, M. (1976). Approximate formulae for the delay in the queueing system GI/G/1. In: *Proceedings of the 8th ITC Congress*, Melbourne, 1976, 235.1-235.8 .

[67] Lancaster, P., Tismenetsky, M. (1985). *The Theory of Matrices* (Academic Press, Orlando).

[68] Langaris, C. (1987). Busy-period analysis of a correlated queue with exponential demand and service. *J. Appl. Prob.* **24**, 476-485.

[69] Laslett, G.M., Pollard, D.M., Tweedie R.L. (1978). Techniques for establishing ergodic and recurrence properties of continuous-valued Markov chains. *Naval Res. Logist. Quart.* **25**, 455-472.

[70] Latouche, G., Ramaswami, V. (1993). A logarithmic reduction algorithm for quasi-birth-death processes. *J. Appl. Prob.* **30**, 650-674.

[71] Loynes, R.M. (1962). The stability of a queue with non-independent interarrival and service times. *Proc. Cambridge Phil. Soc.* **58**, 497-520.

[72] Loynes R. M. (1962). A continuous-time treatment of certain queues and infinite dams. *J. Austr. Math. Soc.* **2**, 484-498.

[73] Lucantoni, D.M. (1991). New results on the single server queue with a batch Markovian arrival process. *Stochastic Models* **7**, 1-46.

[74] Lucantoni, D.M. (1993). The $BMAP/G/1$ queue: A tutorial. In: *Models and Techniques for Performance Evaluation of Computer and Communication Systems,* Eds. L. Donatiello, R. Nelson (Springer, Berlin), 330-358.

[75] Mitrani, I., Mitra, D. (1992). A spectral expansion method for random walks on semi-infinite strips. In: *Iterative Methods in Linear Algebra,* Eds. R. Bauwens, P. de Groen (North-Holland, Amsterdam), 141-149.

[76] Mitrani, I., Chakka, R. (1993). Application and evaluation of the spectral expansion solution method. In: *QMIPS Workshop on Formalisms, Principles and State-of-the-Art*, Eds. N. Götz, U. Herzog, M. Rettelbach (Erlangen), 253-267.

[77] Naor, P., Yechiali, U. (1971). Queueing problems with heterogenous arrivals and services. *Oper. Res.* **19**, 772-734.

[78] Nassehi, M.M. (1989). CRMA: An access scheme for high-speed LANs and MANs. Report IBM Zürich.

[79] Neuts, M.F. (1971). A queue subject to exogeneous phase changes. *Adv. Appl. Prob.* **3**, 78-119.

[80] Neuts, M.F. (1981). *Matrix-Geometric Solutions in Stochastic Models* (The Johns Hopkins University Press, Baltimore).

[81] Neuts, M.F. (1989). *Structured Stochastic Matrices of M/G/1 Type and their Applications* (Dekker, New York).

[82] Ni, L.M., Hwang, K. (1985). Optimal load balancing in a multiple processor system with many job classes. *IEEE Trans. Software Engineering* **SE-11**, 492-496.

[83] Prabhu, N.U. (1965). *Queues and Inventories* (Wiley, New York).

[84] Ramakrishnan, K.K. (1983). *The Design and Analysis of Resource Allocation Policies in Distributed Systems* (Ph.D. Thesis, University of Maryland, Dept. of Computer Science).

[85] Ramaswami, V. (1980). The N/G/1 queue and its detailed analysis. *Adv. Appl. Prob.* **12**, 222-261.

[86] Regterschot, G.J.K., de Smit, J.H.A. (1986). The queue $M/G/1$ with Markov modulated arrivals and services. *Math. of Oper. Res.* **11**, 465-483.

[87] Regterschot, G.J.K. (1987). *Wiener-Hopf Factorization Techniques in Queueing Models* (Ph.D Thesis, University of Twente, Dept. of Applied Mathematics).

[88] Ross, K.W., Yao, D.D. (1991). Optimal load balancing and scheduling in a distributed computer system. *Journal of the ACM* **38**, 676-690.

[89] Shanthikumar, J.G., Buzacott, J.A. (1980). On the approximations to the single server queue. *Int. J. Prod. Res.* **18**, 761-773.

[90] De Smit, J.H.A. (1983). The queue $GI/M/s$ with customers of different types or the queue $GI/H_m/s$. *Adv. Appl. Prob.* **15**, 392-419.

[91] De Smit, J.H.A. (1985). The queue $GI/H_m/s$ in continuous time. *J. Appl. Prob.* **22**, 214-222.

[92] Stewart, W.J. (editor) (1991). *Numerical Solution of Markov Chains* (Marcel Dekker, New York).

[93] Stoyan, D. (1983). *Comparison Methods for Queues and other Stochastic Models* (Wiley, New York, Translated and revised version of German original (1977)).

[94] Takács, L. (1962). *Introduction to the Theory of Queues* (Oxford University Press, New York).

[95] Takahashi, Y. (1971). Queueing systems with gates. *J. Oper. Res. Soc. Jap.* **14**, 103-126.

[96] Titchmarsh, E.C. (1939). *The Theory of Functions* (Oxford University Press, London, 2nd ed.).

[97] Towsley, D. (1993). Providing quality of service in packed switched networks. In: *Models and Techniques for Performance Evaluation of Computer and Communication Systems,* Eds. L. Donatiello, R. Nelson (Springer, Berlin), 560-586.

[98] Tweedie, R.L. (1976). Criteria for classifying general Markov chains. *Adv. Appl. Prob.* **8**, 737-771.

[99] Tang, C.S., van Vliet, M. (1994). Traffic allocation for manufacturing systems. *Eur. J. Oper. Res.* **75**, 171-185.

[100] Wang, Y-T., Morris, R.J.T. (1985). Load sharing in distributed systems. *IEEE Trans. Computers* **C-34**, 204-217.

[101] Wolff, R.W. (1989). *Stochastic Modeling and the Theory of Queues* (Prentice Hall, Englewood Cliffs, NJ).

[102] Yum, T.P. (1981). The design and analysis of a semidynamic deterministic routing rule. *IEEE Trans. Comm.* **29**, 498-504.

[103] Zhao, Y. (1994). Analysis of the $GI^X/M/c$ model. *Queueing Systems* **15**, 347-364.

Samenvatting

De toename in volume en complexiteit van communicatieverkeer geeft naast een behoefte aan technologische vernieuwing ook een dwingende vraag naar kwalitatieve en kwantitatieve methoden voor de bestudering van verkeersstromen in telecommunicatie- en computersystemen. Wachtrijmodellen vormen hiervoor een natuurlijk paradigma, en tegenwoordig speelt de wachtrijtheorie een belangrijke rol bij ontwerp, fine-tuning, besturing en prestatie-analyse van telecommunicatie- en computersystemen.
Dit proefschrift bespreekt modellen en methoden voor de analyse en besturing van verkeersstromen in communicatie- en computersystemen.

In het proefschrift spelen twee thema's een hoofdrol:
(i) afhankelijkheidsstructuren in het aankomstproces van werk
(ii) beheersing van verkeersstromen.
Hieronder worden beide kort besproken.

(i) Afhankelijkheidsstructuren in het aankomstproces van werk
Bij de analyse van wachtrijsystemen is het gebruikelijk om het aankomstprocess met een Poisson proces te modelleren. Echter, aankomstprocessen in moderne wachtrijsystemen zijn moeilijk op een bevredigende manier te beschrijven met een Poisson proces of zelfs een algemeen vernieuwingsproces. Ook de bedieningsduren van klanten laten zich niet altijd karakteriseren als een reeks van onafhankelijke en identiek verdeelde stochastische grootheden.
In het algemeen worden drie vormen van afhankelijkheid beschouwd: tussen opeenvolgende tussenaankomsttijden (de tussenaankomsttijd is de tijd tussen twee opeenvolgende aankomstmomenten van klanten), tussen opeenvolgende bedieningsduren, en afhankelijkheid tussen de tussenaankomsttijd en de bedieningsduur van een klant. Deze eerste twee vormen van afhankelijkheid zijn tamelijk uitvoerig bestudeerd.

In dit proefschrift zijn de hoofdstukken 2,3 en 4 gewijd aan afhankelijkheid tussen tussenaankomsttijd en bedieningsduur van een klant. Deze vorm van afhankelijkheid heeft tot op heden nog niet zo veel aandacht gekregen als de twee andere afhankelijkheidsstructuren. Toch treedt deze vorm van afhankelijkheid op zeer natuurlijke wijze op in wachtrijsystemen. In communicatienetwerken zijn gecorreleerde tussenaankomsttijden en bedieningsduren ten dele inherent aan de structuur van het netwerk. De bedieningsduur van een klant (bericht) in een communicatienetwerk is sterk gekoppeld aan de lengte van het bericht, en deze lengte blijft dikwijls ongewijzigd gedurende het traject dat het bericht doorloopt van oorsprong naar bestemming. Beschouwen we nu een knooppunt in een communicatienetwerk, dan zien we dat de tussenaankomsttijd van een bericht bij dit knooppunt sterk gerelateerd is aan de bedieningsduur van het bericht in het vorige knooppunt, en dus ook aan de bedieningsduur in het huidige knooppunt. Dit creëert een afhankelijkheid tussen tussenaankomsttijd en bedieningsduur. Het optreden van deze afhankelijkheidsstructuur kan ook het gevolg zijn van operationele aspecten. Hierbij denken we met name aan reserveringsprotocollen. Als voorbeeld nemen we een verbindingspunt tussen twee netwerken. Berichten in het ene netwerk met een bestemming in het andere netwerk kunnen met tussenpozen opgehaald worden en afgeleverd worden bij het verbindingspunt. Bij dit ophaalmechanisme is het aantal berichten dat verzameld wordt, en dus ook de totale werklast die deze berichten opleveren voor de bediende, positief gecorreleerd aan de verstreken tijd tussen twee opeenvolgende ophaalmomenten.

(ii) Beheersing van verkeersstromen

In systemen waarin individuele gebruikers capaciteit moeten delen worden besturingsmechanismen gebruikt om deze capaciteit zo te verdelen dat iedere gebruiker de service krijgt waar hij recht op heeft.

Besturingsmechanismen laten zich ruwweg onderverdelen in twee categorieën: toelatingsregels en routeringsstrategieën. Toelatingsregels bepalen wanner een nieuw verzoek tot service in behandeling wordt genomen. Routeringsstrategieën bepalen door welke server(s) (processor, kanaal) een klant(taak, bericht) bediend zal worden.

Besturingsmechanismen leiden tot tal van typische wachtrijsystemen en optimaliseringsvraagstukken. In hoofdstuk 5 bespreken we een model waarin de beslissing over toelating van nieuwe klanten gerelateerd is aan de hoeveelheid werk in het systeem op het tijdstip van aankomst van een klant. Hoofdstuk 6 is gewijd aan het vinden van de optimale toewijzingsstrategie van klanten binnen een bepaalde klasse van strategieën.

Naast de thema's afhankelijkheid in het aankomstproces van werk en beheersing van de verkeersstromen wordt een derde rode draad in dit proefschrift gevormd door het telkens terugkeren van wachtrijmodellen uit de klasse van Markov gemoduleerde wachtrijsystemen. Kort geschetst is een Markov gemoduleerd wachtrijsysteem een model waarin de toestand van een Markovproces

een beschrijving geeft van de tijdsafhankelijke karakteristieken van het wachtrijsysteem. De vrijheid van modelleren die deze wachtrijsystemen bieden, alsmede hun gewillige numerieke eigenschappen blijkt deze klasse van modellen zeer geschikt te maken voor kwalitatieve en kwantitatieve analyse van de typische karakteristieken van moderne wachtrijsystemen.
In hoofdstuk 7 bespreken en vergelijken we aan de hand van een analyse van een Markov gemoduleerd wachtrijmodel een aantal van de meest gangbare methoden voor het analyseren van deze klasse van wachtrijsystemen.

OVERZICHT VAN DE HOOFDSTUKKEN

Hoofdstukken 2,3 en 4 zijn gewijd aan wachtrijsystemen met één bediende waarin de bedieningsduur van een klant positief gecorreleerd is met zijn tussenaankomsttijd. Hoofdstuk 2 bespreekt het basismodel, hoofdstuk 3 en 4 beschouwen twee generalisaties van dit model.
In hoofdstuk 2 behandelen we de afhankelijkheidsstructuur vanuit het perspectief van een ophaalprocedure. In dit basismodel arriveren klanten volgens een Poisson proces bij een bushalte waar enige tijd gewacht moet worden totdat een bus de klanten ophaalt. Ook het aankomstproces van bussen is een Poisson proces. Beschouwen we de som van de bedieningsduren van alle klanten in een bus als de bedieningsduur van één superklant, dan volgt dat de bedieningsduur van de superklant positief gecorreleerd is aan zijn tussenaankomsttijd. Voor dit model bepalen we de Laplace-Stieltjes Transform (LST) van de stationaire gezamenlijke verdeling van de wachttijd en bedieningsduur van een superklant. Dit resultaat leidt tot uitdrukkingen voor de LST's van de wachttijd, verblijftijd en genererende functie voor het aantal klanten, zowel voor superklanten als voor de individuele klanten. Met deze laatste groep worden de klanten bedoeld die arriveren bij de bushalte. Een van de voornaamste resultaten in hoofdstuk 2 is de afleiding van een werkdecompositie eigenschap; deze wordt verkregen door het model te bekijken vanuit het gezichtspunt van wachtrijsystemen waarin de bediende af en toe "vakantie" neemt.

Hoofdstuk 3 breidt het model van hoofdstuk 2 uit tot een meer algemene afhankelijkheidsstructuur. In dit hoofdstuk wordt het perspectief van een wachtrij met een toegangspoort gehanteerd. In dit algemene model is er een aangroei van werk bij de toegangspoort volgens een proces met niet-negatieve onafhankelijke aangroeiingen. Wanneer de toegangspoort open gaat wordt deze hoeveelheid werk samen met nog een extra toegevoegde bedieningsduur (opstart tijd) als één enkele klant bij de wachtrij afgeleverd. In dit model bestaat een bedieningsduur van een klant dus uit twee componenten: een gedeelte dat gecorreleerd is aan de tussenaankomsttijd (en ontstaan is uit het aangroeiproces) en een component die onafhankelijk is van de tussenaankomsttijd (een "gewone $GI/G/1$" bedieningsduur). Dit model herbergt een unificatie van een aantal onlangs verschenen studies naar $M/G/1$ wachtrijmodellen met afhankelijkheid tussen tussenaankomsttijd en bedieningsduur. Analoog aan hoofdstuk 2 bepalen we de LST van de gezamenlijke stationaire verdeling van de wachttijd

en de bedieningsduur van een klant. Door het model te beschouwen als een (stuw)dam model met vakanties voor de bediende wordt een werkdecompositie eigenschap afgeleid.

In moderne wachtrijsystemen is het aankomstproces doorgaans niet Poisson. In het basismodel zoals dat besproken wordt in hoofdstuk 2 zijn zowel het aankomstproces van individuele klanten als het aankomstproces van bussen Poisson processen. In hoofdstuk 4 wordt een generalisatie van het basis model behandeld waarin deze beide processen van meer algemene aard zijn. In dit hoofdstuk ligt de aandacht vooral bij het modelleren van afhankelijkheid tussen tussenaankomsttijden en bedieningsduren met behulp van Markov gemoduleerde aankomstprocessen. In het bijzonder wordt gebruik gemaakt van de structuur van het Batch Markovian Arrival Process (BMAP) en de $BMAP/G/1$ wachtrij. De BMAP is een algemeen groepsaankomstproces dat beschouwd kan worden als een matrix generalisatie van van het Poisson proces.

Hoofdstuk 5 is gewijd aan een wachtrijmodel met ongeduldige klanten. Dit houdt in dat klanten slechts bereid zijn een bepaalde tijd te wachten in de wachtrij; overschrijdt hun wachttijd deze periode dan verlaten de klanten het systeem. Dit model kan ook bezien worden vanuit het gezichtspunt van wachtrijen met een toelatingsmechanisme; afhankelijk van de hoeveelheid werk in het systeem wordt bepaald of een nieuwe klant wordt toegelaten in de wachtrij of wordt geweigerd.
In hoofdstuk 5 bestuderen we dit mechanisme voor een één-bediende systeem met een Markovian Arrival Process (MAP) als aankomstproces voor de klanten. Het geduld van de klanten, d.w.z. de tijd die elke klant bereid is te wachten, wordt beschreven door een stochastische variabele met een exponentiële verdeling. We bepalen de LST voor de stationaire verdeling van de hoeveelheid werk in het systeem. Deze LST leidt tevens tot uitdrukkingen voor de wachttijdverdeling van klanten en de kans dat een klant het systeem voortijdig verlaat.

Hoofdstuk 6 geeft de resultaten van een studie naar optimale statische allocatie van klanten uit een Poisson aankomstproces over een groep van parallelle wachtrijen. Doel is het optimaliseren van prestatiematen zoals kosten verbonden aan de hoeveelheid werk in het systeem of de gemiddelde wachttijd van een klant. De klasse van statische strategieën kan worden omschreven als bestaande uit strategieën die bij de toewijzing van klanten geen gebruik maken van de toestand van het systeem; de toewijzingen zijn gebaseerd op statische karakteristieken. We beschouwen twee strategieën: (i) probabilistische allocatie en (ii) patroon toewijzing. Onder probabilistische allocatie wordt een arriverende klant volgens een vaste kansverdeling aan een van de wachtrijen toegewezen. Bij patroon allocatie worden klanten volgens een vaste tabel aan de wachtrijen toebedeeld. In dit hoofdstuk geven we een optimaliseringsprocedure voor patroon allocatie. Door het toewijzingsprobleem vanuit een enigszins theoretisch gezichtspunt te benaderen worden ook een aantal fundamentele eigenschappen

van toewijzingsproblemen verklaard. Hoofdstuk 6 bespreekt tevens een drietal uitbreidingen van het toewijzingsprobleem.

In hoofdstuk 7 bestuderen we de $BMAP/M/s$ wachtrij: het wachtrijsysteem met één wachtrij en s identieke bedienden waarin klanten arriveren volgens een $BMAP$ aankomstproces en de bedieningsduren van klanten exponentieel verdeeld zijn. De studie naar dit wachtrijmodel ondersteunt een overzicht en vergelijking van de meest gangbare methoden die momenteel gebruikt worden voor het analyseren van Markov gemoduleerde wachtrijsystemen. Met name concentreren we ons op drie methoden die elk expliciet gebruik maken van bepaalde homogeniteits eigenschappen van Markovprocessen in wachtrijsystemen. Met elk der methoden leiden we de genererende functie van de stationaire verdeling van het aantal klanten in de $BMAP/M/s$ wachtrij af. De resultaten worden toegepast op een multi-bedienden generalisatie van de afhankelijkheidsstructuur die bestudeerd wordt in hoofdstuk 2. Hiervoor wordt ook de in hoofdstuk 4 ontwikkelde modelleringstechniek aangewend.

Stellingen

behorende bij het proefschrift

Queueing Models with Dependence Structures

van

Marco Bastiaan Combé

I

Stelling 16 in Prabhu en Zhu[1] is incorrect.
[1] Prabhu, N.U., Zhu, Y. (1989). Markov-modulated queueing systems. *Queueing Systems* **5**, 215-246.

II

Beschouw het toewijzingsprobleem beschreven in Hoofdstuk 6 van dit proefschrift. Optimale probabilistische toewijzingsfrequenties zijn niet altijd goede benaderingen voor de optimale toewijzingsfrequenties bij patroonallocatie. Dit komt voornamelijk doordat bij het bepalen van de optimale probabilistische toewijzingsfrequenties geen rekening wordt gehouden met het feit dat ongelijke toewijzingsfrequenties bij patroonallocatie leiden tot aankomstprocessen met een verschillend karakter, zoals weerspiegeld wordt door de coëfficiënt van variatie in het benaderende Gamma aankomstproces.

III

Een aankomstproces met positieve correlatie tussen tussenaankomsttijd en bedieningsduur van een klant kan goed beschreven worden met het batch Markovian arrival process.
Zie Hoofdstuk 4 van dit proefschrift.

IV

Conflictsituaties van treinen op kruisingen en splitsingen passen goed in het formalisme van wachtrijmodellen; aankomststromen van treinen bij zulke conflictpunten zijn bijvoorbeeld goed te beschrijven met het Markovian arrival process. Toch is wachtrij-analyse vanuit zowel praktisch als modeltheoretisch oogpunt een weinig geschikte methode om zulke conflictsituaties te bestuderen.

V

In Nederland bestaat (nog) geen voedingsbodem voor werkelijk geïnspireerde rap muziek.

VI

Het achteraf verklaren van de uitslag van een voetbalwedstrijd tussen twee min of meer gelijkwaardige teams heeft veel weg van het verklaren van het resultaat van een kop-of-munt experiment met een zuivere munt.

VII

Een stuk bladmuziek is geen algoritme om muziek te maken maar eerder een grafische interface tussen componist en muzikant.

VIII

Na de opheffing van de Anti-apartheidsbeweging is het nu wachten op het opheffen van Amnesty International, Greenpeace, de Kinderbescherming en het Wereldnatuurfonds.

IX

Zogenaamde unplugged music heeft doorgaans minder diepgang te bieden dan gewone studio-opnamen en kan slechts beschouwd worden als commerciële uitbuiting van het gemis aan spiritualiteit in het dagelijks leven.

X

Het succes van de CompactDisc is niet te danken aan de vermeende superieure geluidskwaliteit maar aan de gebruikersvriendelijkheid van de afspeelapparatuur.

XI

Modern voorouderonderzoek laat zich beter karakteriseren als wortelstronk-analyse dan als stamboomonderzoek.

XII

Het schaakeindspel KD-K (wit aan zet) op een cilindrisch schaakbord leidt bij optimaal spel van beide partijen in hooguit 10 zetten tot mat.

XIII

Kennis is onrust.