ISBN 90-646-4667-8

Processor-Sharing Models for Integrated-Services Networks

Processor-Sharing Models for Integrated-Services Networks

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Technische Universiteit Eindhoven,
op gezag van de Rector Magnificus, prof.dr. M. Rem,
voor een commissie aangewezen door het College
voor Promoties in het openbaar te verdedigen
op donderdag 20 januari 2000 om 16.00 uur

door

Rudesindo Núñez Queija

geboren te Heemskerk

os meus pais

Dit proefschrift is goedgekeurd door de promotoren:

prof.dr.ir. O.J. Boxma
en
prof.dr.ir. S.C. Borst

Dankwoord/Agradecimientos (Acknowledgements)

Verscheidene mensen zijn van doorslaggevende betekenis geweest voor de totstandkoming van dit proefschrift en ik wil hen hierbij voor hun hulp bedanken. Allereerst ben ik veel dank verschuldigd aan Onno Boxma, zowel voor zijn uitmuntende begeleiding op het gebied van de wachtrijtheorie alsmede voor de zeer plezierige persoonlijke contacten. Zijn deur stond altijd voor mij open. Na Onno's aanstelling in Eindhoven heeft Sem Borst gedurende het laatste jaar de dagelijkse begeleiding op zich genomen. Ik dank Sem voor de voortreffelijke manier waarop hij dat heeft gedaan. Zowel Onno als Sem wil ik ook bedanken voor hun snelheid en precisie bij het lezen van eerdere versies van het proefschrift. Ik wil ook professor Tijms en Hans van den Berg bedanken voor hun commentaar op het proefschrift. Professor Tijms dank ik verder dat hij mijn interesse in de wachtrijtheorie heeft gewekt en dat hij op een (naar mijn mening) goede dag zei dat ik "maar eens met professor Boxma op het CWI moest gaan praten". Hans van den Berg en Michel Mandjes wil ik graag bedanken voor de goede samenwerking tijdens mijn bezoeken aan KPN Research. Ik ben ook professor Cohen dankbaar voor de plezierige contacten gedurende de afgelopen jaren en de nuttige discussies over de resultaten van Hoofdstuk 3. Verder ben ik het CWI erkentelijk voor de mij ter beschikking gestelde faciliteiten en dank ik alle BS/PNA-ers voor de goede sfeer. Speciaal denk ik hierbij aan mijn kamergenoten, met name aan degenen die het (al) langer dan 6 maanden met mij hebben uitgehouden: Vincent Dumas, Bert Zwart en Miranda van Uitert. Bert wil ik in het bijzonder bedanken voor de vele gesprekken over processor sharing, waar ik veel van heb geleerd.

Ik heb dit proefschrift willen opdragen aan mijn ouders. Als zij ruim 30 jaar geleden niet de moed hadden gehad om het vertrouwde achter zich te laten en voor hun gezin een nieuwe toekomst op te bouwen in het onbekende Nederland, dan was dit proefschrift er zeker niet geweest. Mijn moeder dank ik ook voor de manier waarop ze drie kinderen heeft weten groot te brengen na het overlijden van mijn vader. Daarbij heeft ze mogen rekenen op de steun van mijn broer Isaac die (als vanzelfsprekend) vanaf zijn 19de de rol van vader (van 10 jaar jongere kinderen) op zich heeft genomen. Ik prijs mij verder gelukkig dat Pili het heeft aangedurfd om haar vertrouwde omgeving in te ruilen voor een nieuw bestaan samen met mij (in een land dat zij alleen kende van Cruijff en Koeman).

Zij en José Ramón boden mij, in het bijzonder tijdens het schrijven van het proefschrift, een toevluchtsoord wanneer ik "gejaagd door epsilons en delta's" thuis kwam. Verder ben ik Julia, Pepe, de rest van mijn (schoon)familie en al mijn vrienden dankbaar voor hun steun. Amparo en Maarten dank ik bovendien dat zij mijn paranimfen wilden zijn.

Dedico esta tesis a mis padres. Si hace treinta años no hubieran tenido el valor de dejar atrás su propio ambiente y buscar en la lejana Holanda un nuevo futuro para su familia, esta tesis seguramente no existiría. A mi madre le agradezco, además, la manera en que supo criar tres niños después de haber fallecido mi padre. Para eso pudo contar con la ayuda de mi hermano Isaac que, a los 19 años (dándolo por supuesto), asumió la responsabilidad de hacer de padre de tres niños. A Pili le agradezco el atreverse a empezar una nueva vida conmigo, en un país que sólo "conocía" por Cruijff y Koeman. Sobretodo mientras escribía esta tesis, ella y José Ramón me ofrecían un refugio donde "escapar a los epsilons y deltas". También agradezco el apoyo que me han dado Julia, Pepe, el resto de mi familia (los de Correchouso y los de Móra) y mis amigos. A Amparo y Maarten gracias, también, por querer ser mis paraninfos.

Tenslotte dank ik God voor al het moois in mijn leven en in het bijzonder voor alle bovengenoemde mensen.

Sindo Núñez Queija
november 1999

Contents

Chapter 1

Introduction

The time that telecommunication was synonymous to telephony is long gone. Modern communication networks are emerging from the convergence between traditional telephone systems and computer-communication networks. They are being designed to offer a wide range of services, such as telephony, data transfer, and (interactive) video, on a common infrastructure. In this thesis we develop queueing-theoretic models and techniques to study performance issues in integrated-services (or, multiservice) networks. The focus is on the mathematical analysis of the proposed queueing models. In this introductory chapter we first motivate our modelling approach, giving a description of integrated-services networks and discussing their evolution from traditional telecommunication systems.

1.1 Background

With the integration of different services in a common network, operators aim at (i) responding to the strong demand for new telecommunication services, and (ii) achieving a high utilisation of the network resources. However, the interaction between the different service classes within the network has a significant impact on their performance. We briefly illustrate this with an example. Consider a network link which is used for both telephone connections and data connections (e.g., file transfers). A telephone conversation requires a constant transmission rate to ensure speech quality. Therefore we assume that each voice connection is allocated a certain fixed amount of capacity. The remaining capacity is available for the transmission of data. In comparison with voice, data applications are usually better able to adapt to fluctuations in the transmission capacity. We assume each individual data connection receives an equal portion of the total available capacity. Consequently the transfer time of data typically depends strongly on the characteristics of the voice traffic. In turn, one may want to give some "protection" to on-going data transfers against new voice connections. This may be accomplished by rejecting new telephone calls from the system if otherwise the transmission rate of individual data transfers would drop below some minimum acceptable level. Then the blocking probability of voice calls also depends on the characteristics of data traffic.

The above example plays a central role in this thesis. We study the interaction between two types of traffic — voice and data traffic, or, more generally, "stream" traffic and "elastic" traffic (see Section 1.3) — which share resources in a communication network. We model the above sketched situation by a queueing system in which a server (communication link) serves two types of customers. Customers of the first type (voice) require a certain fixed amount of service capacity for some random period of time. If this capacity is not available, a newly arriving customer is rejected. In contrast, customers of the second type (data) do not need a constant rate, but involve a random amount of work (number of data bits to be transmitted). Customers of the second type share equally in the capacity left over by the first type of customers. This queueing model is composed of two, interacting, elementary models from queueing theory: the Erlang loss model (for voice) and the processor-sharing queue (for data). This thesis concentrates on the mathematical analysis of several variants of this composite queueing model. Particular attention will be devoted to the performance evaluation of data traffic. We therefore focus on the analysis of processor-sharing queueing models with varying service capacity. The variation in service capacity reflects the variation in capacity left over by voice traffic. Processor-sharing models with varying service capacity have not received much attention and it is in this area that the thesis contributes to queueing theory. Our results provide a basis for a careful analysis of performance issues in integrated-services networks. The ultimate goal of such an analysis is to facilitate the design and control of future communication networks, addressing issues such as proper dimensioning and developing adequate capacity allocation strategies.

The remainder of this introductory chapter is organised as follows. The first two sections are devoted to integrated-services networks. Section 1.2 gives an overview of the evolution of these networks. A more detailed description of the various traffic types is given in Section 1.3, providing a basis for a unified approach in modelling the integration of services. In Section 1.4 we turn to queueing theory, describing its basic concepts and the relation to telecommunications. Section 1.5 gives a more detailed description of the above mentioned queueing model, which plays a central role in the thesis. Since the analysis is mainly concerned with processor-sharing models with varying service capacity, Section 1.6 reviews the literature on processor-sharing queues. Section 1.7 gives an overview of the other chapters.

1.2 Evolution of integrated-services networks

Over the past two decades there has been an explosively increasing need for long-distance services other than telephony, such as data transfer and (interactive) video communication. At present the telephone network is already commonly used to connect personal computers via modems for transmission of data files. Since telephone networks were designed to specifically carry voice calls, they are not particularly suited to support data traffic. Hence, there is a strong need for a network such as the current Internet, which was specifically designed for data transfer purposes. In its turn, the Internet is also evolving to offer other services, and is already (experimentally) being used for applications such as voice communication and video-conferencing. It is widely believed that these trends — enabled by technological innovations — eventually will result in future world-wide networks offering a wide range of services on an integrated basis. However, whether these networks should be controlled by central operators (like traditional telephone networks) or in a distributed manner (as is the case in the current Internet) has so far remained a matter of debate.

The development of the current Internet was initiated in the 1960's as a computer-communication network for the U.S. Defense Department. To enable efficient data communication, the Internet was based on the IP (Internet Protocol) concept. After dividing a message into so-called packets, each packet is transported as an independent entity to the destination point, where the message is reassembled. By the 1980's the Internet had evolved into a world-wide network, interconnecting mainly universities and research institutes. This evolution took place despite the fact that the network performance was still poor. Crucial progress was made when TCP (Transmission Control Protocol) was proposed by Jacobson [46]. This protocol enables the users to react dynamically to congestion in the network. The rate at which a TCP-controlled traffic source transmits is high when the load on the network is light and the rate is low when the network is congested. This as opposed to traditional telephone networks, where each user is assigned a fixed transmission capacity (i.e., one channel) for the duration of the connection. A more detailed description of TCP is given in Section 1.3.

After the introduction of TCP helped to control traffic in the network, the growth of the Internet has been impressive. Ten years later, most companies use the Internet for communication and advertisement purposes. Also more and more people are using the Internet at home, mostly for information retrieval (Web browsing) and correspondence (e-mail). Although well-suited for data transfer applications, today's Internet does not provide for interactive applications (such as telephony or video-conferencing). As we saw above, the capacity available to each individual user typically decreases when more users require the same resource. As a consequence, transmission delays may increase and/or transmission quality may degrade. Large delays, however, are unacceptable for interactive applications. Therefore much effort is put into making the Internet suitable for supporting such applications, see White and Crowcroft [117] and Kumar et al. [59].

The integration of different services onto a common platform had long been anticipated by telephone network operators. It was recognised that traditional telephone systems would not be able to meet the rising demand for new services. The discovery that glass-fibre cables provide a means for optical high-speed communication triggered a large world-wide research activity into the applicability of the new technology in future communication networks. The telecommunication community established the ATM (Asynchronous Transfer Mode) concept as the standard for high-speed communication networks. In this community it was commonly believed that, like in traditional telephone systems, centralised traffic control will remain necessary in future networks to provide satisfactory QoS (Quality of Service). In contrast to TCP/IP, packets (or, cells) in ATM are labelled with a connection identification rather than treated as individual entities. This not only enables advanced QoS support mechanisms, but also offers higher transmission capacity because of the limited address space. Because of the higher implementation complexity, however, the deployment of ATM is mainly restricted to high-speed back-bone infrastructures. The use of TCP/IP is widespread, but does not yet allow for QoS support of real-time applications as explained earlier.

Compared to traditional telephone systems, both TCP/IP and ATM offer several advantages other than those due to the use of optical technologies (the latter includes increase of transmission capacity). The two concepts also provide flexibility of resource allocation in two ways. Firstly, in telephone systems each connection gets a fixed transmission capacity, namely one channel. In TCP/IP and ATM the total transmission capacity — henceforth called bandwidth — may be divided arbitrarily. For instance for video-conferencing typically more bandwidth is needed than for telephony (in ATM the amount of bandwidth may be negotiated when a request for a new connection is made). More importantly, the total bandwidth can be used more efficiently due to *statistical multiplexing* gains. These gains are achieved because not all connections constantly need their individual peak bandwidth. In fact, especially when many connections are multiplexed, it will be extremely rare that all connections simultaneously transmit at peak rate. Thus, the sum of the peak rates may exceed the (physical) total bandwidth, while still almost always meeting the actual bandwidth requirements.

Future telecommunication systems are still very much under development. It therefore remains unclear what these networks will precisely look like. However, we propose models that are not concerned with details of network architectures or transmission protocols, making only high-level assumptions on services as experienced by the users. In the next section we further motivate our modelling approach.

1.3 Modelling traffic in integrated-services networks

Modelling hierarchy

To characterise traffic in modern communication networks it is convenient to use a three-level hierarchy. At the highest level — the call level or connection level — connections are being established. For the duration of a connection, information is fragmentised into so-called *packets* which are transmitted through the network. Such networks are therefore called packet-switched networks. Although in IP networks there is no real notion of connections (each packet is transmitted as an independent entity), we use this term to indicate the information flow from a source to a destination. In ATM networks, packets are usually called *cells*, but in this thesis we commonly use the term packet. The original information is recovered at the destination point by reassembling the packets. The level at which individual packets are observed as they flow through the network is called the packet (or cell) level. Packets belonging to the same connection are usually not generated as a constant flow, but rather in *bursts*. This gives rise to an intermediate level — the so-called burst-level — between the connection level and the packet level.

Let us illustrate the notion of the three time scales for telephony and data transfers in a packet-switched network. A telephone connection is established after dialing a phone number and is terminated after hanging up. Packets consist of speech fragments. Bursts of packets then correspond to periods of speech (which alternate with periods of silence). In file transfers each packet contains a segment of the (data) file. The connection duration is equal to the total transfer time of the file. Usually the complete file is instantaneously available for transmission. In that case the connection consists of one large burst of packets.

Communication sessions are not only extremely diverse in traffic characteristics, but also in the QoS requirements. For instance in a telephone connection, packet delays of a few hundred milliseconds imply a severe degradation of speech quality. File transfers on the other hand are more flexible with respect to packet delays, the transfer time of the complete file being of dominant importance.

The models we propose for the performance analysis of integrated-services networks at the connection level abstract from a specific network design or traffic control. Our modelling may apply for instance to both the ATM and the IP concepts. Next we discuss these concepts in some detail because, at present, they are the main candidates for realising integration of services on a common infrastructure. Subsequently we describe a unifying classification of traffic.

ATM versus IP

ATM networks are specifically designed to deal with different traffic types in a different manner. The CBR (Constant Bit Rate) and real-time VBR (Variable Bit Rate) transfer capabilities provide real-time services, such as telephony and interactive video applications. As we already mentioned for telephony, real-time traffic is extremely delay-sensitive (at the packet-level). For these applications QoS is guaranteed through bandwidth reservations for individual connections. More recently the ABR (Available Bit Rate) service was introduced to accommodate data transfers. In addition to a typically small guaranteed MCR (Minimum Cell Rate), ABR traffic is granted the bandwidth left over by real-time services. As stated by the ATM Forum [1], ABR-controlled connections should fairly share the available capacity. The system instructs the ABR connections at which rate to transmit. This is done using the following closed-loop feedback control mechanism. ABR traffic sources periodically transmit so-called resource management cells which are returned by the destination. As they traverse the network these cells are marked if congestion is detected. Upon return, ABR traffic sources adapt their transmission rate to the observed congestion. For more details on ATM's service classes see for instance Kesidis [50].

In Section 1.2 it was mentioned that the current Internet does not specifically support real-time applications. For future IP networks new standards are being developed to overcome this shortcoming, see White and Crowcroft [117] and Kumar et al. [59]. Data is transmitted in the Internet using the TCP protocol. As opposed to the ABR service in ATM networks (where the rate adaptation mechanism is regulated by the system), TCP/IP leaves traffic control to the users. The receiver (destination) of a message sends an acknowledgement for each received packet to the transmitting user (source). If no acknowledgement is received within some period of time, the source concludes that the transmission of the packet has failed due to congestion in the network (the packet is lost), and re-transmits the packet. In addition, a negative acknowledgement (when the destination detects an error in a packet) also causes a re-transmission. The number of packets sent by the source but not (yet) acknowledged by the destination is limited by the *window size*, and is adjusted dynamically. When the acknowledgement of a packet is received, the window size is increased (typically by *adding* a fixed number of packets), and when a packet is assumed to be lost or a negative acknowledgement is received, the window size is decreased (typically by a fixed reduction *factor*). Note that the window size is roughly inversely proportional to the transmission rate. Effectively, a source of TCP traffic transmits at a high rate when few packets are lost (i.e., when the load on the network is light), and at a low rate when relatively many packets are lost (i.e., the network is congested).

Traffic modelling on connection level

We proceed with a unified approach to the modelling (at the connection level) of integrated-services networks, of which ATM and IP networks are particular examples. Based on their QoS requirements, we divide all traffic types into two broad classes: *stream traffic* and *elastic traffic*. A stream traffic connection requires stringent (packet-level) delay guarantees for the duration of its connection time (holding time). Stream traffic may be identified with real-time applications. Elastic traffic on the other hand is more flexible with respect to packet delays, as long as the total transmission delay is "acceptable". An elastic traffic connection involves the transmission of a certain amount of information, typically present in the form of a data file. Therefore an elastic traffic connection will often be referred to as a "data/file transfer". A file transfer is characterised by the file size, and possibly a minimum and a maximum rate between which the actual transmission rate may vary. The minimum rate, for instance the MCR for ABR traffic, ensures a certain maximum transmission delay. We emphasise that in the current Internet there is no minimum rate for TCP flows. The maximum rate may be the consequence of physical limitations, such as the access link rate or the modem rate. If the minimum rate is equal to zero (i.e., there is no guarantee on transmission delays), then elastic traffic is called "best-effort" traffic. This is for instance the case for data transfers in the current Internet.

Bandwidth is allocated to the two traffic types as follows. Connections of both types are each given the (minimum) required bandwidth. The remaining bandwidth is available for elastic traffic, giving each elastic connection an equal share. Under certain assumptions on packet losses and "fairness" criteria, this indeed resembles bandwidth allocation to data transfers in both IP and ATM networks: each data transfer is granted an equal portion of the available bandwidth. To ensure that at all times the minimum required bandwidth is available for each individual connection, there should be connection acceptance control (see Massoulié and Roberts [70] for a discussion). A simple connection acceptance rule is for instance to reject a request for a new connection (of any type) if it would lead to the violation of the bandwidth guarantee for any connection.

The above traffic characterisation and classification underlies our modelling. In our approach we study the performance of traffic at the connection level, assuming a complete separation of time scales. This way we may represent the (discrete) packet flow within a connection by a continuous fluid flow, possibly of a varying rate over time. This approach is justified when traffic fluctuations at the lower time-scale levels are very fast compared to the duration of the connection, i.e., when packets are very small.

The rough classification of stream and elastic traffic was proposed by Roberts [94, 95, 96] and is commonly believed to capture the essential issues of service integration, while allowing thorough mathematical analysis. Furthermore, it is not desirable to analyse and dimension networks based on highly detailed traffic characterisations, since future traffic may again have its own specific characteristics. Networks should thus be flexible enough to be able to cope with changes in the nature of traffic.

1.4 Queueing theory in performance evaluation

Queueing theory and the development of telecommunication systems have had a strong influence on one another. The first queueing-theoretic models were developed by A.K. Erlang in the beginning of the 20^{th} century for the dimensioning of telephone systems. Ever since, queueing theory has played a key role in the design and performance analysis of telecommunication systems. Vice versa, queueing theory has developed partly under the stimulus of new problems encountered in telecommunication systems. For example, the need for computer communications in the 1960's triggered new research into networks of queues, opening up new horizons with the famous papers of Baskett et al. [8] and Kelly [47]. In the 1980's and 1990's the performance analysis of multiservice communication networks further intensified the research activity in this direction. We refer to Cohen and Boxma [21] for a survey of the evolution of queueing theory until the mid-1980's, and to Prabhu [89] for an extensive bibliography of books and survey papers on queueing systems. Next we briefly outline the basic concepts of queueing theory. For a thorough introduction into queueing theory we refer to Kleinrock [54], Cohen [20] and Takagi [110], and to Kelly [48] and Nelson [80] for queueing networks. For further references see [89].

In general, a queueing model describes a situation where limited resources are used to perform certain tasks. The resources are usually called servers (often there is only a single server). The tasks to be performed are viewed as customers that arrive to the server(s), requiring a certain amount of work to be done by the server(s). Characteristic of queueing models is the random nature in which customers arrive, as well as the randomness in the service requirement. Due to the limited capacity, the random fluctuations lead to occasional contention for service among the customers, and hence to congestion effects.

The most elementary queueing model is the single-server system depicted in Figure 1.1. Suppose the customers arrive to the system one at a time. An inter-

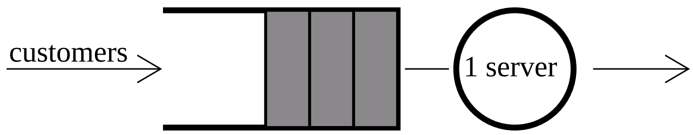

Figure 1.1: The single-server queueing model G/G/1

arrival time is defined to be the time interval between two consecutive arrivals. The arrival process is usually assumed to be such that interarrival times form an i.i.d. (independent and identically distributed) sequence of random variables. The service requirement of a customer is defined as the amount of time that the server needs to serve the customer if the latter would receive the server's complete capacity (it might be the case that the server works on several customers at the same time). Further it is assumed that service requirements are i.i.d., and that customers only leave the system after having received their entire service requirement. The described queueing model is often called the G/G/1 (or GI/GI/1) queue, a notational convention proposed by Kendall [49]. The G's in this notation stand for general probability distributions, the first referring to the distribution of interarrival times and the second to the distribution of service requirements. The alternative notation GI is sometimes used to emphasise that the sequence referred to (either the interarrival times or the service requirements) is an *independent* sequence. The number 1 refers to the single server.

In order to describe the evolution of the queue-length process, we need to specify the *service discipline,* i.e., how the server's capacity is allocated to the customers. Many different service disciplines have been proposed and studied in queueing theory. We mention the two that are most relevant for this thesis. Perhaps the most natural discipline is the FCFS (First Come First Served) discipline. In this discipline the customers are served in order of arrival. In the processor-sharing discipline the service capacity is divided among all customers in the system, each of them receiving an equal share. In the latter case customers with a small service requirement may overtake others with a larger service requirement.

Many variations to the basic G/G/1 model have been studied in the queueing literature. We already noted that different choices of the service discipline lead to different system behaviour. More generally, we may consider the G/G/c/c+d

queue, in which there are c servers (instead of one) and d positions for customers waiting for service. In such a system new customers may be rejected (blocked) from the system, if all the waiting positions are occupied. The last symbol in the notation is omitted when $d = \infty$, as we did in the single-server queue of Figure 1.1. Other modifications which have been studied include the case where customers sometimes — for instance when the queue is large — choose not to enter the system (balking), or abandon the queue if they have to wait too long. An important class of queueing models arises when we assume that the arrival process is a Poisson process, the so called M/G/c/$c + d$ model. The symbol M stands for the Markovian (or Memoryless) nature of the Poisson process. The Poisson arrival process arises naturally in applications when there are many individuals which (at any time) may require service, each with very small probability (independently of each other), see Feller [30, page 355].

1.5 The basic model of the thesis

We propose the following queueing system as a connection-level model of a link in a communication network carrying both stream traffic and elastic traffic. A stream traffic connection is represented by a "stream customer" and an elastic traffic connection (a file transfer) by an "elastic customer". We assume that the two types of customers arrive according to two independent Poisson processes. The mean number of stream customers and elastic customers arriving per unit of time are denoted by $\lambda^{(\mathrm{s})}$ and $\lambda^{(\mathrm{e})}$, respectively. Upon arrival of a customer, the service station decides whether the new customer is taken into service or rejected from the system. A discussion of several issues regarding this decision is given below. A rejected customer never enters the system — the customer is lost — and does not affect the service of other customers. A customer that is accepted, is immediately taken into service until the complete service requirement is fulfilled.

The total service capacity of the station (link bandwidth) equals $C > 0$. A stream customer requires a fixed capacity $r^{(\mathrm{s})}$ for the duration of a random *holding time* (for instance the length of a telephone conversation). The sequence of these holding times is assumed to be i.i.d. with distribution $B_s(t)$, $t \geq 0$. An elastic customer requires a random amount of service. Service requirements of elastic customers — i.e., file sizes in case we model file transfers — are i.i.d. with distribution $B_e(x)$, $x \geq 0$. Each elastic customer gets an equal share of the capacity left over after giving each stream customer the required capacity $r^{(\mathrm{s})}$. However, at any point in time the rate at which an elastic customer is served must be between a minimum rate $r_-^{(\mathrm{e})} \geq 0$ and a maximum rate $r_+^{(\mathrm{e})}$. Elastic customers leave the system upon having received their full service requirement. Finally we assume that holding times, service requirements, and interarrival times of both customer types, are mutually independent.

The above description has immediate implications for allowable acceptance (rejection) policies. If we denote the number of stream customers and elastic customers at a given point in time by $k^{(\mathrm{s})}$ and $k^{(\mathrm{e})}$, respectively, then clearly

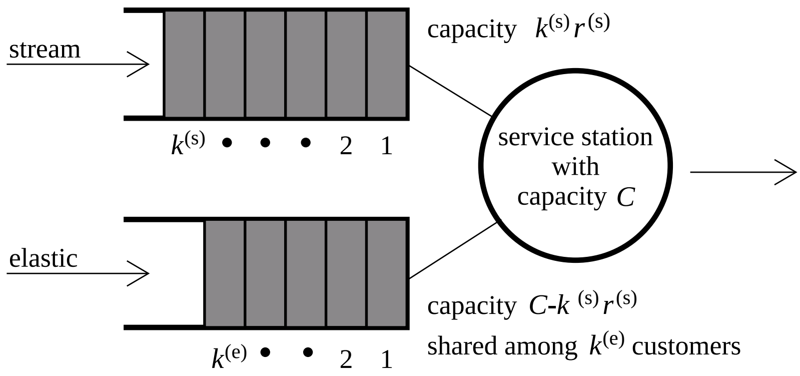

Figure 1.2: Basic model with two customer types

the capacity restriction,

$$k^{(s)}r^{(s)} + k^{(e)}r_{-}^{(e)} \leq C,$$

must be satisfied. If accepting a newly arrived customer would cause this constraint to be violated, the new customer must be rejected (it is not allowed to compromise on the capacity requirement of any customer, or to remove an already accepted customer from the system). In addition, acceptance of customers may be subject to other constraints. For instance, a certain fixed capacity could be reserved for stream traffic and/or for elastic traffic.

As noted previously, the model presented here is a combination of the standard Erlang loss model (for stream customers) and the processor-sharing queue (for elastic customers). We analyse several variants of this hybrid model, focusing on the performance of elastic customers. In the different models we make different assumptions regarding the service requirement and holding-time distributions.

Remark 1.5.1 In general we do not explicitly consider the presence of stream customers, and instead assume that the service capacity (for elastic customers) varies according to some exogenous process. This process may also be dependent on the service process of elastic customers. This way the considered models for elastic traffic are flexible with respect to the precise nature of the service fluctuations.

The above approach to modelling the integration of stream traffic and elastic traffic was used in Núñez Queija and Boxma [86], Blaabjerg et al. [14], Altman et al. [5], and Kulkarni and Li [57]. Other papers in which processor-sharing queues are used for the modelling of elastic traffic are Heyman et al. [43], Roberts and Massoulié [97], Berger and Kogan [11], and Massoulié and Roberts [70, 71]. An experimental investigation of the processor-sharing queue to model TCP traffic is provided by Kumar et al. [58].

1.6 Processor-sharing queues

In view of the central role of processor sharing in this thesis, we give an overview of the relevant literature in this section. In the (egalitarian) processor-sharing service discipline, when there are $n > 0$ customers in the system, all these customers simultaneously get an equal share of the service capacity, i.e., each customer gets a fraction $1/n$ of the capacity. The processor-sharing service discipline became of interest as the idealisation of time-sharing queueing models which arose with the introduction of time-sharing computing in the sixties. As we saw in the previous sections, today processor-sharing models can be applied for the performance analysis of elastic traffic in integrated-services communication networks, in particular for the ABR service class in ATM networks and for TCP traffic in IP networks.

An extensive body of literature on processor-sharing queues was initiated by Kleinrock [52, 53] who studied the M/M/1 case. In particular he showed that the mean sojourn time conditional on the service requirement is proportional to the service requirement. Sakata et al. [99, 100] derived the steady-state queue-length distribution of the M/G/c queue with processor sharing, showing that it is insensitive to the service time distribution except from its first moment. In the multi-server processor-sharing queue with c identical servers, each with capacity 1, the total service capacity is equally shared among all customers present, with the restriction that an individual customer can not be served at a rate higher than 1. As a special case, the queue-length distribution in the M/G/1 processor-sharing queue turned out to be geometric, inheriting the above insensitivity property. Sakata et al. [99, 100] also extended Kleinrock's proportionality result to the M/G/c case, see also Kleinrock [55, Section 4.4]. All the above results were extended by Cohen [19] to a general class of networks, where the rate at which customers at a particular node are served is a function of the node and of the number of customers at that node (there called *generalised* processor sharing). However, determining the sojourn time *distribution* in processor-sharing queues turned out to be a very difficult problem.

For the M/M/1 queue with processor sharing, a closed-form expression for the LST (Laplace-Stieltjes Transform) of the distribution of the sojourn times — conditional on the amount of service required and the number of customers seen upon arrival — was first derived by Coffman et al. [17]. Sengupta and Jagerman [105] found an alternative expression for the LST of the distribution of the sojourn time conditioned only on the number of customers seen upon arrival. In particular they found that the n^{th} moment of the conditional sojourn time is a polynomial of degree n in the number of customers upon arrival. The *distribution* function of the sojourn times, conditioned on the amount of service required, was studied by Morrison [78].

The sojourn time distribution in the M/G/1 processor-sharing queue was first analysed by Yashkov [120]. Schassberger [101] considered the M/G/1 queue with processor sharing as the limit of the round-robin discipline. Ott [87] found the joint LST and generating function of the distribution of the sojourn time and the number of customers left behind by a departing customer. Van den Berg and Boxma [10] exploited the product form structure of an M/M/1 queue with general feedback for an alternative derivation of the sojourn time distribution in the M/G/1 processor-sharing queue. Rege and Sengupta [91] gave a decomposition theorem for the sojourn time distribution for the M/G/1 queue with K classes of customers and *discriminatory* processor sharing. Grishechkin [39, 40] described the M/G/1 queue with batch arrivals and a generalised processor-sharing discipline by means of Crump-Mode-Jagers branching processes. For a more extensive overview of the literature on processor-sharing models we refer to Yashkov's survey papers [122, 123].

An essential difference between the processor-sharing models from the literature cited above and those analysed in this thesis is that in our case the available service capacity varies according to a stochastic process. As described in Section 1.5, the variation in service capacity is motivated by the fluctuation in available bandwidth for elastic traffic in integrated-services networks due to the presence of high-priority stream traffic. Processor-sharing queues with varying capacity have not been studied with any rigour before. This thesis presents the first analytical results for sojourn times in such queues. The fact that the service capacity fluctuates makes the analysis of performance measures, such as the number of customers in the system and their sojourn time, considerably more complicated than in the case of a constant service capacity. As we shall see, several "nice" properties are lost when the service capacity varies, including the earlier mentioned geometric queue-length distribution as well as the proportionality relation between the conditional mean sojourn time and the service requirement of a customer.

1.7 Overview of the thesis

In this first chapter we have motivated the use of processor-sharing queues with variable service capacity for the performance evaluation of (elastic) traffic in integrated-services networks. In the remainder of the thesis we are concerned with the analysis of such queueing models.

In Chapter 2 we give a detailed analysis of the queue length (that is the number of customers in the system) of a processor-sharing model with varying service capacity. The model includes a large class of hybrid models as described in Section 1.5. In particular we study the impact on the performance of elastic customers when stream customers arrive and depart on a very different time scale than elastic customers do. The theoretic analysis of the influence of time-scale differences is illustrated by means of numerical results from the application of the model to a particular telecommunication system. Chapter 2 builds on the analysis presented in Núñez Queija and Boxma [86] and Núñez Queija [82].

In Chapter 3 we turn to *sojourn times* in processor-sharing models with varying service capacity. We present the analysis of Núñez Queija [83] for the extreme case where the service facility alternates between "on-periods", during which the service rate is constant, and "off-periods", during which no service can be rendered. For this on/off-model closed-form expressions are obtained for the first two moments of the distribution of sojourn times conditional on the service requirement. Higher moments can be computed recursively. The conditional sojourn time distribution itself is characterised by means of its LST in terms of the LSTs of two fundamental variables. For the latter a solution is given in terms of an integral equation.

Chapter 4 presents the results of Núñez Queija [84] for sojourn times in models with a more general structure of service capacity fluctuations than the one in Chapter 3. A processor-sharing queue in a Markovian environment is considered, where the service rate depends both on the state of the random environment and on the number of customers in the queue. The evolution of the random environment itself may also depend on the queue-length process. The LST of the sojourn times conditional on the service requirement is found in terms of a matrix-exponential function. We discuss how the conditional mean sojourn time can be computed and propose an efficient approximation for it. The results are validated through the extensive numerical investigation of Núñez Queija et al. [85], where the model is applied to the performance evaluation of a telecommunication system under different integration strategies for elastic and stream traffic.

In Chapter 5 we focus on heavy-tailed service requirement distributions. It is well-known that, if the service requirement distribution is heavy-tailed then, under any non-preemptive service discipline, the tail of the sojourn time distribution is "one degree" heavier: it is as heavy as the *integrated* tail of the service requirement distribution. A service discipline is called non-preemptive if at most one customer is served at any time and no customer's service is interrupted. In contrast, for the processor-sharing discipline it was shown by Zwart and Boxma [128] that the two tails are "equally heavy". We present a new proof of this fact based on Markov's inequality, which we apply to the moments of the distribution of sojourn times conditional on the service requirement. Using the new approach, we extend the result of Zwart and Boxma [128] to the on/off-model of Chapter 3. We also establish the "preservation of tail-heaviness" for two other service disciplines: FBPS (Foreground-Background Processor Sharing) and SRPT (Shortest Remaining Processing Time first).

Chapter 2

Queue length in the case of a varying service capacity

Before focusing on customer sojourn times in processor-sharing queues with varying service capacity we study their queue-length process. Although customers do not queue in processor-sharing systems — they simultaneously share the service capacity — we use the term "queue length" to indicate the number of customers in the system. This slight abuse of terminology is common in the literature of processor-sharing models. This chapter presents a detailed study of the steady-state queue-length distribution of a rather general processor-sharing model with varying service capacity. The model considered here contains many variants of the hybrid model presented in Section 1.5, which allow us to capture different aspects of traffic in integrated-services networks. In our analysis we are concerned with the queue-length performance of elastic customers in hybrid models of the type presented in Section 1.5. For an integral analysis of a large class of such models we abstract from the precise nature of the fluctuation in the service capacity that is available to elastic customers. Instead of specifically modelling stream customers in the system, the capacity available to elastic customers is governed by the state of a general finite Markovian birth and death process. Such a process generalises in a natural way the arrival and departure process of stream customers in (variants of) the model in Section 1.5. After "replacing" stream customers by the birth and death process, the only customers considered are elastic customers.

This chapter extends the analysis of Núñez Queija and Boxma [86] and Núñez Queija [82] to more general models. We briefly describe the model of Núñez Queija and Boxma [86] which we use for illustration purposes, see also Section 1.5. An individual elastic customer does not require a minimum service capacity ($r_-^{(e)} = 0$), and can be served at any positive rate ($r_+^{(e)} = \infty$). A fixed part of the system capacity is reserved for elastic customers and there is a finite waiting room for stream customers. Because $r_-^{(e)} = 0$ there is no bandwidth guarantee for elastic customers. The capacity reserved for elastic traffic, however, protects elastic customers from stream customers taking up all the capacity. A waiting room for stream customers could for instance model re-dialing in telephone calls. The model presented here also allows for impatience in re-dialing: After having re-dialed during a random period of time, a customer becomes impatient and leaves the system.

For the class of models that we consider here, we assume that the birth and death process regulating the amount of capacity available to elastic customers evolves independent of the past arrival and service process of elastic customers. In the light of the hybrid models of Section 1.5 and [86], this assumption corresponds to stream traffic not being affected by the dynamics of elastic traffic. I.e., whether or not a new stream customer is taken into service does not depend on the number of elastic customers in the system. We further assume that the service requirements of elastic customers have an exponential distribution. Moreover, the Markovian nature of the birth and death process regulating the service capacity available to elastic customers corresponds to exponentially distributed holding times (and re-dialing times) of stream customers. These assumptions allow for a detailed study of the queue-length distribution of elastic customers, providing useful qualitative insight into the performance of elastic traffic.

The model gives rise to a two-dimensional Markov process. The two components are (i) the number of (elastic) customers and (ii) the state of the server, that is, the state of the birth and death process regulating the service rate. Since stream customers are not explicitly modelled here, when referring to customers we mean elastic customers (unless otherwise indicated). As a further generalisation we also let the arrival rate depend on the state of the server. With this generalisation the model includes variants of the model in Section 1.5 where both types of customers come from a common finite population.

We determine the steady-state queue-length distribution and compare it to the case with a fixed available service capacity. Both an infinite and a finite queueing capacity are considered. The finite queue can be used to model the case that individual customers have a guaranteed minimum rate ($r_-^{(e)} > 0$). Obviously, in that case there is a maximum to the number of customers in the system. We find the simultaneous steady-state distribution of the queue length and the state of the server. We do so using arguments from the theory of *matrix-geometric* solutions developed by Neuts [81] and the *spectral-expansion* approach, see for instance Mitrani [73], Mitrani and Mitra [77], Mitrani and Chakka [75], or Haverkort and Ost [42]. A third approach for solving the models described in this chapter relies on using generating functions. This method is developed, for instance, in Gail et al. [36]. The three methods are closely related, as we will see in the course of this chapter. We choose using the matrix-geometric approach to enable probabilistic interpretation of various entities. The link with the spectral-expansion approach is made to facilitate the analysis.

The chapter is organised as follows. We present the models in Section 2.1 and discuss related literature in Section 2.2. Section 2.3 provides a starting point for the analysis using the theory of matrix-geometric solutions in combination with the spectral-expansion technique. A detailed spectral analysis is presented in Section 2.4. The steady-state queue-length distributions for both the infinite-queue and the finite-queue model are given in Section 2.5. In Section 2.6 the influence of fast and slow service rate fluctuations is studied. The models are modified in Section 2.7 to include the case where $r_+^{(\mathrm{e})} < \infty$. The results are illustrated by numerical experiments in Section 2.8, using the model of [86]. Conclusions are drawn in Section 2.9.

2.1 Model description

We consider a service station of which the capacity changes according to a birth and death process on $\{1, 2, \ldots, N\}$, N being a positive integer. This process is denoted by $[Y(t)]_{t \geq 0}$. When $Y(t) = i$, the birth rate is $q_i^+ > 0$, $i \in \{1, 2, \ldots, N-1\}$, and the death rate is $q_i^- > 0$, $i \in \{2, 3, \ldots, N\}$. For notational convenience we set $q_1^- = q_N^+ = 0$ and further define $q_i := q_i^- + q_i^+$. The station works at rate $c_i \geq 0$ when $Y(t) = i \in \{1, 2, \ldots, N\}$.

When $Y(t) = i \in \{1, 2, \ldots, N\}$, new customers arrive to the system according to a Poisson process with rate λ_i. We assume that the service requirements of customers are i.i.d., having an exponential distribution with mean $1/\mu$, and are independent of the arrival process. Furthermore, $Y(t)$ — the state of the birth and death process at time t — is assumed to be independent of all interarrival times and service requirements of the customers before time t. The available capacity is shared among all present customers according to the processor-sharing discipline. Because of the exponentially distributed service requirements, the queue-length process is identical to that of the same model with the FCFS (First Come First Served) service discipline.

So far we did not impose any restriction on the number of customers in the system. Suppose that the maximum queue length is $L < \infty$. If a customer arrives at the service station and finds L other customers present, the new customer is rejected and lost. To denote various entities, such as state descriptors and steady-state probabilities, we use the superscript $^{(L)}$ for the finite-queue model and the superscript $^{(\infty)}$ for the infinite-queue model. When concerned with both, no superscripts are attached.

Remark 2.1.1 The model with $L < \infty$ includes variants of the model in Section 1.5 when $r_-^{(\mathrm{e})} > 0$. In that case the number of customers in the queue can be at most

$$L = \left\lfloor \frac{\min_i\{c_i\}}{r_-^{(\mathrm{e})}} \right\rfloor,$$

where $\lfloor x \rfloor$ is the largest integer smaller than or equal to x. Note that this implies that no customers can be accepted when $c_i = 0$, for some $i \in \{1, 2, \ldots, N\}$. In our analysis we consider the more general case where some of the c_i may be equal to 0.

In Section 2.7 we also extend our models to include variants of the model in Section 1.5 where individual customers can not be served at a rate greater than a maximum allowed rate $r_+^{(\mathrm{e})} < \infty$.

Let $X(t)$ be the number of customers present in the system at time t. Then the process $(X(t), Y(t))$ is an irreducible and aperiodic Markovian process. When $X(t) = j$ and $Y(t) = i$ we say that the process $(X(t), Y(t))$ is in state (j, i). By definition, $Y(t)$ is not influenced by $X(t)$. If we define $p_i := \mathbf{P}\{Y = i\} := \lim_{t\to\infty} \mathbf{P}\{Y(t) = i\}$, then

$$\begin{aligned} p_1 &= \left(1 + \sum_{i=2}^{N} \prod_{k=2}^{i} \frac{q_{k-1}^{+}}{q_k^{-}}\right)^{-1}, \\ p_i &= p_1 \prod_{k=2}^{i} \frac{q_{k-1}^{+}}{q_k^{-}}, \qquad i = 2, \ldots, N, \end{aligned} \tag{2.1}$$

see for instance Cohen [20, Section I.4.1]. The (row) vector of these steady-state probabilities is denoted by $\overline{p} = (p_1, p_2, \ldots, p_N)$. We further define the simultaneous equilibrium probabilities

$$\pi_{j,i} := \mathbf{P}\{X = j, Y = i\} := \lim_{t\to\infty} \mathbf{P}\{X(t) = j, Y(t) = i\}, \tag{2.2}$$

and partition them into (row) vectors $\overline{\pi}_j := (\pi_{j,1}, \pi_{j,2}, \ldots, \pi_{j,N})$ of length N. The vector $\overline{\pi}_j$ is associated with level j, that is the set of states in which exactly j customers are present. The partition enables us to write the equilibrium vector as a block vector $\overline{\pi}^{(\infty)} = (\overline{\pi}_0^{(\infty)}, \overline{\pi}_1^{(\infty)}, \overline{\pi}_2^{(\infty)}, \ldots)$ for the infinite-queue model, and $\overline{\pi}^{(L)} = (\overline{\pi}_0^{(L)}, \overline{\pi}_1^{(L)}, \ldots, \overline{\pi}_L^{(L)})$ for the finite-queue model. The corresponding infinitesimal generators of the processes $(X^{(\infty)}(t), Y(t))$ and $(X^{(L)}(t), Y(t))$ — we do not use the superscripts ${}^{(\infty)}$ and ${}^{(L)}$ in the notation of $Y(t)$, because in both cases that process evolves independently of the corresponding X-process — are given by:

$$\mathcal{Q}^{(\infty)} := \begin{bmatrix} Q_d^{(0)} & \Lambda & 0 & \cdots & & & \\ M & Q_d & \Lambda & 0 & \cdots & & \\ 0 & M & Q_d & \Lambda & 0 & \cdots & \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \end{bmatrix}, \tag{2.3}$$

$$\mathcal{Q}^{(L)} := \begin{bmatrix} Q_d^{(0)} & \Lambda & 0 & \cdots & \cdots & 0 \\ M & Q_d & \Lambda & 0 & \cdots & \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & & M & Q_d & \Lambda \\ 0 & \cdots & & 0 & M & Q_d^{(L)} \end{bmatrix}. \tag{2.4}$$

$\mathcal{Q}^{(L)}$ consists of $L+1$ block rows and block columns. The matrices on the diagonal are given by $Q_d^{(0)} = Q^{(Y)} - \Lambda$, $Q_d = Q^{(Y)} - \Lambda - M$, and $Q_d^{(L)} = Q^{(Y)} - M$. The matrices $Q^{(Y)}$, Λ, M and Q_d are all of dimension $N \times N$. Λ is the diagonal matrix $\operatorname{diag}[\lambda_1, \lambda_2, \ldots, \lambda_N]$, M is the diagonal matrix $\operatorname{diag}[\mu c_1, \mu c_2, \ldots, \mu c_N]$, and $Q^{(Y)}$ is the (tri-diagonal) infinitesimal generator of the process $Y(t)$:

$$Q^{(Y)} := \begin{bmatrix} -q_1 & q_1^+ & 0 & \dots & \dots & \dots & 0 \\ q_2^- & -q_2 & q_2^+ & 0 & \dots & \dots & 0 \\ 0 & q_3^- & -q_3 & q_3^+ & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & \dots & q_{N-1}^- & -q_{N-1} & q_{N-1}^+ \\ 0 & \dots & \dots & \dots & \dots & q_N^- & -q_N \end{bmatrix}. \tag{2.5}$$

A two-dimensional Markov process $(X(t), Y(t))$ with a block tri-diagonal generator as in Definitions (2.3) and (2.4) is called a QBD (Quasi Birth and Death) process. We refer to Neuts [81, Chapter 3] for a general discussion of QBD processes. By Theorem 3.1.1 of Neuts [81], we have that the process $(X^{(\infty)}(t), Y(t))$ is ergodic if and only if

$$\rho := \sum_{i=1}^{N} p_i \lambda_i / \mu < \sum_{i=1}^{N} p_i c_i =: c,$$

where ρ is the mean traffic load, that is the amount of work arriving to the system per unit of time, and c is the mean service rate. In the sequel, when addressing the infinite-queue model, we assume that $\rho < c$. For the finite-queue model the steady-state distribution also exists when this is not the case.

In the analysis a special role is played by the number of states of the server for which the arrival rate is zero, and by the number of states for which the service rate is zero. We denote these numbers by

$$\begin{aligned} m_0 &:= \#\{i \,:\, \lambda_i = 0\}, \\ n_0 &:= \#\{i \,:\, c_i = 0\}. \end{aligned}$$

2.2 Related literature

The model of Section 1.5 with $r_+^{(\mathrm{e})} = \infty$, $r_-^{(\mathrm{e})} = 0$, and exponentially distributed file sizes and holding times, is a special case of the one studied in this chapter. To see that, take $\mu = 1/f^{(\mathrm{e})}$ and further, $\forall i$, $\lambda_i = \lambda^{(\mathrm{e})}$, $q_i^+ = \lambda^{(\mathrm{s})}$, $q_i^- = (i-1)/h^{(\mathrm{s})}$, and $c_i = C - (i-1)r^{(\mathrm{s})}$. At time t, $X(t)$ is the number of elastic customers, and $Y(t) - 1$ is the number of stream customers. This model is analysed by Núñez Queija and Boxma [86]. Note that stream customers have preemptive priority over elastic customers. Variants of this priority model were studied by several authors. Mitrani and King [76], and later Gail et al. [35], solved the case where both types of customers have an infinite waiting space and within each customer class the service discipline is FCFS. Lehoczky and Gaver [64] developed a diffusion approximation. Gail et al. [34] also studied the non-preemptive case of this model. Falin et al. [26] analysed the model of [86] with an infinite waiting room for stream customers. A discrete-time variant modelled as an M/G/1-type Markov chain was solved by Gail et al. [33]. An extensive treatment of the spectral analysis of M/G/1-type Markov chains by means of generating functions is given in Gail et al. [36]. Yechiali [125] and Daigle and Lucantoni [23] studied the present model. Here, we are able to carry the analysis somewhat further and, additionally, we discuss the case with a finite queue.

The analysis presented in [86], which is the basis for the analysis in this chapter, was motivated by the introduction of the ABR transfer capability for data transmissions in ATM networks. Several studies concerned with the application of similar models to the performance analysis of ABR have appeared in the literature around the same time as [86, 82]. We mention Altman et al. [5], Blaabjerg et al. [14], and Kulkarni and Li [57].

2.3 Preliminaries

This section provides the starting point for the analysis. A central role is played by both the theory of matrix-geometric solutions developed by Neuts [81] and the closely related spectral-expansion technique, see for instance Mitrani [73] or Mitrani and Mitra [77]. For comparisons of both approaches see Mitrani and Chakka [75] or Haverkort and Ost [42].

It is well known that if $\rho < c$, then the unique probability vector $\overline{\pi}^{(\infty)} = (\overline{\pi}_0^{(\infty)}, \overline{\pi}_1^{(\infty)}, \overline{\pi}_2^{(\infty)}, \ldots)$ satisfying $\overline{\pi}^{(\infty)} \mathcal{Q}^{(\infty)} = 0$ has the matrix-geometric form,

$$\overline{\pi}_{j+1}^{(\infty)} = \overline{\pi}_j^{(\infty)} R, \tag{2.6}$$

or equivalently,

$$\overline{\pi}_j^{(\infty)} = \overline{\pi}_0^{(\infty)} R^j, \tag{2.7}$$

where the "rate matrix" R is the minimal non-negative solution to the quadratic matrix equation,

$$\Lambda + RQ_d + R^2 M = 0, \tag{2.8}$$

see Neuts [81, Theorem 3.1.1].

Remark 2.3.1 For the infinite-queue model, the element $[R]_{k,k'}$ of the matrix R equals $-[Q_d]_{k,k}$ times the expected cumulative time spent in the state $(j+1, k')$ starting from (j,k), before either the first return to the level j or, when $\rho \geq c$, "drifting to infinity", see [81, Section 1.7].

For the finite-queue model a related result holds when $\rho \neq c$ (we come back to this assumption in Remark 2.3.6). In that case the steady-state probability vector can be written as a *sum of two matrix-geometric terms*,

$$\overline{\pi}_j^{(L)} = \overline{x}_0 R^j + \overline{x}_L S^{L-j}, \quad j \in \{0, 1, \ldots, L\}, \tag{2.9}$$

see Naoumov [79, Corollary 5] or Krieger et al. [56, Theorem 1]. This result was first observed by Hajek [41, Remark following Theorem 5], but here we follow the terminology of [56]. The vectors $\overline{x}_0$ and $\overline{x}_L$ are both of dimension

N, and their concatenation $(\overline{x}_0, \overline{x}_L)$ can be found as the solution to a set of linear equations of dimension $2N$, see [56, Equation (21)]. The matrix S is the minimal non-negative solution to

$$S^2\Lambda + SQ_d + M = 0, \tag{2.10}$$

and the matrix R is as above.

Remark 2.3.2 The matrix S may be interpreted as the analogue of the matrix R in the "dual" QBD process with levels $\{\ldots, -2, -1, 0\}$ having generator:

$$\begin{bmatrix} \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \cdots & 0 & M & Q_d & \Lambda & 0 \\ \cdots & \cdots & 0 & M & Q_d & \Lambda \\ \cdots & \cdots & \cdots & 0 & M & Q_d^{(0)'} \end{bmatrix},$$

where $Q_d^{(0)'} = Q^{(Y)} - M$. For this process, the element $[S]_{k,k'}$ is $-[Q_d]_{k,k}$ times the expected time spent in $(j-1, k')$ starting from (j, k) before either returning to the level j or drifting to minus infinity.

Remark 2.3.3 The condition $\rho \neq c$ is equivalent to $S\Lambda + Q_d + RM$ being non-singular as required by Naoumov [79, Corollary 5] and Krieger et al. [56, Theorem 1]. The backward implication is given in [56, Proposition 1.(3)]. The forward implication can be proved using probabilistic arguments by noting that if $\rho \neq c$ then the *two-sided infinite* QBD process is transient. For that process, the matrix $S\Lambda + Q_d + RM$ can be interpreted as the subgenerator containing the transition rates of eventually returning to the level of departure, see also Remark 2.3.4. This matrix is non-singular if and only if it is a true generator, which is not the case when the two-sided infinite QBD process is transient.

Lemma 2.3.1 *The eigenvalues of the matrices R and S lie inside or on the unit circle in the complex plane. Moreover, all the eigenvalues of R lie inside the unit circle if and only if $\rho < c$, and all the eigenvalues of S lie inside the unit circle if and only if $\rho > c$.*

Proof The statements for the matrix R can be found in Neuts [81, Theorems 1.3.1 and 3.1.1]. By symmetry the analogous results follow for the matrix S, see also Naoumov [79, Proposition 12] or Krieger et al. [56, Proposition 1.(2)]. □

In order to study the eigenvalues of the matrices R and S, we define the quadratic matrix polynomial $T(z)$:

$$T(z) := \Lambda + zQ_d + z^2M. \tag{2.11}$$

A treatise of the general theory of matrix polynomials can be found in Gohberg et al. [38]. Here we exploit the tri-diagonal structure of the matrix $T(z)$ for a detailed spectral analysis.

The roots of $\det[T(z)]$ as a (scalar) polynomial function of z are called the nullvalues of $T(z)$. Suppose $\overline{v}$ is an eigenvector of the matrix R corresponding to the eigenvalue ψ. After pre-multiplying both sides of Equation (2.8) by $\overline{v}$, it follows that $\overline{v}$ is in the left nullspace of the matrix $T(\psi)$. Hence, $\det[T(\psi)] = 0$, i.e., ψ is a nullvalue of $T(z)$. Similarly, using Equation (2.10), if $\psi \neq 0$ and $1/\psi$ is an eigenvalue of S then ψ is a nullvalue of $T(z)$. We further explore these properties in the following lemma by factorising the matrix polynomial $T(z)$. Similar factorisations have been established by Zhao et al. [126, Theorem 3.8(a)] for very general Markov processes having generators with repeating block rows. The following lemma is a restatement of the factorisation obtained by Núñez Queija and Boxma [86].

Lemma 2.3.2 *The matrix polynomial $T(z)$ can be factorised as*

$$T(z) = (zI - R)\,(zM + RM + Q_d)\,, \tag{2.12}$$

or alternatively as

$$T(z) = (I - zS)\,(\Lambda + zS\Lambda + zQ_d)\,. \tag{2.13}$$

Proof By writing out the multiplication of individual terms and using Equations (2.8) and (2.10). □

Corollary 2.3.3 *If ψ is an eigenvalue of R with algebraic multiplicity m, then ψ is a root of the polynomial $\det[T(z)]$ and its multiplicity is at least m. And if $\overline{v}$ is an eigenvector of the matrix R corresponding to the eigenvalue ψ, then $\overline{v}$ is a left nullvector of $T(\psi)$.*

If $\psi \neq 0$ and $1/\psi$ is an eigenvalue of S with algebraic multiplicity m, then ψ is a root of $\det[T(z)]$ and its multiplicity is not smaller than m. Moreover, if $\overline{v}$ is an eigenvector of S corresponding to the eigenvalue $1/\psi$, then $\overline{v}$ is a left nullvector of $\det[T(\psi)]$.

Proof Directly from Lemma 2.3.2. □

The previous corollary does not account for a zero eigenvalue of the matrix S. This problem will be tackled by Corollary 2.3.5, which follows from the next lemma.

Lemma 2.3.4 *The left nullspaces of R and Λ coincide, and so do the left nullspaces of S and M.*

Proof We prove the statement for R and Λ. The statement for S and M follows from an analogous argument. Suppose R is singular with left nullvector $\overline{v}$. After pre-multiplying Equation (2.8) by $\overline{v}$ it is seen that $\overline{v}$ is also a left nullvector of Λ. Conversely, suppose λ_i equals 0, and so the vector $\overline{1}_i$ with the i^{th} entry equal to 1 and all other entries equal to 0 is a left nullvector of Λ. Then in the infinite-queue model the level can not increase (i.e., no customers can arrive) when the state of the server is i. Therefore starting from (j,i), for any j, no state of the level $j+1$ can ever be visited before another state in level j is visited. Hence, from Remark 2.3.1, all entries in row i of the matrix R equal 0, so that $\overline{1}_i$ is a left nullvector of R. □

Corollary 2.3.5 *The matrix S is singular if and only if $n_0 > 0$, i.e., if there is at least one c_i equal to 0. In that case $z = 0$ is a zero of the polynomial* $\det\left[z^2 T(\frac{1}{z})\right]$*, the multiplicity of this zero is not smaller than the algebraic multiplicity of the eigenvalue 0 of S.*

Proof From Lemma 2.3.4 and Expression (2.13). □

Remark 2.3.4 The factorisation of $T(z)$ given by Expression (2.12) — and analogously for Expression (2.13) — can be further elaborated on to obtain:

$$T(z) = (R - zI)\,(Q_d + RM)\,(zG - I)\,, \tag{2.14}$$

where the matrix G is the minimal non-negative solution to $\Lambda G^2 + Q_d\, G + M = 0$. The element $[G]_{i,i'}$ of the matrix G is the probability that in the infinite-queue model, for any j, starting in state $(j+1,i)$ the process enters the level j for the first time through the state (j,i'), see Neuts [81, Section 3.3]. In the ergodic case the matrix G obviously is stochastic. Using probabilistic arguments it can be argued that $\Lambda G = RM$. Both sides of this equality contain the transition rates of returning to a level from the level above it. For a technical proof of this identity (in the discrete-time case) see Latouche et al. [61, Theorem 2.1]. With this identity Equation (2.14) is readily verified. This factorisation leads to the identification of the eigenvalues of G as the inverses of roots of $\det[T(z)]$, see Núñez Queija [82].

In Section 2.4 we show that both for R and S the set of eigenvectors spans $\mathbb{R}^N$. If $\rho < c$ we can thus rewrite Expression (2.7) in the "spectral-expansion" form:

$$\overline{\pi}_j^{(\infty)} = \sum_{k=1}^{N} \alpha_k\,(\psi_k)^j\,\overline{v}_k, \quad j = 0,1,2,\ldots, \tag{2.15}$$

with $\psi_1,\ldots,\psi_N$ the eigenvalues of the matrix R and $\overline{v}_1,\ldots,\overline{v}_N$ the corresponding left eigenvectors, i.e., $\overline{v}_k R = \psi_k \overline{v}_k$, $k \in \{1,\ldots,N\}$. Similarly, if $\rho \neq c$ and $m_0 = n_0 = 0$ then, for $j = 0,1,\ldots,L$, Expression (2.9) can be rewritten as:

$$\overline{\pi}_j^{(L)} = \sum_{k=1}^{N} \beta_k\,(\psi_k)^j\,\overline{v}_k + \sum_{k=N+1}^{2N} \beta_k \left(\frac{1}{\psi_k}\right)^{L-j} \overline{v}_k. \tag{2.16}$$

Here, $\psi_1,\ldots,\psi_N$ and $\overline{v}_1,\ldots,\overline{v}_N$ are as before, and $1/\psi_{N+1},\ldots,1/\psi_{2N}$ are the eigenvalues of S with corresponding left eigenvectors $\overline{v}_{N+1},\ldots,\overline{v}_{2N}$.

In Section 2.5 we show how the coefficients α_k in Expression (2.15) and the coefficients β_k in Expression (2.16) can be found once the required ψ_k and $\overline{v}_k$ are determined.

Remark 2.3.5 If $m_0 > 0$ or $n_0 > 0$ (or both) then Expressions (2.15) and (2.16) essentially remain valid. If S has a zero eigenvalue, then we set the corresponding ψ_k equal to ∞ and we write $1/\psi_k = 0$. Moreover, by convention we set $0^0 = 1$.

Remark 2.3.6 If $\rho = c$ then both R and S have an eigenvalue 1, since $\psi_N = \psi_{N+1} = 1$, and $\overline{p}$ — of which the entries are defined by Expression (2.1) — is the (unique) corresponding left eigenvector for both matrices. The steady-state distribution for the finite-queue model can not be written as in Expression (2.9). However, Expression (2.16) can be modified to include this case. When $\rho = c$, the steady-state queue-length distribution can be written as

$$\overline{\pi}_j^{(L)} = \beta_N \overline{p} + \beta_{N+1} \left[\overline{u} + j\overline{p}\right] + \sum_{k=1}^{N-1} \beta_k \left(\psi_k\right)^j \overline{v}_k + \sum_{k=N+2}^{2N} \beta_k \left(\frac{1}{\psi_k}\right)^{L-j} \overline{v}_k. \quad (2.17)$$

Here $\overline{u}$ is a vector satisfying $\overline{u}Q^{(Y)} = \overline{p}\left[\Lambda - M\right]$. Note that such a vector indeed exists, since the row space of the matrix $Q^{(Y)}$ is exactly the hyperplane that is perpendicular to the column vector $\overline{1}$ consisting only of ones. Note that when $\rho = c$, indeed $\overline{p}\left[\Lambda - M\right]\overline{1} = 0$. The vector $\overline{u}$ is unique up to a translation along the vector $\overline{p}$. We choose the unique $\overline{u}$ which is perpendicular to $\overline{p}$, i.e., $\overline{u}$ is such that $\langle \overline{u}, \overline{p} \rangle = 0$.

Remark 2.3.7 Using generating functions, the steady-state probabilities for the infinite-queue model are obtained from the equation:

$$\overline{\pi}^{(\infty)}(z)\left[z^2\Lambda + zQ_d + M\right] = (1-z)\overline{\pi}_0^{(\infty)}M.$$

Hence, the poles of the generating function $\overline{\pi}^{(\infty)}(z) := \sum_{j=0}^{\infty} z^j \overline{\pi}_j^{(\infty)}$ correspond to the inverses of the nullvalues of $T(z)$. See Gail et al. [36] for an extensive treatment of this method.

2.4 Spectral analysis

We investigate the roots of the polynomial $\det\left[T(z)\right]$. We show that all these roots are real and positive. Yechiali [125] and Daigle and Lucantoni [23] present a related spectral analysis for the finite-queue model when $\rho < c$. We extend their results for that case and also analyse the case with $\rho \geq c$. In particular, in [23] it is not proved but *assumed* that the zeros of $\det\left[T(z)\right]$ are different to ensure the validity of Expression (2.15). Except for the matrix symmetrisation used in the proof of Lemma 2.4.1, the techniques used in our analysis are different from the ones in [125, 23]. We mainly use continuity arguments to prove our results, whereas for instance in [23] the authors use the fact that the matrices Λ, $Q^{(Y)}$, and M are semidefinite.

Lemma 2.4.1 *For real $z \neq 0$ the matrix $T(z)$ has N different, real, eigenvalues.*

Proof Note that $T(z)$ is a tri-diagonal matrix with off-diagonal elements:

$$\begin{aligned} T(z)_{i-1,i} &= q_{i-1}^{+} z, \\ T(z)_{i,i-1} &= q_i^{-} z, \end{aligned}$$

where $i = 2, 3, \ldots, N$. We denote the i^{th} diagonal element $T(z)_{i,i}$ by $t_i(z)$,

$$t_i(z) := \lambda_i - \{q_i + \lambda_i + \mu c_i\} z + \mu c_i z^2, \quad i = 1, 2, \ldots, N.$$

For real z the matrix $T(z)$ is *similar* to a real symmetric matrix, i.e., there exists a non-singular matrix D such that $DT(z)D^{-1}$ is a real symmetric matrix. Take D to be the diagonal matrix diag $[d_1, d_2, \ldots, d_N]$ with $d_i = \sqrt{\frac{p_i}{p_1}}$. The p_i are given in Expression (2.1). Define the symmetric matrix $S(z) := DT(z)D^{-1}$. The eigenvalues of $T(z)$ and $S(z)$ coincide, and hence it remains to prove the assertions for $S(z)$. The entries of $S(z)$ are given by $[S(z)]_{i,i} = t_i(z)$, $[S(z)]_{i-1,i} = [S(z)]_{i,i-1} = z\sqrt{q_{i-1}^{+} q_i^{-}}$ and are zero in all other positions. The fact that, for real $z \neq 0$, $S(z)$ has N different real eigenvalues, can be seen as follows (see also Parlett [88, Section 7.7]): First, every eigenvalue of a real symmetric matrix is real. Second, any real symmetric matrix has a full set of eigenvectors, therefore if $S(z)$ has an eigenvalue θ with (algebraic) multiplicity larger than 1 then there must be (at least) two independent eigenvectors corresponding to θ. But $S(z)$ is tri-diagonal with non-zero elements directly above and directly below the diagonal, and so each eigenvalue has a unique corresponding eigenvector (up to multiplication by a scalar). □

The fact that the eigenvalues of $T(z)$ are real for real z, simplifies the analysis considerably. In the sequel we only consider the eigenvalues of $T(z)$ as real functions of the real variable z. Therefore, using Lemma 2.4.1, for real $z \neq 0$, we may denote the eigenvalues of $T(z)$ by

$$\theta_1(z) < \theta_2(z) < \ldots < \theta_N(z). \tag{2.18}$$

Obviously,

$$\{\theta_1(0), \theta_2(0), \ldots, \theta_N(0)\} = \{\lambda_1, \lambda_2, \ldots, \lambda_N\}. \tag{2.19}$$

In general the eigenvalues of a matrix are not continuous functions of the entries, see for instance Gail et al. [33]. However, using Lemma 2.4.1 the following is true for real values of z:

Lemma 2.4.2 *All eigenvalues $\theta_k(z)$, $k = 1, 2, \ldots, N$, are continuous functions of $z \in \mathbb{R}$.*

Proof By Lemma 2.4.1 the roots of $\det[T(z)-\theta I]$ as a polynomial of θ are real, for real z, and we may order them as in Relation (2.18). The roots of a polynomial depend continuously upon the coefficients of the polynomial, except for the coefficient of the leading term at the value 0, see Horn and Johnson [44, Appendix D]. The leading term of the characteristic polynomial is θ^N, hence its coefficient is 1, independent of z. Hence, using the established ordering, the eigenvalues are continuous *functions* of the coefficients of the characteristic polynomials. The coefficients of the non-leading terms are themselves polynomials of z and therefore continuous in z. □

Recall that m_0 and n_0 are the numbers of states $i \in \{1, 2, \ldots, N\}$ for which $\lambda_i = 0$ and $c_i = 0$, respectively. The next lemma localises all but one of the $2N - n_0$ nullvalues of $T(z)$.

Lemma 2.4.3 *If $m_0 > 0$ then $\theta_1(0) = \theta_2(0) = \ldots = \theta_{m_0}(0) = 0$. Moreover, $\theta_N(1) = 0$, for $k = m_0+1, m_0+2, \ldots, N-1$ the equation $\theta_k(z) = 0$ has (at least) one solution for $z \in (0,1)$, and for $k = n_0+1, n_0+2, \ldots, N-1$ the equation $\theta_k(z) = 0$ has (at least) one solution for $z \in (1,\infty)$.*

Proof It is clear that $\det[T(1)] = 0$, since the rows of $T(1)$ sum to 0. Furthermore, by Gerŝgorin's theorem all the eigenvalues of $T(1)$ are non-positive, since each eigenvalue must be in at least one of the N Gerŝgorin discs (see for instance Marcus and Minc [69, Section III.2.2]). Consider the Gerŝgorin discs in the complex plane corresponding to the rows of the matrix. Each row determines such a disc in the following way: the diagonal element in the row is the center of the disc and the radius of the disc is equal to the sum of the absolute values of the off-diagonal elements in the row. Since (i) the diagonal elements of $T(1)$ are negative reals, (ii) the off-diagonal elements are non-negative reals, (iii) all rows sum to 0, and (iv) the eigenvalues are real, all eigenvalues must be non-positive. This combined with $\det[T(1)] = 0$ and Relation (2.18) gives

$$\theta_0(1) < \theta_1(1) \ldots < \theta_N(1) = 0.$$

The roots in $[0,1)$ now follow immediately from Equation (2.19) and the continuity of the $\theta_k(z)$, since each of the $\theta_k(z)$, for $k = m_0+1, m_0+2, \ldots, N-1$, must cross the horizontal axis at least once, somewhere in $(0,1)$.

The remaining zeros can be found by repeating the above argument for the matrix $z^2T(\frac{1}{z})$. Note that, for $z \neq 0$, the eigenvalues of this matrix are given by $\vartheta_k(z) := \theta_k(\frac{1}{z})$, $k = 1, 2, \ldots, N$. By continuity we may define the $\vartheta_k(0)$:

$$\{\vartheta_1(0), \vartheta_2(0), \ldots, \vartheta_N(0)\} = \{\mu c_1, \mu c_2, \ldots, \mu c_N\}.$$

Note that $\vartheta_1(0) = \vartheta_2(0) = \ldots = \vartheta_{n_0}(0) = 0 < \vartheta_{n_0+1}(0) < \ldots < \vartheta_N(0)$, and each of the $\vartheta_k(z)$, for $k = n_0+1, n_0+2, \ldots, N-1$, must cross the horizontal axis at least once, somewhere in (0,1). Hence, the $\theta_k(z)$ for $k = n_0+1, n_0+2, \ldots, N-1$, must cross the horizontal axis somewhere in $(1,\infty)$. □

Lemma 2.4.4 *If $\rho < c$, then $\theta_N(z) = 0$ for some $z \in (0,1)$. If $\rho > c$, then $\theta_N(z) = 0$ for some $z \in (1,\infty)$. If $\rho = c$, then the zero of $\theta_N(z)$ at $z = 1$ is of multiplicity 2.*

Proof See Appendix 2.A. □

Theorem 2.4.5 *The polynomial* $\det[T(z)]$ *has a root of multiplicity m_0 located at $z = 0$. All remaining $2N - m_0 - n_0$ roots are positive reals. For their location we distinguish three cases:*

(i) If $\rho < c$ then all non-zero roots are single, $N - m_0$ of them lie in $(0,1)$, one at $z = 1$, and $N - n_0 - 1$ in $(1,\infty)$.

(ii) If $\rho > c$ then all non-zero roots are single, $N - m_0 - 1$ of them lie in $(0,1)$, one at $z = 1$, and $N - n_0$ in $(1,\infty)$.

(iii) If $\rho = c$ then the zero at $z = 1$ is of multiplicity 2 and all other non-zero roots are single. $N - m_0 - 1$ of them lie in $(0,1)$ and $N - n_0 - 1$ lie in $(1,\infty)$.

Proof From the definition of $T(z)$ it follows that the degree of $\det[T(z)]$ is $2N - n_0$. By Lemmas 2.4.3 and 2.4.4 we have found all roots of $\det[T(z)]$ with the required locations. □

Corollary 2.4.6 *The matrices R and S have a full set of eigenvectors, all their non-zero eigenvalues are single, and their left nullspaces correspond to the left nullspaces of Λ and M, respectively.*

(i) If $\rho < c$ then the non-zero eigenvalues of R correspond to the zeros of $\det[T(z)]$ *in $(0,1)$ and the nonzero eigenvalues of S correspond to the inverses of zeros of* $\det[T(z)]$ *in $[1,\infty)$.*

(ii) If $\rho > c$ then the non-zero eigenvalues of R correspond to the zeros of $\det[T(z)]$ *in $(0,1]$ and the non-zero eigenvalues of S correspond to the inverses of zeros of* $\det[T(z)]$ *in $(1,\infty)$.*

(iii) If $\rho = c$ then 1 is an eigenvalue of both matrices R and S. The other non-zero eigenvalues of R and S correspond to the zeros of $\det[T(z)]$ *in $(0,1)$ and to the inverses of zeros of* $\det[T(z)]$ *in $(1,\infty)$.*

Moreover, if $\psi \neq 0$ is an eigenvalue of R or $1/\psi \neq 0$ an eigenvalue of S then the corresponding left eigenvector (of R or S) equals the unique left nullvector of $T(\psi)$.

Proof The Corollary follows from Lemmas 2.3.1 and 2.3.4, Corollary 2.3.3, and Theorem 2.4.5. In particular the full dimension of the eigenspaces is ensured by Lemma 2.3.4 for the zero eigenvalues (since Λ and M are diagonal), and by the fact that all other eigenvalues are single. □

2.5 Queue length in steady state

Corollary 2.4.6 ensures that the matrix R has a full set of eigenvalues, hence, if $\rho < c$ then the equilibrium distribution for the infinite-queue model can be written as in Expression (2.15). As before, let m_0 and n_0 be the number of λ_i and the number of c_i, respectively, that are equal to 0. We order the eigenvalues of R, which are the roots of $\det[T(z)]$ in $(0,1)$, as $\psi_k = 0$, for $k = 1, 2, \ldots, m_0$, and $0 < \psi_{m_0+1} < \psi_{m_0+2} < \ldots < \psi_N < 1$, and construct the diagonal matrix $\Psi = \mathrm{diag}[\psi_1, \psi_2, \ldots, \psi_N]$. The corresponding (normalised) left eigenvectors $\overline{v}_1, \overline{v}_2, \ldots, \overline{v}_N$ constitute the matrix V, $\overline{v}_k$ being the k^{th} row of V. We have the obvious Jordan decomposition $R = V^{-1}\Psi V$.

Having determined the ψ_k and $\overline{v}_k$ as the roots inside $[0,1)$ of $\det[T(\cdot)]$ and the corresponding left nullvectors of $T(\cdot)$, respectively, it remains to find the coefficients α_k in Equation (2.15). Writing $\overline{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_N)$, and using $\sum_{j=0}^{\infty} \overline{\pi}_j^{(\infty)} = \overline{p}$, we have,

$$\overline{\alpha}\,(I - \Psi)^{-1}\, V = \overline{p}. \tag{2.20}$$

Indeed, the inverse of $I - \Psi$ exists, since 1 is not an eigenvalue of R when $\rho < c$, cf. Lemma 2.3.1. The entries of the probability vector $\overline{p}$ are given by Expression (2.1). Equation (2.20) uniquely determines $\overline{\alpha}$, since V is non-singular. Hence, the steady-state probability vectors $\overline{\pi}_j^{(\infty)}$, $j = 0, 1, 2, \ldots$, are determined through (2.15). In particular, the marginal queue-length distribution is given by

$$\mathbf{P}\left\{X^{(\infty)} = j\right\} = \overline{\alpha}\Psi^j V\overline{1} = \sum_{k=0}^{N} \alpha_k\,(\psi_k)^j\,\overline{v}_k\overline{1}. \tag{2.21}$$

Remark 2.5.1 From Expression (2.21) the moments of the number of customers in the infinite-queue model are easily determined, in particular the mean $\mathbf{E}\left[X^{(\infty)}\right]$ and the variance $\mathbf{Var}\left[X^{(\infty)}\right]$. Using Little's formula we immediately obtain the mean processing time (or sojourn time) from $\mathbf{E}\left[X^{(\infty)}\right]$.

Now we turn to the steady-state queue-length distribution for the finite-queue model, for the case that $\rho \neq c$. Analogous to the infinite-queue case, we order the zeros of $\det[T(z)]$ as $\psi_k = 0$, for $k = 1, 2, \ldots, m_0$, and

$$0 < \psi_{m_0+1} < \ldots < \psi_N \leq 1 \leq \psi_{N+1} < \ldots < \psi_{2N-n_0}.$$

The corresponding (normalised) left nullvectors of $T(z)$ are again denoted by $\overline{v}_k$. It was already noted below Expression (2.9) that the vectors $\overline{x}_0$ and $\overline{x}_L$ are uniquely determined by a set of linear equations of dimension $2N$, see Naoumov [79, Corollary 5] or Krieger et al. [56, Equation (21)]. Here we derive an alternative set of equations which uniquely determines the coefficients β_k by combining Expression (2.16) with the boundary equations,

$$\begin{aligned}
\overline{\pi}_0^{(L)}\left[Q^{(Y)} - \Lambda\right] + \overline{\pi}_1^{(L)} M &= \overline{0}, \\
\overline{\pi}_{L-1}^{(L)}\Lambda + \overline{\pi}_L^{(L)}\left[Q^{(Y)} - M\right] &= \overline{0},
\end{aligned}$$

and the normalisation condition $\sum_{j=0}^{L} \overline{\pi}_j^{(L)} \overline{1} = 1$. The resulting equations for the coefficients β_k are

$$
\begin{aligned}
&(\beta_1, \ldots, \beta_N, \beta_{N+1}, \ldots, \beta_{2N}) \\
&\times \begin{bmatrix} V\left[Q^{(Y)} - \Lambda\right] + \Psi V M & \Psi^{L-1} V \Lambda + \Psi^L V \left[Q^{(Y)} - M\right] \\ \Phi^L W \left[Q^{(Y)} - \Lambda\right] + \Phi^{L-1} W M & \Phi W \Lambda + W \left[Q^{(Y)} - M\right] \end{bmatrix} \\
&= \left(\overline{0}, \overline{0}\right),
\end{aligned}
\tag{2.22}
$$

with the normalisation

$$
\sum_{k=1}^{N} \beta_k \frac{1 - \psi_k^{L+1}}{1 - \psi_k} \overline{v}_k \overline{1} + \sum_{k=N+1}^{2N} \beta_k \frac{1 - \left(\frac{1}{\psi_k}\right)^{L+1}}{1 - \frac{1}{\psi_k}} \overline{v}_k \overline{1} = 1,
$$

with $\frac{1-x^{L+1}}{1-x} := L + 1$ when $x = 1$. The matrix Ψ is, as before, the diagonalisation of R, i.e., $\Psi = VRV^{-1}$. Similarly $\Phi = WSW^{-1}$ is the diagonalisation of S. Thus, Φ is the diagonal matrix containing the eigenvalues $1/\psi_{N+1}, 1/\psi_{N+2}, \ldots, 1/\psi_{2N}$, of S and the k^{th} row of W is $\overline{v}_{N+k}$, which is the left eigenvector of S corresponding to $1/\psi_{N+k}$. As in Remark 2.3.5, if the k^{th} eigenvalue of S equals zero we write $\psi_{N+k} = \infty$ and $1/\psi_{N+k} = 0$.

Remark 2.5.2 Since there is a unique equilibrium distribution $\overline{\pi}^{(L)}$ for the Markov process, a vector $(\beta_1, \beta_2, \ldots, \beta_{2N})$ must exist such that Equation (2.22) is satisfied. Recall that any such vector would lead to a solution of the internal balance equations, that is a solution to the equilibrium equations corresponding to the levels $j = 1, 2, \ldots, L-1$. Two different solutions to Equation (2.22) can not lead to the same steady-state probability distribution $\overline{\pi}^{(L)}$, using Expression (2.16), since $\rho \neq c$. Hence, the vector $(\beta_1, \beta_2, \ldots, \beta_{2N})$ solving Equation (2.22) is unique up to multiplication by a scalar.

Remark 2.5.3 If $\rho = c$, the steady-state queue-length distribution for the finite-queue model is given by Expression (2.17). In the same way as for Equation (2.22) we can derive a set of $2N$ equations for the β_k, by substituting Expression (2.17) into the boundary equations. We can argue in the same way as in the previous remark that this defines the β_k uniquely (up to multiplication by a common scalar).

2.6 Fast and slow fluctuations of the service rates

If the service capacity fluctuates very fast compared to the rate at which customers arrive (and depart), we may expect the queue-length processes to behave as those in the standard M/M/1/L queues (including $L = \infty$ for the infinite-queue model) with *constant* service capacity. An intuitive argument for this is that customers stay so long in the system, with respect to the fluctuations of the service rate, that the average service capacity during the residence time of a customer in the system is close to the (overall) mean service capacity.

When the service capacity changes very slowly, we expect the queues to behave as if the customers with probability p_i arrive to an M/M/1/L queue with service rate c_i, $i = 1, 2, \ldots, N$. Similar to the intuitive argument above, this can be explained by arguing that the service requirements of customers are so small that they typically do not see a change of service rate. In this argument special care should be given to the case that for at least one $i \in \{1, 2, \ldots, N\}$ the average amount of work arriving to the system, $\sum_{i=1}^{N} p_i \lambda_i / \mu$, is larger than or equal to the service rate c_i. Then with $L = \infty$ the "limiting" model is non-ergodic.

The limiting behaviour of the models is often referred to as nearly complete decomposability of the Markov chain under consideration. The first rigorous treatment of nearly completely decomposable Markov chains with a finite state space was given by Simon and Ando [107]. The authors argued that in many previous studies nearly complete decomposability was used to develop approximations of, for instance, the steady-state distribution of Markov chains. The results in [107] provide a theoretical basis for such an approximation. Courtois [22] further developed the theory giving an error analysis of the approximation, and applied it to various models, including queueing networks. For another application of nearly completely decomposable Markov chains in the performance analysis of multiservice communication networks see Reiman and Schmitt [92]. In this section we formalise the above statements about the limiting behaviour for fast and slow fluctuations of the service rate. We do this using a direct analytic approach based on the spectral analysis of Section 2.4. The finite-queue model could be analysed using the results of [107], but for the infinite-queue model that approach is not suitable. For completeness we choose to include the analysis of the finite-queue model too.

We study the above mentioned nearly complete decomposability by introducing a *time scale parameter* $\epsilon \in (0, \infty)$. For fixed ϵ, we define a Markovian birth and death process $Y_\epsilon(t)$ with infinitesimal generator $Q_\epsilon^{(Y)} := \epsilon Q^{(Y)}$. Let $Y(t) = Y_\epsilon(t)$ be the state of the server at time t. Note that the generator $Q_\epsilon^{(Y)}$ has the same structure as the generator Q_Y of the process $Y(t)$, see Expression (2.5), only the time scale has changed: transitions of the birth and death process occur ϵ times faster. The steady-state probability distribution of the process $Y_\epsilon(t)$ is independent of ϵ, see Expression (2.1). Therefore ρ and c are independent of ϵ too, and in particular the ergodicity condition for the infinite-queue model, $\rho < c$, remains unchanged. We extend previous definitions by using a subscript ϵ. For instance, we define the matrix $T_\epsilon(z)$ as the analogue of $T(z)$. It will be convenient to write

$$T_\epsilon(z) := (1 - z)\Lambda + zQ_\epsilon^{(Y)} + z(z - 1)M,$$

where we used the definition of $T(z)$ in Expression (2.11), and substituted $Q_d = Q_\epsilon^{(Y)} - \Lambda - M$. All results proved in Section 2.4 for $T(z)$ remain true for $T_\epsilon(z)$ as a function of z, when $\epsilon \in (0, \infty)$ is fixed. The case $\epsilon = 0$ is not included. For

$\epsilon \in (0, \infty)$ we also define the eigenvalues of $T_\epsilon(z)$

$$\theta_{\epsilon,1}(z) \leq \theta_{\epsilon,2}(z) \leq \ldots \leq \theta_{\epsilon,N}(z), \quad z \in \mathbb{R},$$

and the roots of det $[T_\epsilon(z)]$

$$\psi_{\epsilon,1} \leq \psi_{\epsilon,2} \leq \cdots \leq \psi_{\epsilon,2N}.$$

Remark 2.6.1 When $\rho = c$, the steady-state queue-length distribution of the finite-queue model is given by

$$\begin{aligned} \overline{\pi}_{\epsilon j}^{(L)} &= \beta_{\epsilon,N}\overline{p} + \beta_{\epsilon,N+1}\left[\frac{1}{\epsilon}\overline{u} + j\overline{p}\right] + \sum_{k=1}^{N-1} \beta_{\epsilon,k} \left(\psi_{\epsilon,k}\right)^j \overline{v}_{\epsilon,k} \\ &\quad + \sum_{k=N+2}^{2N} \beta_{\epsilon,k} \left(\frac{1}{\psi_{\epsilon,k}}\right)^{L-j} \overline{v}_{\epsilon,k}, \end{aligned}$$

cf. Expression (2.17) for the case $\epsilon = 1$. The vector $\overline{u}$ is defined as in Remark 2.3.6. Obviously the vector $\frac{1}{\epsilon}\overline{u}$ is continuous in $\epsilon \in (0, \infty)$.

Using the previous remark we can analyse the case $\rho = c$ by similar arguments as for $\rho \neq c$. For ease of presentation we make the following assumption throughout this section.

Assumption 2.6.1 *The mean arrival rate and the mean service rate are not equal, i.e.,* $\rho \neq c$.

Lemma 2.6.1 *For* $z \in \mathbb{R}$, *the eigenvalues* $\theta_{\epsilon,k}(z)$, $k = 1, 2, \ldots, N$, *are real and continuous functions of* $\epsilon \in [0, \infty)$. *Moreover,*

$$\{\theta_{0,k}(z) : k = 1, 2, \ldots, N\} = \{(1 - z)(\lambda_k - \mu c_k z) : k = 1, 2, \ldots, N\},$$

and,

$$\lim_{\epsilon \to \infty} \frac{\theta_{\epsilon,k}(z)}{\epsilon} = z\theta_{1,k}(1).$$

Proof For $\epsilon > 0$ we know from Lemma 2.4.1 that the eigenvalues of $T_\epsilon(z)$ are real. As in the proof of Lemma 2.4.2, the continuity of the $\theta_{\epsilon,k}(z)$ as functions of ϵ then follows from Horn and Johnson [44, Appendix D]. Because of the continuity in ϵ, the limit $\lim_{\epsilon\downarrow 0} \theta_{\epsilon,k}(z)$ is found by setting $\epsilon = 0$ in $T_\epsilon(z)$, which then becomes a diagonal matrix. Applying the same arguments to $T_\epsilon(z)/\epsilon$ as a matrix polynomial in $1/\epsilon$, using $\lim_{\epsilon\to\infty} T_\epsilon(z)/\epsilon = zQ^{(Y)}$, and $Q^{(Y)} = T_1(1)$, we find the second limit. $\square$

Lemma 2.6.2 *For* $k = 1, 2, \ldots, 2N - n_0$, *the nullvalues* $\psi_{\epsilon,k}$ *and the corresponding left nullvectors* $\overline{v}_{\epsilon,k}$ *of the matrix* $T_\epsilon(z)$ *are continuous functions of* $\epsilon \in (0, \infty)$.

Proof For $\epsilon \neq 0$, the coefficient of the leading term of det $[T_\epsilon(z)]$, as a polynomial of z, is non-zero. Therefore the roots $\psi_{\epsilon,k}$ are continuous in $\epsilon \neq 0$, see also the proof of Lemma 2.4.2.

The continuity of the nullvectors follows from the following construction of $\overline{v}_{\epsilon,k}$, for $k = 1, 2, \dots, 2N - n_0$. Let $\overline{v} = (v_1, v_2, \dots, v_N)$ be the vector with components $v_1 := 1$, $v_2 := -\left[T_\epsilon(\psi_{\epsilon,k})\right]_{1,1} / \left[T_\epsilon(\psi_{\epsilon,k})\right]_{2,1}$, and for $i = 3, 4, \dots, N$,

$$v_i := \frac{-v_{i-1}\left[T_\epsilon(\psi_{\epsilon,k})\right]_{i-1,i-1} - v_{i-2}\left[T_\epsilon(\psi_{\epsilon,k})\right]_{i-2,i-1}}{\left[T_\epsilon(\psi_{\epsilon,k})\right]_{i,i-1}}.$$

Since the nullvalues are continuous in $\epsilon > 0$, so are the v_i. Normalising the nullvectors such that their (Euclidean) norm equals 1, we have $\overline{v}_{\epsilon,k} = \frac{1}{|\overline{v}|}\overline{v}$. □

Corollary 2.6.3 *The coefficients $\alpha_{\epsilon,k}$, $k = 1, \dots, N$, and $\beta_{\epsilon,k}$, $k = 1, \dots, 2N$, are continuous functions of $\epsilon \in (0, \infty)$.*

Proof For $\epsilon > 0$, the matrix V_ϵ is non-singular and, from Lemma 2.6.2, continuous in ϵ. Hence, so is the inverse matrix V_ϵ^{-1}. The continuity of the coefficients $\alpha_{\epsilon,k}$ now follows from $\overline{\alpha}_\epsilon = \overline{p} V_\epsilon^{-1}\left[I - \Psi_\epsilon\right]$, see Equation (2.20).

The proof of the continuity of the coefficients $\beta_{\epsilon,k}$ is somewhat more involved but follows along the same lines. As in Remark 2.3.5, for $k = 2N - n_0 + 1, \dots, 2N$, we set by convention $\psi_{\epsilon,k} = +\infty$ and $\frac{1}{\psi_{\epsilon,k}} \equiv 0$ is a constant (and continuous) function of ϵ. For $\epsilon > 0$ we formulate the analogue of Equation (2.22):

$$\begin{aligned}
&(\beta_{\epsilon,1}, \dots, \beta_{\epsilon,N}, \beta_{\epsilon,N+1}, \dots, \beta_{\epsilon,2N}) \\
&\times \begin{bmatrix} \Psi_\epsilon V_\epsilon M + V_\epsilon\left[Q_\epsilon^{(Y)} - \Lambda\right] & \Psi_\epsilon^{L-1} V_\epsilon \Lambda + \Psi_\epsilon^L V_\epsilon \left[Q_\epsilon^{(Y)} - M\right] \\ \Phi_\epsilon^{L-1} W_\epsilon M + \Phi_\epsilon^L W_\epsilon \left[Q_\epsilon^{(Y)} - \Lambda\right] & \Phi_\epsilon W_\epsilon \Lambda + W_\epsilon \left[Q_\epsilon^{(Y)} - M\right] \end{bmatrix} \\
&= \left(\overline{0}, \overline{0}\right).
\end{aligned} \tag{2.23}$$

For fixed $\epsilon > 0$ the coefficients $\beta_{\epsilon,k}$ are unique. Hence, Equation (2.23) together with the normalisation condition defines a non-singular system. We conclude that these coefficients are continuous in $\epsilon > 0$. □

Let $\rho_i := \lambda_i / \mu$, $i \in \{1, 2, \dots, N\}$, and define the numbers $l_1 := \#\{i : \rho_i < c_i\}$ and $l_2 := \#\{i : \rho_i > c_i\}$. Furthermore, the permutation σ of $\{1, 2, \dots, N\}$ is such that it orders the ratios ρ_i / c_i in *non-decreasing* order:

$$\frac{\rho_{\sigma(1)}}{c_{\sigma(1)}} \leq \frac{\rho_{\sigma(2)}}{c_{\sigma(2)}} \leq \dots \leq \frac{\rho_{\sigma(N)}}{c_{\sigma(N)}}.$$

The permutation σ is not necessarily unique. Recall that by convention we set $\frac{1}{0}$ equal to $+\infty$, see Remark 2.3.5. To exclude the case $\rho_i = c_i = 0$, for some i, we make the following assumption for the remainder of this section.

Assumption 2.6.2 *For all $i = 1, 2, \dots, N$, either $\lambda_i > 0$ or $c_i > 0$.*

This assumption is not essential for our analysis, but facilitates the presentation. Moreover, the general case where we allow $\rho_i = c_i = 0$, for some i, can be derived from the results under the assumption that it is not allowed. In the following remark we indicate how this can be done.

Remark 2.6.2 When Assumption 2.6.2 is not satisfied, the states of the server that violate the assumption may be "eliminated" in the following way. Suppose $\lambda_i = c_i = 0$, for some $i \in \{1, 2, \dots, N\}$. Consider the length of the queue and the state of the server (for either model) only at times t that $Y(t) \neq i$. The resulting model is one with the same structure as before, only now the state of the server can be in a total number of $N-1$ states $\{1, 2, \dots, i-1, i+1, \dots, N\}$. From state $i-1$ the server moves to $i+1$ with rate $q_{i-1}^+ \times \left(q_i^+/q_i\right)$, and from state $i+1$ to $i-1$ with rate $q_{i+1}^- \times \left(q_i^-/q_i\right)$. All other transition rates for the state of the server, as well as the arrival and service rates of customers, are as before. The steady-state queue-length probabilities in the translated model correspond to those in the original model *conditional* on the state of the server not being equal to i.

Lemma 2.6.4 *The limits $\psi_{0,k} := \lim_{\epsilon\downarrow 0} \psi_{\epsilon,k}$ and $\overline{v}_{0,k} := \lim_{\epsilon\downarrow 0} \overline{v}_{\epsilon,k}$ exist for $k = 1, 2, \dots, 2N$. For $k \in \{1, 2, \dots, N\}$, it holds that*

$$\begin{aligned} \psi_{0,k} &= \min\left\{\frac{\rho_{\sigma(k)}}{c_{\sigma(k)}}, 1\right\}, \\ \psi_{0,k+N} &= \max\left\{\frac{\rho_{\sigma(k)}}{c_{\sigma(k)}}, 1\right\}. \end{aligned}$$

For $k \in \{1, 2, \dots, N\}$, if $\rho_{\sigma(k)} \leq c_{\sigma(k)}$ then $\overline{v}_{0,k}\left[\Lambda - \frac{\rho_{\sigma(k)}}{c_{\sigma(k)}} M\right] = \overline{0}$, otherwise

$$\overline{v}_{0,k}\left(\kappa_k\left[\Lambda - M\right] + Q^{(Y)}\right) = \overline{0},$$

where $0 \leq \kappa_N < \kappa_{N-1} < \dots < \kappa_{N-l_2+1} < \infty$ are all the l_2 non-negative and finite nullvalues of the matrix polynomial $\kappa\left[\Lambda - M\right] + Q^{(Y)}$.

For $k \in \{N+1, \dots, 2N\}$, if $\rho_{\sigma(k-N)} \geq c_{\sigma(k-N)}$ then $\overline{v}_{0,k}\left[\Lambda - \frac{\rho_{\sigma(k-N)}}{c_{\sigma(k-N)}} M\right] = \overline{0}$, otherwise

$$\overline{v}_{0,k}\left(\kappa_k\left[\Lambda - M\right] + Q^{(Y)}\right) = \overline{0},$$

where $0 \geq \kappa_{N+1} > \kappa_{N+2} > \dots > \kappa_{N+l_1} > -\infty$ are all the l_1 non-positive and finite nullvalues of the matrix polynomial $\kappa\left[\Lambda - M\right] + Q^{(Y)}$.

The matrices $V_0 := \lim_{\epsilon\downarrow 0} V_\epsilon$ and $W_0 := \lim_{\epsilon\downarrow 0} W_\epsilon$ are non-singular. It further holds that if $\rho_{\sigma(k)} = c_{\sigma(k)}$ then $\overline{v}_{0,k} = \overline{v}_{0,N+k}$.

Proof See Appendix 2.B. □

Corollary 2.6.5 *For $k = 1, 2, \ldots, N$,*

$$\lim_{\epsilon \downarrow 0} \alpha_{\epsilon,k} = p_{\sigma(k)} \left(1 - \min\left\{\frac{\rho_{\sigma(k)}}{c_{\sigma(k)}}, 1\right\}\right) \gamma_k,$$

where

$$\gamma_k := \sum_{i: \frac{\rho_{\sigma(i)}}{c_{\sigma(i)}} = \frac{\rho_{\sigma(k)}}{c_{\sigma(k)}}} \left[V_0^{-1}\right]_{i,k},$$

i.e., γ_k is the sum of the entries i in the k^{th} column of V_0^{-1} for which $\rho_{\sigma(i)} = \rho_{\sigma(k)}$.

Proof Write the vector $\overline{p}$ as a combination of the rows of V_0, this is possible since V_0 is non-singular, see Lemma 2.6.4. Then the limits follow from Equation (2.20) and Lemma 2.6.4. □

Corollary 2.6.6 *Let U_0 be the non-singular matrix with its successive rows equal to $\overline{v}_{0,i}$, $i \in \{1, 2, \ldots, N - l_2\} \cup \{2N - l_2 + 1, 2N - l_2 + 2, \ldots, 2N\}$. For $k = 1, 2, \ldots, N$, it holds that*

(i) if $\rho_{\sigma(k)} < c_{\sigma(k)}$ then

$$\lim_{\epsilon \downarrow 0} \beta_{\epsilon,k} = p_{\sigma(k)} \frac{1 - \rho_{\sigma(k)}/c_{\sigma(k)}}{1 - \left(\rho_{\sigma(k)}/c_{\sigma(k)}\right)^{L+1}} \gamma_k, \quad \lim_{\epsilon \downarrow 0} \beta_{\epsilon,N+k} = 0,$$

(ii) if $\rho_{\sigma(k)} > c_{\sigma(k)}$ then

$$\lim_{\epsilon \downarrow 0} \beta_{\epsilon,N+k} = p_{\sigma(k)} \frac{1 - \left(c_{\sigma(k)}/\rho_{\sigma(k)}\right)}{1 - \left(c_{\sigma(k)}/\rho_{\sigma(k)}\right)^{L+1}} \gamma_k, \quad \lim_{\epsilon \downarrow 0} \beta_{\epsilon,k} = 0,$$

(iii) if $\rho_{\sigma(k)} = c_{\sigma(k)}$ then

$$\lim_{\epsilon \downarrow 0} \beta_{\epsilon,k} + \beta_{\epsilon,N+k} = p_{\sigma(k)} \frac{1}{L+1} \gamma_k,$$

where

$$\gamma_k := \sum_{i: \frac{\rho_{\sigma(i)}}{c_{\sigma(i)}} = \frac{\rho_{\sigma(k)}}{c_{\sigma(k)}}} \left[U_0^{-1}\right]_{i,k}.$$

Proof See Appendix 2.C. □

Now we derive the limiting steady-state queue-length distributions, when $\epsilon \downarrow 0$, for both the infinite-queue and the finite-queue models. Note that in the infinite-queue model, if at least one of the ratios ρ_k/c_k is not smaller than 1, then the queue-length process becomes unstable in the limit.

Theorem 2.6.7 *For $i \in \{1, \ldots, N\}$, $j \in \{0, \ldots, L\}$,*

$$\lim_{\epsilon \downarrow 0} \mathbf{P}\left\{X_\epsilon^{(L)} = j, Y_\epsilon = i\right\} = p_i \frac{1 - \rho_i/c_i}{1 - (\rho_i/c_i)^{L+1}} \left(\frac{\rho_i}{c_i}\right)^j,$$

where $\frac{1-x}{1-x^{L+1}} := \frac{1}{L+1}$, when $x = 1$.
If $\rho < c$ then for $i \in \{1, \ldots, N\}$, $j \in \{0, 1, \ldots\}$,

$$\lim_{\epsilon \downarrow 0} \mathbf{P}\left\{X_\epsilon^{(\infty)} = j, Y_\epsilon = i\right\} = p_i \left(1 - \min\left\{\frac{\rho_i}{c_i}, 1\right\}\right) \left(\min\left\{\frac{\rho_i}{c_i}, 1\right\}\right)^j.$$

Proof The limiting distributions follow directly from Lemma 2.6.4 and Corollaries 2.6.5 and 2.6.6. □

Having determined the limiting steady-state queue-length distributions for slowly fluctuating service rates, we now turn to the case where the fluctuations in the service rates are fast with respect to the mean arrival and departure rates of customers.

Lemma 2.6.8 *The limits $\psi_{\infty,k} := \lim_{\epsilon \to \infty} \psi_{\epsilon,k}$ and $\overline{v}_{\infty,k} := \lim_{\epsilon \to \infty} \overline{v}_{\epsilon,k}$ exist for $k = 1, 2, \ldots, 2N$. The limits of the nullvalues are given by $\psi_{\infty,k} = 0$, for $k = 1, 2, \ldots, N-1$, by $\psi_{\infty,k} = \infty$, for $k = N+2, N+3, \ldots, 2N$, and by*

$$\begin{aligned} \psi_{\infty,N} &= \min\left\{\frac{\rho}{c}, 1\right\}, \\ \psi_{\infty,N+1} &= \max\left\{\frac{\rho}{c}, 1\right\}. \end{aligned}$$

Moreover, for $k = 1, 2, \ldots, N$,

$$\begin{aligned} &\overline{v}_{\infty,k}\left[Q^{(Y)} + \xi_k \Lambda\right] = 0, \\ &\overline{v}_{\infty,N+k}\left[Q^{(Y)} - \zeta_k M\right] = \overline{0}. \end{aligned}$$

Here $\infty \geq \xi_1 \geq \xi_2 \geq \ldots \geq \xi_N = 0$ are the nullvalues of the matrix polynomial $Q^{(Y)} + \xi\Lambda$, and $0 = \zeta_1 \geq \zeta_2 \geq \ldots \geq \zeta_N \geq -\infty$ are the nullvalues of the matrix polynomial $Q^{(Y)} - \zeta M$. In particular, $\overline{v}_{\infty,N} = \overline{v}_{\infty,N+1} = \overline{p}$. The matrices $V_\infty := \lim_{\epsilon \to \infty} V_\epsilon$ and $W_\infty := \lim_{\epsilon \to \infty} W_\epsilon$ exist and are non-singular.

Proof See Appendix 2.D. □

Corollary 2.6.9 *If $\rho < c$, then*

$$\begin{aligned} \lim_{\epsilon \to \infty} \alpha_{\epsilon,k} &= 0, \quad k = 1, 2, \ldots, N-1, \\ \lim_{\epsilon \to \infty} \alpha_{\epsilon,N} &= 1 - \frac{\rho}{c}. \end{aligned}$$

Proof Lemma 2.6.8 states that V_∞ exists and is non-singular. The vector $\overline{p}$ is (up to multiplication by a scalar) equal to the last row of V_∞ and, hence,

$$\lim_{\epsilon\to\infty} \overline{p}V_\epsilon^{-1} = (0,\ldots,0,1)\,.$$

The desired limits now follow directly from Equation (2.20). □

Corollary 2.6.10 *For $k = 1,2,\ldots,N-1,N+2,\ldots,2N$, it holds that*

$$\lim_{\epsilon\to\infty} \beta_{\epsilon,k} = 0.$$

If $\rho < c$ then $\lim_{\epsilon\to\infty} \beta_{\epsilon,N+1} = 0$ and

$$\lim_{\epsilon\to\infty} \beta_{\epsilon,N} = \frac{1-\rho/c}{1-(\rho/c)^{L+1}}.$$

If $\rho > c$ then $\lim_{\epsilon\to\infty} \beta_{\epsilon,N} = 0$ and

$$\lim_{\epsilon\to\infty} \beta_{\epsilon,N+1} = \frac{1-\rho/c}{1-(\rho/c)^{L+1}}.$$

Proof See Appendix 2.E. □

In the next theorem we use the above results to find the limiting steady-state queue-length distributions, when $\epsilon\to\infty$, for both the infinite-queue and the finite-queue models. Indeed, the limiting distributions correspond to the case where the arrival and service rates are fixed and equal to $\lambda := \sum_{k=1}^N p_k\lambda_k$ and c, respectively.

Theorem 2.6.11 *For $i\in\{1,\ldots,N\}$, $j\in\{0,\ldots,L\}$,*

$$\lim_{\epsilon\to\infty} \mathbf{P}\left\{X_\epsilon^{(L)} = j, Y_\epsilon = i\right\} = p_i \frac{1-\rho/c}{1-(\rho/c)^{L+1}}\left(\frac{\rho}{c}\right)^j.$$

If $\rho < c$, then for $i\in\{1,\ldots,N\}$, $j\in\{0,1,\ldots\}$,

$$\lim_{\epsilon\to\infty} \mathbf{P}\left\{X_\epsilon^{(\infty)} = j, Y_\epsilon = i\right\} = p_i\left(1-\frac{\rho}{c}\right)\left(\frac{\rho}{c}\right)^j.$$

Proof Follows from Lemma 2.6.8 and Corollaries 2.6.9 and 2.6.10. □

Remark 2.6.3 In this section we concentrated on finding the limiting distribution. A next step would be to investigate the speed of convergence to the limiting distributions, as ϵ tends to zero or to infinity. The speed of convergence is an indication of the error when approximating the probabilities $\pi_{\epsilon,(j,i)}$ by their limits, as ϵ tends to 0 or to ∞, see also Courtois [22, Chapter 2]. Partial results were obtained in the proofs of Lemmas 2.6.4 and 2.6.8 and in the proof of Corollary 2.6.6. In particular we found that if $\rho_i \neq c_i$ for all i, then the error of the approximation, $\pi_{\epsilon,(j,i)} - \pi_{0,(j,i)}$, is of the order $\mathrm{O}(\epsilon)$ when $\epsilon\downarrow 0$, for both the infinite- and finite-queue models.

2.7 Incorporating a maximum service rate

We show how the models presented in Section 2.1 can be modified to incorporate the case where the maximum rate $r_+^{(\mathrm{e})}$ at which one customer can be served is smaller than the maximum service rate of the station $\max_i \{c_i\}$. When the number of customers in the system is j and the state of the server is i, the departure rate of customers is equal to $\min\left\{\mu c_i, j\mu r_+^{(\mathrm{e})}\right\}$. We define

$$J := \left\lceil \frac{\max\{c_i : i = 1, 2, \ldots, N\}}{r_+^{(\mathrm{e})}} \right\rceil - 1,$$

which is not smaller than 1 by assumption. For any $x \in \mathbb{R}$, $\lceil x \rceil$ is the smallest integer which is larger than or equal to x. Note that $J+1$ is the minimum number of customers needed such that in any state of the server the total service capacity can be employed to serve customers. We illustrate the effect of the maximum service rate on the analysis of the infinite-queue model. Similar arguments hold for the finite-queue model.

The generator of the Markov chain describing the state of the server and the number of customers in the infinite-queue model can now be written as

$$Q^{(\infty)} := \begin{bmatrix} Q_d^{(0)} & \Lambda & & & & & & \\ M^{(1)} & Q_d^{(1)} & \Lambda & & & & & \\ & \ddots & \ddots & \ddots & & & & \\ & & M^{(J)} & Q_d^{(J)} & \Lambda & & & \\ & & & M & Q_d & \Lambda & & \\ & & & & M & Q_d & \Lambda & \\ & & & & & \ddots & \ddots & \ddots \end{bmatrix}.$$

$M^{(j)}$ is the diagonal matrix with i^{th} diagonal entry equal to $\min\left\{\mu c_i, j\mu r_+^{(\mathrm{e})}\right\}$ and $Q_d^{(j)} = Q^{(Y)} - \Lambda - M^{(j)}$, for $j = 1, 2, \ldots, J$. All other matrices are as before.

Remark 2.7.1 The discussion below also holds (with minor modifications) in the case that the matrix Λ in row $j = 0, 1, 2, \ldots, J-1$, is replaced with a matrix $\Lambda^{(j)}$ which may be different for different j.

Since the modification of the rates – with respect to the original model of Section 2.1 – only concerns a finite number of states, the ergodicity condition $\rho < c$ does not change. Moreover, the equilibrium equations for the levels $j > J$ remain exactly the same as in the original model, hence, in the ergodic case we still have Relation (2.6) for $j \geq J$, and with *exactly the same matrix* R *as before*. Therefore, the analysis with respect to the homogeneous part of the state space (all levels $j \geq J$) remains unchanged. However, once the eigenvalues and eigenvectors of the matrix R have been determined, we need to solve the following finite set of equations to find the $\overline{\pi}_0^{(\infty)}, \overline{\pi}_1^{(\infty)}, \ldots, \overline{\pi}_J^{(\infty)}$ (up to multiplication by a scalar):

$$\begin{aligned} \overline{\pi}_0^{(\infty)} Q_d^{(0)} + \overline{\pi}_1^{(\infty)} M^{(1)} &= \overline{0}, \\ \overline{\pi}_{j-1}^{(\infty)} \Lambda + \overline{\pi}_j^{(\infty)} Q_d^{(j)} + \overline{\pi}_{j+1}^{(\infty)} M^{(j+1)} &= \overline{0}, \qquad j = 1, \ldots, J-1, \\ \overline{\pi}_{J-1}^{(\infty)} \Lambda + \overline{\pi}_J^{(\infty)} Q_d^{(J)} + \overline{\pi}_J^{(\infty)} RM &= \overline{0}. \end{aligned} \tag{2.24}$$

Here, $\overline{\pi}_{J+1}^{(\infty)}$ in the last equation has been replaced with $\overline{\pi}_J^{(\infty)} R$ according to Relation (2.6), which is now valid for $j \geq J$. Having solved these equations, we can find the coefficients α_k in the following modification of Expression (2.15)

$$\overline{\pi}_j^{(\infty)} = \sum_{k=1}^{N} \alpha_k \left(\psi_k\right)^{j-J} \overline{v}_k, \quad j = J, J+1, J+2, \ldots, \tag{2.25}$$

by solving

$$\overline{\alpha} V = \overline{\pi}_J^{(\infty)}.$$

Note that $\overline{\pi}_J^{(\infty)}$ — as well as $\overline{\pi}_0^{(\infty)}, \overline{\pi}_1^{(\infty)}, \ldots, \overline{\pi}_{J-1}^{(\infty)}$ — was found up to multiplication by a scalar, hence, so is $\overline{\alpha}$. The common scalar can then be found by requiring the resulting distribution $\overline{\pi}^{(\infty)}$ to sum up to 1.

2.8 Numerical experiments

In this section we present numerical results to illustrate the influence of a varying service rate on the performance of (elastic) customers. We are particularly interested in the effect of fast and slow fluctuations of the service rate, which we studied extensively in Section 2.6. We use the variant of the model of Section 1.5 presented by Núñez Queija and Boxma [86]. This model was described in some detail in the introduction of this chapter and in Section 2.2. It arises as a special case of the infinite-queue model in the following way. The customers in the general model of Section 2.1 are *elastic* customers, and in this section we denote the queue length at time t by $X_\epsilon^{(\mathrm{e})}(t) = X_\epsilon^{(\infty)}(t)$. To avoid confusion, we note that here we replace the superscript $^{(\infty)}$ — in the numerical experiments we only consider an infinite queue for elastic customers — by the superscript $^{(\mathrm{e})}$. The number of stream customers $X_\epsilon^{(\mathrm{s})}(t)$ determines the state of the server: $Y_\epsilon(t) = X_\epsilon^{(\mathrm{s})}(t) + 1$. Here it will be more convenient to work with the number of stream customers than with the state of the server. In steady state we use $X_\epsilon^{(\mathrm{e})}$ and $X_\epsilon^{(\mathrm{s})}$ instaed of $X_\epsilon^{(\mathrm{e})}(t)$ and $X_\epsilon^{(\mathrm{s})}(t)$. As in Section 2.6, $\epsilon > 0$ is a time scale parameter. Stream customers arrive according to a Poisson process with rate $\epsilon\lambda^{(\mathrm{s})}$. Each stream customer requires a fixed capacity $r^{(\mathrm{s})}$ for the total duration of the holding time, which is exponentially distributed with mean $h^{(\mathrm{s})}/\epsilon$. In [86] $r^{(\mathrm{s})}$ is equal to the capacity of one server. We further denote the work *offered* to the system by stream customers by $\rho^{(\mathrm{s})} := \epsilon\lambda^{(\mathrm{s})} \times h^{(\mathrm{s})}/\epsilon = \lambda^{(\mathrm{s})} h^{(\mathrm{s})}$, which obviously is independent of ϵ. In the numerical experiments we assume that the total service capacity of the system C is available for stream customers, that stream customers have preemptive resume priority over elastic customers, and that there is no waiting room for stream customers. We further assume that C is a multiple of $r^{(\mathrm{s})}$: $C = L^{(\mathrm{s})} r^{(\mathrm{s})}$. Here $L^{(\mathrm{s})}$ is the maximum number of stream customers in the system. Note that $N = L^{(\mathrm{s})} + 1$, since the number of stream customers can vary from 0 to $L^{(\mathrm{s})}$. Furthermore, the (marginal) process $X_\epsilon^{(\mathrm{s})}(t)$ evolves as the queue-length process of the M/M/$L^{(\mathrm{s})}$ queue. The departure rate of stream customers is $q_{j+1}^- = j\epsilon/h^{(\mathrm{s})}$ when $X_\epsilon^{(\mathrm{s})}(t) = Y_\epsilon(t) - 1 = j$, $j = 1, 2, \ldots, L^{(\mathrm{s})}$, and the arrival rate is $q_{j+1}^+ = \epsilon\lambda^{(\mathrm{s})}$, $j = 0, 1, \ldots, L^{(\mathrm{s})} - 1$. The steady-state probabilities of the number of stream customers (the state of the server) given by Expression (2.1) indeed become the well known truncated Poisson distribution of Erlang's loss system in equilibrium:

$$\mathbf{P}\left\{X_\epsilon^{(\mathrm{s})} = j\right\} = p_{j+1} = \frac{(\rho^{(\mathrm{s})})^j/j!}{\sum_{k=0}^{L^{(\mathrm{s})}} (\rho^{(\mathrm{s})})^k/k!}, \quad j = 0, 1, \ldots, L^{(\mathrm{s})}.$$

The mean service requirement of elastic customers is $f^{(\mathrm{e})} = 1/\mu$ and their arrival rate is independent of the number of stream customers, i.e., $\lambda_i = \lambda^{(\mathrm{e})}$, for all i. The work load of elastic customers (previously ρ) is denoted by $\rho^{(\mathrm{e})}$. An individual elastic customer does not require a minimum service capacity ($r_-^{(\mathrm{e})} = 0$), and can be served at any positive rate ($r_+^{(\mathrm{e})} = \infty$). Hence, with j stream customers and at least 1 elastic customer in the system, the (total) service rate of elastic customers (c_{j+1}) is $C - jr^{(\mathrm{s})}$.

For normalisation purposes we choose $\mu = 1$ and $h^{(\mathrm{s})} = 1$ (this is no restriction). These values are such that stream customers and elastic customers have the same mean service requirement for $\epsilon = 1$. In all cases we take $L^{(\mathrm{s})} = 17$. In our first experiment we choose $\rho^{(\mathrm{s})} = 10$. We note that $\rho^{(\mathrm{s})}$ is not equal to the amount of work processed for stream customers, since new stream customers are rejected from the system if there are $L^{(\mathrm{s})}$ other stream customers present. The value of $\rho^{(\mathrm{s})}$ determines the steady-state distribution of the number of stream customers and, in particular, the mean service rate for elastic customers

$$c = \sum_{j=0}^{L^{(\mathrm{s})}} \left(C - jr^{(\mathrm{s})}\right) \mathbf{P}\left\{X_\epsilon^{(\mathrm{s})} = j\right\},$$

which inherits the independence of ϵ from the distribution of $X_\epsilon^{(\mathrm{s})}$.

In Figure 2.1 we have plotted for $\epsilon = \frac{1}{5}, 1$ and ∞, the mean number of elastic customers for increasing $\rho^{(\mathrm{e})}$. On the horizontal axis we indicate $\rho^{(\mathrm{e})}/c$. The lowest curve, denoted by $\epsilon = \infty$, corresponds to the ordinary M/M/1 queue with fixed service capacity c. In Figure 2.2 we have done the same for the variance of the number of elastic customers. We observe a significant performance degradation for elastic customers when the service fluctuations are very slow, particularly when $\rho^{(\mathrm{e})}$ is close to c.

Next we investigate the impact of the chosen value of $\rho^{(\mathrm{s})}$ on the observed service degradation for small ϵ. For this we consider the system for a fixed

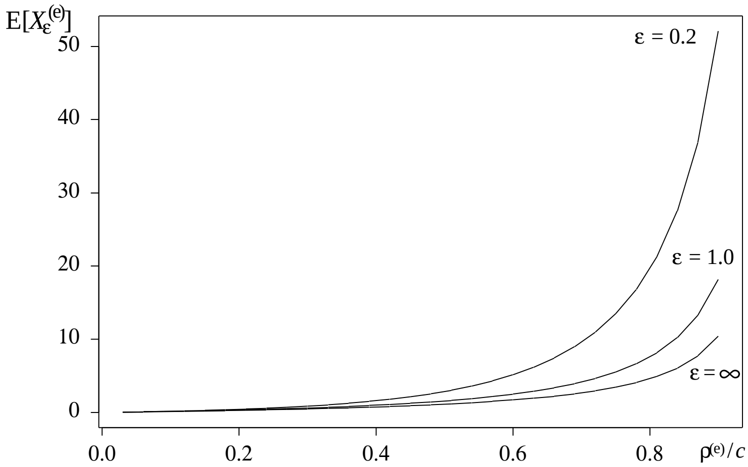

Figure 2.1: Mean queue length; $\rho^{(s)} = 10$

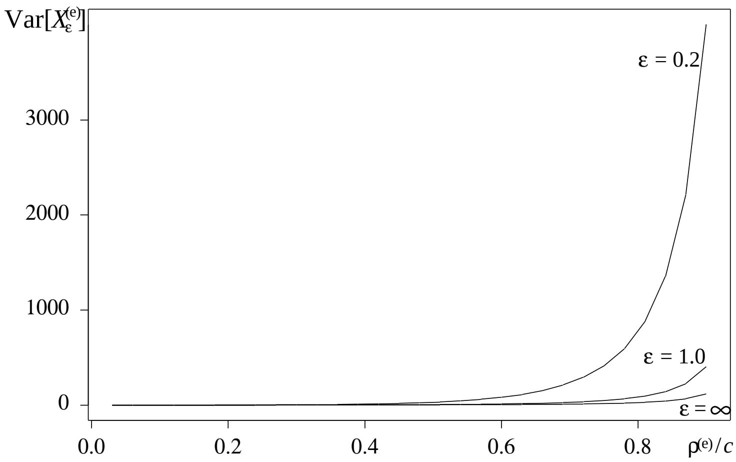

Figure 2.2: Queue-length variance; $\rho^{(s)} = 10$

amount of *carried traffic*. Clearly, since all elastic customers are eventually served, the amount of elastic traffic carried equals the amount of elastic traffic offered, which is $\rho^{(e)}$. Stream customers are not always accepted into the system. Using the definition of c, the mean capacity available to elastic customers, we may write for the mean amount of stream traffic carried per unit of time:

$$r^{(s)}\mathbf{E}\left[X_\epsilon^{(s)}\right] = C - c.$$

Note that c is completely determined by $\rho^{(s)}$, which we indicate by writing $c = c(\rho^{(s)})$. For a fixed amount of carried traffic $v := \rho^{(e)} + C - c(\rho^{(s)})$, we compute the mean and the variance of the number of elastic customers in the system for varying $\rho^{(e)}$ and $\rho^{(s)}$. We do this for $\epsilon = \frac{1}{5}, 1, 5$, and ∞. Obviously,

$$\mathbf{P}\left\{X_{\epsilon=\infty}^{(e)} = j\right\} = \left(1 - \frac{\rho^{(e)}}{c}\right)\left(\frac{\rho^{(e)}}{c}\right)^j, \quad j = 0, 1, 2, \ldots,$$

hence,

$$\mathbf{E}\left[X_{\epsilon=\infty}^{(e)}\right] = \frac{\rho^{(e)}}{C - v}, \tag{2.26}$$

$$\mathbf{Var}\left[X_{\epsilon=\infty}^{(e)}\right] = \frac{\rho^{(e)}}{C - v}\left(1 + \frac{\rho^{(e)}}{C - v}\right). \tag{2.27}$$

In Figures 2.3 and 2.4 the results are shown for $v = 0.7C$, and in Figures 2.5 and 2.6 for $v = 0.9C$. As $\rho^{(e)}$ increases from 0 to v, $\rho^{(s)}$ decreases such that at all times $\rho^{(e)} + C - c(\rho^{(s)})$ remains equal to the chosen value of v. On the horizontal axis $\rho^{(e)}$ is normalised as $\rho^{(e)}/v$. In all of the Figures 2.3 – 2.6, the top curve corresponds to the case $\epsilon = \frac{1}{5}$, the second to $\epsilon = 1$, the third to $\epsilon = 5$, and the bottom curve to $\epsilon = \infty$. In accordance with Expressions (2.26) and (2.27), in Figures 2.3 and 2.5, the bottom curve is a straight line, and in Figures 2.4 and 2.6 the bottom curve has a quadratic shape. Further note from Figures 2.4 and 2.6 that for "slow" stream traffic ($\epsilon = \frac{1}{5}$), the queue-length variance is not a monotone function of $\rho^{(e)}$. This may be explained by the fact that, as $\rho^{(e)}$ increases, a trade-off between opposite effects takes place. On one hand the load of elastic traffic increases as $\rho^{(e)}$ increases; this expectedly has an increasing effect on the variance. On the other hand the load of stream traffic (and the variability of the capacity available to elastic traffic) decreases, such that the total amount of carried traffic v remains constant. This has a decreasing effect on the variance of the number of elastic customers in the system.

We observe that the mean and the variance of the number of elastic customers in the system are particularly sensitive to ϵ when the amount of traffic carried of either type is of the same order ($\rho^{(e)} \approx C - c$) and when the system utilisation is large ($v \approx C$).

Remark 2.8.1 The numerical evaluation of this model proved to be fast and stable in many experiments over a wide range of parameter values. In general,

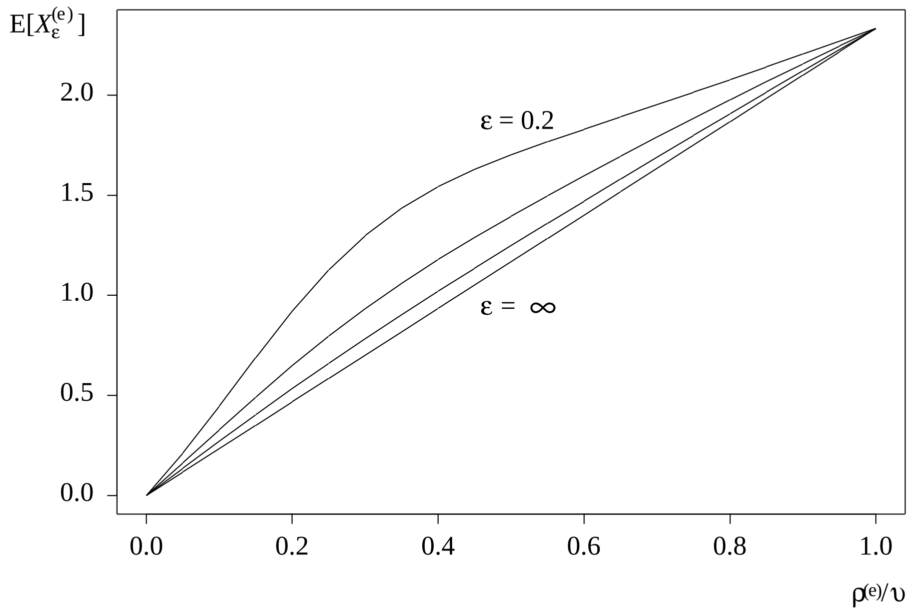

Figure 2.3: Mean queue length; $\upsilon = 0.7C$

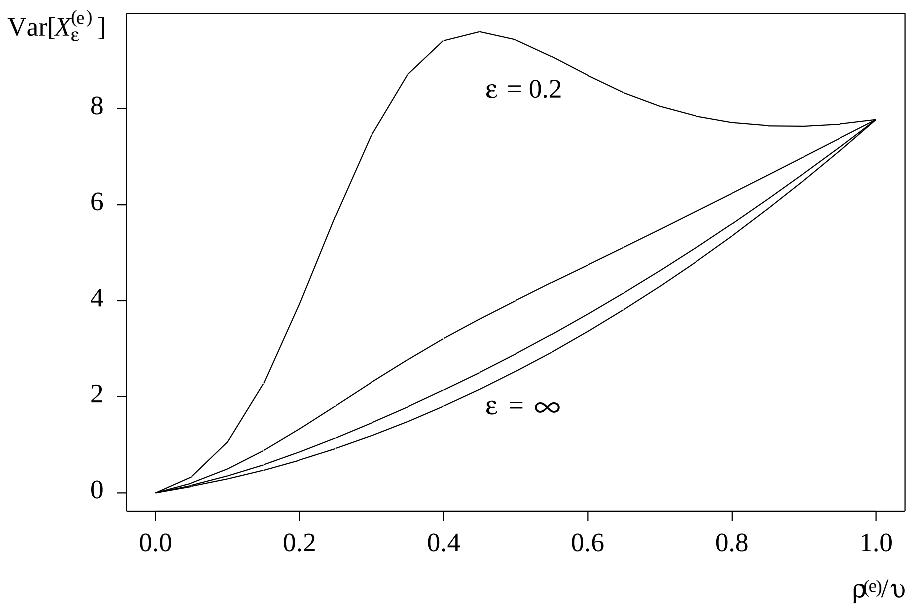

Figure 2.4: Queue-length variance; $\upsilon = 0.7C$

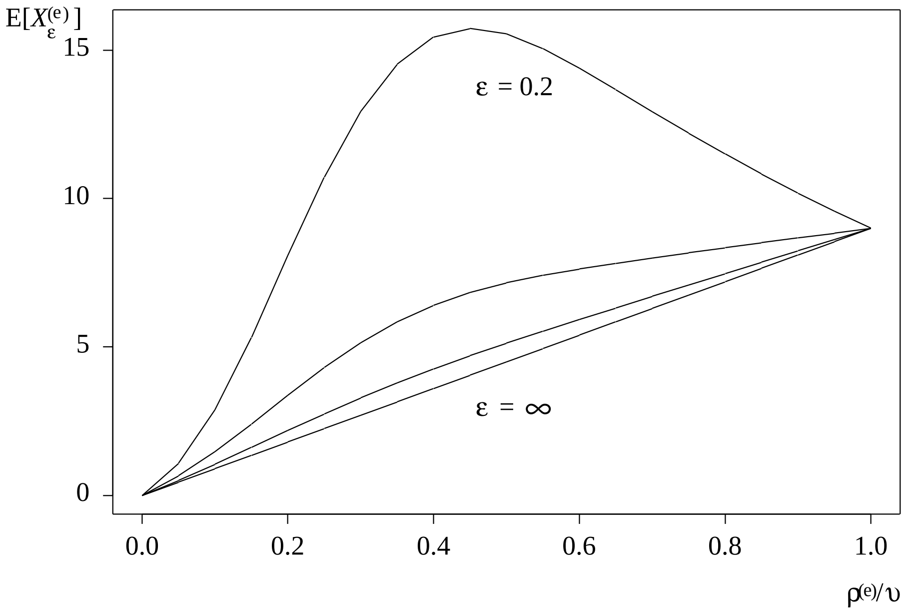

Figure 2.5: Mean queue length; $\upsilon = 0.9C$

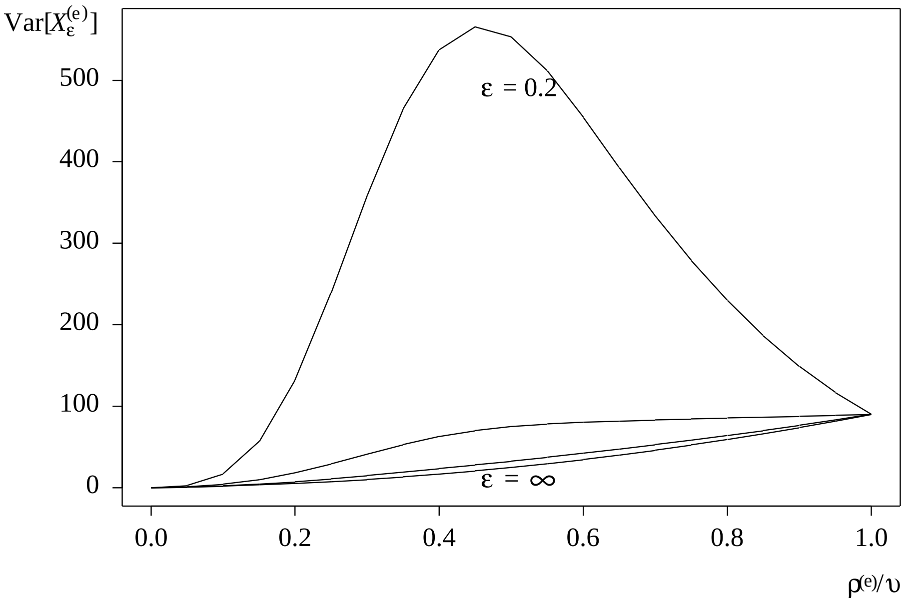

Figure 2.6: Queue-length variance; $\upsilon = 0.9C$

use of spectral methods instead of matrix-geometric routines, to find the matrix R, improves the computational efficiency. With matrix-geometric techniques the computation time increases with $\rho = \rho^{(\mathrm{e})}$, while the computational effort using spectral methods is insensitive to this parameter. Based on this observation, the use of the spectral-expansion technique is advocated by Mitrani [73], Mitrani and Chakka [75] and Haverkort and Ost [42]. However, Latouche and Ramaswami [62] show that matrix-geometric routines tend to be more robust, due to the fact that the matrix iterates in the algorithms are stochastic matrices. Thus, at each intermediate step, the matrices can be re-normalised such that the rows sum to 1. In our case, numerical evaluation is facilitated by the fact that the roots of $T(z)$ are positive real numbers. We have computed these roots using essentially a bisection search routine, and no serious problems were encountered. Finding the left nullvectors $\overline{v}_k$ and the coefficients α_k did not lead to complications, even when the dimension of $T(z)$ was of the order of several hundreds. We emphasise that the numerical stability of this straightforward procedure is due to the tri-diagonal structure of $T(z)$. For more general models this procedure may not give satisfactory results.

2.9 Concluding remarks

In this chapter we have studied a queueing model with a server that changes its service rate according to a finite birth and death process. Both the cases of an infinite and a finite maximum queue size were considered. The models constitute an important subclass of the basic model presented in Section 1.5. Under exponentiality assumptions on the service requirements and the process regulating the available service capacity (a finite birth and death process), we were able to give a detailed analysis of the distribution of the number of (elastic) customers in the system. Although the assumption of exponentially distributed service requirements is restrictive, the model is useful to gain qualitative insights into the performance of elastic traffic.

Particular attention was devoted to the effect of the varying service rate on the queue-length distribution. In numerical experiments we observed that, depending on time scale differences, ignoring the service variability and fixing the capacity at the mean, can be a very bad approximation.

From the mean queue length we also obtain the mean sojourn time (processing time) through Little's formula. In the remainder of this thesis we investigate the sojourn time in greater detail. Special attention will be given to the sojourn time conditional on the service requirement.

Appendix

2.A Proof of Lemma 2.4.4

Lemma *If $\rho < c$, then $\theta_N(z) = 0$ for some $z \in (0, 1)$. If $\rho > c$, then $\theta_N(z) = 0$ for some $z \in (1, \infty)$. If $\rho = c$, then the zero of $\theta_N(z)$ at $z = 1$ is of multiplicity*

2.

Proof First we single out the known root at $z = 1$ by factorising the determinant of $T(z)$,

$$\det[T(z)] = (1-z)g(z), \tag{2.28}$$

with $g(z)$ the determinant of the matrix obtained by replacing the last column of $T(z)$ by the sum of all columns and then dividing that column by $1-z$:

$$g(z) = \begin{vmatrix} t_1(z) & q_1^+ z & & & & \lambda_1 - \mu c_1 z \\ q_2^- z & t_2(z) & q_2^+ z & & & \lambda_2 - \mu c_2 z \\ & \ddots & \ddots & \ddots & & \vdots \\ & & \ddots & \ddots & \ddots & \vdots \\ & & & q_{N-1}^- z & t_{N-1}(z) & \lambda_{N-1} - \mu c_{N-1} z \\ & & & & q_N^- z & \lambda_N - \mu c_N z \end{vmatrix}.$$

All non-specified entries are zero. We evaluate $g(1)$ by manipulating the above matrix evaluated at $z = 1$. First add to each column, except for the first and the last one, all columns to the left of it. We now have:

$$\begin{aligned} g(1) &= \begin{vmatrix} -q_1^+ & 0 & & & & \lambda_1 - \mu c_1 \\ q_2^- & -q_2^+ & 0 & & & \lambda_2 - \mu c_2 \\ & \ddots & \ddots & \ddots & & \vdots \\ & & q_{N-2}^- & -q_{N-2}^+ & 0 & \vdots \\ & & & q_{N-1}^- & -q_{N-1}^+ & \lambda_{N-1} - \mu c_{N-1} \\ & & & & q_N^- & \lambda_N - \mu c_N \end{vmatrix} \\ &= \sum_{i=1}^{N} (-1)^{N+i} (\lambda_i - \mu c_i) \prod_{k=1}^{i-1} (-q_k^+) \prod_{k=i+1}^{N} q_k^-, \end{aligned}$$

with the empty product being equal to 1. The last equality follows by expanding the determinant with respect to its last column. Using the probabilities p_i given by Expression (2.1) we rewrite this to

$$\begin{aligned} g(1) &= (-1)^{N-1} \sum_{i=1}^{N} (\lambda_i - \mu c_i) \frac{p_i}{p_1} \prod_{k=2}^{N} q_k^- \\ &= (-1)^{N-1} \frac{\prod_{k=2}^{N} q_k^-}{p_1} \left(\sum_{i=1}^{N} p_i \lambda_i - \sum_{i=1}^{N} p_i \mu c_i \right). \end{aligned}$$

If $\rho < c$ then $\mathrm{sign}[g(1)] = (-1)^N$. By using $\det[T(z)] = \prod_{k=1}^{N} \theta_k(z)$, Equation (2.28), and Lemmas 2.4.2 and 2.4.3 it follows that $\theta_N(z) < 0$ for z in some left neighbourhood of 1, i.e., there is an $\varepsilon > 0$ such that $\theta_N(z) < 0$ for $z \in (1-\varepsilon, 1)$.

Therefore $\theta_N(z)$ must cross the horizontal axis somewhere in the interval (0, 1). Similarly, if $\rho > c$ then $\theta_N(z) < 0$ in a right neighbourhood of 1 and, hence, $\theta_N(z)$ must cross the horizontal axis somewhere in the interval $(1, \infty)$ (see the proof of Lemma 2.4.3). Finally, if $\rho = c$ then $\lim_{z\to 1} \frac{\theta_N(z)}{1-z} = 0$. □

2.B Proof of Lemma 2.6.4

Lemma *The limits $\psi_{0,k} := \lim_{\epsilon\downarrow 0} \psi_{\epsilon,k}$ and $\overline{v}_{0,k} := \lim_{\epsilon\downarrow 0} \overline{v}_{\epsilon,k}$ exist for $k = 1, 2, \ldots, 2N$. For $k \in \{1, 2, \ldots, N\}$, it holds that*

$$\begin{aligned} \psi_{0,k} &= \min\left\{\frac{\rho_{\sigma(k)}}{c_{\sigma(k)}}, 1\right\}, \\ \psi_{0,k+N} &= \max\left\{\frac{\rho_{\sigma(k)}}{c_{\sigma(k)}}, 1\right\}. \end{aligned}$$

For $k \in \{1, 2, \ldots, N\}$, if $\rho_{\sigma(k)} \le c_{\sigma(k)}$ then $\overline{v}_{0,k}\left[\Lambda - \frac{\rho_{\sigma(k)}}{c_{\sigma(k)}} M\right] = \overline{0}$, otherwise

$$\overline{v}_{0,k}\left(\kappa_k\left[\Lambda - M\right] + Q^{(Y)}\right) = \overline{0},$$

where $0 \le \kappa_N < \kappa_{N-1} < \ldots < \kappa_{N-l_2+1} < \infty$ are all the l_2 non-negative and finite nullvalues of the matrix polynomial $\kappa\left[\Lambda - M\right] + Q^{(Y)}$.

For $k \in \{N+1, \ldots, 2N\}$, if $\rho_{\sigma(k-N)} \ge c_{\sigma(k-N)}$ then $\overline{v}_{0,k}\left[\Lambda - \frac{\rho_{\sigma(k-N)}}{c_{\sigma(k-N)}} M\right] = \overline{0}$, otherwise

$$\overline{v}_{0,k}\left(\kappa_k\left[\Lambda - M\right] + Q^{(Y)}\right) = \overline{0},$$

where $0 \ge \kappa_{N+1} > \kappa_{N+2} > \ldots > \kappa_{N+l_1} > -\infty$ are all the l_1 non-positive and finite nullvalues of the matrix polynomial $\kappa\left[\Lambda - M\right] + Q^{(Y)}$.

The matrices $V_0 := \lim_{\epsilon\downarrow 0} V_\epsilon$ and $W_0 := \lim_{\epsilon\downarrow 0} W_\epsilon$ are non-singular. It further holds that if $\rho_{\sigma(k)} = c_{\sigma(k)}$ then $\overline{v}_{0,k} = \overline{v}_{0,N+k}$.

Proof Under Assumption 2.6.2 the continuity of the nullvalues, as functions of ϵ, can be extended to the point $\epsilon = 0$, since in that case the coefficient of the leading term of $\det\left[T_\epsilon(z)\right]$ does not vanish at $\epsilon = 0$. The continuity at $\epsilon = 0$ implies the limits of the nullvalues. The existence of the limits of the corresponding left nullvectors then follows from the construction of these in the proof of Lemma 2.6.2: each entry of a left nullvector is a bounded rational function of the corresponding nullvalue. From that construction of the left nullvectors we also directly find $\overline{v}_{0,k}$ when $\rho_{\sigma(k)} < c_{\sigma(k)}$, and $\overline{v}_{0,k+N}$ when $\rho_{\sigma(k)} > c_{\sigma(k)}$. The nullvalues corresponding to the remaining left nullvector tend to 1 as $\epsilon \downarrow 0$. Suppose $\psi_{\epsilon,i} \to 1$ as $\epsilon \downarrow 0$. Then $\lim_{\epsilon\downarrow 0} T_\epsilon(\psi_{\epsilon,i})$ is a matrix with all entries equal to 0. Therefore the argument above does not apply. We consider the matrix $\kappa\left[\Lambda - \delta M\right] + \delta Q^{(Y)}$ as a polynomial in $\kappa \in \mathbb{R}$ with parameter

$\delta \in (0,\infty)$. For fixed δ smaller than but close enough to 1 (such that $\lambda_i \neq \delta\mu c_i$ for all i), the matrix polynomial $\kappa\,[\Lambda - \delta M] + Q^{(Y)}$ has nullvalues

$$\kappa_{N+l_1}(\delta) < \ldots < \kappa_{N+1}(\delta) \leq 0 \leq \kappa_N(\delta) < \ldots < \kappa_{l_1+1}(\delta).$$

For fixed δ larger than but close to 1, the nullvalues are

$$\kappa_{2N-l_2}(\delta) < \ldots < \kappa_{N+1}(\delta) \leq 0 \leq \kappa_N(\delta) < \ldots < \kappa_{N-l_2+1}(\delta).$$

For $\delta = 1$ we have

$$\kappa_{N+l_1}(1) < \ldots < \kappa_{N+1}(1) \leq 0 \leq \kappa_N(1) < \ldots < \kappa_{N-l_2+1}(1),$$

and we define $\kappa_{N-l_2}(1) = \ldots = \kappa_{l_1+1}(1) = +\infty$ as the limits of the corresponding nullvalues as $\delta \uparrow 1$, and $\kappa_{2N-l_2}(1) = \ldots = \kappa_{N+l_1+1}(1) = -\infty$ as the limits of the corresponding nullvalues as $\delta \downarrow 1$. In all cases $\kappa_N(\delta) < 0$ if and only if $\rho < c$, and $\kappa_{N+1}(\delta) > 0$ if and only if $\rho > c$.

Note that if $\psi_{\epsilon,i} \to 1$ as $\epsilon \downarrow 0$, then

$$\kappa_i(\psi_{\epsilon,i}) = \frac{1-\psi_{\epsilon,i}}{\epsilon},$$

and $\overline{v}_{\epsilon,i}$ is the left nullvector of the matrix $\kappa_i(\psi_{\epsilon,i})\,[\Lambda - \psi_{\epsilon,i}M] + \psi_{\epsilon,i}Q^{(Y)}$. The desired limit of $\overline{v}_{\epsilon,i}$ now follows from continuity arguments. Note that if $\rho_{\sigma(i)} = c_{\sigma(i)}$ then $\kappa_i(1) = +\infty$ and $\kappa_{N+i}(1) = -\infty$. The corresponding limiting nullvectors are indeed nullvectors of $\Lambda - M$.

The non-singularity of the matrix V_0 is a consequence of the fact that it contains the full set of left eigenvectors of a matrix R_0. The same argument applies for W_0 with S_0 instead of R_0.

Finally to see that if $\rho_{\sigma(k)} = c_{\sigma(k)}$ then $\overline{v}_{0,k} = \overline{v}_{0,N+k}$, we note that both are equal to the unique limit of the left nullvector of the matrix $\frac{1}{\delta}\Lambda - M + \frac{1}{\kappa_k(\delta)}Q^{(Y)}$ as $\delta \to 1$ (from the left or from the right, respectively). Note that $\lim_{\delta\to 1}\frac{1}{\kappa_k(\delta)} = 0$. □

2.C Proof of Corollary 2.6.6

Corollary *Let U_0 be the non-singular matrix with its successive rows equal to $\overline{v}_{0,i}$, $i \in \{1,2,\ldots,N-l_2\} \cup \{2N-l_2+1, 2N-l_2+2,\ldots,2N\}$. For $k = 1,2,\ldots,N$, it holds that*

(i) if $\rho_{\sigma(k)} < c_{\sigma(k)}$ then

$$\lim_{\epsilon\downarrow 0}\beta_{\epsilon,k} = p_{\sigma(k)}\frac{1-\rho_{\sigma(k)}/c_{\sigma(k)}}{1-\left(\rho_{\sigma(k)}/c_{\sigma(k)}\right)^{L+1}}\gamma_k, \quad \lim_{\epsilon\downarrow 0}\beta_{\epsilon,N+k} = 0,$$

(ii) if $\rho_{\sigma(k)} > c_{\sigma(k)}$ *then*

$$\lim_{\epsilon \downarrow 0} \beta_{\epsilon,N+k} = p_{\sigma(k)} \frac{1 - \left(c_{\sigma(k)}/\rho_{\sigma(k)}\right)}{1 - \left(c_{\sigma(k)}/\rho_{\sigma(k)}\right)^{L+1}} \gamma_k, \quad \lim_{\epsilon \downarrow 0} \beta_{\epsilon,k} = 0,$$

(iii) if $\rho_{\sigma(k)} = c_{\sigma(k)}$ *then*

$$\lim_{\epsilon \downarrow 0} \beta_{\epsilon,k} + \beta_{\epsilon,N+k} = p_{\sigma(k)} \frac{1}{L+1} \gamma_k,$$

where

$$\gamma_k := \sum_{i : \frac{\rho_{\sigma(i)}}{c_{\sigma(i)}} = \frac{\rho_{\sigma(k)}}{c_{\sigma(k)}}} \left[U_0^{-1}\right]_{i,k}.$$

Proof Note that since $\rho \neq c$, for all $j = 0, 1, \ldots, L$,

$$(0,1) \ni \overline{\pi}^{(L)}_{\epsilon,j} = \sum_{k=1}^{N} \beta_{\epsilon,k} \left(\psi_{\epsilon,k}\right)^j \overline{v}_{\epsilon,k} + \sum_{k=N+1}^{2N} \beta_{\epsilon,k} \left(\frac{1}{\psi_{\epsilon,k}}\right)^{L-j} \overline{v}_{\epsilon,k}, \tag{2.29}$$

and after summing over j we have, $\forall \epsilon > 0$,

$$\overline{p} = \sum_{k=1}^{N} \beta_{\epsilon,k} \frac{1 - \left(\psi_{\epsilon,k}\right)^{L+1}}{1 - \psi_{\epsilon,k}} \overline{v}_{\epsilon,k} + \sum_{k=N+1}^{2N} \beta_{\epsilon,k} \frac{1 - \left(\frac{1}{\psi_{\epsilon,k}}\right)^{L+1}}{1 - \frac{1}{\psi_{\epsilon,k}}} \overline{v}_{\epsilon,k}, \tag{2.30}$$

with $\frac{1-x^{L+1}}{1-x} := L+1$ when $x = 1$.

First we prove the limits under the assumption that the coefficients $\beta_{\epsilon,i}$, $i = 1, 2, \ldots, 2N$, remain bounded as $\epsilon \downarrow 0$. Below we justify this assumption. Consider row i of the matrix in the left-hand side of Equation (2.23) and let $\epsilon \downarrow 0$. If $i \in \{1, 2, \ldots, N\}$ and $\rho_{\sigma(i)} \leq c_{\sigma(i)}$, or $i \in \{N+1, N+2, \ldots, 2N\}$ and $\rho_{\sigma(i)} \geq c_{\sigma(i)}$ then row i vanishes. Under the boundedness assumption we obtain from Equation (2.23)

$$\lim_{\epsilon \downarrow 0} \sum_{k=N+1-l_2}^{N+l_1} \beta_{\epsilon,k} \overline{v}_{0,k} \left[M - \Lambda\right] = \overline{0}$$

where, as before, $l_1 := \#\{i : \rho_i < c_i\}$ and $l_2 := \#\{i : \rho_i > c_i\}$. Recall that $\overline{v}_{0,k}$, for $k \in \{N+1-l_2, N+2-l_2, \ldots, N+l_1\}$, are the left nullvectors of the matrix polynomial $\kappa\left[\Lambda - M\right] + Q^{(Y)}$ corresponding to the finite nullvalues, see Lemma 2.6.4. No non-zero combination of these can be a left nullvector of $\Lambda - M$, since the latter correspond to the nullvalue $\kappa = \infty$, either $+\infty$ or $-\infty$. Therefore $\lim_{\epsilon \downarrow 0} \beta_{\epsilon,k} = 0$ for $k \in \{N+1-l_2, N+2-l_2, \ldots, N+l_1\}$. Again using Lemma 2.6.4, and in particular that if $\rho_{\sigma(k)} = c_{\sigma(k)}$ then $\overline{v}_{0,k} = \overline{v}_{0,N+k}$, we have that the $\overline{v}_{0,k}$, for $k \in \{1, 2, \ldots, N-l_2\} \cup \{2N-l_2+1, 2N-l_2+2, \ldots, 2N\}$ constitute a basis for $\mathbb{R}^N$. Therefore we may write $\overline{p}$ as a combination of these, and from Equation (2.30) we note that this uniquely determines $\lim_{\epsilon\downarrow 0}\beta_{\epsilon,k}$ for $k \in \{1,2,\ldots,l_1\} \cup \{2N-l_2+1, 2N-l_2+2,\ldots,2N\}$, and $\lim_{\epsilon\downarrow 0}\beta_{\epsilon,k}+\beta_{\epsilon,N+k}$ for $k \in \{l_1+1, l_1+2,\ldots,N-l_2\}$. The precise form of these limits can be obtained from Equation (2.30) by elementary linear algebra.

It remains to be shown that the coefficients $\beta_{\epsilon,i}$, $i = 1,2,\ldots,2N$, remain bounded as $\epsilon \downarrow 0$. Suppose this is not the case. Then we can select a sequence $\epsilon_n > 0$, $n \in \{1,2,\ldots\}$, such that $\lim_{n\to\infty}\epsilon_n = 0$ and for some $k \in \{1,2,\ldots,2N\}$, $\lim_{n\to\infty}\beta_{\epsilon_n,k}$ is equal to $+\infty$ or $-\infty$, and $|\beta_{\epsilon_n,i}| \leq |\beta_{\epsilon_n,k}|$, when n is large enough, for all i. Note that because of Equation (2.30) it must be true that $\beta_{\epsilon_n,k} \neq 0$. We assume $k \in \{1,2,\ldots,N\}$; the alternative case $k \in \{N+1, N+2,\ldots,2N\}$ can be treated by analogous arguments. Divide both sides of Equation (2.23) by $\beta_{\epsilon,k}$. Note that the right-hand side is unaffected. Substitute ϵ_n for ϵ and consider row i of the matrix in the left-hand side of that equation as $n \to \infty$. Since $\beta_{\epsilon_n,i}/\beta_{\epsilon_n,k}$ is bounded for all i, we conclude, using the same arguments as above, that $\lim_{\epsilon\downarrow 0}\beta_{\epsilon_n,i}/\beta_{\epsilon_n,k} = 0$ for $i = N+1-l_2, N+2-l_2,\ldots,N+l_1$. In particular note that k can not be in the set $\{N+1-l_2, N+2-l_2,\ldots,N+l_1\}$. Now divide Equation (2.30) by $\beta_{\epsilon,k}$ and note that, after having substituted ϵ_n for ϵ, the left-hand side vanishes as $n \to \infty$. This uniquely determines $\lim_{n\to\infty}\beta_{\epsilon_n,i}/\beta_{\epsilon_n,k}$ for $i \in \{1,2,\ldots,l_1\} \cup \{2N-l_2+1, 2N-l_2+2,\ldots,2N\}$, and $\lim_{n\to\infty}(\beta_{\epsilon_n,i}+\beta_{\epsilon_n,N+i})/\beta_{\epsilon_n,k}$ for $i \in \{l_1+1, l_1+2,\ldots,N-l_2\}$. All these limits are equal to 0 because the left-hand side vanishes. This can only be true if $k \in \{l_1+1, l_1+2,\ldots,N-l_2\}$, i.e., if $\rho_{\sigma(k)} = c_{\sigma(k)}$. Note that, in particular, the ratio $\beta_{\epsilon_n,N+k}/\beta_{\epsilon_n,k}$ tends to -1 as $n \to \infty$. We show that is not possible either. We can not apply the same argument as above since the rows k and $N+k$ of the matrix in the left-hand side of Equation (2.23) vanish as $\epsilon \downarrow 0$.

From the proof of Lemma 2.6.4 we have that if $\rho_{\sigma(i)} > c_{\sigma(i)}$ then

$$\lim_{\epsilon\downarrow 0}\frac{1-\psi_{\epsilon,i}}{\epsilon} = \kappa_i(1) \in [0,\infty),$$

and if $\rho_{\sigma(i)} < c_{\sigma(i)}$ then

$$\lim_{\epsilon\downarrow 0}\frac{1-\psi_{\epsilon,N+i}}{\epsilon} = \kappa_{N+i}(1) \in (-\infty,0].$$

Using similar arguments we now show that if $\rho_{\sigma(i)} \leq c_{\sigma(i)}$ then

$$\lim_{\epsilon\downarrow 0}\frac{\left(\rho_{\sigma(i)}/c_{\sigma(i)} - \psi_{\epsilon,i}\right)(1-\psi_{\epsilon,i})}{\epsilon} =: K_i \in (0,\infty),$$

and if $\rho_{\sigma(i)} \geq c_{\sigma(i)}$ then

$$\lim_{\epsilon\downarrow 0}\frac{\left(\rho_{\sigma(i)}/c_{\sigma(i)} - \psi_{\epsilon,N+i}\right)(1-\psi_{\epsilon,N+i})}{\epsilon} =: K_{N+i} \in (-\infty,0).$$

To see this, we consider the matrix $K[\Lambda - \delta M] + (\delta - \delta_0)\delta Q^{(Y)}$ as a matrix polynomial in $K \in \mathbb{R}$ for fixed parameters $\delta \in \mathbb{R}$ and $\delta_0 \in \mathbb{R}$. Suppose

$\rho_{\sigma(i)} \le c_{\sigma(i)}$ and set $\delta := \rho_{\sigma(i)}/c_{\sigma(i)}$. Then, for δ close to but not equal to δ_0 the matrix polynomial $K[\Lambda - \delta M] + (\delta - \delta_0)\delta Q^{(Y)}$ has N distinct and finite nullvalues which we denote by $K_i(\delta, \delta_0)$, $i \in \{1, 2, \dots, N\}$. The indices of these nullvalues are such that it is the i^{th} largest eigenvalue of the matrix polynomial (which are all distinct for $\delta \neq 0$ and $\delta \neq \delta_0$) which is equal to zero at the point $K = K_i(\delta, \delta_0)$. Clearly,

$$\frac{K_i(\psi_{\epsilon,i}, \rho_{\sigma(i)}/c_{\sigma(i)})}{\psi_{\epsilon,i} - \rho_{\sigma(i)}/c_{\sigma(i)}} = \frac{1 - \psi_{\epsilon,i}}{\epsilon}.$$

Note that if $\rho_k = \delta_0 c_k$ then all entries in row k of the matrix polynomial have a common factor $\delta - \delta_0$. For each of these rows take this common factor out of the determinant of the matrix polynomial. The roots of the resulting (scalar) polynomial are still equal to the nullvalues $K_i(\delta, \delta_0)$. Passing $\delta \to \delta_0$ we see that the nullvalues have a finite limit. Only the nullvalue(s) for which the largest (if $\delta \uparrow \delta_0$) or the smallest (if $\delta \downarrow \delta_0$) eigenvalue is zero can have a zero limit. This gives the desired limits for $\psi_{\epsilon,i}$ in the case $\rho_{\sigma(i)} \le c_{\sigma(i)}$. Similarly the limits for $\psi_{\epsilon,N+i}$ when $\rho_{\sigma(i)} \ge c_{\sigma(i)}$ can be obtained.

We further note that the entries of $\overline{v}_{\epsilon,i}$ are rational functions of ϵ and $\psi_{\epsilon,i}$ with a proper limit as $\epsilon \downarrow 0$. Hence, also the limits

$$\lim_{\epsilon \downarrow 0} \frac{1}{\psi_{\epsilon,i} - \psi_{0,i}} \left(\overline{v}_{\epsilon,i} - \overline{v}_{0,i}\right)$$

are well defined. Moreover, these limiting vectors coincide for the indices i and $N+i$ when $\rho_{\sigma(i)} = c_{\sigma(i)}$; this can be seen in the same way as we proved the equality $\overline{v}_{0,i} = \overline{v}_{0,N+i}$ in Lemma 2.6.4.

It was already noted that if $\rho_{\sigma(k)} = c_{\sigma(k)}$ then in Equation (2.23) the entire k^{th} column of the matrix in the left-hand side vanishes as $\epsilon \downarrow 0$. From the above limits we have that if we divide this column by $\sqrt{\epsilon}$ then, if also $\rho_{\sigma(i)} = c_{\sigma(i)}$, the entries in row i and $N+i$ have a finite non-zero limit, and all other entries still vanish (for $\epsilon \downarrow 0$). We obtain

$$\lim_{n\to\infty} \sum_{i:\rho_{\sigma(i)} = c_{\sigma(i)}} \left\{ K_i \frac{\beta_{\epsilon_n,i}}{\beta_{\epsilon_n,k}} + K_{N+i} \frac{\beta_{\epsilon_n,N+i}}{\beta_{\epsilon_n,k}} \right\} \overline{v}_{0,i} M = \overline{0}.$$

This defines a non-singular system from which we conclude

$$\lim_{n\to\infty} \left\{ K_i \frac{\beta_{\epsilon_n,i}}{\beta_{\epsilon_n,k}} + K_{N+i} \frac{\beta_{\epsilon_n,N+i}}{\beta_{\epsilon_n,k}} \right\} = 0,$$

for all i such that $\rho_{\sigma(i)} = c_{\sigma(i)}$, and in particular for $i = k$. Since $K_k > 0$ and $K_{N+k} < 0$, the coefficients $\beta_{\epsilon_n,N+k}$ and $\beta_{\epsilon_n,k}$ must be of the same sign for large n, which is a contradiction with the earlier observation that their ratio tends to -1. Hence, $\beta_{\epsilon_n,k}$ must be bounded for all n. □

2.D Proof of Lemma 2.6.8

Lemma *For $k = 1, 2, \ldots, 2N$, the limits $\psi_{\infty,k} := \lim_{\epsilon\to\infty} \psi_{\epsilon,k}$ and $\overline{v}_{\infty,k} := \lim_{\epsilon\to\infty} \overline{v}_{\epsilon,k}$ exist. The limits of the nullvalues are given by $\psi_{\infty,k} = 0$, for $k = 1, 2, \ldots, N-1$, by $\psi_{\infty,k} = \infty$, for $k = N+2, N+3, \ldots, 2N$, and by*

$$\begin{aligned}\psi_{\infty,N} &= \min\left\{\frac{\rho}{c}, 1\right\},\\ \psi_{\infty,N+1} &= \max\left\{\frac{\rho}{c}, 1\right\}.\end{aligned}$$

Moreover, for $k = 1, 2, \ldots, N$,

$$\begin{aligned}&\overline{v}_{\infty,k}\left[Q^{(Y)} + \xi_k\Lambda\right] = 0,\\ &\overline{v}_{\infty,N+k}\left[Q^{(Y)} - \zeta_k M\right] = \overline{0}.\end{aligned}$$

Here $\infty \geq \xi_1 \geq \xi_2 \geq \ldots \geq \xi_N = 0$ are the nullvalues of the matrix polynomial $Q^{(Y)} + \xi\Lambda$, and $0 = \zeta_1 \geq \zeta_2 \geq \ldots \geq \zeta_N \geq -\infty$ are the nullvalues of the matrix polynomial $Q^{(Y)} - \zeta M$. In particular, $\overline{v}_{\infty,N} = \overline{v}_{\infty,N+1} = \overline{p}$. The matrices $V_\infty := \lim_{\epsilon\to\infty} V_\epsilon$ and $W_\infty := \lim_{\epsilon\to\infty} W_\epsilon$ exist and are non-singular.

Proof If $\rho < c$ then $\psi_{\epsilon,N+1} = 1$ for all $\epsilon > 0$, and if $\rho > c$ then $\psi_{\epsilon,N} = 1$ for all $\epsilon > 0$. In both cases $\overline{p}$ is the corresponding left nullvector for all $\epsilon > 0$. Thus, it remains to prove the statements for the nullvalues $\psi_{\epsilon,i}$ which are not identical to 1 for all $\epsilon > 1$, and their corresponding left nullvectors.

The limits of the nullvalues $\psi_{\epsilon,k}$ for $k \in \{1, 2, \ldots, N-1, N+2, \ldots, 2N\}$ follow from Lemma 2.6.1. All but the largest eigenvalue of $T_\epsilon(z)$ converge pointwise to a linear function in z which goes through the origin and has a negative slope. Each of these eigenvalues is continuous in z and has a root in $[0, 1)$ and one in $(1, \infty)$. Therefore, as $\epsilon \to \infty$, the root in $[0, 1)$ must go to 0 and the root in $(1, \infty)$ must go to ∞.

To investigate the limiting behaviour of the left nullvectors corresponding to the nullvalues inside the unit interval, we consider

$$\frac{\xi}{1-\delta} T_{\frac{1-\delta}{\xi\delta}}(\delta) = \xi\Lambda + Q^{(Y)} - \xi\delta M$$

as a matrix polynomial in ξ. Note that $\xi\Lambda + Q^{(Y)} - \xi\delta M$ is well defined for all $\delta \geq 0$. Mimicking the arguments in the proofs of Lemmas 2.4.3 and 2.4.4 it can be shown that if $\delta > 0$ then $\xi\Lambda + Q^{(Y)} - \xi\delta M$ has at least $N - m_0 - 1$ nullvalues for $\xi \in (0, \infty)$, at least one nullvalue at $\xi = 0$, and at least $N - n_0 - 1$ for $\xi \in (-\infty, 0)$. The remaining zero lies in $(0, \infty)$ and is single if $\rho < \delta c$, it is single and lies in $(-\infty, 0)$ if $\rho > \delta c$, and it is equal to 0 if $\rho = \delta c$. In the last case the nullvalue $\xi = 0$ is of multiplicity 2. Let these nullvalues be denoted by $\infty > \xi_{m_0+1}(\delta) \geq \ldots > \xi_{N-n_0}(\delta)$. As in Remark 2.3.5 we set $\xi_k = +\infty$, for $k = 1, \ldots, m_0$, and $\xi_k = -\infty$, for $k = N - n_0, \ldots, N$. Note that for $\epsilon > 0$ and

$i = 1, 2, \ldots, 2N$, it must be true that $\xi_i(\psi_{\epsilon,i}) = \epsilon\psi_{\epsilon,i}$. For $i = 1, 2, \ldots, N-1$, we noted above that $\lim_{\epsilon\downarrow 0}\psi_{\epsilon,i} = 0$, hence, by the continuity of $\xi_i(\delta)$ in $\delta \geq 0$,

$$\lim_{\epsilon\to\infty} \epsilon\psi_{\epsilon,i} \lim_{\epsilon\to\infty} \frac{1-\psi_{\epsilon,i}}{\xi_i(\psi_{\epsilon,i})} = \frac{1}{\xi_i(0)}.$$

Also, $\overline{v}_{\epsilon,i}$ is the left nullvector of $\xi\Lambda + Q^{(Y)} - \xi\delta M$ for $\xi = \frac{1-\psi_{\epsilon,i}}{\epsilon\psi_{\epsilon,i}}$ and $\delta = \psi_{\epsilon,i}$. Again by continuity arguments the $\overline{v}_{\infty,i}$ follow, for $i = 1, 2, \ldots, N-1$.

Suppose $\rho < c$. Then also $\psi_{\epsilon,N}$ lies in $(0, 1)$ for all $\epsilon > 0$. Now we argue that $\psi_{\epsilon,N} > \rho/c$ for all $\epsilon > 0$. From Lemma 2.6.4 it follows that this is certainly true for ϵ close enough to 0. Using the relation $\xi_N(\psi_{\epsilon,N}) = \epsilon\psi_{\epsilon,N}$, and the fact that $\xi_N(\rho/c) = 0$, we must conclude that indeed it can not be that $\psi_{\epsilon,N} = \rho/c$ for some $\epsilon > 0$. Then the desired inequality $\psi_{\epsilon,N} > \rho/c$ follows from the continuity of $\psi_{\epsilon,N}$ in ϵ. Combining the relation $\overline{v}_{\epsilon,N}T_\epsilon(\psi_{\epsilon,N}) \equiv \overline{0}$, for all $\epsilon > 0$, with the fact that $\psi_{\epsilon,N}$ can not vanish, it is clear from the construction of the left nullvectors in the proof of Lemma 2.6.2, that $\overline{v}_{\epsilon,N}$ tends to the unique left nullvector $\overline{p}$ of $Q^{(Y)}$, as $\epsilon \to \infty$. The analogue of Equation (2.8) for $\epsilon \neq 1$ is given by

$$\Lambda + R_\epsilon\left[Q_\epsilon^{(Y)} - \Lambda - M\right] + R_\epsilon^2 M = 0.$$

Pre-multiply this equation by $\overline{v}_{\epsilon,N}$, and post-multiply by $\overline{1}$, the column vector consisting only of ones. What remains is a quadratic (scalar) function in $\psi_{\epsilon,N}$:

$$\overline{v}_{\epsilon,N}\Lambda\overline{1} + \psi_{\epsilon,N}\left(\overline{v}_{\epsilon,N}\left[Q_\epsilon^{(Y)} - \Lambda - M\right]\overline{1}\right) + \psi_{\epsilon,N}^2\left(\overline{v}_{\epsilon,N}M\overline{1}\right) = 0.$$

Since $Q_\epsilon^{(Y)}\overline{1}$ is the transpose of $\overline{0}$, $\psi_{\epsilon,N} = 1$ is a solution of the above equation. However, $\psi_{\epsilon,N} < 1$ for all $\epsilon > 0$, hence,

$$\psi_{\epsilon,N} = \frac{\overline{v}_{\epsilon,N}\Lambda\overline{1}}{\overline{v}_{\epsilon,N}M\overline{1}} \longrightarrow \frac{\rho}{c},$$

as $\epsilon \to \infty$.

V_∞ is non-singular since different nullvalues of $\lim_{\xi\to\infty}\frac{1}{\xi}\left[\xi\Lambda + Q^{(Y)} - \xi\delta M\right]$ yield independent nullvectors. Furthermore, the nullvalues which are equal to $+\infty$ also have a complete set of nullvectors, since $\lim_{\xi\to\infty}\frac{1}{\xi}\left[\xi\Lambda + Q^{(Y)} - \xi\delta M\right]$ is a diagonal matrix.

The statements about the nullvalues outside the unit interval and their corresponding left nullvectors follow by analogous reasoning, using the matrix $\frac{\delta^2\zeta}{\delta-1}T_{\frac{\delta-1}{\delta\zeta}}(1/\delta) = \zeta\delta\Lambda + Q^{(Y)} - \zeta M$. In the proof, the role of ξ above is now played by ζ. □

2.E Proof of Corollary 2.6.10

Corollary *For $k = 1, 2, \ldots, N-1, N+2, \ldots, 2N$, it holds that*

$$\lim_{\epsilon\to\infty}\beta_{\epsilon,k} = 0.$$

If $\rho < c$ *then* $\lim_{\epsilon\to\infty} \beta_{\epsilon,N+1} = 0$ *and*

$$\lim_{\epsilon\to\infty} \beta_{\epsilon,N} = \frac{1-\rho/c}{1-(\rho/c)^{L+1}}.$$

If $\rho > c$ *then* $\lim_{\epsilon\to\infty} \beta_{\epsilon,N} = 0$ *and*

$$\lim_{\epsilon\to\infty} \beta_{\epsilon,N+1} = \frac{1-\rho/c}{1-(\rho/c)^{L+1}}.$$

Proof We assume that $\lim_{\epsilon\to\infty} \beta_{\epsilon,i}$ exists (possibly equal to $+\infty$ or $-\infty$), for all $i = 1, 2, \ldots, 2N$. This assumption can be dropped by considering appropriate sequences of ϵ_n, $n \in \{1, 2, \ldots\}$, that tend to $+\infty$ as $n \to \infty$, as we did in the proof of Corollary 2.6.6. Select $k \in \{1, 2, \ldots, 2N\}$ such that $\beta_{\epsilon,i}/\beta\epsilon, k$ remains bounded, as $\epsilon \to \infty$, for all $i \in \{1, 2, \ldots, 2N\}$. Such a k always exists (but is not necessarily unique). Assume that $k \in \{1, 2, \ldots, N\}$; the alternative case $k \in \{N+1, N+2, \ldots, 2N\}$ can be treated by analogous arguments.

Note that in Equation (2.23), if we divide the matrix in the left-hand side by ϵ and then let $\epsilon \to \infty$, it becomes

$$\begin{bmatrix} V_\infty Q^{(Y)} & 0 \\ 0 & W_\infty Q^{(Y)} \end{bmatrix},$$

where the (block) entries denoted by 0 represent matrices with all entries equal to zero. Here we use Lemma 2.6.8. In particular, the only non-zero entries of the matrices Ψ_∞ and Φ_∞ are $[\Psi_\infty]_{N,N} = \psi_{\infty,N}$ and $[\Phi_\infty]_{1,1} = 1/\psi_{\infty,N+1}$, respectively. Furthermore, the last row of V_∞ and the first row of W_∞ are equal to $\overline{p}$, which is the left nullvector of $Q^{(Y)}$. Now, if we divide Equation (2.23) by $\epsilon\beta_{\epsilon,k}$ and let $\epsilon \to \infty$, we conclude that $k = N$. This follows from the fact that in the above matrix all rows $i \in \{1, 2, \ldots, N-1\}$ have non-zero entries, and all entries in row N are zero. Similarly, if $k \in \{N+1, N+2, \ldots, 2N\}$ then $k = N+1$.

Now we show that $\beta_{\epsilon,N}$ and $\beta_{\epsilon,N+1}$ must be bounded as $\epsilon \to \infty$. Suppose $\beta_{\epsilon,N}$ is unbounded for large ϵ. If we divide Equation (2.29) by $\beta_{\epsilon,N}$ and let $\epsilon \to \infty$ we conclude that

$$\lim_{\epsilon\to\infty} \frac{\beta_{\epsilon,N+1}}{\beta_{\epsilon,N}} = -(\frac{\rho}{c})^j, \quad j = 0, 1, 2, \ldots, L.$$

This can not be true since $\rho \neq c$. We conclude that all coefficients $\beta_{\epsilon,i}$, $i = 1, 2, \ldots, 2N$ remain bounded as $\epsilon \to \infty$.

Using the boundedness of the coefficients and letting $\epsilon \to \infty$ in Equation (2.22) yields $\beta_{\epsilon,k} \to 0$, $k \in \{1, \ldots, N-1, N+2, \ldots, 2N\}$. Finally, we use the equality $(\overline{\pi}_\epsilon^{(L)})_j \Lambda \overline{1} = (\overline{\pi}_\epsilon^{(L)})_{j+1} M \overline{1}$, or equivalently,

$$\begin{aligned} &\sum_{k=1}^{N} \beta_{\epsilon,k} (\psi_{\epsilon,k})^j (\overline{v}_{\epsilon,k} \Lambda \overline{1}) + \sum_{k=N+1}^{2N} \beta_{\epsilon,k} (1/\psi_{\epsilon,k})^{L-j} (\overline{v}_{\epsilon,k} \Lambda \overline{1}) \\ &= \sum_{k=1}^{N} \beta_{\epsilon,k} (\psi_{\epsilon,k})^{j+1} (\overline{v}_{\epsilon,k} M \overline{1}) + \sum_{k=N+1}^{2N} \beta_{\epsilon,k} (1/\psi_{\epsilon,k})^{L-j-1} (\overline{v}_{\epsilon,k} M \overline{1}), \end{aligned}$$

for $j = 0, 1, \ldots, L$. These equations result from summing the balance equations — in $\overline{\pi}_\epsilon^{(L)} \mathcal{Q}_\epsilon = \overline{0}$ — corresponding to all levels $l = 0, 1, 2, \ldots, j$. For $j = 0, 1, \ldots, L-1$, we have that

$$\left(\beta_{\epsilon,N}\left(\min\left\{\frac{\rho}{c},1\right\}\right)^{j+1}\sum_{k=1}^{N} p_k\mu c_k + \beta_{\epsilon,N+1}\left(1/\max\left\{\frac{\rho}{c},1\right\}\right)^{L-j-1}\sum_{k=1}^{N} p_k\mu c_k\right)$$
$$-\left(\beta_{\epsilon,N}\left(\min\left\{\frac{\rho}{c},1\right\}\right)^{j}\sum_{k=1}^{N} p_k\lambda_k + \beta_{\epsilon,N+1}\left(1/\max\left\{\frac{\rho}{c},1\right\}\right)^{L-j}\sum_{k=1}^{N} p_k\lambda_k\right)$$

tends to 0, as $\epsilon \to \infty$. If $\rho < c$ this leads to $\lim_{\epsilon\to\infty} \beta_{\infty,N+1}(\sum_{k=1}^{N} p_k\mu c_k)(1 - \rho/c) = 0$, and hence, $\beta_{\epsilon,N+1} \to 0$, as $\epsilon \to \infty$. From Equation (2.30) we then obtain $\lim_{\epsilon\to\infty} \beta_{\epsilon,N}$. The limits of $\beta_{\epsilon,N}$ and $\beta_{\epsilon,N+1}$ in case $\rho > c$ are found similarly. $\square$

Chapter 3

Sojourn times in the case of service interruptions

In the previous chapter we analysed the queue-length distribution of a large class of processor-sharing queues with varying service capacity. In the remainder of the thesis we study the *sojourn times* of customers in such queueing models. In this chapter we start by presenting the analysis of Núñez Queija [83], where the extreme case is considered that the available capacity alternates between a positive value and zero. In Chapter 4 we consider a more general structure of the service fluctuations. A model of which the service capacity alternates between a positive value and zero is obtained by taking $N = 2$, $c_1 > 0$, and $c_2 = 0$ in the models described in Section 2.1. Periods during which service is available will be called on-periods (or availability periods). When no service is available we speak of service interruptions, breakdowns, or off-periods. In the above mentioned example from the models of Section 2.1 (with $N = 2$, $c_1 > 0$, and $c_2 = 0$) the lengths of the on- and off-periods are exponentially distributed with means $1/q_1$ and $1/q_2$, respectively. In this chapter we study the more general case where the off-periods may have an arbitrary distribution. However, we do require that the service requirements of customers and the lengths of the on-periods are exponentially distributed. The case of generally distributed service requirements is studied in Chapter 5. We only consider the case where there is no restriction on the number of customers in the queue. Thus, we are dealing with an M/M/1 processor-sharing queue with a server that is subject to breakdowns, with exponentially distributed availability periods and generally distributed breakdowns.

Remark 3.0.1 Because of the generally distributed off-periods, the model also includes the following cases of the infinite-queue model described in Section 2.1:

- $c_i > 0$ for some $i \in \{1, 2, \ldots, N\}$ and $c_j = 0$ for all $j \neq i$,
- $c_1 = c_N > 0$, $c_j = 0$ for all $j \in \{2, 3, \ldots, N-1\}$, and $q_i^+ = q_{N+1-i}^-$ for all $i \in \{1, 2, \ldots, N-1\}$.

The assumption of exponentially distributed service requirements may be relaxed for some parts of our analysis. This will be done in Chapter 5. For instance, the decomposition result in Section 3.2 may be obtained for generally distributed service requirements (see Theorem 5.3.2). Also, the results of Sections 3.3 and 3.4 may be generalised for that case, using the Laplace Transform method for solving differential equations (see Section 5.4). In this chapter the reasons for presenting the results for exponentially distributed service requirements are twofold. Firstly, the fundamental ideas are the same as for general service requirements, while the presentation is more transparent. Secondly, some of the closed-form results obtained under the exponentiality assumption for service requirements — which allow to carry the analysis further — could not be extended to general service requirement distributions.

Queueing models with unreliable servers have received much attention in the literature for the case that the service discipline is FCFS (First Come First Served). The first ones to consider queueing models with service interruptions (and their connection with priority models) were White and Christie [116]. Gaver [37] obtained the steady-state queue-length distribution of the $\mathrm{M}^X/\mathrm{G}/1$ queue with exponentially distributed on-times and generally distributed off-times. We further mention the early work of Mitrani and Avi-Itzhak [74] on a queueing model with multiple servers which are subject to breakdowns, and the work of Neuts [81, Chapter 6] concerned with queues in a random environment. Bounds and approximations for queue lengths and sojourn times when the on-times have a general distribution as well, are studied by Federgruen and Green [27, 28] and Sengupta [103]. Takine and Sengupta [111] consider an unreliable server with a Markovian Arrival Process, possibly dependent on the on/off process. Li et al. [66] give a transient analysis of the model of this chapter (with the FCFS service discipline) and Lee [63] studies a discrete time variant. For an extensive overview of the literature on queueing models with service interruptions we refer to Federgruen and Green [27, 28]. More recent references can be found in Takine and Sengupta [111].

To the author's knowledge, the first analytic results for *processor-sharing queues* with service interruptions — which we present in this chapter — were derived in Núñez Queija [83]. Recursive computational schemes to evaluate queue-length probabilities and the (unconditional) mean sojourn time may be found in Almási [3] and Almási and Sztrik [4], see also references therein.

The chapter is organised as follows. In Section 3.1 we describe the model and give the joint steady-state distribution of the state of the server and the number of customers in the system. In Section 3.2 we represent the sojourn time of a customer conditional on his service requirement, by a branching process. We characterise the distributions of two fundamental random variables in the branching process in Section 3.3, by deriving differential equations for the LSTs (Laplace-Stieltjes Transforms) of their distributions and then solving these in terms of a single integral equation. We derive the first two moments of the two fundamental random variables in Section 3.4, and give the general form of higher moments. In Section 3.5 we use these results to obtain closed-form expressions for the first two moments of the sojourn time of a customer conditioned on his service requirement, the state of the server upon arrival and the number of other customers in the system upon arrival. In particular we extend a result of Sengupta and Jagerman [105], proving that the k^{th} moment of the conditional sojourn time is a polynomial of degree k in the number of customers upon arrival. In Section 3.6 we give the LST of the distribution of the sojourn time of a customer conditioned only on the customer's own service requirement, assuming that the system is in steady state upon arrival. In particular we see that — unlike the case without server breakdowns — the mean sojourn time of a customer is not proportional to the service requirement. The next two sections are devoted to an asymptotic analysis of the model. In Section 3.7 we study sojourn times of customers with large service requirements (tending to infinity), and in Section 3.8 we consider the heavy-traffic case. We conclude the chapter in Section 3.9 with some final remarks.

3.1 Model description

We consider a server which alternates between an *on*-state and an *off*-state. The on-periods are assumed to be exponentially distributed with mean $1/\nu$, independent of everything else. The off-periods are i.i.d. random variables (generically denoted by T_{off}) having probability distribution $F(t) := \mathbf{P}\{T_{off} \le t\}$, $t \ge 0$. The LST of this distribution will be denoted by

$$\phi(s) := \int_{t=0}^{\infty} \mathrm{e}^{-st} \mathrm{d}F(t), \qquad \mathrm{Re}(s) \ge 0,$$

and the k^{th} moment of $F(t)$ by

$$m_k := \int_{t=0}^{\infty} t^k \mathrm{d}F(t).$$

Throughout this chapter we assume that $m_1 < \infty$.

Customers arrive to the server according to a Poisson process with rate λ, requiring an exponentially distributed amount of service with mean $1/\mu$. There is room for infinitely many customers at the server. When the server is on, all customers present are simultaneously served according to the processor-sharing discipline. Thus, because of the exponentially distributed service requirements, the service of any of the customers is completed within the next Δt time units with probability $\frac{1}{n}\mu\Delta t + \mathrm{o}(\Delta t)$. During off-periods the service of all customers is interrupted until the server becomes active again.

We define the random variable $X(t)$ to be the number of customers at the server at time $t \ge 0$. The random variable $Y(t)$ is equal to 1 if at time $t \ge 0$ the server is on, and $Y(t)$ is equal to 0 otherwise. Under the ergodicity condition,

$$\rho := \frac{\lambda}{\mu} < \frac{1}{1 + \nu m_1} =: c, \tag{3.1}$$

the pair $(X(t), Y(t))$ has a non-trivial limiting distribution. In Condition (3.1) ρ is the traffic load (the average amount of work that arrives to the system per unit of time), and c is the average service capacity per unit of time, which is equal to the fraction of time that the server is available. Let (X, Y) be a pair of random variables having the limiting distribution of $(X(t), Y(t))$, under Condition (3.1). Below we show that the joint distribution of (X, Y) satisfies, for $|z| \leq 1$,

$$\mathbf{E}\left[z^X \,|Y=1\right] = \frac{\mu - \lambda(1+\nu m_1)}{\mu - \lambda z - \nu z \frac{1-\phi(\lambda(1-z))}{1-z}}, \tag{3.2}$$

$$\mathbf{E}\left[z^X \,|Y=0\right] = \frac{1-\phi(\lambda(1-z))}{m_1\lambda(1-z)}\mathbf{E}\left[z^X \,|Y=1\right], \tag{3.3}$$

and $\mathbf{P}\{Y=1\} = 1 - \mathbf{P}\{Y=0\} = c$. For later use, we give the means of the above conditional distributions:

$$\mathbf{E}[X\,|Y=1] = \frac{\rho}{c-\rho}\left(1 + (1-c)\lambda\frac{m_2}{2m_1}\right), \tag{3.4}$$

$$\mathbf{E}[X\,|Y=0] = \lambda\frac{m_2}{2m_1} + \frac{\rho}{c-\rho}\left(1 + (1-c)\lambda\frac{m_2}{2m_1}\right). \tag{3.5}$$

By deconditioning Expressions (3.2) and (3.3), we find the p.g.f. (probability generating function) of the marginal distribution of X:

$$\mathbf{E}\left[z^X\right] = c\left(1 + \nu\frac{1-\phi(\lambda(1-z))}{\lambda(1-z)}\right)\frac{\mu - \lambda(1+\nu m_1)}{\mu - \lambda z - \nu z\frac{1-\phi(\lambda(1-z))}{1-z}}. \tag{3.6}$$

In the remainder of this section we give an informal discussion of the derivation of Expressions (3.2) and (3.3). In particular, in Remark 3.1.1 we discuss the equivalence of the queue-length process in our model with the queue-length process of two queueing models with the FCFS queue discipline. Expression (3.2) can be found by considering the queue-length process only during on-periods. For this, we "delete" all off-periods and interpret the arrivals during an off-period as a batch arrival. In the resulting transformed model there are three types of events: (i) Departures of customers at rate μ when there is at least one customer present, (ii) single arrivals according to a Poisson process with rate λ, and (iii) batch arrivals according to a Poisson process with rate ν and batch sizes having p.g.f. $\phi(\lambda(1-z))$, which is the p.g.f. of the number of arrivals during an off-period. Note that batches are "empty" with probability $\phi(\lambda)$. To avoid this, we may consider only non-empty batches which arrive with rate $\nu(1-\phi(\lambda))$, having p.g.f. $\frac{\phi(\lambda(1-z))-\phi(\lambda)}{1-\phi(\lambda)}$. The balance equations for this transformed model readily lead to Equation (3.2).

The factor $\frac{1-\phi(\lambda(1-z))}{m_1\lambda(1-z)}$ in Equation (3.3) is the p.g.f. of the number of customers that arrive during the backward recurrence time of an off-period. This can be explained as follows. At an arbitrary time instant at which the server is off, the number of customers in the system is the sum of the number of customers that were at the server when the server turned off and the number of customers that have arrived since that time. The elapsed time since the server turned off is distributed as the backward recurrence time of an off-period. Moreover, because of the exponentially distributed on-periods, we may use the PASTA (Poisson Arrivals See Time Averages) property — see Wolff [119] — to show that the number of customers present when the server turns off has the same distribution as X *given* that $Y = 1$.

Remark 3.1.1 Because of the exponentially distributed services, the queue-length process remains unchanged if we replace the processor-sharing service discipline by the FCFS discipline. Expression (3.6) can therefore be obtained from Gaver [37, Formula 8.4], where the p.g.f. of the number of customers in the system at arbitrary points in time is given for the case of a general service time distribution. The analysis is based on *completion times* of customers, see Gaver [37, Section 4.2]. In our case the distribution of the completion times has LST

$$\widetilde{\beta}(s) = \frac{\mu}{\mu + s + \nu\,(1 - \phi\,(s))}, \quad \mathrm{Re}(s) \geq 0. \tag{3.7}$$

These "inflated" service times are the sum of the actual time it takes to serve a customer (exponentially distributed with mean $1/\mu$) and all off-periods that occur during such a service. It can be shown that the first customer in a busy period has to wait before his service begins (this corresponds to the server being in the off-state in the original model with breakdowns) with probability $p = \frac{\nu(1-\phi(\lambda))}{\lambda+\nu(1-\phi(\lambda))}$, in which case the distribution of the residual off-period has LST

$$\delta(s) = \frac{\lambda}{1 - \phi(\lambda)} \times \frac{\phi(s) - \phi(\lambda)}{\lambda - s}, \quad \mathrm{Re}(s) \geq 0.$$

Accordingly, Expression (3.6) can also be verified using the LST of the queue-length distribution in an M/G/1 queue with exceptional first service, see Welch [115, Theorem 2]). In that queue the distribution of the regular services has LST $\widetilde{\beta}(s)$ and that of the exceptional first services has LST $(1 - p + p\delta(s))\,\widetilde{\beta}(s)$.

Remark 3.1.2 If the breakdowns (off-periods) are exponentially distributed too, i.e., $\phi(s) = (1 + m_1 s)^{-1}$, the probabilities $\mathbf{P}\{X = i, Y = j\}$, $j \in \{0, 1\}$, $i = 0, 1, \ldots$, can be found in Neuts [81, Theorem 6.3.1].

Remark 3.1.3 Expression (3.6) implies the decomposition $X \stackrel{d}{=} X_1 + X_2$ of the steady-state queue length, where $\stackrel{d}{=}$ means equality in distribution and the random variables X_1 and X_2 are independent. Furthermore, the p.g.f. of the random variable X_1 is given by Expression (3.2), i.e., X_1 is distributed as the steady-state queue length in a (particular) M^G/M/1 batch-arrival queueing model without service interruptions (see the discussion above). The random variable X_2 is equal to 0 with probability c, and with probability $1 - c$ it is distributed as the number of Poisson arrivals, with intensity λ, during a backward recurrence time of the off-periods. This decomposition does not fit into the framework of Fuhrmann and Cooper [32]. Even if we assume a FCFS service discipline — as in Remark 3.1.1 — Assumption 4 (non-preemptive service) of [32] is not satisfied.

3.2 A branching process representation

We show how the sojourn time of a customer (that is the total time spent in the system) can be studied by means of a branching process. For this purpose we will observe the process on a *transformed time scale*. The first to use this time-transformation method — commonly called the method of random time-change — for the analysis of processor-sharing queues apparently was Yashkov [121]. In its most essential form, but without transformation of time, this method was already used for the analysis of the M/G/1 processor-sharing queue in Yashkov [120]. Foley and Klutke [31] studied the queue-length process and the process of accumulated work after applying the random time change to a processor-sharing model in which the total service capacity may depend on the number of customers in the system, see also the model of the next chapter and particularly Section 4.3. Grishechkin [39, 40] further exploited the method by reformulating it in terms of Crump-Mode-Jagers branching processes and applying it to the analysis of queues with a general class of service disciplines, including processor sharing. For more references on the time-transformation method and its use in the analysis of processor-sharing queues we refer to Yashkov [123, Section 2.4].

We present a direct use of the time-transformation technique to analyse sojourn times in the processor-sharing queue with service interruptions presented in Section 3.1. However, the same approach is applicable to more general models, for instance those in Grishechkin [39, 40]. In Remark 3.2.1 we illustrate how the analysis in this section may be extended to the case with generally distributed service requirements. Restricting ourself to the model of Section 3.1 makes the presentation more transparent, while the fundamental ideas are the same as in the more general cases. Furthermore, we are able to carry the analysis further, and in particular in Sections 3.5 and 3.6 we obtain closed-form results.

In our presentation we first assume there is a permanent customer which never leaves the system. For this customer we study the accumulation of received service. All other ("non-permanent") customers — which arrive according to a Poisson process with rate λ — have an exponentially distributed service requirement with mean $1/\mu$. Let $Z(t)$ be the number of *non-permanent* customers at the server at time $t \geq 0$. For the service process we use the same notation as before: $Y(t)$ is 1 if the server is on at time t and 0 otherwise. This is justified since the *marginal* distribution of the process $Y(t)$, for all $t \geq 0$ is the same as before (when there was no permanent customer in the queue). Then, at time t, the permanent customer receives service at rate

$$\frac{Y(t)}{1+Z(t)}.$$

Let the random variable $R(t)$ be the amount of service received by the permanent customer during the time interval $[0,t]$:

$$R(t) := \int_{u=0}^{t} \frac{Y(u)}{1+Z(u)} \mathrm{d}u.$$

We define for $\tau \geq 0$:

$$V(\tau) := \inf\{t \geq 0 \ : \ R(t) \geq \tau\}.$$

Thus, $V(\tau)$ is the moment that the amount of service received by the permanent customer reaches the level τ. In Figure 3.1 a typical realisation of $R(t)$ and $V(\tau)$ is depicted. In that picture, at time $t_0 = 0$ there are two other customers in

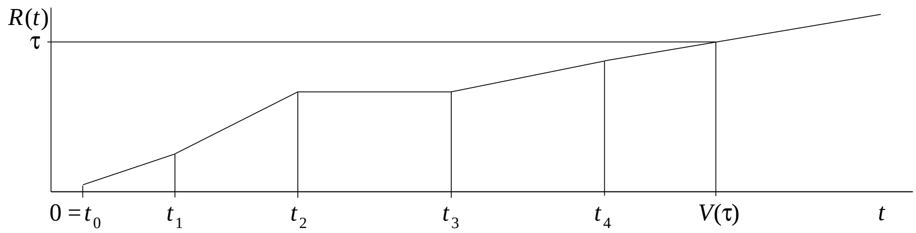

Figure 3.1: $R(t)$ and $V(\tau)$.

the system along with the permanent customer, therefore $R(t)$ increases at rate $1/3$ immediately after time t_0. At time t_1 one of the customers leaves and the rate increases to $1/2$. From t_2 until t_3 the server is off, and during this period 3 customers arrive, leading to a rate $1/5$ immediately after t_3. At time t_4 another customer arrives, etc. $V(\tau)$ is the moment that the service received by the permanent customer reaches the level τ.

If at time $t = 0$ the permanent customer is replaced by a customer requiring an amount of service τ, then $V(\tau)$ is the time at which this customer leaves the system, i.e. $V(\tau)$ is the sojourn time of that customer. Our goal is to determine the distribution of the random variable $V(\tau)$, for arbitrary $\tau > 0$.

We distinguish between the cases where $Y(0) = 1$ (start with a working server) and $Y(0) = 0$ (start with a server in the off-state). For $n \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$, we denote by $V_{n,1}(\tau)$ the process $V(\tau)$ conditional on $Y(0) = 1$ and $Z(0) = n$, or equivalently: $V_{n,1}(\tau) := V(\tau)\,|\{Y(0) = 1, Z(0) = n\}$. Similarly we define the conditional processes $Z_{n,1}(t) := Z(t)\,|\{Y(0) = 1, Z(0) = n\}$ and $Y_{n,1}(t) := Y(t)\,|\{Y(0) = 1, Z(0) = n\}$. First we concentrate on $V_{n,1}(\tau)$, the conditional sojourn time of a customer that arrives when the server is working and the number of customers is n. At the end of this section we derive the results for the sojourn time of a customer that arrives when the server is off.

In the sequel we use the notation $x(y+) := \lim_{u \downarrow y} x(u)$ and $x(y-) := \lim_{u \uparrow y} x(u)$ for any function $x(y)$ for which these limits exist.

Lemma 3.2.1 *For arbitrary $n \in \mathbb{N}_0$, it holds with probability 1 that $V_{n,1}(0+) = 0$.*

Proof The lemma follows immediately from the fact that, for small τ, $V_{n,1}(\tau)$ is equal to $(n+1)\tau$ with probability $1 - \left(\lambda + \frac{n}{n+1}\mu + \nu\right)(n+1)\tau + \mathrm{o}(\tau)$. □

Consider the model with one permanent customer and suppose that at time 0 the server is working and n non-permanent customers are present, i.e., $Y(0) = 1$ and $Z(0) = n$. Denote the number of times that the server turned off during the period $(0,t)$ by the random variable $N_{n,1}(t)$, and the length of the i^{th} off-period started after time 0 by D_i, $i \in \{1, 2, \ldots\}$. Note that $\{D_1, D_2, \ldots\}$ is an i.i.d. sequence with distribution $F(t)$, and that $N_{n,1}(t)$ does not depend on n. Furthermore, define for $\tau > 0$:

$$N'_{n,1}(\tau) := N_{n,1}\left(V_{n,1}(\tau)\right).$$

The random variable $N'_{n,1}(\tau)$ is well defined because $V_{n,1}(\tau)$ — also a random variable — is strictly increasing in τ (with probability 1). Note that $N'_{n,1}(\tau+) - N'_{n,1}(\tau) = 1$ if and only if at time $t = V_{n,1}(\tau)$ the server turns into the off-state. Otherwise $N'_{n,1}(\tau+) - N'_{n,1}(\tau) = 0$.

Similar to $N'_{n,1}(\tau)$, we define for $\tau > 0$ the processes $Z'_{n,1}(\tau) := Z_{n,1}(V_{n,1}(\tau))$ and $Y'_{n,1}(\tau) := Y_{n,1}(V_{n,1}(\tau)+)$.

Lemma 3.2.2 *The process $V_{n,1}(\tau)$ is related to $Z'_{n,1}(\tau)$, $N'_{n,1}(\tau)$, and the sequence D_i, $i \in \{1, 2, \ldots\}$, through the equation*

$$V_{n,1}(\tau) = \int_{\sigma=0}^{\tau} \left[1 + Z'_{n,1}(\sigma)\right] \mathrm{d}\sigma + \sum_{i=1}^{N'_{n,1}(\tau)} D_i, \tag{3.8}$$

with the empty sum being equal to zero (when $N'_{n,1}(\tau) = 0$).

Proof Consider any realisation of the arrival process, the sequence of required services, and the process $\{Y(t), t \geq 0\}$. In Figure 3.1 a particular realisation is depicted. Note that if $N'_{n,1}(\tau+) - N'_{n,1}(\tau) = 0$, then

$$\frac{\mathrm{d}V_{n,1}(\tau)}{\mathrm{d}\tau} = 1 + Z'_{n,1}(\tau),$$

and if $N'_{n,1}(\tau+) - N'_{n,1}(\tau) = 1$, then $V_{n,1}(\tau+) - V_{n,1}(\tau) = D_{N'_{n,1}(\tau+)}$. □

From the definition of the processes $\left\{Z'_{n,1}(\tau), \tau \geq 0\right\}$ and $\left\{N'_{n,1}(\tau), \tau \geq 0\right\}$, and with the aid of Figure 3.1, we make the following observation:

Lemma 3.2.3 *The transformed process* $\left(Z'_{n,1}(\tau), N'_{n,1}(\tau)\right)$ *is Markovian, with transition rates given in the following table for* n, k *and* $j \in \mathbb{N}_0$,

from state	*to state*	*transition rate*
(n,k)	$(n+1,k)$	$(n+1)\lambda$
(n,k)	$(n-1,k)$	$n\mu$
(n,k)	$(n+j,k+1)$	$(n+1)\nu p_j$

where p_j *is the probability that during an off-period,* j *new customers arrive:*

$$\sum_{j=0}^{\infty} z^j p_j = \phi(\lambda(1-z)).$$

Proof In words, the transformation from $(Z_{n,1}(t), N_{n,1}(t), Y_{n,1}(t))$ to the process $\left(Z'_{n,1}(\tau), N'_{n,1}(\tau)\right)$ consists in (i) shrinking the time scale by a factor $n+1$ when $Z_{n,1}(t) = n$ and $Y_{n,1}(t) = 1$, and (ii) replacing off-periods by batch arrivals of customers. From this construction it is clear that in the transformed process the residence time in any state is exponentially distributed and that the transition rates are as stated. □

In Equation (3.8), $V_{n,1}(\tau)$ also depends on $D_1, \ldots, D_{N'_{n,1}(\tau)}$. We emphasise that, if $N'_{n,1}(\tau) - N'_{n,1}(\tau-) = 1$ then $Z'_{n,1}(\tau) - Z'_{n,1}(\tau-)$ and $D_{N'_{n,1}(\tau)}$ are *not* independent: $D_{N'_{n,1}(\tau)}$ is the length of an off-period in the original process and $Z'_{n,1}(\tau) - Z'_{n,1}(\tau-)$ is the number of customers that arrived during that period:

$$\begin{aligned}
&\mathbf{E}\left[\mathrm{e}^{-sD_{N'_{n,1}(\tau)}} z^{Z'_{n,1}(\tau) - Z'_{n,1}(\tau-)} \,|N'_{n,1}(\tau) - N'_{n,1}(\tau-) = 1\right] \\
&= \phi(s + \lambda(1-z)).
\end{aligned}$$

In order to study the distribution of $V_{n,1}(\tau)$, we construct a branching process that is "equivalent" with $\left(Z'_{n,1}(\tau), N'_{n,1}(\tau); D_1, \ldots, D_{N'(\tau)}\right)$, and associate a reward structure with this branching process that will turn out to be useful. Consider a population $\mathcal{P}$ of elements which evolves in the following way: The life time of an element of the population has an exponential distribution with mean duration $1/\mu$. During its life time an element receives a reward at rate 1 (per time unit). An element generates children in two ways, independent from all other living elements. According to a Poisson process with rate λ an element gives birth to children, one at a time. In addition, according to another (independent) Poisson process with rate ν, an element generates nests of children (possibly empty nests), and receives an immediate reward which depends on the number of children in the nest in a stochastic way. The joint LST and p.g.f. of the simultaneous distribution of A children in the nest and the immediate reward D, is given by

$$\mathbf{E}\left[\mathrm{e}^{-sD} z^A\right] = \phi(s + \lambda(1-z)).$$

Finally, there is a permanent element in the population which generates children — and receives rewards — in the same way as the other elements (but never dies).

Suppose that at time 0 there are n non-permanent elements in the population $\mathcal{P}$. Denote the number of non-permanent elements in the population at time $\tau \geq 0$ by $Z''_n(\tau)$, the number of nest-births between time 0 and time τ by $N''_n(\tau)$, and the reward of the i^{th} nest by D''_i.

Lemma 3.2.4 *For all $n \in \mathbb{N}_0$ and $\tau \geq 0$, it holds that*

$$\left(Z'_{n,1}(\tau), N'_{n,1}(\tau); D_1, \ldots, D_{N'_{n,1}(\tau)}\right) \stackrel{d}{=} \left(Z''_n(\tau), N''_n(\tau); D''_1, \ldots, D''_{N''_n(\tau)}\right),$$

where $\stackrel{d}{=}$ means equality in distribution.

Moreover, $V_{n,1}(\tau)$ is distributed as the reward of the population $\mathcal{P}$ from time 0 until time τ, starting with n individuals in the population at time 0.

Proof The lemma follows from Lemma 3.2.3 and by comparing the transition rates of both processes. □

In the next theorem we formulate the main result of this section. For this, we need to introduce the random variables $C_i(\tau)$, $i \in \{0, 1, 2, \ldots\}$. $C_0(\tau)$ is the reward for the permanent element and his offspring between time instants 0 and τ. Similarly, $C_i(\tau)$, $i = 1, 2, \ldots, n$, is the reward for the i^{th} non-permanent individual, who was present at time 0, plus the reward for his offspring between time instants 0 and τ. Note that all $C_i(\tau)$, $i \geq 1$ have the same distribution.

The decomposition of the sojourn time given in the theorem was established by Yashkov [120, Expression (3.4)] for the ordinary M/G/1 processor-sharing queue, and by Rege and Sengupta [91, Theorem 6] for the M/G/1 queue with discriminatory processor sharing.

Theorem 3.2.5 *The conditional sojourn time $V_{n,1}(\tau)$ of a customer who finds the server working upon arrival, with n other customers present, can be decomposed as,*

$$V_{n,1}(\tau) \stackrel{d}{=} C_0(\tau) + \sum_{i=1}^{n} C_i(\tau),$$

where $\stackrel{d}{=}$ means equality in distribution. All random variables involved in the right-hand side are mutually independent. In particular, $C_0(\tau) \stackrel{d}{=} V_{0,1}(\tau)$.

Proof Using the reward-interpretation of $V_{n,1}(\tau)$ given in Lemma 3.2.4, we can split $V_{n,1}(\tau)$ into the individual rewards of all elements. By construction, the elements of the population $\mathcal{P}$ behave independently of each other. Therefore, the $C_i(\tau)$ – including $C_0(\tau)$ – form an independent sequence. □

We now turn to the sojourn time of a customer that arrives to the system during a service interruption. As before, suppose that the number of customers present is n, and that the service requirement of the arriving customer is τ. Let D_0 be the *residual* off-period at time zero and A_0 be the number of arrivals during D_0. The LST of the distribution of D_0 will be denoted by $\phi_0(s)$.

If $V_{n,0}(\tau)$ is the sojourn time of the arriving customer, then by conditioning on the length of D_0 and the number of arrivals A_0:

$$V_{n,0}(\tau)\,|\{D_0 = d_0, A_0 = k\} \stackrel{d}{=} d_0 + V_{n+k,1}(\tau). \tag{3.9}$$

Corollary 3.2.6 *$V_{n,0}(\tau)$, the conditional sojourn time of a customer who finds the server in the off-state upon arrival with n other customers present, can be written as:*

$$V_{n,0}(\tau) \stackrel{d}{=} D_0 + C_0(\tau) + \sum_{i=1}^{n+A_0} C_i(\tau).$$

All random variables on the right-hand side are mutually independent, except for the pair (D_0, A_0) which has the joint distribution,

$$\mathbf{E}\left[\mathrm{e}^{-sD_0} z^{A_0}\right] = \phi_0(s + \lambda(1-z)), \quad \mathrm{Re}(s) \geq 0, |z| \leq 1.$$

Proof The corollary follows from Theorem 3.2.5 and Relation (3.9). □

We define the LSTs of the distributions of $C_0(\tau)$ and $C_i(\tau)$, $i \in \{1, 2, \ldots\}$, by $g_0(\tau; s)$ and $g_1(\tau; s)$: For $\mathrm{Re}(s) \geq 0$,

$$g_0(\tau; s) := \mathbf{E}\left[\mathrm{e}^{-sC_0(\tau)}\right], \quad g_1(\tau; s) := \mathbf{E}\left[\mathrm{e}^{-sC_i(\tau)}\right], \quad i = 1, 2, \ldots.$$

From Theorem 3.2.5 and Corollary 3.2.6 we have, for $\mathrm{Re}(s) \geq 0$,

$$\mathbf{E}\left[\mathrm{e}^{-sV(\tau)}\,|Y(0) = 1, Z(0) = n\right] = g_0(\tau; s)\,\{g_1(\tau; s)\}^n, \tag{3.10}$$

$$\mathbf{E}\left[\mathrm{e}^{-sV(\tau)}\,|Y(0) = 0, Z(0) = n\right] = g_0(\tau; s)\,\{g_1(\tau; s)\}^n \times\phi_0(s + \lambda(1 - g_1(\tau; s))). \tag{3.11}$$

In Section 3.3 we characterise $g_0(\tau; s)$ and $g_1(\tau; s)$ by means of a set of differential equations, which we solve in terms of an integral equation.

We conclude this section with the following remark, which indicates how the representation of the sojourn time by a branching process can be extended to the case of general service time distributions.

Remark 3.2.1 The generalisation of this representation by branching processes to general service time distributions, $B(x)$, $x \geq 0$, can be obtained by using the method of supplementary variables. We extend the state space representation with the vector $(x_1, x_2, \ldots, x_n)$ when there are n customers in the system. We again assume that a newly arrived customer with service requirement τ finds the server available, and we further condition on the number of customers in the system upon arrival (n) and the residual service requirement of each of those customers (x_i, $i = 1, 2, \ldots, n$). If we denote the conditional sojourn time of the new customer by $V_{n,1}(\tau; x_1, \ldots, x_n)$ then

$$V_{n,1}(\tau; x_1, \ldots, x_n) \stackrel{d}{=} C_0(\tau) + \sum_{i=1}^{n} C_i(\tau; x_i),$$

where the $C_0(\tau)$ and $C_i(\tau; x_i)$, $i = 1, 2, \ldots$, are the analogues of the earlier $C_0(\tau)$ and $C_i(\tau)$ for the population model with life time distribution $B(x)$. Thus $C_i(\tau; x_i)$ is the reward for a family until time τ, starting with one individual with a remaining life time x_i. This generalisation is studied in greater detail in Section 5.3. See also Yashkov [120] for a related analysis of the case without service interruptions.

3.3 Characterisation of $g_0(\tau; s)$ and $g_1(\tau; s)$

We derive a set of differential equations which uniquely determine $g_0(\tau; s)$ and $g_1(\tau; s)$, the LSTs of the distributions of $C_0(\tau)$ and $C_1(\tau)$. We then express $g_0(\tau; s)$ in terms of $g_1(\tau; s)$, and — for real $s > 0$ — derive a useful integral equation for $g_1(\tau; s)$.

Lemma 3.3.1 *For* $\mathrm{Re}(s) \geq 0$ *and* $\tau \geq 0$, $g_0(\tau; s)$ *and* $g_1(\tau; s)$ *are uniquely determined by the following set of differential equations,*

$$\begin{aligned}
\frac{\partial}{\partial \tau} g_1(\tau; s) &= -(s + \lambda + \mu + \nu)\, g_1(\tau; s) + \lambda \{g_1(\tau; s)\}^2 + \mu && (3.12)\\
&\quad + \nu g_1(\tau; s) \phi(s + \lambda(1 - g_1(\tau; s))), \\
\frac{\partial}{\partial \tau} g_0(\tau; s) &= -(s + \lambda + \nu)\, g_0(\tau; s) + \lambda g_0(\tau; s) g_1(\tau; s) && (3.13)\\
&\quad + \nu g_0(\tau; s) \phi(s + \lambda(1 - g_1(\tau; s))),
\end{aligned}$$

and initial conditions,

$$g_0(0; s) = g_1(0; s) = 1. \tag{3.14}$$

Proof See Appendix 3.A. □

Theorem 3.3.2 *We can express* $g_0(\tau; s)$ *in terms of* $g_1(\tau; s)$ *as,*

$$g_0(\tau; s) = g_1(\tau; s) \exp\left\{\mu \left(\tau - \int_{u=0}^{\tau} g_1(u; s)^{-1} \mathrm{d}u\right)\right\}. \tag{3.15}$$

Proof From Equations (3.13) and (3.14) we immediately obtain $g_0(\tau; s)$ in terms of $g_1(\tau; s)$:

$$g_0(\tau; s) = \exp\left\{-(s + \lambda + \nu)\tau + \int_{u=0}^{\tau} [\lambda g_1(u; s) + \nu \phi(s + \lambda(1 - g_1(u; s)))]\, \mathrm{d}u\right\}. \tag{3.16}$$

If we also use Equation (3.12) we may rewrite this as,

$$g_0(\tau; s) = \exp\left\{\int_{u=0}^{\tau} \frac{\frac{\partial}{\partial u} g_1(u; s) - \mu\,(1 - g_1(u; s))}{g_1(u; s)} \mathrm{d}u\right\},$$

which leads to Relation (3.15). □

The remainder of this section is devoted to finding the solution of Equation (3.12) for real $s > 0$. We first define the *clearing period* of the model of Section 3.1 as the time it takes for the system to become empty, starting with one customer and a working server. If there were no off-periods, the clearing period would be equal to the busy period. We generically denote the clearing period by the random variable CP and the LST of its distribution by $r_1(s) = \mathbf{E}\left[\mathrm{e}^{-sCP}\right]$.

Lemma 3.3.3 *The clearing period has the same distribution as the busy period of an ordinary M/G/1 queue with arrival rate λ and LST of the service time distribution $\widetilde{\beta}(.)$ given by Expression (3.7).*

As a consequence, for $\mathrm{Re}(s) \geq 0$, *$x = r_1(s)$ is the unique root — inside (or on) the unit circle in the complex plane — of the equation:*

$$(s + \lambda + \mu + \nu)\, x = \lambda x^2 + \mu + \nu x \phi(s + \lambda(1 - x)). \tag{3.17}$$

Proof Note that for the model with the FCFS queue discipline — described in Remark 3.1.1 — we may define the clearing period as we did above for the model of Section 3.1. Moreover, the clearing periods of both models have the same distribution. It is easily seen that the clearing period of the model in Remark 3.1.1 has the same distribution as the busy period of an ordinary M/G/1 queue with arrival rate λ and LST of the service time distribution $\widetilde{\beta}(.)$. This proves the first statement of the Lemma. Furthermore, we immediately have that for $\mathrm{Re}(s) \geq 0$, $r_1(s)$ is equal to the (unique) root inside (or on) the unit circle of the equation,

$$x = \widetilde{\beta}(s + \lambda(1 - x)),$$

see for instance Cohen [20, p. 250]. This equation readily leads to Relation (3.17). □

Lemma 3.3.4 *For $s \geq 0$ and $\tau \geq 0$,*

$$r_1(s) \leq g_1(\tau; s) \leq 1.$$

Proof Fix $s \geq 0$. Obviously, $C_1(\tau)$ is non-decreasing in τ with probability 1, and so $g_1(\tau; s)$ is non-increasing in τ. Therefore, the right-hand side of Equation (3.12) is negative for $\tau \geq 0$. Indeed, for $\tau = 0$ this is easily verified because $g_1(0; s) = 1$. If $r_1(s) > g_1(\tau; s)$, for some $\tau > 0$, it follows from Lemma 3.3.3 that the right-hand side of Equation (3.12) is positive for this τ, since the zero $r_1(s)$ has multiplicity 1. □

Theorem 3.3.5 *For real $s > 0$, the solution to Equation (3.12) satisfying Condition (3.14), is obtained from,*

$$\int_{x=1}^{g_1(\tau;s)} \frac{1}{\mu - (s+\lambda+\mu+\nu)x + \lambda x^2 + \nu x \phi(s+\lambda-\lambda x)} \mathrm{d}x = \tau. \tag{3.18}$$

Proof The integral in Relation (3.18) is well defined, because the denominator of the integrand has no zeroes in $(r_1(s), 1)$ for $s > 0$, see Lemma 3.3.3. The integral is taken for x from 1 to $g_1(\tau; s)$ so that the initial Condition (3.14) is satisfied. By differentiating with respect to τ, it is readily seen that Equation (3.12) is also satisfied. $\square$

In Section 3.7 we use Relation (3.18) to study the asymptotics of $g_1(\tau; s)$ as $\tau \to \infty$. This in turn enables us to prove the convergence in probability of $\frac{C_0(\tau)}{\tau}$ and (more importantly) $\frac{V(\tau)}{\tau}$ for $\tau \to \infty$. Relation (3.18) is not very practical for determining moments of $C_1(\tau)$ (and $C_0(\tau)$). In Section 3.4 we study these moments directly.

Computational issues

In the remainder of this section we show how the distribution of $C_1(\tau)$ can be computed in the case that the distribution of the off-periods has a rational LST. We also discuss some difficulties regarding the computation of the distribution of $C_0(\tau)$.

Suppose for the moment that — for *real* (and positive) values of s — $g_1(\tau; s)$ can be evaluated from Relation (3.18). Then we use the Gaver-Stehfest algorithm, see Abate and Whitt [2, pp. 52–55], to compute the distribution function of $C_1(\tau)$. We emphasise that the (generally more stable) algorithms Euler and Poisson, see Abate and Whitt [2, pp. 48–51], can not be used because in Relation (3.18) we assumed s to be real. To evaluate the n^{th} Gaver-Stehfest approximant, one typically needs $2n$-digit precision in the calculations. In general taking $n = 15$ gives good results — relative errors are typically less than 3% for tail probabilities of the order 10^{-3} — and comparison with the results using $n = 20$ provides a useful accuracy check.

To illustrate how $g_1(\tau; s)$ can be evaluated, let us first consider the case that the off-periods have a hyper-exponential distribution. In that case the LST of the distribution of the off-periods is of the form:

$$\phi(s) = \sum_{i=1}^{k} \frac{w^{(i)}}{1 + m_1^{(i)} s},$$

with $w^{(i)} > 0$, $\sum_{i=1}^{k} w^{(i)} = 1$, $m_1^{(1)} > m_1^{(2)} > \ldots > m_1^{(k)} > 0$, and $\mathrm{Re}(s) > -1/m_1^{(1)}$. Note that $m_1 = \sum_{i=1}^{k} w^{(i)} m_1^{(i)}$. After multiplying the numerator and the denominator of the integrand in Relation (3.18) by

$$\prod_{i=1}^{k} \left\{1 + m_1^{(i)} \left(s + \lambda(1-x)\right)\right\},$$

it becomes a rational function in x with the degree of the denominator equal to $k+2$, and that of the numerator equal to k. It can be seen that the denominator is positive for $x = 0$ and for $x = \left(s + \lambda + 1/m_1^{(i)}\right)/\lambda$ when i is odd, and the denominator is negative for $x = 1$ and for $x = \left(s + \lambda + 1/m_1^{(i)}\right)/\lambda$ when i is even. Moreover, if $x \to \infty$ then the denominator tends to $+\infty$ when k is even, and to $-\infty$ when k is odd. Therefore, for $s > 0$ and $i = 1, 2, \ldots, k+2$, the roots $r_i(s)$ of the denominator satisfy:

$$0 < r_1(s) < 1 < r_2(s) < \frac{s + \lambda + \frac{1}{m_1^{(1)}}}{\lambda} < r_3(s) < \ldots < \frac{s + \lambda + \frac{1}{m_1^{(k)}}}{\lambda} < r_{k+2}(s).$$

This relation enables an efficient computation of the roots, for instance using the Newton-Raphson method (combined with the bisection method) on each of the above intervals containing exactly one root. By partial fraction expansion, Relation (3.18) can now be written as:

$$\tau = \int_{x=1}^{g_1(\tau;s)} \sum_{i=1}^{k+2} \frac{a_i(s)}{r_i(s) - x} \mathrm{d}x = -\sum_{i=1}^{k+2} a_i(s) \ln\left(\frac{r_i(s) - g_1(\tau; s)}{r_i(s) - 1}\right). \qquad (3.19)$$

The functions $a_i(s)$ are given by:

$$a_i(s) = \frac{\prod_{j=1}^{k} \left(1 + m_1^{(j)} \left\{s + \lambda(1 - r_i(s))\right\}\right)}{\lambda^{k+1} \prod_{j=1}^{k} m_1^{(j)} \prod_{j \neq i} (r_j(s) - r_i(s))}.$$

Note that, for $s > 0$, $r_1(s) < g_1(\tau; s) < 1$ and $a_1(s) > 0$, whereas $r_i(s) > 1$ and $a_i(s) < 0$, $i \in \{2, 3, \ldots, k+2\}$. After computing the roots $r_i(s)$ and the coefficients $a_i(s)$, $g_1(\tau; s)$ can be found from Expression (3.19), again using the Newton-Raphson method.

We tested the above procedure to compute the distribution function of $C_1(\tau)$ for the case of no service interruptions, and for the case of exponentially distributed off-periods. In the first case a closed-form expression for $g_1(\tau; s)$ can be found in Coffman et al. [17, Equation (16)]. Using this expression, the Euler algorithm — see Abate and Whitt [2, Section 7] — gives a reliable alternative to compare the results. In general, the outcomes of both methods agreed up to a relative difference of at most 3% for tail probabilities of the order 10^{-3}. In the case of exponentially distributed off-periods we compared our results to those generated by simulation, and again found that the relative differences were at most 3%.

We saw above that for hyper-exponential off-periods the roots $r_i(s)$ are all real and positive, and we found disjoint intervals on the positive real line, each containing exactly one root. When the distribution of the off-periods has a rational LST, but is not a hyper-exponential distribution, the analysis proceeds along the same lines. However, in general some of the roots may be complex. This is for instance the case when the off-periods have an Erlang distribution.

Serious complications arise when the distribution of the off-periods does not have a rational LST. In principle, the left-hand side of Relation (3.18) can be computed using for instance Simpson's rule (or a higher order Newton-Cotes method) for numerical integration. However, like any other inversion method, the Gaver-Stehfest algorithm is highly sensitive to small errors in the computation of the LST that is to be inverted. Therefore, computation of the integral in Relation (3.18) requires exceedingly long computation times due to the typical accuracy problems with numerical integration.

The same difficulties are encountered in the computation of $g_0(\tau; s)$ using Equation (3.15). Even if $g_1(\tau; s)$ has been computed accurately, for instance using the above procedure for the case that the distribution of the off-periods has a rational LST, evaluating the right-hand side of Equation (3.15) requires an additional numerical integration leading to prohibitively long computation times (poor results were obtained even after 2 hours on a Sun Sparc 4 station).

3.4 Moments of $C_0(\tau)$ and $C_1(\tau)$

In Section 3.3 we saw that $g_0(\tau; s)$ and $g_1(\tau; s)$, the LSTs of the distributions of $C_0(\tau)$ and $C_1(\tau)$, are determined by a set of differential equations. The solution for these differential equations is given by Relations (3.15) and (3.18). However, this solution is not very practical for determining moments of $C_0(\tau)$ and $C_1(\tau)$. In this section we show how the moments of $C_0(\tau)$ and $C_1(\tau)$ can be found by directly solving an alternative system of differential equations. Yashkov [120] also remarks that, in the M/G/1 processor-sharing queue, such an approach leads to a more tractable derivation of moments. First we state the following theorem which is a consequence of a result of De Meyer and Teugels [72, Lemma 3].

Theorem 3.4.1 *If the k^{th} moment of the off-periods, m_k, exists, then so do the k^{th} moments of $C_0(\tau)$ and $C_1(\tau)$ exist.*

Proof See Appendix 3.B. □

We start by illustrating the derivation of the first and second moments of $C_0(\tau)$ and $C_1(\tau)$. We then formulate and prove Theorem 3.4.2 which reveals the structure of the higher moments, as a function of τ.

By differentiating both sides of Equations (3.12) and (3.13) with respect to s and then setting $s = 0$ we get,

$$\frac{\partial}{\partial\tau}\mathbf{E}\left[C_1(\tau)\right] = 1 + \nu m_1 - \{\mu - \lambda(1 + \nu m_1)\}\,\mathbf{E}\left[C_1(\tau)\right], \tag{3.20}$$

$$\frac{\partial}{\partial\tau}\mathbf{E}\left[C_0(\tau)\right] = 1 + \nu m_1 + \lambda(1 + \nu m_1)\mathbf{E}\left[C_1(\tau)\right]. \tag{3.21}$$

Formally it should first be verified that interchanging the order of differentiation is allowed. However, in our case, we can also find Equations (3.20) and (3.21) by directly applying the argument of conditioning on the events in a time interval of length Δ to $\mathbf{E}[C_0(\tau)]$ and $\mathbf{E}[C_1(\tau)]$, and then letting $\Delta \downarrow 0$. Using the initial conditions $C_0(0) = C_1(0) = 0$, we find:

$$\mathbf{E}[C_1(\tau)] = \frac{1/\mu}{c-\rho}\left(1 - \mathrm{e}^{-\mu\,(1-\rho/c)\,\tau}\right), \tag{3.22}$$

$$\mathbf{E}[C_0(\tau)] = \frac{\tau}{c-\rho} - \frac{\rho/\mu}{(c-\rho)^2}\left(1 - \mathrm{e}^{-\mu\,(1-\rho/c)\,\tau}\right). \tag{3.23}$$

If $m_2 < \infty$, we can repeat this procedure to find $\mathbf{E}\left[C_0(\tau)^2\right]$ and $\mathbf{E}\left[C_1(\tau)^2\right]$. Differentiating Equations (3.12) and (3.13) twice w.r.t. s and then setting $s = 0$ (or again by a direct conditioning argument) we find

$$\begin{aligned}\frac{\partial}{\partial\tau}\mathbf{E}\left[C_1(\tau)^2\right] &= -\{\mu - \lambda(1+\nu m_1)\}\,\mathbf{E}\left[C_1(\tau)^2\right] + 2(1+\nu m_1)\mathbf{E}[C_1(\tau)] \\ &\quad + 2\lambda(1+\nu m_1)\mathbf{E}[C_1(\tau)]^2 + \nu m_2\{1+\lambda\mathbf{E}[C_1(\tau)]\}^2, \qquad (3.24)\\ \frac{\partial}{\partial\tau}\mathbf{E}\left[C_0(\tau)^2\right] &= 2(1+\nu m_1)\mathbf{E}[C_0(\tau)] + 2\lambda(1+\nu m_1)\mathbf{E}[C_0(\tau)]\,\mathbf{E}[C_1(\tau)] \\ &\quad + \lambda(1+\nu m_1)\mathbf{E}\left[C_1(\tau)^2\right] + \nu m_2\{1+\lambda\mathbf{E}[C_1(\tau)]\}^2. \qquad (3.25)\end{aligned}$$

We can solve this using Expressions (3.22) and (3.23):

$$\begin{aligned}\mathbf{E}\left[C_1(\tau)^2\right] &= -(a_1+2a_2)\tau\mathrm{e}^{-\mu\,(1-\rho/c)\,\tau} \\ &\quad + \frac{ca_1 + (1-c)\frac{m_2}{m_1}}{\mu\,(c-\rho)}\left(1 - \mathrm{e}^{-\mu\,(1-\rho/c)\,\tau}\right) \qquad (3.26)\\ &\quad + \frac{ca_2}{\mu\,(c-\rho)}\left(1 - \mathrm{e}^{-2\mu\,(1-\rho/c)\,\tau}\right), \\ \mathbf{E}\left[C_0(\tau)^2\right] &= b_1\tau + b_2\tau^2 + b_3\tau\mathrm{e}^{-\mu\,(1-\rho/c)\,\tau} - b_4\left(1 - \mathrm{e}^{-\mu\,(1-\rho/c)\,\tau}\right) \\ &\quad - b_5\left(1 - \mathrm{e}^{-2\mu\,(1-\rho/c)\,\tau}\right), \qquad (3.27)\end{aligned}$$

where

$$\begin{aligned}a_1 &= 2(1+\nu m_1 + \lambda\nu m_2)\frac{1/\mu}{c-\rho}, \\ a_2 &= \lambda(2(1+\nu m_1)+\lambda\nu m_2)\left(\frac{1/\mu}{c-\rho}\right)^2, \\ b_1 &= \nu m_2\left(\frac{c}{c-\rho}\right)^3, \\ b_2 &= \left(\frac{1}{c-\rho}\right)^2,\end{aligned}$$

$$\begin{aligned} b_3 &= 2\rho\frac{2+\rho(1+\nu m_1)+\lambda c\nu m_2}{\mu\,(c-\rho)^3}, \\ b_4 &= \frac{2\rho}{\mu\,(c-\rho)^4}\left(\frac{2c-\rho}{\mu}+\frac{1}{2}c\nu m_2\left(3c^2-\rho^2\right)\right), \\ b_5 &= \frac{\rho^2}{(c-\rho)^4}\left(\frac{2}{\mu^2}+\frac{1}{2}\nu m_2 c\frac{2\rho-c}{\mu}\right). \end{aligned}$$

The same approach can be applied to determine higher moments. In Theorem 3.4.2 this is done to reveal the structure of these moments.

Theorem 3.4.2 *For $k \geq 1$, provided that $m_k < \infty$, and thus $\mathbf{E}\left[C_1(\tau)^k\right] < \infty$ and $\mathbf{E}\left[C_0(\tau)^k\right] < \infty$,*

$$\mathbf{E}\left[C_1(\tau)^k\right] = \alpha_0^{(k)} + \sum_{m=1}^{k} \mathrm{e}^{-m\mu\,(1-\rho/c)\,\tau} \sum_{n=0}^{k-m} \alpha_{m,n}^{(k)}\tau^n, \tag{3.28}$$

$$\mathbf{E}\left[C_0(\tau)^k\right] = \sum_{m=0}^{k} \mathrm{e}^{-m\mu\,(1-\rho/c)\,\tau} \sum_{n=0}^{k-m} \beta_{m,n}^{(k)}\tau^n, \tag{3.29}$$

where the $\alpha_0^{(k)}$, $\alpha_{m,n}^{(k)}$ and $\beta_{m,n}^{(k)}$ are coefficients that are independent of τ.

Proof See Appendix 3.C. □

3.5 Moments of the conditional sojourn time

In this section we study the moments of the sojourn time of a customer conditioned on the service requirement, the state of the server upon arrival, and the number of other customers in the system. We give these moments in terms of the moments of $C_0(\tau)$ and $C_1(\tau)$. In particular, using the expressions for the first two moments of $C_1(\tau)$ and $C_0(\tau)$ found in Section 3.4, we find closed-form expressions for the first two moments of the conditional sojourn time. From the definition of $V_{n,i}(\tau)$ in Section 3.2 we obviously have:

$$\begin{aligned} \mathbf{E}\left[V_{n,1}(\tau)^k\right] &= \mathbf{E}\left[V(\tau)^k|\{Y(0)=1,Z(0)=n\}\right], \\ \mathbf{E}\left[V_{n,0}(\tau)^k\right] &= \mathbf{E}\left[V(\tau)^k|\{Y(0)=0,Z(0)=n\}\right]. \end{aligned}$$

Lemma 3.5.1 *For $k, n \in \mathbb{N}$ and $\tau \geq 0$,*

$$\mathbf{E}\left[V_{n,1}(\tau)^k\right] = \sum_{j=0}^{k}\binom{k}{j}\mathbf{E}\left[C_1(\tau)^{k-j}\right]\,\mathbf{E}\left[V_{n-1,1}(\tau)^j\right], \tag{3.30}$$

$$\mathbf{E}\left[V_{n,0}(\tau)^k\right] = \sum_{j=0}^{k}\binom{k}{j}\mathbf{E}\left[\left(D_0+\sum_{i=1}^{A_0}C_i(\tau)\right)^j\right]\mathbf{E}\left[V_{n,1}(\tau)^{k-j}\right], \tag{3.31}$$

with $\mathbf{E}\left[V_{0,1}(\tau)^k\right] = \mathbf{E}\left[C_0(\tau)^k\right]$ *and*

$$\mathbf{E}\left[\left(D_0 + \sum_{i=1}^{A_0} C_i(\tau)\right)^j\right] = (-1)^j \frac{\partial^j}{\partial s^j}\phi_0(s + \lambda - \lambda g_1(\tau; s))\Bigg|_{s=0}.$$

D_0 is the residual off-period at time zero, $\phi_0(.)$ denotes the LST of the distribution of D_0, and A_0 is the number of arrivals during D_0.

Proof From Theorem 3.2.5, we have for $k, n \in \mathbb{N}$, and $\tau \geq 0$,

$$\begin{aligned}
\mathbf{E}\left[V_{n,1}(\tau)^k\right] &= \mathbf{E}\left[(C_0(\tau) + \ldots C_n(\tau))^k\right] \\
&= \sum_{j=0}^{k} \binom{k}{j} \mathbf{E}\left[C_n(\tau)^{k-j}\right] \mathbf{E}\left[(C_0(\tau) + \ldots C_{n-1}(\tau))^j\right] \\
&= \sum_{j=0}^{k} \binom{k}{j} \mathbf{E}\left[C_1(\tau)^{k-j}\right] \mathbf{E}\left[V_{n-1,1}(\tau)^j\right],
\end{aligned}$$

and $\mathbf{E}\left[V_{0,1}(\tau)^k\right] = \mathbf{E}\left[C_0(\tau)^k\right]$. Moreover, combining Theorem 3.2.5 and Corollary 3.2.6, we find Expression (3.31). □

Remark 3.5.1 The derivatives $\frac{\partial^j}{\partial s^j}\phi_0(s + \lambda - \lambda g_1(\tau; s))\Big|_{s=0}$ can be found by using Lemma 1 of De Meyer and Teugels [72] to expand $\phi_0(s + \lambda - \lambda g_1(\tau; s))$ in a Taylor series, analogous to Equation (3.53) below.

From Relation (3.30) we can compute the conditional moments $\mathbf{E}\left[V_{n,1}(\tau)^k\right]$ recursively, once we have the moments of $C_0(\tau)$ and $C_1(\tau)$. The moments of $V_0(\tau)$ are then found from Equation (3.31). In particular we have for $k = 1$, see also Equations (3.10) and (3.11),

$$\mathbf{E}\left[V_{n,1}(\tau)\right] = \mathbf{E}\left[C_0(\tau)\right] + n\mathbf{E}\left[C_1(\tau)\right], \tag{3.32}$$

$$\mathbf{E}\left[V_{n,0}(\tau)\right] = \mathbf{E}\left[D_0\right] + \mathbf{E}\left[C_0(\tau)\right] + (n + \lambda\mathbf{E}\left[D_0\right])\mathbf{E}\left[C_1(\tau)\right], \tag{3.33}$$

and for $k = 2$,

$$\begin{aligned}
\mathbf{E}\left[V_{n,1}(\tau)^2\right] = {} & \mathbf{E}\left[C_0(\tau)^2\right] + n\mathbf{E}\left[C_1(\tau)^2\right] + 2n\mathbf{E}\left[C_0(\tau)\right]\mathbf{E}\left[C_1(\tau)\right] \\
& + n(n-1)\mathbf{E}\left[C_1(\tau)\right]^2,
\end{aligned} \tag{3.34}$$

$$\begin{aligned}
\mathbf{E}\left[V_{n,0}(\tau)^2\right] = {} & \mathbf{E}\left[D_0^2\right] + 2\mathbf{E}\left[D_0\right](\mathbf{E}\left[C_0(\tau)\right] + n\mathbf{E}\left[C_1(\tau)\right]) \\
& + 2\lambda\mathbf{E}\left[D_0^2\right]\mathbf{E}\left[C_1(\tau)\right] + \mathbf{E}\left[C_0(\tau)^2\right] \\
& + 2(n + \lambda\mathbf{E}\left[D_0\right])\mathbf{E}\left[C_0(\tau)\right]\mathbf{E}\left[C_1(\tau)\right] \\
& + (n + \lambda\mathbf{E}\left[D_0\right])\mathbf{E}\left[C_1(\tau)^2\right] \\
& + \left(n(n-1) + (2n-1)\lambda\mathbf{E}\left[D_0\right] + \lambda^2\mathbf{E}\left[D_0^2\right]\right)\mathbf{E}\left[C_1(\tau)\right]^2.
\end{aligned} \tag{3.35}$$

Using Expressions (3.22), (3.23), (3.26) and (3.27) we have closed-form formulas for these first and second moments.

Theorem 3.5.2 *Let $k \in \mathbb{N}$ be fixed. If $m_k < \infty$ and $\mathbf{E}\left[D_0^k\right] < \infty$ then $\mathbf{E}\left[V_{n,1}(\tau)^k\right]$ and $\mathbf{E}\left[V_{n,0}(\tau)^k\right]$ are polynomials in n of degree k:*

$$\mathbf{E}\left[V_{n,i}(\tau)^k\right] = \sum_{l=0}^{k} c_{k,l}^{(i)}(\tau) n^l, \quad i \in \{0,1\}. \tag{3.36}$$

The coefficients $c_{k,l}^{(1)}(\tau)$ are recursively defined by

$$\begin{aligned} c_{k,0}^{(1)}(\tau) &= \mathbf{E}\left[C_0(\tau)^k\right], \\ c_{k,l+1}^{(1)}(\tau) &= \frac{1}{l+1}\left\{\sum_{i=l+2}^{k}(-1)^{i-l}\binom{i}{l} c_{k,i}^{(1)}(\tau)\right. \\ &\quad \left. + \sum_{j=l}^{k-1}\sum_{i=l}^{j}(-1)^{i-l}\binom{i}{l}\binom{k}{j}\mathbf{E}\left[C_1(\tau)^{k-j}\right] c_{j,i}^{(1)}(\tau)\right\}, \end{aligned} \tag{3.37}$$

with $k \in \mathbb{N}$, and $l = 0, 1, \ldots, k-1$. The empty sum (when $l+2 = k+1$) is equal to zero.

For $k \in \mathbb{N}$, and $l = 0, 1, \ldots, k$, the $c_{k,l}^{(0)}(\tau)$ are given by

$$c_{k,l}^{(0)}(\tau) = \sum_{j=l}^{k}\binom{k}{j} c_{j,l}^{(1)}(\tau)\mathbf{E}\left[\left(D_0 + \sum_{i=1}^{A_0} C_i(\tau)\right)^{k-j}\right]. \tag{3.38}$$

Hence, for $i \in \{0,1\}$, $k \in \mathbb{N}$, and $l \in \{0,1,\ldots,k\}$, the functions $c_{k,l}^{(i)}(\tau)$ are of the same form as $\mathbf{E}\left[C_0(\tau)^k\right]$ in Theorem 3.4.2.

Proof To prove Expression (3.36), for $i = 1$, note that Recursion (3.30) uniquely determines the $\mathbf{E}\left[V_{n,1}(\tau)^k\right]$ for $k, n \in \mathbb{N}$, and that Expression (3.36), for $i = 1$, with the $c_{k,l}^{(1)}(\tau)$ defined by Equation (3.37), satisfies the recursion. Expression (3.36), for $i = 0$, and Relation (3.38) then follow from Relation (3.31).

The last statement follows from the fact that a product of two functions of the class defined by Relation (3.29), one with $k = l_1$, and the other with $k = l_2$, gives a function of the same class, with $k = l_1 + l_2$. □

Sengupta and Jagerman [105, Theorem 1] proved that, in the M/M/1 processor-sharing queue without server breakdowns, the k^{th} moment of the sojourn time conditional on starting with n competing customers, is a polynomial in n of degree k. As a corollary of Theorem 3.5.2 we have that the result of Sengupta and Jagerman is also true for the M/M/1 processor-sharing queue with (generally distributed) server breakdowns.

Corollary 3.5.3 *If $m_k < \infty$ and $\mathbf{E}\left[D_0^k\right] < \infty$ then, for $i \in \{0,1\}$,*

$$\mathbf{E}\left[(V_{n,i})^k\right] := \int_{\tau=0}^{\infty} \mathbf{E}\left[V_{n,i}(\tau)^k\right] \mu \mathrm{e}^{-\mu\tau} \mathrm{d}\tau = \sum_{l=0}^{k} n^l \int_{\tau=0}^{\infty} c_{k,l}^{(i)}(\tau)\mu \mathrm{e}^{-\mu\tau} \mathrm{d}\tau.$$

Proof From the last statement of Theorem 3.5.2, $\int_{\tau=0}^{\infty} c_{k,l}^{(i)}(\tau)\mu \mathrm{e}^{-\mu\tau}\mathrm{d}\tau < \infty$, for $i \in \{0,1\}$. The corollary then follows from Expression (3.36). □

3.6 Sojourn times in steady state

We study the sojourn time distribution of a customer with service requirement τ, arriving to the system in steady state. As before, we denote the number of competing customers in the system and the state of the server at the beginning of the sojourn time by $Z(0)$ and $Y(0)$, respectively. Obviously, in steady state,

$$(Z(0), Y(0)) \stackrel{d}{=} (X, Y),$$

and the distribution of (X, Y) is given by Expressions (3.2) and (3.3).

Theorem 3.6.1 *For $\mathrm{Re}(s) \geq 0$, the LST of the distribution of $V(\tau)$ is given by,*

$$\mathbf{E}\left[\mathrm{e}^{-sV(\tau)} \,|Y(0) = 1\right] = \frac{(\mu - \lambda(1 + \nu m_1))\, g_0(\tau; s)}{\mu - \lambda g_1(\tau; s) - \nu g_1(\tau; s)\frac{1-\phi(\lambda(1-g_1(\tau;s)))}{1-g_1(\tau;s)}}, \quad (3.39)$$

$$\mathbf{E}\left[\mathrm{e}^{-sV(\tau)} \,|Y(0) = 0\right] = \mathbf{E}\left[\mathrm{e}^{-sV(\tau)} \,|Y(0) = 1\right] \frac{1 - \phi(s + \lambda - \lambda g_1(\tau; s))}{m_1\,(s + \lambda - \lambda g_1(\tau; s))} \times \frac{1 - \phi(\lambda - \lambda g_1(\tau; s))}{m_1 \lambda(1 - g_1(\tau; s))}. \quad (3.40)$$

Proof Expression (3.39) is found from Expressions (3.2) and (3.10). To find the sojourn times that start with an off-period, we remark that the residual length of that off-period is distributed as the forward recurrence time of the off-periods, i.e. $\phi_0(s) = \frac{1-\phi(s)}{m_1 s}$. Then using Expressions (3.3) and (3.11) we get Expression (3.40). □

Corollary 3.6.2 *The mean sojourn time is given by*

$$\mathbf{E}\left[V(\tau)\right] = \frac{\tau}{c-\rho} + (1-c)\frac{m_2}{2m_1} + (1-c)\rho\frac{m_2}{2m_1} \times \frac{2c-\rho}{(c-\rho)^2}\left(1 - \mathrm{e}^{-\mu\,(1-\rho/c)\,\tau}\right). \quad (3.41)$$

Proof From Theorem 3.6.1, by differentiating w.r.t. s and putting $s = 0$, we find

$$\begin{aligned}
\mathbf{E}\left[V(\tau)\,|Y(0)=1\right] &= \frac{\tau}{c-\rho} + (1-c)\frac{m_2}{2m_1}\left(\frac{\rho}{c-\rho}\right)^2\left(1-\mathrm{e}^{-\mu\,(1-\rho/c)\,\tau}\right),\\
\mathbf{E}\left[V(\tau)\,|Y(0)=0\right] &= \frac{m_2}{2m_1} + \frac{\tau}{c-\rho} + \frac{m_2}{2m_1} \times \frac{\rho}{c-\rho}\left(2+(1-c)\frac{\rho}{c-\rho}\right)\\
&\qquad \times \left(1-\mathrm{e}^{-\mu\,(1-\rho/c)\,\tau}\right).
\end{aligned}$$

Alternatively, we may find $\mathbf{E}\left[V(\tau)\,|Y(0)=1\right]$ more directly by substituting Expression (3.4) for n in Expression (3.32), and using Expressions (3.22) and (3.23). Similarly, we can find $\mathbf{E}\left[V(\tau)\,|Y(0)=0\right]$ by substituting $\mathbf{E}\left[X\,|Y=0\right]$, given by Expression (3.5), for n in Expression (3.33), and using $\mathbf{E}\left[D_0\right] = \frac{m_2}{2m_1}$. Finally after deconditioning, using $\mathbf{P}\left\{Y=1\right\} = c$ and $\mathbf{P}\left\{Y=0\right\} = 1-c$, we get $\mathbf{E}\left[V(\tau)\right]$. □

As pointed out in Section 1.6, it is well-known that in "standard" processor-sharing queues the conditional mean sojourn time, $\mathbf{E}\left[V(\tau)\right]$, is proportional to the service requirement τ. From Expression (3.41) we conclude that this is not the case with an unreliable server. If we replace the unreliable server by one that works with *constant* capacity c, i.e. the average service capacity of the unreliable server, $\mathbf{E}\left[V(\tau)\right]$ will be equal to $\frac{\tau}{c-\rho}$. This corresponds to the linear term in Expression (3.41). Note that for fixed τ, ρ, and c, Expression (3.41) is fully determined by $\frac{m_2}{2m_1}$, the mean backward recurrence time of the off-periods. $\mathbf{E}\left[V(\tau)\right]$ is minimal for deterministic off-periods, i.e. when $m_2 = (m_1)^2$, and can become arbitrarily large for increasing $\frac{m_2}{2m_1}$.

We conclude this section with two remarks, discussing two cases in which the conditional mean sojourn time is approximately linear in τ.

Remark 3.6.1 $\mathbf{E}\left[V(\tau)\right]$ is "almost linear" in τ when the on- and off-periods alternate rapidly. As in Section 2.6, to make this statement formal, we construct a new sequence of on- and off-periods by dividing each on- and off-period by a scalar $\epsilon \in (0,\infty)$. In the new sequence, the on-periods are exponentially distributed with mean $1/(\epsilon\nu)$, and the distribution of the new off-periods, which are generically denoted by $T_{off}^{(\epsilon)}$, has LST

$$\mathbf{E}\left[\mathrm{e}^{-sT_{off}^{(\epsilon)}}\right] = \phi(s/\epsilon).$$

In particular, the first two moments of $T_{off}^{(\epsilon)}$ are $m_1^{(\epsilon)} = m_1/\epsilon$ and $m_2^{(\epsilon)} = m_2/\epsilon^2$. Obviously, $\epsilon\nu m_1^{(\epsilon)} = \nu m_1$ is independent of ϵ, and so is the probability that the server is on (with the new sequence of on- and off-periods). Therefore the ergodicity condition remains unchanged. With the new on- and off-periods, let $V^{(\epsilon)}(\tau)$ be the sojourn time of a customer with service requirement τ, then

$$\lim_{\epsilon \to \infty} \mathbf{E}\left[V^{(\epsilon)}(\tau)\right] = \frac{\tau}{c-\rho}.$$

Recall that this limiting case ($\epsilon \to \infty$) corresponds to the case where the server is always available and works at the constant speed c, see the discussion in Section 2.6.

On the other hand, when the server alternates very slowly, the expected sojourn time can become arbitrarily large (irrespective of the service requirement of the customer):

$$\lim_{\epsilon \downarrow 0} \epsilon \mathbf{E}\left[V^{(\epsilon)}(\tau)\right] = (1-c)\frac{m_2}{2m_1}\left(1+\rho\frac{2c-\rho}{(c-\rho)^2}\left(1-\mathrm{e}^{-\mu\,(1-\rho/c)\,\tau}\right)\right).$$

Remark 3.6.2 From Expression (3.41) we also conclude that $\mathbf{E}[V(\tau)]$ is approximately linear for large τ. This can intuitively be explained by noting that if τ is large, then also the sojourn time will be large. Over a long period of time, the fluctuations in the server availability average out, and for large τ an additional amount of work $\Delta\tau$ requires $\frac{1}{c-\rho}\Delta\tau$ time units. The term $c-\rho$ can be seen as the average speed at which the permanent customer receives service, when the system *with the permanent customer* is in steady state: The average service capacity is c per time unit, and on average an amount of capacity ρ per time unit is required to serve other customers (since the system with a permanent customer is ergodic, all non-permanent customers eventually leave the system). In the next section we study the case with $\tau \to \infty$ in greater detail.

3.7 Asymptotic analysis for $\tau \to \infty$.

We study the behaviour of $g_1(\tau; s)$ as $\tau \to \infty$. Then we use these asymptotics to show the convergence of $\frac{V(\tau)}{\tau}$ for $\tau \to \infty$. Our starting point is Relation (3.18). By partial fraction expansion,

$$\frac{1}{\mu-(s+\lambda+\mu+\nu)x+\lambda x^2+\nu x\phi(s+\lambda-\lambda x)} = \frac{k_1(s)}{x-r_1(s)} + k_2(x;s), \quad (3.42)$$

where

$$k_1(s) := \lim_{x \to r_1(s)} \frac{x-r_1(s)}{\mu-(s+\lambda+\mu+\nu)x+\lambda x^2+\nu x\phi(s+\lambda-\lambda x)}, \quad (3.43)$$

exists and the function $k_2(x;s)$ is analytic in x, for $|x| \le 1$ and $\mathrm{Re}(s) \ge 0$. Using Equation (3.42) in Relation (3.18) we get, for $s > 0$,

$$k_1(s)\ln(g_1(\tau;s)-r_1(s)) + \int_{x=1}^{g_1(\tau;s)} k_2(x;s)\mathrm{d}x = k_1(s)\ln(1-r_1(s)) + \tau. \quad (3.44)$$

If we let $\tau \to \infty$ in Equation (3.44), we may conclude that

$$\lim_{\tau\to\infty} g_1(\tau; s) = r_1(s), \qquad s > 0. \tag{3.45}$$

This is an immediate consequence of the analyticity of $k_2(x; s)$ in x and the boundedness of $g_1(\tau; s)$, which imply that the second term on the left-hand side of Equation (3.44) is bounded. In Remark 3.7.1 we discuss how this limiting property can be obtained probabilistically in our model.

Remark 3.7.1 If we concentrate on a non-permanent element of the population model of Section 3.2 *and* his offspring (we call this a *family*), then under the ergodicity condition $\rho < c$, this family dies out with probability 1. Consider the reward that this family generates until its extinction. This reward is equal to the sum of the life times of all the members of this family *plus* the reward of all nests in this family. By assigning the reward of a nest to the individual that generated it, and concatenating the life times of all family members, it can be seen that the total reward of this family is distributed as a clearing period of the model of Section 3.1:

$$\lim_{\tau\to\infty} C_1(\tau) \stackrel{d}{=} CP.$$

This corresponds to Equation (3.45).

Further exploiting Equation (3.44), we can carry our asymptotic analysis one step further: For $s > 0$,

$$\lim_{\tau\to\infty} \left\{ k_1(s) \ln \left(\frac{g_1(\tau; s) - r_1(s)}{1 - r_1(s)} \right) - \tau \right\} = -\int_{x=1}^{r_1(s)} k_2(x; s) \mathrm{d}x. \tag{3.46}$$

Using Equation (3.46) we can prove the following Lemma:

Lemma 3.7.1 *For $s > 0$,*

$$\lim_{\tau\to\infty} \int_{u=0}^{\tau} \left(g_1(u; \frac{s}{\tau}) - r_1(\frac{s}{\tau}) \right) \mathrm{d}u = 0,$$

and consequently,

$$\lim_{\tau\to\infty} \int_{u=0}^{\tau} \left(\phi(\frac{s}{\tau} + \lambda - \lambda g_1(u; \frac{s}{\tau})) - \phi(\frac{s}{\tau} + \lambda - \lambda r_1(\frac{s}{\tau})) \right) \mathrm{d}u = 0.$$

Proof See Appendix 3.D. □

Theorem 3.7.2 *For $s \geq 0$,*

$$\lim_{\tau\to\infty} g_0(\tau; \frac{s}{\tau}) = \mathrm{e}^{-\frac{s}{c-\rho}},$$

and hence,

$$\frac{C_0(\tau)}{\tau} \xrightarrow{\mathrm{P}} \frac{1}{c - \rho},$$

as $\tau \to \infty$. Here $\xrightarrow{\mathrm{P}}$ denotes convergence in probability.

Proof Using the first part of Lemma 3.7.1 we can write, for $s \geq 0$,

$$\lim_{\tau\to\infty} \int_{u=0}^{\tau} \Big(1 - g_1(u; \frac{s}{\tau})\Big)\, \mathrm{d}u = \lim_{\tau\to\infty} \tau \Big(1 - r_1(\frac{s}{\tau})\Big) = \frac{s}{\mu(c-\rho)},$$

where we use that $\lim_{s\downarrow 0} \frac{1-r_1(s)}{s} = \mathbf{E}\,[CP]$. We can find $\mathbf{E}\,[CP] = \frac{1/\mu}{c-\rho}$ from Relation (3.17). Similarly, using the second part of Lemma 3.7.1 we have, again for $s \geq 0$,

$$\begin{aligned} \lim_{\tau\to\infty} \int_{u=0}^{\tau} \Big(1 - \phi(\frac{s}{\tau} + \lambda - \lambda g_1(u; \frac{s}{\tau}))\Big)\, \mathrm{d}u &= \lim_{\tau\to\infty} \tau \Big(1 - \phi(\frac{s}{\tau} + \lambda - \lambda r_1(\frac{s}{\tau}))\Big) \\ &= \frac{s m_1 c}{c-\rho}. \end{aligned}$$

Using this in Relation (3.16), gives the convergence in distribution by the continuity theorem for LSTs of probability distributions, see Feller [30, Theorem 2, p. 408]. The convergence in probability then follows immediately, because the limit is a constant. □

Using Formulas (3.10) and (3.11), Theorem 3.7.2 immediately gives the following corollary. The result is in agreement with Remark 3.6.2.

Corollary 3.7.3 *The sojourn time $V(\tau)$ of a customer with service requirement τ, satisfies*

$$\frac{V(\tau)}{\tau} \xrightarrow{\mathrm{P}} \frac{1}{c-\rho},$$

as $\tau \to \infty$.

Remark 3.7.2 Using the Renewal Reward Theorem, see for instance Tijms [112, Theorem 1.3.1], it can be shown that the convergence of $\frac{V(\tau)}{\tau}$, and $\frac{C_0(\tau)}{\tau}$, is in fact convergence with probability 1. To see this, note that $N''(\tau)$, the process counting the number of elements in the population $\mathcal{P}$ at time τ, is regenerative. The regeneration points can be taken to be the times at which the permanent element *becomes* the only element of the population. It can then be shown that the lengths of the regeneration cycles have a finite expectation.

Remark 3.7.3 In addition to Theorem 3.7.2 and Corollary 3.7.3, it can be shown that

$$\frac{V(\tau) - C_0(\tau)}{\tau} \xrightarrow{\mathrm{P}} 0, \quad \tau \to \infty.$$

This is a consequence of Theorem 3.2.5, Corollary 3.2.6, and Remark 3.7.1.

3.8 Heavy traffic

In Section 3.3 we have seen that the distributions of $C_0(\tau)$ and $C_1(\tau)$ are not easy to compute. Hence, the same holds for the distribution of $V(\tau)$. We now study the sojourn time of a customer conditional on the service requirement in heavy traffic, that is for $\rho \uparrow c$. It turns out that $V(\tau)$ scaled by a factor $1-\rho/c$ has a proper limiting distribution as $\rho \uparrow c$. This limiting distribution provides a natural approximation of the distribution of $V(\tau)$ when ρ is close to c. The investigation of the quality of this approximation is a topic for further research. Before stating the main result of this section, we establish the following limits:

Lemma 3.8.1 *For* $\mathrm{Re}(s) \geq 0$,

(i) $\lim_{\rho\uparrow c} g_1(\tau;(1-\rho/c)s) = 1$ *and* $\lim_{\rho\uparrow c} g_0(\tau;(1-\rho/c)s) = 1$,

(ii) $\lim_{\rho\uparrow c} \frac{1-g_1(\tau;(1-\rho/c)s)}{1-\rho/c} = (1+\nu m_1)s\tau$.

Proof See Appendix 3.E. □

Note that statement (ii) in the lemma can be rewritten in terms of the LST of the distribution of the backward (or forward) recurrence time of $C_1(\tau)$:

$$\lim_{\rho\uparrow c} \frac{1-g_1(\tau;(1-\rho/c)s)}{\mathbf{E}\left[C_1(\tau)\right](1-\rho/c)s} = 1,$$

see Formula (3.22).

With Lemma 3.8.1 we can prove the main result of this section which is stated in the next theorem.

Theorem 3.8.2 *Provided that the second moment of the off-periods, m_2, is finite,*

$$\lim_{\rho\uparrow c} \mathbf{E}\left[\mathrm{e}^{-(1-\rho/c)sV(\tau)}\right] = \frac{1}{1+(1+\nu m_1+\frac{1}{2}\nu\lambda m_2)s\tau}, \qquad \mathrm{Re}(s) \geq 0.$$

Proof Using Lemma 3.8.1, the theorem can be proved by substituting $(1-\rho/c)s$ for s in Expressions (3.39) and (3.40), and letting $\rho \uparrow c$. □

As $\rho \uparrow c$, the distribution of $(1-\rho/c)V(\tau)$ converges to the exponential distribution with mean $\left(1+\nu m_1+\frac{1}{2}\nu\lambda m_2\right)\tau$. For the ordinary M/G/1 processor-sharing queue (without service interruptions) it is already known that the heavy-traffic limit is an exponential distribution, see Sengupta [104] and Yashkov [124].

Remark 3.8.1 In Remarks 3.6.1 and 3.6.2 we observed that the conditional mean sojourn time in steady state is "approximately linear" when the on- and off-periods alternate rapidly and when τ is large. From Theorem 3.8.2 we conclude that $(1-\rho/c)\mathbf{E}\left[V(\tau)\right]$ is also approximately linear in τ for ρ close to c. This can also be derived from Expression (3.41).

3.9 Concluding remarks

We studied the sojourn times of customers in the M/M/1 queue with processor-sharing service discipline, and the server alternating between exponentially distributed on-periods and generally distributed off-periods. By using a time-scale transformation, we formulated the problem in terms of a branching process with a reward structure associated with it. The sojourn time $V(\tau)$ of a customer, conditional on his service requirement τ, was decomposed into a sum of independent "fundamental" random variables. We indicated how the same transformation can be applied to the case with generally distributed service requirements which allows to generalise the decomposition result known for the standard M/G/1 queue with processor sharing, see also Theorem 5.3.2. However, for generally distributed service requirements the steady-state distribution of the queue length — and the attained or remaining service requirements of the customers in the queue — is not known. Therefore in that case the analysis of sojourn times in steady state is more complicated, see Chapter 5.

For exponentially distributed service requirements, the LSTs of the distributions of the fundamental random variables were characterised through an integral equation. We computed the first two moments of the fundamental random variables, and identified the structure of higher moments. We used these to find the moments of $V(\tau)$, conditional on the number of competing customers, and generalised a result of Sengupta and Jagerman [105, Theorem 1]. We gave a closed-form expression for the LST of the sojourn time distribution in steady state, in terms of the LSTs of the distributions of the fundamental random variables. The mean of the steady-state sojourn times was found in terms of the input parameters. We further studied asymptotics of the queueing model. First we analysed the case for $\tau \to \infty$, proving that $V(\tau)/\tau$ converges (with probability 1) to a constant. Then we proved that under heavy-traffic conditions, that is for the traffic load $\rho \uparrow c$, the scaled sojourn time $(1-\rho/c)V(\tau)$ converges in distribution to an exponential one, of which the mean is linear in τ.

A crucial observation is that $\mathbf{E}[V(\tau)]$ is not proportional to τ, unlike in processor-sharing queues without service interruptions. We saw that $\mathbf{E}[V(\tau)]$ is approximately (asymptotically) linear in three cases: (i) when the on- and off-periods alternate rapidly, (ii) when τ is large, and (iii) in heavy traffic. An intuitive explanation for this linearity in all three cases is that the sojourn times are large compared to the lengths of the on- and off-periods, so that fluctuations in the service availability average out.

The obtained closed-form results are the basis for the analysis in the next chapter, where we consider more general service rate fluctuations. In particular we will be interested in extending the asymptotic (structural) properties of the sojourn times obtained in this chapter. Special attention will also be devoted to the fact that, when the service rate is not constant, there is no proportionality between the conditional mean sojourn time and the service requirement.

Appendix

3.A Proof of Lemma 3.3.1

Lemma *For* $\mathrm{Re}(s) \geq 0$ *and* $\tau \geq 0$*,* $g_0(\tau; s)$ *and* $g_1(\tau; s)$ *are uniquely determined by the following set of differential equations,*

$$\begin{aligned}
\frac{\partial}{\partial \tau} g_1(\tau; s) &= -(s + \lambda + \mu + \nu)\, g_1(\tau; s) + \lambda \{g_1(\tau; s)\}^2 + \mu \\
&\quad + \nu g_1(\tau; s)\phi(s + \lambda(1 - g_1(\tau; s))), \\
\frac{\partial}{\partial \tau} g_0(\tau; s) &= -(s + \lambda + \nu)\, g_0(\tau; s) + \lambda g_0(\tau; s) g_1(\tau; s) \\
&\quad + \nu g_0(\tau; s)\phi(s + \lambda(1 - g_1(\tau; s))),
\end{aligned}$$

and initial conditions,

$$g_0(0; s) = g_1(0; s) = 1.$$

Proof By conditioning on the number of "single" children and the number of nests that a non-permanent element in the population model generates in a time interval of length Δ, as well as on the survival probability of the element itself in that interval, we get,

$$\begin{aligned}
g_1(\tau + \Delta; s) &= \int_{t=0}^{\Delta} \mu \mathrm{e}^{-\mu t} \mathrm{e}^{-st} \sum_{m=0}^{\infty} \mathrm{e}^{-\lambda t} \frac{(\lambda t)^m}{m!} \left(\frac{1}{t} \int_{u=0}^{t} g_1(\tau + u; s) \mathrm{d}u \right)^m \\
&\quad \times \sum_{n=0}^{\infty} \mathrm{e}^{-\nu t} \frac{(\nu t)^n}{n!} \left(\frac{1}{t} \int_{u=0}^{t} \phi(s + \lambda(1 - g_1(\tau + u; s))) \mathrm{d}u \right)^n \mathrm{d}t \\
&+ \mathrm{e}^{-(\mu + \lambda + s)\Delta} g_1(\tau; s) \sum_{m=0}^{\infty} \frac{(\lambda \Delta)^m}{m!} \left(\int_{u=0}^{\Delta} \frac{g_1(\tau + u; s)}{\Delta} \mathrm{d}u \right)^m \\
&\quad \times \sum_{n=0}^{\infty} \mathrm{e}^{-\nu \Delta} \frac{(\nu \Delta)^n}{n!} \left(\int_{u=0}^{\Delta} \frac{\phi(s + \lambda(1 - g_1(\tau + u; s)))}{\Delta} \mathrm{d}u \right)^n.
\end{aligned}$$

Here we use the fact that "Poisson arrivals occur homogeneously in time', see for instance Tijms [112, Theorem 1.2.5]. Note that $\phi(s + \lambda(1 - g_1(\tau; s)))$ is the LST of the distribution of the reward of a nest plus the rewards of all children in that nest and their offspring, until time τ. Equivalently we may write,

$$\begin{aligned}
g_1(\tau + \Delta; s) &= \\
&\int_{t=0}^{\Delta} \mu \exp\left\{ -\mu t - st - \lambda \left(t - \int_{u=0}^{t} g_1(\tau + u; s) \mathrm{d}u \right) \right. \\
&\qquad \left. - \nu \left(t - \int_{u=0}^{t} \phi(s + \lambda(1 - g_1(\tau + u; s))) \mathrm{d}u \right) \right\} \mathrm{d}t \\
&+ g_1(\tau; s) \exp\left\{ -\mu \Delta - s\Delta - \lambda \left(\Delta - \int_{u=0}^{\Delta} g_1(\tau + u; s) \mathrm{d}u \right) \right.
\end{aligned}$$

$$-\nu\left(\Delta-\int_{u=0}^{\Delta}\phi(s+\lambda(1-g_1(\tau+u;s)))\mathrm{d}u\right)\Bigg\}. \quad (3.47)$$

By similar arguments we also find,

$$\begin{aligned}g_0(\tau+\Delta;s) \;=&\\ &g_0(\tau;s)\exp\left\{-s\Delta-\lambda\left(\Delta-\int_{u=0}^{\Delta}g_1(\tau+u;s)\mathrm{d}u\right)\right.\\ &\left.-\nu\left(\Delta-\int_{u=0}^{\Delta}\phi(s+\lambda(1-g_1(\tau+u;s)))\mathrm{d}u\right)\right\}. \quad (3.48)\end{aligned}$$

From Equations (3.47) and (3.48) we can show that, for $\Delta\downarrow 0$,

$$\begin{aligned}g_1(\tau+\Delta;s) &= (1-(s+\lambda+\mu+\nu)\,\Delta)\,g_1(\tau;s)+\lambda\Delta\,\{g_1(\tau;s)\}^2+\mu\Delta\\ &\quad+\nu\Delta g_1(\tau;s)\phi(s+\lambda(1-g_1(\tau;s)))+\mathrm{o}(\Delta), \quad (3.49)\\ g_0(\tau+\Delta;s) &= (1-(s+\lambda+\nu)\,\Delta)\,g_0(\tau;s)+\lambda\Delta g_0(\tau;s)g_1(\tau;s)\\ &\quad+\nu\Delta g_0(\tau;s)\phi(s+\lambda(1-g_1(\tau;s)))+\mathrm{o}(\Delta). \quad (3.50)\end{aligned}$$

With Equations (3.49) and (3.50) it is immediate that $g_1(\tau;s)$ and $g_0(\tau;s)$ are continuous from the right in τ. If we replace τ in Equations (3.49) and (3.50) by $\tau-\Delta$, the continuity from the left in τ also easily follows. Subsequently it can be shown that, for $i\in\{0,1\}$,

$$\lim_{\Delta\downarrow 0}\frac{g_i(\tau+\Delta;s)-g_i(\tau;s)}{\Delta}=\lim_{\Delta\downarrow 0}\frac{g_i(\tau;s)-g_i(\tau-\Delta;s)}{\Delta},$$

so that $\frac{\partial}{\partial\tau}g_1(\tau;s)$ and $\frac{\partial}{\partial\tau}g_0(\tau;s)$ exist and satisfy the differential equations stated in the lemma. The initial condition follows from $C_0(0)=C_1(0)=0$. □

3.B Proof of Theorem 3.4.1

Theorem *If the k^{th} moment of the off-periods, m_k, exists, then so do the k^{th} moments of $C_0(\tau)$ and $C_1(\tau)$ exist.*

Proof It is known for the M/G/1 queue that the k^{th} moment of the busy period distribution exists if and only if the k^{th} moment of the service time distribution exists, see De Meyer and Teugels [72, Lemma 3]. With Lemma 3.3.3, this implies that the k^{th} moment of the clearing period exists, if and only if $m_k<\infty$. Since $C_1(\tau)$ is non-decreasing in τ with probability 1, and $C_1(\tau)$ converges to the clearing period CP, as $\tau\to\infty$, it must be that

$$\mathbf{E}\left[C_1(\tau)^k\right]\le\mathbf{E}\left[CP^k\right],$$

($C_1(\tau)$ is stochastically smaller than CP), and hence the k^{th} moment of $C_1(\tau)$ exists when $m_k<\infty$.

To prove the result for $C_0(\tau)$, we first write the following identity:

$$C_0(\tau) = \sum_{i=1}^{N^{(\lambda)}(\tau)} C_i(\tau - T_i^{(\lambda)}) + \sum_{j=1}^{N^{(\nu)}(\tau)} D_j + \sum_{n=1}^{N^{(\lambda)}(D_j)} C_{j,n}(\tau - T_j^{(\nu)}).$$

Here, $N^{(\lambda)}(\tau)$ is the number of "regular" children that the permanent element, in the population $\mathcal{P}$, generates (at rate λ) over a time span of length τ. $T_i^{(\lambda)}$ is the time at which the i^{th} regular child is born, and $C_i(\tau - T_i^{(\lambda)})$ is the reward of this child and his offspring until time τ. Similarly, $N^{(\nu)}(\tau)$ is the number of batches of children of the permanent element (generated at rate ν) until time τ. D_j is the direct reward of the j^{th} batch, $N^{(\lambda)}(D_j)$ is the number of children in the j^{th} batch, $T_j^{(\nu)}$ is the time at which the j^{th} batch is generated, and $C_{j,n}(\tau - T_j^{(\nu)})$ is the reward of the n^{th} child in the j^{th} batch and his offspring, until time τ. The above identity was given in terms of LSTs in Relation (3.16).

If we replace each of the rewards until time τ associated with a child of the permanent customer and his offspring, by the reward of the family of that child over a total time-span of length τ, we clearly have an upper bound for $C_0(\tau)$:

$$C_0(\tau) \leq \overline{C_0}(\tau) := \sum_{i=1}^{N^{(\lambda)}(\tau)} C_i(\tau) + \sum_{j=1}^{N^{(\nu)}(\tau)} \left(D_j + \sum_{n=1}^{N^{(\lambda)}(D_j)} C_{j,n}(\tau) \right).$$

For $\text{Re}(s) > 0$, the LST of the distribution of $\overline{C_0}(\tau)$ is given by,

$$\begin{aligned} \mathbf{E}\left[\mathrm{e}^{-s\overline{C_0}(\tau)}\right] &= \exp\{-\tau\,(s + \lambda\,(1 - g_1(\tau; s)) \\ &\quad + \nu\,[1 - \phi(s + \lambda - \lambda g_1(\tau; s))])\}. \end{aligned} \tag{3.51}$$

If $m_k < \infty$, and hence by the first part of the theorem $\mathbf{E}\left[C_1(\tau)^k\right] < \infty$, we can write, for $s \downarrow 0$,

$$\begin{aligned} \phi(s) &= 1 + \sum_{i=1}^{k} m_i \frac{(-s)^i}{i!} + \mathrm{o}(s^k), \\ g_1(\tau; s) &= 1 + \sum_{j=1}^{k} \mathbf{E}\left[C_1(\tau)^j\right] \frac{(-s)^j}{j!} + \mathrm{o}(s^k), \end{aligned} \tag{3.52}$$

see De Meyer and Teugels [72, Lemma 1]. Combining these, we get,

$$\begin{aligned} &\phi(s + \lambda - \lambda g_1(\tau; s)) = \\ &\quad 1 + \sum_{i=1}^{k} m_i \frac{\left(-s + \lambda \sum_{j=1}^{k} \mathbf{E}\left[C_1(\tau)^j\right] \frac{(-s)^j}{j!}\right)^i}{i!} + \mathrm{o}(s^k). \end{aligned} \tag{3.53}$$

From Equation (3.51) it is now straightforward to see that the LST of the distribution of $\overline{C_0}(\tau)$ has a finite k^{th} derivative in $s = 0$. Therefore, the k^{th} moment of $\overline{C_0}(\tau)$, and hence the k^{th} moment of $C_0(\tau)$, exists. □

3.C Proof of Theorem 3.4.2

Theorem *For $k \geq 1$, provided that $m_k < \infty$, and thus $\mathbf{E}\left[C_1(\tau)^k\right] < \infty$ and $\mathbf{E}\left[C_0(\tau)^k\right] < \infty$,*

$$\begin{aligned}
\mathbf{E}\left[C_1(\tau)^k\right] &= \alpha_0^{(k)} + \sum_{m=1}^{k} \mathrm{e}^{-m\mu\,(1-\rho/c)\,\tau} \sum_{n=0}^{k-m} \alpha_{m,n}^{(k)} \tau^n, && (3.54)\\
\mathbf{E}\left[C_0(\tau)^k\right] &= \sum_{m=0}^{k} \mathrm{e}^{-m\mu\,(1-\rho/c)\,\tau} \sum_{n=0}^{k-m} \beta_{m,n}^{(k)} \tau^n, && (3.55)
\end{aligned}$$

where the $\alpha_0^{(k)}$, $\alpha_{m,n}^{(k)}$ and $\beta_{m,n}^{(k)}$ are coefficients that are independent of τ.

Proof Let T_{off} be as before and $N(T_{off})$ be the number of Poisson arrivals (with rate λ) during the period T_{off}. If $C_1(\tau), C_2(\tau), \ldots$ is an i.i.d. sequence with LST of its distribution $g_1(\tau; s)$, then using Equations (3.52) and (3.53),

$$\begin{aligned}
&\mathbf{E}\left[\mathrm{e}^{-s\left(T_{off} + C_1(\tau) + \ldots C_{1+N(T_{off})}(\tau)\right)}\right]\\
&= g_1(\tau;s)\phi(s+\lambda-\lambda g_1(\tau;s))\\
&= \left(\sum_{l=0}^{k}(-s)^l \frac{\mathbf{E}\left[C_1(\tau)^l\right]}{l!} + \mathrm{o}(s^k)\right)\\
&\quad \times \left(\sum_{n=0}^{k} \frac{m_n}{n!}\left(s - \lambda \sum_{l=1}^{k}(-s)^l \frac{\mathbf{E}\left[C_1(\tau)^l\right]}{l!}\right)^n + \mathrm{o}(s^k)\right)\\
&= 1 + \frac{s}{\lambda} + \frac{1}{\lambda}\sum_{i=1}^{k}\left((s+\lambda)\frac{m_i}{i!} + \frac{m_{i-1}}{(i-1)!}\right)\left(-s + \lambda\sum_{j=1}^{k}\mathbf{E}\left[C_1(\tau)^j\right]\frac{(-s)^j}{j!}\right)^i\\
&\quad +\mathrm{o}(s^k). && (3.56)
\end{aligned}$$

We write out the terms in the summation as,

$$\begin{aligned}
&\left(-s + \lambda\sum_{j=1}^{k}\mathbf{E}\left[C_1(\tau)^j\right]\frac{(-s)^j}{j!}\right)^i\\
&= \sum_{i_0+i_1+\ldots+i_k=i} \binom{i}{i_0,\ldots,i_k} (-s)^{i_0} \prod_{j=1}^{k}\left(\lambda\mathbf{E}\left[C_1(\tau)^j\right]\frac{(-s)^j}{j!}\right)^{i_j}\\
&= \sum_{n=0}^{k}(-s)^n \sum_{(i_0,\ldots,i_k)\in S_{k,i,n}} \binom{i}{i_0,\ldots,i_k} \prod_{j=1}^{k}\left(\lambda\mathbf{E}\left[C_1(\tau)^j\right]\frac{1}{j!}\right)^{i_j}\\
&\quad +\mathrm{o}(s^k), && (3.57)
\end{aligned}$$

where,

$$S_{k,i,n} := \left\{ (i_0, i_1, \ldots, i_k) \in \mathbb{N}_0^{k+1} : \sum_{j=0}^{k} i_j = i, i_0 + \sum_{j=1}^{k} j i_j = n \right\}.$$

Note that there are combinations of $k, i, n \in \mathbb{N}$, for which $S_{k,i,n} = \emptyset$.

We now prove the theorem by induction on k. From Equations (3.56) and (3.57) it follows that if $\mathbf{E}\left[C_1(\tau)^j\right]$ has the form of Expression (3.54) for $j = 1, 2, \ldots, k-1$, then:

$$\begin{aligned}
&\mathbf{E}\left[\left(T_{off} + C_1(\tau) + \ldots + C_{1+N(T_{off})}(\tau)\right)^k\right] \\
&= (\lambda m_1 + 1)\mathbf{E}\left[C_1(\tau)^k\right] + \gamma_0^{(k)} + \sum_{m=1}^{k} \mathrm{e}^{-m\mu(1-\rho/c)\tau} \sum_{n=0}^{k-m} \gamma_{m,n}^{(k)} \tau^n.
\end{aligned}$$

This can be verified by noting that the only contribution of $\mathbf{E}\left[C_1(\tau)^k\right]$ to the coefficient of $(-s)^k$ in Equation (3.56), is through the term with $i = 1$. All other contributions to the coefficient of $(-s)^k$ are either zero, or come from products of the $\mathbf{E}\left[C_1(\tau)^j\right]$, for $j = 1, 2, \ldots, k-1$. Apart from a constant in τ, they all consist of terms of the form $\tau^n \mathrm{e}^{-m\mu(1-\rho/c)\tau}$, with $m \geq 1$, $n \geq 0$ and $m+n \leq k$. Writing out the terms, it is seen that $\gamma_{1,k-1}^{(k)} = 0$. This is a consequence of the fact that for $l_1, l_2 = 1, 2, \ldots$, the product $\mathbf{E}\left[C_1(\tau)^{l_1}\right] \times \mathbf{E}\left[C_1(\tau)^{l_2}\right]$ is of the same form as $\mathbf{E}\left[C_1(\tau)^{l_1+l_2}\right]$ in Expression (3.54), except for the terms containing $\tau^n \mathrm{e}^{-\mu(1-\rho/c)\tau}$, with $n \geq \max(l_1, l_2)$, which do not appear. The other coefficients $\gamma_{m,n}^{(k)}$ can be found from the $\alpha_{m,n}^{(j)}$ for $j < k$, by use of Equations (3.56) and (3.57).

As before, we can derive a differential equation for $\mathbf{E}\left[C_1(\tau)^k\right]$:

$$\begin{aligned}
\frac{\partial}{\partial\tau}\mathbf{E}\left[C_1(\tau)^k\right] &= -(\lambda+\mu+\nu)\mathbf{E}\left[C_1(\tau)^k\right] + k\mathbf{E}\left[C_1(\tau)^{k-1}\right] \\
&\quad + \lambda\mathbf{E}\left[\left(C_1(\tau) + C_2(\tau)\right)^k\right] \\
&\quad + \nu\mathbf{E}\left[\left(T_{off} + C_1(\tau) + \ldots + C_{1+N(T_{off})}(\tau)\right)^k\right] \\
&= -\{\mu - \lambda(1+\nu m_1)\}\mathbf{E}\left[C_1(\tau)^k\right] + k\mathbf{E}\left[C_1(\tau)^{k-1}\right] \\
&\quad + \lambda\sum_{l=1}^{k-1} \binom{k}{l} \mathbf{E}\left[C_1(\tau)^l\right]\mathbf{E}\left[C_1(\tau)^{k-l}\right] \\
&\quad + \nu\gamma_0^{(k)} + \nu\mathrm{e}^{-\mu(1-\rho/c)\tau} \sum_{n=0}^{k-2} \gamma_{1,n}^{(k)} \tau^n \\
&\quad + \nu\sum_{m=2}^{k} \mathrm{e}^{-m\mu(1-\rho/c)\tau} \sum_{n=0}^{k-m} \gamma_{m,n}^{(k)} \tau^n.
\end{aligned}$$

Note that in the right-hand side of this differential equation, no term with $\mathrm{e}^{-\mu(1-\rho/c)\tau}\tau^{k-1}$ appears. Solving for $\mathbf{E}\left[C_1(\tau)^k\right]$, indeed leads to the form of Relation (3.54). The coefficients $\alpha_0^{(k)}$ and $\alpha_{m,n}^{(k)}$ are recursively determined by the $\alpha_0^{(j)}$ and $\alpha_{m,n}^{(j)}$ for $j<k$.

To prove the second part of the theorem we use the differential equation for $\mathbf{E}\left[C_0(\tau)^k\right]$:

$$\begin{aligned}\frac{\partial}{\partial\tau}\mathbf{E}\left[C_0(\tau)^k\right] &= -(\lambda+\nu)\mathbf{E}\left[C_0(\tau)^k\right]+\lambda\mathbf{E}\left[(C_0(\tau)+C_1(\tau))^k\right]\\ &\quad +k\mathbf{E}\left[C_0(\tau)^{k-1}\right]+\nu\mathbf{E}\left[\left(T_{off}+C_0(\tau)+\sum_{i=1}^{N(T_{off})}C_i(\tau)\right)^k\right]\\ &= k\mathbf{E}\left[C_0(\tau)^{k-1}\right]+\lambda\sum_{l=0}^{k-1}\binom{k}{l}\mathbf{E}\left[C_0(\tau)^l\right]\mathbf{E}\left[C_1(\tau)^{k-l}\right]\\ &\quad +\nu\sum_{l=0}^{k-1}\binom{k}{l}\mathbf{E}\left[C_0(\tau)^l\right]\mathbf{E}\left[\left(T_{off}+\sum_{i=1}^{N(T_{off})}C_i(\tau)\right)^{k-l}\right].\end{aligned}$$

By similar arguments as before, we find Relation (3.55). □

3.D Proof of Lemma 3.7.1

Lemma *For $s>0$,*

$$\lim_{\tau\to\infty}\int_{u=0}^{\tau}\left(g_1(u;\frac{s}{\tau})-r_1(\frac{s}{\tau})\right)\mathrm{d}u=0,$$

and consequently,

$$\lim_{\tau\to\infty}\int_{u=0}^{\tau}\left(\phi(\frac{s}{\tau}+\lambda-\lambda g_1(u;\frac{s}{\tau}))-\phi(\frac{s}{\tau}+\lambda-\lambda r_1(\frac{s}{\tau}))\right)\mathrm{d}u=0.$$

Proof Using Relation (3.44) we may write:

$$\begin{aligned}&\int_{u=0}^{\tau}\left(g_1(u;\frac{s}{\tau})-r_1(\frac{s}{\tau})\right)\mathrm{d}u\\ &\quad=\left(1-r_1(\frac{s}{\tau})\right)\int_{u=0}^{\tau}\exp\left\{\frac{1}{k_1(\frac{s}{\tau})}\left(u-\int_{x=1}^{g_1(u;\frac{s}{\tau})}k_2(x;\frac{s}{\tau})\mathrm{d}x\right)\right\}\mathrm{d}u.\end{aligned}$$

It is clear from Definition (3.43) that $k_1(s)<0$, for $s>0$: for $x=0$ the numerator on the right-hand side of (3.43) is negative and the denominator is positive, and as $x\uparrow r_1(s)$ neither the numerator nor the denominator changes sign.

For $s > 0$, let $M(s) \in [r_1(s), 1]$ be such that $\int_{x=1}^{M(s)} k_2(x; s)\mathrm{d}x$ is maximal. Then we may write,

$$\begin{aligned}
0 &\leq \int_{u=0}^{\tau} \left(g_1(u; \frac{s}{\tau}) - r_1(\frac{s}{\tau})\right) \mathrm{d}u \\
&\leq \left(1 - r_1(\frac{s}{\tau})\right) \mathrm{e}^{\frac{-1}{k_1(\frac{s}{\tau})} \int_{x=1}^{M(\frac{s}{\tau})} k_2(x; \frac{s}{\tau})\mathrm{d}x} k_1(\frac{s}{\tau}) \left(\mathrm{e}^{\frac{\tau}{k_1(\frac{s}{\tau})}} - 1\right). \quad (3.58)
\end{aligned}$$

Now, if we take $\tau \to \infty$ then $r_1(\frac{s}{\tau})$ and $M(\frac{s}{\tau})$ go to 1, $k_2(x; \frac{s}{\tau})$ remains bounded for $r_1(\frac{s}{\tau}) \leq x \leq 1$, and

$$\lim_{s \downarrow 0} k_1(s) = \frac{-1/\mu}{1 - \rho/c}.$$

Thus, if we let $\tau \to \infty$ in Relation (3.58) then its upper bound goes to 0.

The second part of the lemma follows from the first part by noting that, since it is a LST, $\phi(s)$ is a decreasing and convex function for $s \geq 0$, and $\frac{\mathrm{d}}{\mathrm{d}s}\phi(s)|_{s=0} = -m_1$. Therefore it holds that $\phi(s_1) - \phi(s_2) \leq m_1(s_2 - s_1)$, whenever $0 \leq s_1 \leq s_2$. □

3.E Proof of Lemma 3.8.1

Lemma *For* $\mathrm{Re}(s) \geq 0$,

(i) $\lim_{\rho\uparrow c} g_1(\tau; (1 - \rho/c)s) = 1$ *and* $\lim_{\rho\uparrow c} g_0(\tau; (1 - \rho/c)s) = 1$,

(ii) $\lim_{\rho\uparrow c} \frac{1-g_1(\tau;(1-\rho/c)s)}{1-\rho/c} = (1 + \nu m_1)s\tau$.

Proof Part (i): Substitute $(1 - \rho/c)s$ for s in Equations (3.47) and (3.48), and let $\rho \uparrow c$. Assuming that $h_1(\tau; s) := \lim_{\rho\uparrow c} g_1(\tau; (1 - \rho/c)s)$ and $h_0(\tau; s) := \lim_{\rho\uparrow c} g_0(\tau; (1-\rho/c)s)$ exist we find (using the Dominated Convergence Theorem for the interchange of limit and integrals),

$$\begin{aligned}
h_1(\tau + \Delta; s) &= \int_{t=0}^{\Delta} \mu \exp\left\{-\mu t - \lambda\left(t - \int_{u=0}^{t} h_1(\tau + u; s)\mathrm{d}u\right)\right. \\
&\qquad \left. -\nu\left(t - \int_{u=0}^{t} \phi(\lambda(1 - h_1(\tau + u; s)))\mathrm{d}u\right)\right\} \mathrm{d}t \\
&\quad + h_1(\tau; s) \exp\left\{-\mu\Delta - \lambda\left(\Delta - \int_{u=0}^{\Delta} h_1(\tau + u; s)\mathrm{d}u\right)\right. \\
&\qquad \left. -\nu\left(\Delta - \int_{u=0}^{\Delta} \phi(\lambda(1 - h_1(\tau + u; s)))\mathrm{d}u\right)\right\}, \\
h_0(\tau + \Delta; s) &= h_0(\tau; s) \exp\left\{-\lambda\left(\Delta - \int_{u=0}^{\Delta} h_1(\tau + u; s)\mathrm{d}u\right)\right. \\
&\qquad \left. -\nu\left(\Delta - \int_{u=0}^{\Delta} \phi(\lambda(1 - h_1(\tau + u; s)))\mathrm{d}u\right)\right\}.
\end{aligned}$$

From this we can derive the following set of differential equations (as in Section 3.3):

$$
\begin{aligned}
\frac{\partial}{\partial \tau} h_1(\tau;s) &= \mu + h_1(\tau;s)\left\{\lambda h_1(\tau;s) - (\lambda+\nu+\mu) + \nu\phi(\lambda(1-h_1(\tau;s)))\right\},\\
\frac{\partial}{\partial \tau} h_0(\tau;s) &= h_0(\tau;s)\left\{\lambda h_1(\tau;s) - (\lambda+\nu) + \nu\phi(\lambda(1-h_1(\tau;s)))\right\}.
\end{aligned}
$$

Together with the boundary conditions $h_1(0;s) = h_0(0;s) = 1$, these differential equations uniquely determine $h_1(\tau;s)$ and $h_0(\tau;s)$. Part (i) is now proved by noting that $h_1(\tau;s) \equiv 1$ and $h_0(\tau;s) \equiv 1$ satisfy these equations. A comment should however be made about the assumption on the existence of $h_1(\tau;s)$ and $h_0(\tau;s)$: Since, for any $\mathrm{Re}(s) \geq 0$, $|g_1(\tau;s)| \leq 1$, we can find a sequence $\left(\rho^{(k)}\right)_{k\in\mathbf{N}}$ in the interval $[0,c]$ such that $\lim_{k\to\infty}\rho^{(k)} = c$ and $\overline{h}_1(\tau;s) := \lim_{k\to\infty} g_1(\tau;(1-\rho^{(k)}/c)s)$ exists. For $\overline{h}_1(\tau;s)$ we can formulate the differential equations, leading to $\overline{h}_1(\tau;s) \equiv 1$. Since the limit is the same for all convergent sequences, $h_1(\tau;s)$ exists. In the same way it can be argued that $h_0(\tau;s)$ exists.

Part (ii): The proof proceeds along the same lines as for Part (i). We assume the existence of

$$
l_1(\tau;s) := \lim_{\rho\uparrow c} \frac{1-g_1(\tau;(1-\rho/c)s)}{1-\rho/c}.
$$

Again this existence can be shown by following the subsequent steps for the limit of a convergent sequence

$$
\frac{1-g_1(\tau;(1-\rho^{(k)}/c)s)}{1-\rho^{(k)}/c}.
$$

Such a sequence exists because $|1-g_1(\tau;\omega)| \leq |\omega|\mathbf{E}\left[C_1(\tau)\right]$ for any $\mathrm{Re}(\omega) \geq 0$, and $\mathbf{E}\left[C_1(\tau)\right]$ is bounded in $\rho/c \in [0,1]$, see Formula (3.22).

Substitute $(1-\rho/c)s$ for s in Equation (3.47), subtract both sides of this equation from 1, and use,

$$
\begin{aligned}
&\lim_{\rho\uparrow c} \frac{1}{1-\rho/c}\left(1 - \exp\left\{-(1-\rho/c)sx - \lambda\int_{u=0}^{x}(1-g_1(\tau+u;(1-\rho/c)s))\,\mathrm{d}u\right.\right.\\
&\qquad \left.\left. -\nu\int_{u=0}^{x}(1-\phi((1-\rho/c)s+\lambda(1-g_1(\tau+u;(1-\rho/c)s))))\,\mathrm{d}u\right\}\right)\\
&\qquad = sx + \lambda\int_{u=0}^{x} l_1(\tau+u;s)\mathrm{d}u + \nu\int_{u=0}^{x} m_1\left(s+\lambda l_1(\tau+u;s)\right)\mathrm{d}u,
\end{aligned}
$$

(again with the Dominated Convergence Theorem to interchange limit and integrals), to find,

$$
\begin{aligned}
&l_1(\tau+\Delta;s)\\
&\quad = \int_{t=0}^{\Delta}\mu\mathrm{e}^{-\mu t}\left((1+\nu m_1)st + \lambda(1+\nu m_1)\int_{u=0}^{t} l_1(\tau+u;s)\mathrm{d}u\right)\mathrm{d}t\\
&\quad +\mathrm{e}^{-\mu\Delta}\left(l_1(\tau;s) + (1+\nu m_1)s\Delta + \lambda(1+\nu m_1)\int_{u=0}^{\Delta} l_1(\tau+u;s)\mathrm{d}u\right).
\end{aligned}
$$

For $\Delta \downarrow 0$ we may now write:

$$\begin{aligned} l_1(\tau+\Delta;s) &= l_1(\tau;s) - \Delta\mu l_1(\tau;s) + \Delta(1+\nu m_1)s + \Delta\lambda(1+\nu m_1)l_1(\tau;s) + \mathrm{o}(\Delta) \\ &= l_1(\tau;s) + \Delta(1+\nu m_1)s + \mathrm{o}(\Delta), \end{aligned}$$

where for the last equality we have used that $\mu = \lambda(1+\nu m_1)$ when $\rho/c = 1$. Using the boundary condition $l_1(0;s) \equiv 0$ we readily find $l_1(\tau;s) = (1+\nu m_1)s\tau$. $\square$

Chapter 4

Sojourn times in a Markovian random environment

In the previous chapter we studied sojourn times in a processor-sharing queue with an unreliable server. This chapter presents the analysis of Núñez Queija [84] for a model with a more general structure for the service fluctuations. Here the service rate of the processor-sharing queue depends on the state of an underlying Markov chain. The model is more general than the on/off model of the previous chapter in the sense that different positive service rates are possible, and that the (total) service capacity may also depend on the number of customers in the queue. To be more precise: As before, let $X(t)$ be the number of customers in the system at time $t \geq 0$. Also, $Y(t)$ denotes the state of some (yet to be specified) *random environment* (of the queue), at time $t \geq 0$. In Chapter 3 the random environment determined the state of the server which alternated between an "on" and an "off" state, the on-periods being exponentially distributed and the off-periods having a general distribution. Now $\{Y(t), t \geq 0\}$ is a general Markov process on a finite state space $\{1, 2, \ldots, N\}$. The (total) service rate in the processor-sharing queue depends on the state of the random environment, but may also depend on the number of customers in the queue. If $X(t) = k$ and $Y(t) = i$ then the total service rate at which customers in the processor-sharing queue are served, is $c_i^{(k)} \geq 0$. Thus, if $k \geq 0$, each of the customers in the queue receives a service capacity $\frac{1}{k}c_i^{(k)}$. Under the assumption of Poisson arrivals (the arrival rate at time t possibly depending on $X(t)$ and $Y(t)$) and exponentially distributed service requirements, the two-dimensional process $\{(X(t), Y(t)), t \geq 0\}$ is a non-homogeneous — or level-dependent — QBD (Quasi Birth and Death) process, see Chapter 2 or Definition (4.2) below. The QBD structure is not essential to the analysis, but has computationally attractive properties. The analysis is presented under the above assumptions to show how the QBD structure is preserved throughout the analysis and reflected in the results. In Section 4.8 we show how the analysis can be extended to the case when service requirements have a phase-type distribution. This destroys the QBD structure, but qualitative properties of sojourn times are preserved.

As pointed out in Section 1.6, it is well known for processor-sharing systems with constant service rate that the mean sojourn time conditional on the amount of service required is proportional to the service requirement. In Section 3.6 we showed that this is not true for the on/off model. However, an asymptotic linearity (for the amount of service required tending to infinity) was revealed. In this chapter we show that this asymptotic result is also valid for the present model, in which the service rate may assume different positive values. Using a time-scale transformation which extends the branching-process approach of Section 3.2, it is shown that the problem may be viewed in the context of a Markov-Reward process.

The remainder of the chapter is organised as follows. The model under consideration is presented in detail in Section 4.1. In Section 4.2 the sojourn times of customers are studied. As in the previous chapter, we concentrate on sojourn times conditional on the state upon arrival and on the service requirement. An explicit expression for the LST (Laplace-Stieltjes Transform) of the conditional sojourn-time distribution is derived. Particular attention is paid to the conditional mean sojourn time as a function of the service requirement, and we prove the existence of an asymptote, as the amount of required service tends to infinity. In Section 4.3 we extend the method of random time change to the present model. We used this method in Section 3.2 for the construction of a branching process that enabled the analysis of sojourn times in the model with an unreliable server. By means of the random time-change method, we “translate” sojourn times in the queueing system into rewards in a Markov-Reward process. In Section 4.5 we explain the proportionality property between conditional mean sojourn time and the service requirement in processor-sharing queues without random environment. We show in Section 4.6 how the conditional mean sojourn times may be computed. In view of the complexity in computing the exact conditional mean sojourn time, we propose an approximation that only depends on steady-state characteristics, and hence, can efficiently be computed. Section 4.7 presents the numerical results of Núñez Queija et al. [85] for an application of the analysis to a variant of the telecommunication system described in Section 1.5. The numerical results validate the proposed approximation. In Section 4.8 it is shown that phase-type services and discriminatory processor sharing essentially fall within the framework of the model. We also discuss the extension to infinite state spaces. Concluding remarks are made in Section 4.9.

4.1 The model

The model that we study in this chapter is more general than the one of Chapter 2, where the random environment (that is, the process regulating the available service rate) evolved independent of the queue-length process. Here we do not make that assumption and, furthermore, allow arrivals and departures of customers to cause an instantaneous change in the random environment. Let us first describe the model in detail.

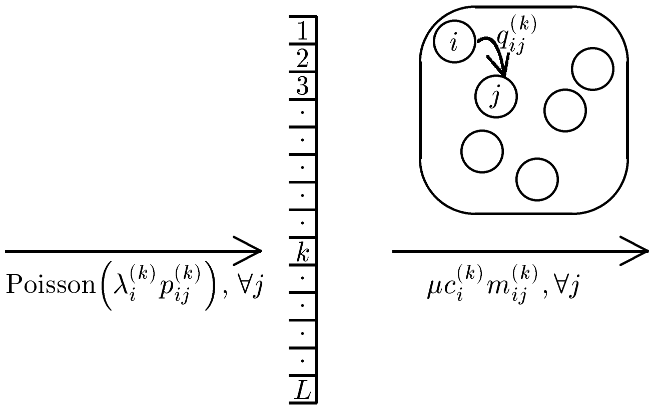

Figure 4.1: The queueing model

Consider a processor-sharing queue in a random environment as depicted in Figure 4.1. In the queue at most $L \in \mathbb{N}$ customers can be present. We assume that the random environment may be modelled as a Markov process with state space $\{1, 2, \ldots, N\}$, with $N \in \mathbb{N}$. Changes in the random environment may be dependent on the arrival and departure process of customers. The set of possible states of the random environment when the number of customers is equal to $k \in \{0, 1, 2, \ldots, L\}$ is denoted by the subset $E^{(k)} \subseteq \{1, 2, \ldots, N\}$. We say that the queueing system of Figure 4.1 is in state (k, i) when there are $k \in \{0, 1, \ldots, L\}$ customers present and the state of the random environment is $i \in E^{(k)} \subseteq \{1, 2, \ldots, N\}$. The set of all possible system states is denoted by:

$$\mathbf{S} := \left\{(k, i) \ : \ k = 0, 1, \ldots, L; \ i \in E^{(k)}\right\}. \tag{4.1}$$

The arrival rate of new customers and the service rate of customers in the queue are determined by both the queue length and the state of the random environment. For the time being (we come back to this in Section 4.8.1), it is assumed that customers have an exponentially distributed service requirement with mean $1/\mu$ (independent of other service requirements, the arrival process, and the random environment). If the state of the system is (k, i), then new customers arrive according to a Poisson process with rate $\lambda_i^{(k)}$. Upon such an arrival, the number of customers in the system is increased by one, and the random environment changes (immediately) to state $j \in E^{(k+1)}$ with probability $p_{ij}^{(k)}$, where $\sum_{j \in E^{(k+1)}} p_{ij}^{(k)} = 1$ for $0 \leq k \leq L-1$. If $j = i$ then $p_{ij}^{(k)}$ is the probability that the random environment does not change state. In state (k, i) with $k > 0$, the server works at rate $c_i^{(k)} \geq 0$. This service capacity is equally divided among all customers present (processor sharing). Hence, each customer leaves in an interval of length Δ with probability $\frac{1}{k}\mu c_i^{(k)}\Delta + \mathrm{o}(\Delta)$, for $\Delta \downarrow 0$. The total departure rate of customers is therefore $\mu c_i^{(k)}$. Upon such a departure, the random environment changes to state $j \in E^{(k-1)}$ with probability $m_{ij}^{(k)}$, where $\sum_{j \in E^{(k-1)}} m_{ij}^{(k)} = 1$ for $1 \leq k \leq L$. Finally, in state (k,i) the random environment may change to state $j \in E^{(k)}$ – without changing the number of customers – at rate $q_{ij}^{(k)}$, $j \neq i$. For $(k,i) \in \mathbf{S}$ it is convenient to define $p_{ij}^{(k)} = 0,\ j \notin E^{(k+1)}$, $m_{ij}^{(k)} = 0,\ j \notin E^{(k-1)}$, and $q_{ij}^{(k)} = 0,\ j \notin E^{(k)}$. Note that $E^{(-1)}$ and $E^{(L+1)}$ are not defined, therefore we further set $\lambda_i^{(L)} := c_i^{(0)} := 0$, for all $i \in \{1, 2, \ldots, N\}$.

At time $t \geq 0$, $X(t)$ is the number of customers in the system and $Y(t)$ is the state of the random environment. The Markovian process $\{(X(t), Y(t)), t \geq 0\}$ is a non-homogeneous QBD process. Its infinitesimal generator can be written as:

$$\mathcal{G} := \begin{bmatrix} Q_d^{(0)} & \Lambda^{(0)} & 0 & \ldots & \ldots & 0 \\ M^{(1)} & Q_d^{(1)} & \Lambda^{(1)} & 0 & \ldots & \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & & M^{(L-1)} & Q_d^{(L-1)} & \Lambda^{(L-1)} \\ 0 & \ldots & & 0 & M^{(L)} & Q_d^{(L)} \end{bmatrix}. \tag{4.2}$$

The submatrices in this generator are given by:

$$\begin{aligned} \Lambda^{(k)} &= \left[\lambda_i^{(k)} p_{ij}^{(k)}\right]_{i \in E^{(k)}, j \in E^{(k+1)}}, \\ M^{(k)} &= \left[\mu c_i^{(k)} m_{ij}^{(k)}\right]_{i \in E^{(k)}, j \in E^{(k-1)}}, \\ Q_d^{(k)} &= \left[q_{ij}^{(k)}\right]_{i,j \in E^{(k)}}, \end{aligned}$$

where the $q_{ii}^{(k)}$ are such that (4.2) is a true generator (all rows sum to 0).

The state space of the process $(X(t), Y(t))$ is given by $\mathbf{S}$ in Definition (4.1). The components k and i of the state $(k,i) \in \mathbf{S}$ are called the level and the phase of the QBD process, respectively. The level of the process corresponds to the number of customers in the system, and the phase of the process corresponds to the state of the random environment.

It will be assumed throughout this chapter that $\mathcal{G}$ is irreducible (all states in the corresponding Markov process communicate). The process is called a homogeneous QBD process if, as is the case in Chapter 2, for all k: $M^{(k)} = M$, $Q_d^{(k)} = Q_d$ and $\Lambda^{(k)} = \Lambda$. The state space is finite, since we assumed that L, the maximum number of customers in the system, and N, the number of states of the random environment, are both finite. The submatrices $Q_d^{(k)}$, $\Lambda^{(k)}$, and $M^{(k)}$ are all of finite – but not necessarily the same – dimension. Generalisations to infinite state spaces are possible, but require specific attention regarding ergodicity issues. We briefly address these issues in Section 4.8.3.

We denote the steady-state probability vector by $\overline{\pi}$:

$$\overline{\pi}\, \mathcal{G} = \overline{0}, \quad \overline{\pi}\, \overline{1} = 1,$$

with $\overline{0}$ being the vector with all entries equal to zero and $\overline{1}$ the vector with all entries equal to one. Throughout this chapter, for any vector $\overline{v}$ its entries $v_{k,i}$ are ordered lexicographically, i.e. $v_{k,i}$ precedes $v_{l,j}$ if $k < l$, or if $k = l$ *and* $i < j$. Another notational convention we adopt, is that any vector multiplying a matrix from the left (right) is a row (column) vector. Furthermore we use the symbol I to denote the identity matrix. Whenever used, the vectors $\overline{1}$ and $\overline{0}$, and the matrix I are of the appropriate dimension.

Usually the service discipline considered for queueing systems which can be modelled as a QBD process is FCFS (First Come First Served). In the present queueing system the service discipline is processor sharing. Because of the exponentially distributed service requirements, the queue length process obeys the same probabilistic law for all work-conserving service disciplines that do not take into account actual service requirements (including FCFS and processor sharing). The queue length in non-homogeneous QBD-processes has been studied extensively in the literature, see for instance De Nitto Personè and Grassi [24] where an algorithm is described for the computation of the steady-state queue-length distribution in non-homogeneous QBD processes with a finite state space. Here we do not discuss the computation of the steady-state probability vector $\overline{\pi}$.

For the sojourn times of customers, the service discipline *does* matter. Sojourn times in QBD processes under the FCFS discipline are discussed in Neuts [81, Section 3.9]. For non-homogeneous QBD processes an analogous treatment is possible. The distribution in terms of LSTs may be found in Li and Sheng [65]. Here, our concern is with the sojourn time distribution under the processor-sharing service discipline.

4.2 Sojourn times

In this section we study the sojourn time of a customer conditioned on the number of customers and the state of the random environment upon his arrival. Particular attention is paid to the case where we also condition on the amount of work brought into the system. It will be useful to define the following generator:

$$\mathcal{H} := \left[\begin{array}{c|cccccc}
0 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ \hline
M^{(1)}\overline{1} & Q_d^{(1)} & \Lambda^{(1)} & & & & \\
\frac{1}{2}M^{(2)}\overline{1} & \frac{1}{2}M^{(2)} & Q_d^{(2)} & \Lambda^{(2)} & & & \\
\vdots & & \ddots & \ddots & \ddots & & \\
\frac{1}{k}M^{(k)}\overline{1} & & & \frac{k-1}{k}M^{(k)} & Q_d^{(k)} & \Lambda^{(k)} & \\
\vdots & & & & \ddots & \ddots & \ddots \\
\frac{1}{L}M^{(L)}\overline{1} & & & & & \frac{L-1}{L}M^{(L)} & Q_d^{(L)}
\end{array}\right]. \tag{4.3}$$

The state space of a Markov process with generator $\mathcal{H}$ may be denoted by all pairs (k,i), with $k = 1, 2, \ldots, L$ and $i \in E^{(k)}$, and an absorbing state 0. Note that there are no states (k,i) with $k=0$, and that all states (k,i), $k=1,2,\ldots,L$ are transient. The latter statement follows from the irreducibility of $\mathcal{G}$ given by Definition (4.2). In the first column of $\mathcal{H}$ we find the transition (absorption) rates from all other states into state 0. From any state (k,i) the absorption rate (into state 0) equals $\frac{1}{k}\mu c_i^{(k)}$.

Theorem 4.2.1 *The sojourn time of a customer who enters the system with $k-1$ other customers present and the random environment being in state i, is distributed as the absorption time in a Markov process $\mathcal{M}_{\mathcal{H}}$ with generator $\mathcal{H}$ defined by (4.3), starting from state (k,i).*

Proof The proof can be given by comparing the evolution of the queueing system of Figure 4.1, from the moment that the tagged customer arrives (and finds $k-1$ other customers and the random environment in state i), with the evolution of the Markov process $\mathcal{M}_{\mathcal{H}}$, starting in state (k,i), until absorption in state 0.

In particular, at any moment that the tagged customer is in service with $l-1$ other customers and the random environment in state j, the rate at which he is served is $\frac{1}{l}c_j^{(l)}$, and his "departure rate" is therefore $\mu\frac{1}{l}c_j^{(l)}$. The departure of the tagged customer from the queueing system corresponds to absorption in state 0 in the Markov process $\mathcal{M}_{\mathcal{H}}$ (see the first column of $\mathcal{H}$). □

Remark 4.2.1 For the computation of the moments of the absorption time in $\mathcal{M}_{\mathcal{H}}$ (and hence of the sojourn time in the queueing system) from any initial state we refer to Li and Sheng [65].

We further concentrate on the sojourn time of a customer with service requirement $\tau>0$. For $k=1,2,\ldots,L$ and $i\in E^{(k)}$, let again $V_{k,i}(\tau)$ be the (remaining) sojourn time of a (tagged) customer, starting with $k-1$ other customers present, the random environment in state i, and the tagged customer having a (remaining) service requirement of τ. Define the LST (Laplace-Stieltjes Transform) of the distribution of $V_{k,i}(\tau)$ by

$$v_{k,i}(s;\tau) \;:=\; \mathbf{E}\left[\mathrm{e}^{-sV_{k,i}(\tau)}\right], \quad \mathrm{Re}(s)\geq 0,$$

and let $\overline{v}(s;\tau)$ be the vector with the $v_{k,i}(s;\tau)$ ordered lexicographically. In the following we derive an explicit expression for $\overline{v}(s;\tau)$.

Remark 4.2.2 In this section and in Section 4.3 we concentrate on the case where the $c_i^{(k)}$ are all *strictly positive*. In Section 4.4 we extend the analysis to the case where some of the $c_i^{(k)}$ may be zero.

As in Section 3.2 we study the sojourn times conditional on the service requirement using the model with one *permanent* customer. Suppose we consider the queueing system of Figure 4.1, with the modification that there is one customer that never leaves the system, but shares in the service rate as if he were an ordinary customer. With that modification, the number of customers k ranges from 1 to L. Having placed a permanent customer in the system at time 0, we denote the total number of customers in the system (including the permanent customer) at time $t \geq 0$ by $X^*(t)$ and the state of the random environment by $Y^*(t)$. Recall that in Section 3.2 we considered the number of non-permanent customers $Z(t) = X^*(t) - 1$ instead of $X^*(t)$, and that the evolution of the random environment was unaffected by the presence of the permanent customer. The process $\{(X^*(t), Y^*(t)),\ t \geq 0\}$ is again a non-homogeneous QBD process with generator $\mathcal{G}^*$ defined by:

$$\mathcal{G}^* := \begin{bmatrix} \widetilde{Q_d}^{(1)} & \Lambda^{(1)} & & & & & & \\ \frac{1}{2}M^{(2)} & \widetilde{Q_d}^{(2)} & \Lambda^{(2)} & & & & & \\ & \ddots & \ddots & \ddots & & & & \\ & & \frac{k-1}{k}M^{(k)} & \widetilde{Q_d}^{(k)} & \Lambda^{(k)} & & & \\ & & & \ddots & \ddots & \ddots & & \\ & & & & \frac{L-2}{L-1}M^{(L-1)} & \widetilde{Q_d}^{(L-1)} & \Lambda^{(L-1)} \\ & & & & & \frac{L-1}{L}M^{(L)} & \widetilde{Q_d}^{(L)} \end{bmatrix}. \tag{4.4}$$

The matrices $\widetilde{Q_d}^{(k)}$ differ from the matrices $Q_d^{(k)}$ only in their diagonal elements: These are such that each row of $\mathcal{G}^*$ sums to 0. The state space of the process $(X^*(t), Y^*(t))$ will be denoted by:

$$\mathbf{S}^* := \{(k,i) \in \mathbf{S} \ : \ k > 0\}.$$

Additionally, we define the diagonal matrix

$$\mathcal{R} := \text{diag}\left[\frac{1}{k}c_i^{(k)}\right]_{(k,i)\in\mathbf{S}^*}.$$

The entries of $\mathcal{R}$ along the diagonal are ordered lexicographically in (k,i).

We follow a different approach than in Section 3.2. Here we first analytically derive an expression for the vector of LSTs $\overline{v}(s;\tau)$, exploiting that in the present case the model is Markovian. In Section 4.3 we relate the solution to the permanent-customer model using the method of random time change.

Theorem 4.2.2 *If $c_i^{(k)} > 0$ for all $(k,i) \in \mathbf{S}^*$, then for $\tau \geq 0$ and* $\text{Re}(s) \geq 0$*,*

$$\frac{\partial}{\partial\tau}\overline{v}(s;\tau) = \mathcal{R}^{-1}\left[\mathcal{G}^* - s\ I\right]\overline{v}(s;\tau), \tag{4.5}$$

$$\overline{v}(s;0) = \overline{1}; \tag{4.6}$$

and hence,

$$\overline{v}(s;\tau) = \exp\left\{\tau\mathcal{R}^{-1}\left[\mathcal{G}^* - s\ I\right]\right\}\ \overline{1}. \tag{4.7}$$

Proof The proof can be given by marginal analysis: When the state of the queueing system is $(k,i) \in \mathbf{S}^*$, the customer with a remaining amount of work τ (as well as all other customers) is served at rate $\frac{1}{k}c_i^{(k)}$. Consider a small time interval of length $\frac{k\Delta}{c_i^{(k)}}$ and condition on the possible events occurring in this interval:

$$
\begin{aligned}
&v_{k,i}(s;\tau+\Delta) \\
&= \quad \mathrm{e}^{-s\,\frac{k\Delta}{c_i^{(k)}}} \times \\
&\quad \left\{ \lambda_i^{(k)} \frac{k\Delta}{c_i^{(k)}} \sum_j p_{ij}^{(k)} v_{k+1,j}(s;\tau) + \mu \frac{k-1}{k} c_i^{(k)} \frac{k\Delta}{c_i^{(k)}} \sum_j m_{ij}^{(k)} v_{k-1,j}(s;\tau) \right. \\
&\quad \left. + \sum_{j \neq i} q_{ij}^{(k)} \frac{k\Delta}{c_i^{(k)}} v_{k,j}(s;\tau) + \left(1 + \widetilde{q}_{ii}^{(k)} \frac{k\Delta}{c_i^{(k)}}\right) v_{k,i}(s;\tau) \right\} + \mathrm{o}(\Delta),
\end{aligned}
$$

with $\widetilde{q}_{ii}^{(k)} = -\mu\frac{k-1}{k}c_i^{(k)} - \lambda_i^{(k)} - \sum_{j\neq i} q_{ij}^{(k)}$. This leads to the differential equation (4.5). The initial conditions (4.6) follow from the fact that all $c_i^{(k)}$ are positive, and hence $\mathbf{E}\left[V_{k,i}(0+)\right] = 0$. It can then be verified that (4.7) is a solution to (4.5) and (4.6). Since the solution must be unique, we are done. $\square$

In the proof of the following corollary we use standard results for Markov-Reward processes. These processes fall within the framework of Markov decision theory (with the restriction that here no decisions are to be made). The first to present a systematic treatment of Markov-Reward processes on a finite state space seems to have been Howard [45]. In particular the results on continuous-time Markov-Reward processes (pp. 99–104) are of interest to us. The close relationship between the continuous-time case and the discrete-time case is exploited by Tijms [112, Section 3.5]. In the proof of the following corollary we further rely on Zijm [127].

We use the symbol $\overline{1}_{k,i}$ to denote the vector with the entry in position (k,i) equal to 1, and all other entries equal to 0.

Corollary 4.2.3 *If $c_i^{(k)} > 0$ for all $(k,i) \in \mathbf{S}^*$, then for $\tau \geq 0$,*

$$
\mathbf{E}\left[V_{k,i}(\tau)\right] = \frac{\tau}{c^* - \rho^*} \;+\; \overline{1}_{k,i} \;\left[I - \exp\left\{\tau \mathcal{R}^{-1}\mathcal{G}^*\right\}\right] \overline{\gamma}, \tag{4.8}
$$

where

$$
\begin{aligned}
c^* &= \sum_{(k,i)\in\mathbf{S}^*} \pi_{k,i}^* c_i^{(k)}, \\
\rho^* &= \sum_{(k,i)\in\mathbf{S}^*} \pi_{k,i}^* \lambda_i^{(k)} \frac{1}{\mu},
\end{aligned}
$$

with $\overline{\pi}^* = (\pi^*_{k,i})_{(k,i)\in\mathbf{S}^*}$ *the steady-state probability vector of the model with one permanent customer:*

$$\overline{\pi}^*\ \mathcal{G}^* = \overline{0}, \quad \overline{\pi}^*\ \overline{1} = 1.$$

The vector $\overline{\gamma}$ *satisfies*

$$-\mathcal{G}^*\ \overline{\gamma} = \overline{1} - \frac{1}{c^* - \rho^*}\ \mathcal{R}\overline{1}, \tag{4.9}$$

and is unique up to translation by the vector $\overline{1}$*. Expression (4.8) is, however, invariant with respect to such a translation. We may normalise* $\overline{\gamma}$ *such that* $\overline{\pi}^*\mathcal{R}\overline{\gamma} = 0$*.*

Proof The result can be obtained by differentiating (4.7) with respect to s, and setting $s = 0$. However, we give a more direct proof. In the same way as we derived (4.5) and (4.6), we may find the following set of differential equations and initial conditions:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\tau}\left(\mathbf{E}\left[V_{k,i}(\tau)\right]\right)_{k,i} &= \mathcal{R}^{-1}\overline{1} + \mathcal{R}^{-1}\mathcal{G}^*\left(\mathbf{E}\left[V_{k,i}(\tau)\right]\right)_{k,i}, & (4.10)\\ \mathbf{E}\left[V_{k,i}(0)\right] &= 0, \quad \forall\ (k,i) \in \mathbf{S}^*. & (4.11) \end{aligned}$$

By $(\cdot)_{k,i}$ we mean the vector with the entries between brackets ordered lexicographically in $(k,i) \in \mathbf{S}^*$. It is left to the reader to verify that there is at most one solution to this set of differential equations and initial conditions.

Suppose for the moment that a vector $\overline{\gamma}$ exists, satisfying (4.9) and normalised as required. By substitution of (4.8) into (4.10) and (4.11), we may verify that these differential equations and initial conditions are satisfied, and hence (4.8) is the unique solution. Note that $\mathcal{G}^*\overline{1} = \overline{0}$, since $\mathcal{G}^*$ is the generator of a Markov process.

From Equation (4.9) we note that if the vector $\overline{\gamma}$ exists, it may be interpreted as the "relative reward" vector in a Markov-Reward process. This vector contains for each state of the process the long-run difference in accumulated rewards when starting in that state relative to those when starting in steady state, see Tijms [112, pp. 187–188] for a discussion. The generator of this Markov-Reward process is $\mathcal{G}^*$ and rewards are generated at rate $1 - \frac{1}{k}c_i^{(k)}\frac{1}{c^*-\rho^*}$ when the process is in state $(k,i) \in \mathbf{S}^*$.

In order for Equation (4.9) to have a solution, it is necessary that this Markov-Reward process has average reward per time unit equal to 0, because the left-hand side is equal to zero if we pre-multiply by $\overline{\pi}^*$. Indeed,

$$\begin{aligned} \sum_{(k,i)\in\mathbf{S}^*} \pi^*_{k,i}\frac{1}{k}c_i^{(k)} &= c^* - \frac{1}{\mu}\sum_{(k,i)\in\mathbf{S}^*} \pi^*_{k,i}\frac{k-1}{k}c_i^{(k)}\mu \\ &= c^* - \frac{1}{\mu}\sum_{(k,i)\in\mathbf{S}^*} \pi^*_{k,i}\lambda_i^{(k)} \\ &= c^* - \rho^*, \end{aligned} \tag{4.12}$$

where the one-but-last equality sign is due to the fact that the average number of customers leaving the system per time unit equals the average number of customers entering the system per time unit. Since the state space is finite, the existence of a vector $\overline{\gamma}$, and its uniqueness up to translation along the vector $\overline{1}$, is guaranteed by Zijm [127, Theorem 4.5].

Note that translation along the vector $\overline{1}$ of a vector $\overline{\gamma}$ satisfying (4.9) does not alter the solution for $\mathbf{E}\left[V_{k,i}(\tau)\right]$, since all rows of the matrix $I - \exp\left\{\tau\mathcal{R}^{-1}\mathcal{G}^*\right\}$ in (4.8) sum to 0 (because all rows of $\mathcal{G}^*$ do). Therefore, if $\overline{\gamma} = \overline{v}$ satisfies (4.9), then so does

$$\overline{\gamma} := \overline{v} - \frac{\overline{\pi}^*\mathcal{R}\overline{v}}{c^* - \rho^*}\overline{1},$$

which is normalised as required. □

Remark 4.2.3 The entities c^* and ρ^* have the following interpretation. In the queueing system with one permanent customer, the service capacity not given to other customers is assigned to the permanent customer. The average capacity per unit of time available for all customers (including the permanent one) is c^*. Per unit of time, on average $\sum_{(k,i)\in\mathbf{S}^*} \lambda_i^{(k)}$ customers enter the system, each requiring an expected amount of work $1/\mu$. Therefore ρ^* is the average amount of work entering the system per unit of time.

Corollary 4.2.4 *If $c_i^{(k)} > 0$ for all $(k,i) \in \mathbf{S}^*$, then for $\tau \to \infty$ we have:*

$$\mathbf{E}\left[V_{k,i}(\tau)\right] - \frac{\tau}{c^* - \rho^*} \longrightarrow \gamma_{k,i}, \quad (k,i) \in \mathbf{S}^*.$$

Proof Note that $\mathcal{R}^{-1}\mathcal{G}^*$ is the infinitesimal generator of an irreducible Markov process on a finite state space. Its largest eigenvalue is therefore equal to 0 and of multiplicity 1. The left and right eigenvectors corresponding to the eigenvalue 0 are $\frac{1}{c^*-\rho^*}\overline{\pi}^*\mathcal{R}$ and $\overline{1}$ (after proper scaling). As a consequence, the matrix $\lim_{\tau\to\infty} \exp\left\{\tau\mathcal{R}^{-1}\mathcal{G}^*\right\}$ exists and has all rows equal to the probability vector $\frac{1}{c^*-\rho^*}\overline{\pi}^*\mathcal{R}$. The corollary now follows from Expression (4.8). □

Remark 4.2.4 For the special case of a constant service rate, $c_i^{(k)} = 1,\ \forall\,(k,i)$, and state independent arrivals (Poisson with rate λ), the model reduces to the M/M/1/L queue with processor sharing. It can be shown that for this case (the second subscripts are omitted since there is no random environment):

$$\gamma_{k+1} - \gamma_k = \frac{1}{k\lambda\rho^k}\sum_{j=1}^{k}\left(\frac{1}{1-\rho^*} - j\right)\rho^j > 0, \quad k = 1, 2, \ldots, L-1.$$

Here, $\rho := \lambda/\mu$, and

$$\rho^* = \frac{\sum_{l=1}^{L-1} l\ \rho^{l-1}}{\sum_{l=1}^{L} l\ \rho^{l-1}}\rho.$$

Passing $L \to \infty$, we find in case $\rho < 1$:

$$\gamma_{k+1} - \gamma_k \longrightarrow \frac{\rho}{\lambda(1-\rho)}, \quad L \to \infty$$

which indeed corresponds to the M/M/1 queue with processor sharing. For that model we may even explicitly find:

$$\mathbf{E}\left[V_k(\tau)\right] = \frac{\tau}{1-\rho} \; + \; \frac{1}{\mu-\lambda}\left(k - \frac{1}{1-\rho}\right)\left(1 - \mathrm{e}^{-\tau\,\{\mu-\lambda\}}\right),$$

cf. Coffman et al. [17, Formula (33)] (there the *delay* is studied instead of the sojourn time, which gives a term $\frac{\rho\tau}{1-\rho}$ instead of $\frac{\tau}{1-\rho}$).

4.3 Random time change

In Section 3.2 we studied sojourn times in the on/off model by translating these in terms of appropriately defined rewards in a branching process. That approach can also be applied to the analysis of sojourn times in the model of this chapter. However, here it is less convenient to work with the branching-process interpretation. In the previous chapter, the branches of the resulting branching process evolved independently. We will see that this is not the case for the present model. Exploiting that here the model is Markovian, we work towards the branching process in a different way than we did in Section 3.2. The precise relation with the analysis in Section 3.2 is discussed in Section 4.4, where we allow for periods of service unavailability. In the proof of Corollary 4.2.3 we mentioned the interpretation of the coefficients $\gamma_{k,i}$ as relative rewards in a Markov-Reward process. In this section we explore such an interpretation further and link this to the method of random time change, which was introduced for the analysis of processor-sharing systems by Yashkov [120, 121]. See Section 3.2 for more references on the application of the random time-change method.

Recall that we are still restricting ourself to the case where all the service rates $c_i^{(k)}$ are strictly positive, see Remark 4.2.2. In Section 4.4 we extend the analysis to the case where service rates may be equal to zero.

Our starting point is the Markov process $(X^*(t), Y^*(t))$, i.e. the queue length and the state of the random environment in the queueing model of Figure 4.1, when there is one permanent customer in the system. This permanent customer shares in the service capacity as any other customer, but never leaves the system. We already saw that $\mathcal{G}^*$ is the infinitesimal generator of the process $(X^*(t), Y^*(t))$. We make a random time change in the following way. When $(X^*(t), Y^*(t))$ is in state (k, i), all transitions out of this state are "sped up" by a factor $k/c_i^{(k)}$. For instance, the new arrival rate of customers in state (k, i) is $\lambda_i^{(k)} \times k/c_i^{(k)}$. More importantly, the new departure rate of customers is exactly $(k-1)\mu$ in all states (k, i). Note that $k-1$ is the number of non-permanent

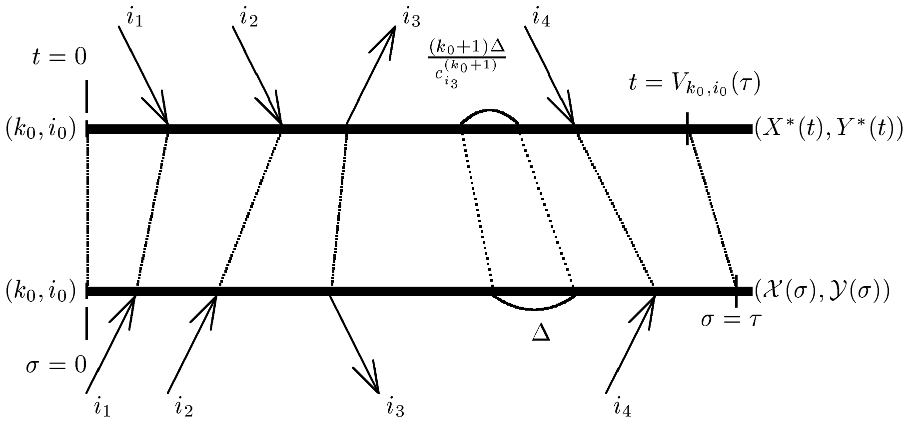

Figure 4.2: Coupling of the jump-chains of $(X^*(t), Y^*(t))$ and $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$.

customers in state $(k, i) \in \mathbf{S}^*$. Apparently, *in the new time scale*, each customer receives one unit of service per "time" unit. By $\{(\mathcal{X}(\sigma), \mathcal{Y}(\sigma)),\ \sigma \geq 0\}$ we denote the process of queue length and state of the random environment in the new time scale. The generator of the Markov process $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$ is $\mathcal{R}^{-1}\mathcal{G}^*$. The inverse of the matrix $\mathcal{R}$ exists since we assumed all service rates $c_i^{(k)}$ to be non-zero. Note that the processes $(X^*(t), Y^*(t))$ and $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$ have the same *jump-chain*. By the jump-chain of a Markov process we mean the embedded Markov chain at transition epochs.

We now explain how the process $\{(X^*(t), Y^*(t)),\ t \geq 0\}$ is related to the process $\{(\mathcal{X}(\sigma), \mathcal{Y}(\sigma)),\ \sigma \geq 0\}$, using a coupling argument on their jump-chains: Suppose that at time $t = 0$, the process $(X^*(t), Y^*(t))$ is in state (k_0, i_0) and observe the process as it evolves over time. For a given path of the process $(X^*(t), Y^*(t))$, we may "perform" the random time change as indicated above: For any period of time that $(X^*(t), Y^*(t))$ resides in a state (k, i), we "shrink" the length of this period by a factor $k/c_i^{(k)}$, i.e. we divide the length of the period by this number. We may so construct a path for the process $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$, starting in (k_0, i_0) for $\sigma = 0$. In Figure 4.2 such a construction is depicted. Two horizontal axes are drawn. The upper axis corresponds to the "normal" time-axis on which we observe the process $(X^*(t), Y^*(t))$ for $t \geq 0$. The lower-axis corresponds to the new "time" scale, after the random time change. On this axis we observe the process $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$, $\sigma \geq 0$.

In the realisation depicted in Figure 4.2 the following events happen successively: The process starts in (k_0, i_0), then a customer arrives and the random environment changes to state i_1, another customer arrives and the random environment changes to state i_2, a customer departs and the random environment changes to state i_3, and finally another customer arrives and the random environment moves to state i_4. Of course, the random environment may change without changing the number of customers, but for transparency of the picture no such event is drawn. Note that since both processes $(X^*(t), Y^*(t))$ and $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$ have the same jump-chain, any such realisation (indeed) occurs with the same probability in both processes. Now concentrate on the indicated time-interval of length $(k_0+1)\Delta/c_{i_3}^{(k_0+1)}$ on the upper axis. This interval lies between the moment of the first departure and the moment of the third arrival. At any point in this interval the number of customers $X^*(t)$ (including the permanent one) is k_0+1, and the random environment $Y^*(t)$ is in state i_3. During this interval of time, the amount of service received by the permanent customer (and any other customer in the system), equals Δ. This argument can be used for any time interval during which the state does not change. It is seen that the amount of service received by the permanent customer between time $t=0$ and the time point (on the upper axis) which corresponds to the point $\sigma=\tau$ (on the lower axis), is exactly τ. Therefore, the point on the upper axis corresponding to $\sigma=\tau$ on the lower axis is exactly $V_{k_0,i_0}(\tau)$: It is the amount of time that a customer must stay in the system before he has received an amount of service τ, starting at time $t=0$ with no service received, k_0-1 other customers, and the random environment in state i_0.

We introduce the following reward structure in the process $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$: In state (k,i) reward is earned at rate $k/c_i^{(k)}$. The accumulation of rewards in this process can now be related to sojourn times in the processor-sharing queue (with exclusively positive service rates). For the on/off model of the previous chapter a similar result was obtained in Lemma 3.2.4. For the present model with possible service interruptions, i.e., when not all service rates are positive, the analogous result is derived in the next section.

Theorem 4.3.1 *The sojourn time $V_{k,i}(\tau)$ of a customer in the queueing system of Figure 4.1, arriving when there are $k-1$ other customers in the system, the random environment being in state i, and bringing an amount of work τ, is distributed as the cumulative reward in the process $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$ over the interval $\sigma \in (0,\tau)$, starting at $\sigma=0$ in state (k,i).*

Proof From our construction of the coupled (jump-)processes above, it follows that the accumulated reward in the process $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$ over the interval $\sigma \in (0,\tau)$ on the lower axis, is equal to $V_{k_0,i_0}(\tau)$ on the upper axis (Figure 4.2). As we already remarked, any such realisation has the same probability for both processes $(X^*(t), Y^*(t))$ and $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$. □

From Theorem 4.3.1 we may obtain the result of Corollary 4.2.4, which is restated in terms of the transformed process in the following corollary:

Corollary 4.3.2 *With probability 1:*

$$\lim_{\tau\to\infty} \frac{V_{k,i}(\tau)}{\tau} = g^* := \mathbf{E}\left[\frac{\mathcal{X}}{\mu c_{\mathcal{Y}}^{(\mathcal{X})}}\right], \tag{4.13}$$

where the distribution of $(\mathcal{X}, \mathcal{Y})$ *is the equilibrium distribution of* $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$. *Furthermore, the limit*

$$\lim_{\tau\to\infty} \mathbf{E}\left[V_{k,i}(\tau)\right] - g^*\tau, \tag{4.14}$$

exists and is finite.

Proof Relation (4.13) is standard for irreducible Markov-Reward processes with a finite state space, see for instance Ross [98, Corollary 6.20] or Tijms [112, Theorem 3.1.1]. We use Zijm [127, Theorem 4.5] to establish the convergence in Relation (4.14). □

Remark 4.3.1 Equation (4.13) holds under much more general assumptions. In fact, with $L < \infty$, $N < \infty$, and all $c_i^{(k)} > 0$, it can be proved using the Renewal Reward theorem (e.g. Ross [98, Theorem 3.16] or Tijms [112, Theorem 1.3.1]) under the sole assumption that the original process $(X(t), Y(t))$ is regenerative with finite expected regeneration time. However, the limit in (4.14) may not exist under such general assumptions.

Remark 4.3.2 Corollaries 4.2.4 and 4.3.2 imply that $g^* = \frac{1}{c^*-\rho^*}$, or equivalently:

$$\mathbf{E}\left[\frac{\mathcal{X}}{c_{\mathcal{Y}}^{(\mathcal{X})}}\right] = \left(\mathbf{E}\left[\frac{c_{Y^*}^{(X^*)}}{X^*}\right]\right)^{-1},$$

where we use Equation (4.12). This can be verified by noting that the distribution of $(\mathcal{X}, \mathcal{Y})$ is given by the vector $\frac{1}{c^*-\rho^*}\overline{\pi}^*\mathcal{R}$, and that the reward vector in that process is given by $\mathcal{R}^{-1}\overline{1}$.

Remark 4.3.3 Similar to the discussion in Remark 3.6.2, we can argue that

$$\mathbf{E}\left[\frac{c_{Y^*}^{(X^*)}}{X^*}\right] = c^* - \rho^*.$$

As we saw in Remark 4.2.3, c^* is the average capacity per unit of time available for all customers, and ρ^* is the average amount of work entering the system per unit of time. Since all non-permanent customers eventually leave the system, ρ^* is also the average amount of service capacity assigned to non-permanent customers (in the long run). Hence, $c^* - \rho^*$ is the average capacity per unit of time assigned to the permanent customer.

In the on/off model of the previous chapter, the arrival rate and the service rate did not depend on the number of customers in the system. Moreover, the random environment — there an alternating renewal process — evolved independent of the queue length. Therefore in the on/off model, the average total service capacity and the average amount of work entering the system per unit of time are the same for the original model and the model with one permanent customer, i.e., $c^* = c$ and $\rho^* = \rho$.

Remark 4.3.4 The process $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$ is closely related to the branching process constructed in Section 3.2. The branching-process approach has been successfully used in the literature to study sojourn times in traditional processor-sharing systems with constant service capacity (i.e., $c_i^{(k)} = 1$; there is no random environment), constant arrival rate (λ), and infinite space for customers ($L = \infty$), see for instance Grishechkin [39, 40]. Let us briefly recall the essentials of this approach. After the random time change, customers may be seen as individuals in a population, one of them having an infinite life time (corresponding to the permanent customer), and all others having an exponentially distributed life time with mean $1/\mu$ (independent of everything else). Thus, with $k \geq 1$ individuals in the population (including the permanent one) the total "death" rate is $(k-1)\mu$. Each of the k individuals gives birth to new individuals at rate λ, the total birth rate is thus $k\lambda$. Clearly, the evolution of each individual and all his descendants is independent of all other individuals, which makes this branching process very suitable for analysis. In fact, the approach is applicable to other service disciplines, including *discriminatory processor sharing*, see Grishechkin [40] and Section 4.8.2.

In our case, the branching process is governed by a random environment. The life time of non-permanent individuals is still exponentially distributed with mean $1/\mu$. If the state of the random environment is i and the population size is k (including the permanent one), then each of the k individuals gives birth to new individuals with rate $\lambda_i^{(k)}/c_i^{(k)}$. This birth rate depends both on the state of the random environment, and on the number of individuals in the population. The random environment also evolves dependent on the number of individuals. With k living individuals, the random environment may change from state i to state j with rate $k \times q_{ij}^{(k)}/c_i^{(k)}$. The mutual dependence of the branching process and the random environment and the dependence among individuals make this approach less suitable for the analysis of sojourn times in the present model.

4.4 Server unavailability

In this section we extend the analysis of Sections 4.2 and 4.3 to the case where some of the $c_i^{(k)}$ may be equal to zero, i.e. there are periods during which no service is provided to the customers. In the setting of our model, unavailability periods are exponentially distributed or, more generally, have a phase-type distribution (when two or more states of the random environment for which the service rate is zero, communicate directly with each other). In the previous chapter the unavailability periods were allowed to have a general distribution. In Remark 4.4.1 we point out how this can also be allowed in the current model.

We define the subset of states

$$\mathbf{S}_0^* := \left\{ (l,j) \in \mathbf{S}^* \;:\; c_j^{(l)} = 0 \right\}.$$

In applications, the fact whether $c_i^{(k)} = 0$ will typically only depend on i, but for generality of the presentation we do not assume this. Partition the state space $\mathbf{S}^*$ into $\mathbf{S}_0^*$ and its complement $\mathbf{S}_+^* := \mathbf{S}^* - \mathbf{S}_0^*$, and "re-order" the rows and the columns of the generator $\mathcal{G}^*$ accordingly:

$$\mathcal{G}^* = \begin{bmatrix} \mathcal{G}_+^* & \mathcal{G}_{+0}^* \\ \mathcal{G}_{0+}^* & \mathcal{G}_0^* \end{bmatrix}.$$

Some reflection shows that if the states within $\mathbf{S}_+^*$ and those within $\mathbf{S}_0^*$ are ordered lexicographically, then $\mathcal{G}_+^*$ and $\mathcal{G}_0^*$ are the generators of (possibly reducible) transient QBD processes. We also re-order the entries of $\overline{\pi}^* = \left(\overline{\pi}_+^*, \overline{\pi}_0^*\right)$, with $\overline{\pi}_+^*$ and $\overline{\pi}_0^*$ vectors with their entries ordered lexicographically. Starting from any $(l,j) \in \mathbf{S}_0^*$, let $U_{l,j}$ be the amount of time the process remains in the set $\mathbf{S}_0^*$, and $W_{l,j} \in \mathbf{S}_+^*$ the first state that is visited after leaving $\mathbf{S}_0^*$. Note that $U_{l,j}$ is the sojourn time (or time until exit) in a transient QBD process, for which an efficient routine to compute moments of the distribution can be found in Li and Sheng [65]. Furthermore, for $\mathrm{Re}(s) \geq 0$, define the matrix $\mathcal{U}(s)$ of dimension $|\mathbf{S}_0^*| \times |\mathbf{S}_+^*|$ with entries:

$$\mathcal{U}_{(l,j),(k,i)}(s) := \mathbf{E}\left[\mathrm{e}^{-sU_{l,j}} \mathbf{1}_{\{W_{l,j}=(k,i)\}}\right], \quad (l,j) \in \mathbf{S}_0^*,\ (k,i) \in \mathbf{S}_+^*.$$

Here $\mathbf{1}_{\{\cdot\}}$ is the indicator function. Note that, in particular, $\mathcal{U}(0)$ is a probability matrix, and that $-\frac{\mathrm{d}}{\mathrm{d}s}\mathcal{U}(s)|_{s=0}\overline{1} = \left(\mathbf{E}\left[U_{l,j}\right]\right)_{(l,j)\in\mathbf{S}_0^*}$.

Lemma 4.4.1 *The matrix* $\mathcal{U}(s)$ *is given by*

$$\mathcal{U}(s) = -\left[\mathcal{G}_0^* - sI\right]^{-1}\ \mathcal{G}_{0+}^*, \quad \mathrm{Re}(s) \geq 0,$$

and hence

$$\left(\mathbf{E}\left[U_{l,j}\right]\right)_{(l,j)\in\mathbf{S}_0^*} = \left[-\mathcal{G}_0^*\right]^{-1}\overline{1}.$$

Proof By conditioning on the possible transitions in an interval Δ when we start from any state in $\mathbf{S}_0^*$ we find for $\Delta \downarrow 0$:

$$\begin{aligned} \mathcal{U}(s) &= \mathrm{e}^{-\Delta s}\left([I + \Delta\mathcal{G}_0^*]\mathcal{U}(s) + \Delta\mathcal{G}_{0+}^*\right) + \mathrm{o}(\Delta) \\ &= \left(I + \Delta\left[\mathcal{G}_0^* - sI\right]\right)\ \mathcal{U}(s) + \Delta\mathcal{G}_{0+}^* + \mathrm{o}(\Delta), \end{aligned}$$

where $\mathrm{o}(\Delta)$ applies to each entry in the matrix equations. Canceling terms, dividing by Δ, and taking $\Delta \downarrow 0$ we have:

$$-\left[\mathcal{G}_0^* - sI\right]\ \mathcal{U}(s) = \mathcal{G}_{0+}^*, \quad \mathrm{Re}(s) \geq 0. \tag{4.15}$$

Since $\mathcal{G}_0^*$ is a transient generator, $\mathcal{G}_0^* - sI$ is invertible for all $\mathrm{Re}(s) \geq 0$, and hence the first statement of the lemma follows.

Differentiating (4.15) with respect to s, setting $s = 0$, and using the fact that $\mathcal{U}(0)$ is a probability matrix (so that $\mathcal{U}(0)\overline{1} = \overline{1}$), we may prove the second statement of the lemma. □

As before, denote by $v_{k,i}(s;\tau)$ the LST of $V_{k,i}(\tau)$, the (remaining) sojourn time of a customer with a (remaining) amount of work τ, starting in state $(k,i) \in \mathbf{S}^*$. Construct the vectors

$$\overline{v}_0(s;\tau) = (v_{l,j}(s;\tau))_{(l,j)\in\mathbf{S}_0^*} \quad \text{and} \quad \overline{v}_+(s;\tau) = (v_{k,i}(s;\tau))_{(k,i)\in\mathbf{S}_+^*},$$

according to the partitioning $\mathbf{S}^* = \mathbf{S}_0^* \cup \mathbf{S}_+^*$. The following lemma gives the relation between the two vectors.

Lemma 4.4.2 *For* $\tau \geq 0$ *and* $\mathrm{Re}(s) \geq 0$*:*

$$\overline{v}_0(s;\tau) = \mathcal{U}(s)\overline{v}_+(s;\tau), \tag{4.16}$$

and in particular,

$$(\mathbf{E}\,[V_{l,j}(\tau)])_{(l,j)\in\mathbf{S}_0^*} = (\mathbf{E}\,[U_{l,j}])_{(l,j)\in\mathbf{S}_0^*} + \mathcal{U}(0)\,(\mathbf{E}\,[V_{k,i}(\tau)])_{(k,i)\in\mathbf{S}_+^*}\,. \tag{4.17}$$

Proof The proof of the first part is immediate by noting that (i) as long as the system is in $\mathbf{S}_0^*$ no service is received, and (ii) the LST of the joint distribution of the first state visited in $\mathbf{S}_+^*$ and the time until that moment is given by the matrix $\mathcal{U}(s)$. The second part follows by differentiating with respect to s and putting $s = 0$. □

With the aid of the two preceding lemmas we are able to prove the following theorem, which generalises Theorem 4.2.2 to the case $\mathbf{S}_0^* \neq \emptyset$. Before proceeding, we define the matrix

$$\mathcal{R}_+ := \mathrm{diag}\left[\frac{1}{k}c_i^{(k)}\right]_{(k,i)\in\mathbf{S}_+^*},$$

with the entries along the diagonal ordered lexicographically in $(k,i) \in \mathbf{S}_+^*$. Note that $\mathcal{R}_+^{-1}$ is well defined.

Theorem 4.4.3 *For* $\tau \geq 0$ *and* $\mathrm{Re}(s) \geq 0$*,*

$$\begin{aligned} \frac{\partial}{\partial\tau}\overline{v}_+(s;\tau) &= \mathcal{R}_+^{-1}\left[\mathcal{G}_+^* + \mathcal{G}_{+0}^*\,\mathcal{U}(s) - sI\right]\overline{v}_+(s;\tau), \\ \overline{v}_+(s;0) &= \overline{1}; \end{aligned}$$

and hence,

$$\overline{v}_+(s;\tau) = \exp\left\{\tau\mathcal{R}_+^{-1}\left[\mathcal{G}_+^* + \mathcal{G}_{+0}^*\,\mathcal{U}(s) - sI\right]\right\}\overline{1}.$$

Proof The proof may proceed as that for Theorem 4.2.2: For any $(k,i) \in \mathbf{S}_+^*$ derive the differential equation for $v_{k,i}(s;\tau)$ by conditioning on the possible events in a time interval $\frac{k}{c_i^{(k)}}\Delta$, and then take $\Delta \downarrow 0$. Substituting (4.16) for $\overline{v}_0(s;\tau)$ readily leads to the desired result. □

Consequently we have the following corollary, which generalises Corollaries 4.2.3 and 4.2.4:

Corollary 4.4.4 *For $\tau \geq 0$,*

$$\left(\mathbf{E}\left[V_{k,i}(\tau)\right]\right)_{(k,i)\in\mathbf{S}_+^*} = \frac{\tau}{c^* - \rho^*}\overline{1} + \left[I - \exp\left\{\tau\mathcal{R}_+^{-1}\left[\mathcal{G}_+^* + \mathcal{G}_{+0}^*\mathcal{U}(0)\right]\right\}\right]\overline{\gamma}.$$

The vector $\overline{\gamma}$ satisfies

$$-\left[\mathcal{G}_+^* + \mathcal{G}_{+0}^*\mathcal{U}(0)\right]\ \overline{\gamma} = \overline{1} + \mathcal{G}_{+0}^*\left(\mathbf{E}\left[U_{l,j}\right]\right)_{(l,j)\in\mathbf{S}_0^*} - \frac{1}{c^* - \rho^*}\ \mathcal{R}_+\overline{1},$$

and is uniquely determined by normalising such that $\overline{\pi}_+^*\mathcal{R}_+\overline{\gamma} = 0$. Consequently we have for $(k,i) \in \mathbf{S}_+^*$:

$$\mathbf{E}\left[V_{k,i}(\tau)\right] - \frac{\tau}{c^* - \rho^*} \longrightarrow \gamma_{k,i}.$$

Proof Similar to (4.10) and (4.11) we may derive differential equations for $\mathbf{E}\left[V_{k,i}(\tau)\right]$, $(k,i) \in \mathbf{S}_+^*$. Using (4.17) we get:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\tau}\left(\mathbf{E}\left[V_{k,i}(\tau)\right]\right)_{(k,i)\in\mathbf{S}_+^*} &= \mathcal{R}_+^{-1}\left[\overline{1} + \mathcal{G}_{+0}^*\left(\mathbf{E}\left[U_{l,j}\right]\right)_{(l,j)\in\mathbf{S}_0^*}\right] \\ &\quad + \mathcal{R}_+^{-1}\left[\mathcal{G}_+^* + \mathcal{G}_{+0}^*\mathcal{U}(0)\right]\left(\mathbf{E}\left[V_{k,i}(\tau)\right]\right)_{(k,i)\in\mathbf{S}_+^*}.\end{aligned}$$

Of course, $\mathbf{E}\left[V_{k,i}(0)\right] = 0$, for $(k,i) \in \mathbf{S}_+^*$. To see that the solution given in the corollary satisfies this set of differential equations and initial conditions, note that the vector $\overline{\gamma}$ may be interpreted as the vector of relative rewards in a Markov-Reward process with generator $\mathcal{G}_+^* + \mathcal{G}_{+0}^*\mathcal{U}(0)$, rewards being earned according to the vector $\overline{1} + \mathcal{G}_{+0}^*\left(\mathbf{E}\left[U_{l,j}\right]\right)_{(l,j)\in\mathbf{S}_0^*} - \frac{1}{c^*-\rho^*}\ \mathcal{R}_+\overline{1}$. It remains to be shown that the average rewards in this Markov-Reward process equal zero. The steady-state probability vector of this process is given by $\frac{1}{\overline{\pi}_+^*\overline{1}}\overline{\pi}_+^*$, and using (cf. Lemma 4.4.1),

$$\overline{\pi}_+^*\mathcal{G}_{+0}^*\left(\mathbf{E}\left[U_{l,j}\right]\right)_{(l,j)\in\mathbf{S}_0^*} = \overline{\pi}_+^*\mathcal{G}_{+0}^*\left[-\mathcal{G}_0^*\right]^{-1}\overline{1} = \overline{\pi}_0^*\overline{1},$$

we indeed find that the reward per unit time in steady state equals 0. The limit as $\tau \to \infty$ can be obtained as in the proof of Corollary 4.2.4. □

Corollary 4.4.5 *For $\tau \to \infty$,*

$$\left(\mathbf{E}\left[V_{l,j}(\tau)\right]\right)_{(l,j)\in\mathbf{S}_0^*} - \frac{\tau}{c^* - \rho^*}\overline{1} \longrightarrow \left(\mathbf{E}\left[U_{l,j}\right]\right)_{(l,j)\in\mathbf{S}_0^*} + \mathcal{U}(0)\overline{\gamma}.$$

Proof By Lemma 4.4.2, Corollary 4.4.4, and using the fact that $\mathcal{U}(0)$ is a probability matrix. □

The remainder of this section is devoted to the method of random time change in the case that $\mathbf{S}_0^* \neq \emptyset$, unifying the approaches in Sections 3.2 and 4.3. We use the same arguments as in Section 4.3, with the following modifications: All transitions which occur when the process $(X^*(t), Y^*(t))$ is in the set $\mathbf{S}_0^*$ are "collapsed" into one single event when constructing the process $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$. More precisely: When the process $(X^*(t), Y^*(t))$ is in some state in $\mathbf{S}_+^*$, we change the time scale as before, speeding up all transitions out of state $(k,i) \in \mathbf{S}_+^*$ by a factor $k/c_i^{(k)}$. The difference with the case $\mathbf{S}_0^* = \emptyset$ is that when the state is $(l,j) \in \mathbf{S}_0^*$ this time transformation can not be done since $c_j^{(l)} = 0$. Suppose that at some time $t \geq 0$ the process $(X^*(t), Y^*(t))$ changes from state $(k,i) \in \mathbf{S}_+^*$ to some state in $\mathbf{S}_0^*$. Suppose further that the first state within $\mathbf{S}_+^*$ visited thereafter is $(l,j) \in \mathbf{S}_+^*$. If $\sigma \geq 0$ is the point on the transformed time-scale corresponding to time t, then the process $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$ makes a (direct) transition at the point σ from $(k,i) \in \mathbf{S}_+^*$ to $(l,j) \in \mathbf{S}_+^*$. Thus $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$ *is not observed* on the states in $\mathbf{S}_0^*$. When $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$ makes such a transition corresponding to a visit of $(X^*(t), Y^*(t))$ to the set $\mathbf{S}_0^*$, an immediate reward is earned which is equal to the time that $(X^*(t), Y^*(t))$ spends within the set $\mathbf{S}_0^*$.

In the process $\{(\mathcal{X}(\sigma), \mathcal{Y}(\sigma)), \sigma \geq 0\}$ with state space $\mathbf{S}_+^*$ and generator $\mathcal{R}_+^{-1}\left[\mathcal{G}_+^* + \mathcal{G}_{+0}^*\mathcal{U}(0)\right]$ there are two types of transitions and two types of rewards. "Ordinary" transitions occur according to the transient generator $\mathcal{G}_+^*$, and "ordinary" rewards are earned at rate $\frac{k}{c_i^{(k)}}$ in state $(k,i) \in \mathbf{S}_+^*$. The other transitions and rewards are related as follows. The entry in row $(k,i) \in \mathbf{S}_+^*$ and column $(l,j) \in \mathbf{S}_0^*$ of the matrix $\mathcal{G}_{+0}^*$ gives the rate with which an (l,j)-event occurs. An (l,j)-event has two consequences: (i) a transition is made, and (ii) an instantaneous reward is earned. The instantaneous reward, and the state after an (l,j)-event are jointly distributed as the pair $(U_{l,j}, W_{l,j})$, and (the LST of) their joint distribution is given by the matrix $\mathcal{U}(s)$ for $\mathrm{Re}(s) \geq 0$. Note that the state after the (l,j) event may be the same as the state before the event. We emphasise that the jump-chains of the process $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$ and the process $(X^*(t), Y^*(t))$ *restricted to the set* $\mathbf{S}_+^*$ (often called a *censored process*) are identical.

Theorem 4.3.1 remains true with the above modifications, and so does Corollary 4.3.2 if we redefine the constant g^* as

$$g^* := \frac{1}{\overline{\pi}_+^* \mathcal{R}_+ \overline{1}}.$$

Remark 4.4.1 The discussion in Remark 4.3.1 also applies to this section. The relation $g^* = \frac{1}{c^* - \rho^*}$ discussed in Remark 4.3.2 is true for the redefined constant g^*. As in Remark 4.3.3, in the queueing system with one permanent customer, $c^* - \rho^*$ is the average capacity per unit of time assigned to the permanent customer. Remark 4.3.4 only needs to be modified to account for the transitions with instantaneous rewards. This can be done, as in Section 3.2, by "attaching" the instantaneous rewards to the birth of a nest of children. This way, the analysis can proceed even if the periods of service unavailability are generally distributed.

4.5 The proportionality result

We now discuss the proportionality between the conditional mean sojourn time and the amount of work brought into the system, in processor-sharing systems without random environment. This result is well known for the M/G/1 queue with processor sharing, see for instance Sakata et al. [99, Formula (10)], Sakata et al. [100, Formula (49)], or Kleinrock [55, Formula (4.17)]. Cohen [19, Formula (7.27)] found the proportionality property for the M/G/1/L queue with processor sharing and queue-dependent total service capacity (there called generalised processor sharing).

In this section we explain *why* this proportionality property holds, using the results from the random time-change method of Section 4.3. Note that since there is no random environment, this discussion only applies to the case with $\mathbf{S}_0^* = \emptyset$: If $c^{(k)} = 0$ for some $k \geq 1$, then the states with less than k customers are transient. For the M/G/1 queue with queue-dependent service rates the same arguments were used by Foley and Klutke [31]. We show that the arguments also apply to the M/G/1/L processor-sharing system with queue-dependent total service rates. A related discussion for the M/G/1 queue is given in Van den Berg [9, Remark 5.10, p. 115], and Van den Berg and Boxma [10, Remark 8.2].

In the absence of a random environment and with queue-independent arrivals (at rate λ), the queue length process $\{X(t),\ t \geq 0\}$ is an ordinary birth-death process. The queueing models of Remark 4.2.4 (M/M/1/L and M/M/1) possess these properties. Note however, that (unlike the M/M/1/L and M/M/1 models) the service rates may be queue-dependent, i.e. the $c^{(k)}$ may be different for different $k = 1, 2, \ldots, L$. The steady-state probabilities π_k, $k = 0, 1, \ldots, L$ of the process $X(t)$, and the steady-state probabilities π_k^*, $k = 1, 2, \ldots, L$ — *not including* $k = 0$ — of the process $X^*(t)$, satisfy:

$$\pi_k^* \frac{1}{k} c^{(k)} \sim \pi_{k-1}, \quad k = 1, 2, \ldots, L,$$

where the symbol $\sim$ means equality up to multiplication by a constant (independent of k). We already saw in Remark 4.3.2 that the steady-state distribution of the process $(\mathcal{X}(\sigma), \mathcal{Y}(\sigma))$ is given by the vector $\frac{1}{c^* - \rho^*} \overline{\pi}^* \mathcal{R}$. For the present case, in the absence of a random environment, we thus have for $k = 1, 2, \ldots, L$, that $\mathbf{P}\{\mathcal{X} = k\} = \frac{1}{c^* - \rho^*} \pi_k^* \frac{c^{(k)}}{k}$, and hence, $\mathbf{P}\{\mathcal{X} = k\} = \frac{\pi_{k-1}}{1 - \pi_L}$. This property has an interesting consequence: Suppose the queueing system under consideration is in steady state, and let the random variable X have this distribution: $\mathbf{P}\{X = k\} = \pi_k$. Since we assumed Poisson arrivals, from the PASTA (Poisson Arrivals See Time Averages) property, the number of customers seen by a newly arrived customer is distributed as X. Condition on the fact that the new customer is accepted, which occurs with probability $\mathbf{P}\{X < L\} = 1 - \pi_L$. Let the amount of work of the new (tagged) customer be $\tau > 0$, and denote his sojourn time by the random variable $V(\tau)$. Theorem 4.3.1 tells us that $V(\tau)$ is distributed as

$$\int_{\sigma=0}^{\tau} \frac{\mathcal{X}(\sigma)}{c^{(\mathcal{X}(\sigma))}} \mathrm{d}\sigma,$$

with $\mathcal{X}_0$ distributed as $X + 1$ *given that* $X < L$. However, this distribution is the steady-state distribution of the process $\mathcal{X}(\sigma)$, and so $\mathbf{P}\{\mathcal{X}(\sigma) = k\} = \frac{1}{1-\pi_L}\pi_{k-1}$, $k = 1, 2, \ldots, L$, for any $\sigma \in [0, \tau]$. Therefore, in steady state we find for the mean of the sojourn time $V(\tau)$ (of an *accepted* customer with service requirement τ):

$$\mathbf{E}[V(\tau)] = g^* \ \tau.$$

So, for the model with exponentially distributed service requirements we have explained why this proportionality occurs, namely because the stationary distribution of $\mathcal{X}(\sigma)$ is the same as that of X *given that* $X < L$. We can generalise our arguments to the M/G/1/L queue with processor sharing and queue-dependent service rates, cf. Cohen [19, Formula (7.27)] (for $L = \infty$ this was done in Foley and Klutke [31]). We give a brief outline of the proof: If the service requirements are distributed according to the distribution $B(x)$, $x \geq 0$, then

$$p_k(x_1, \ldots, x_k) = p_0 \frac{\lambda^k}{\prod_{j=1}^k c^{(j)}} \prod_{j=1}^k (1 - B(x_j)), \quad k = 1, 2, \ldots, L,$$

is the density function of there being k customers in the system with respective remaining service requirements $x_1, \ldots, x_k$, see Cohen [19, Formula (5.9)]. For this model we may apply the random time change to the system with one permanent customer, as described above: we "shrink" the time-scale by a factor $k/c^{(k)}$ when there are k customers in the system. Viewing the resulting process as a branching process, then $p_k(x_1, \ldots, x_k)$, for $k < L$, is also the density function (up to normalisation) of there being $k+1$ living individuals, the k non-permanent ones having respective remaining life times $x_1, \ldots, x_k$. It is beyond our purposes to work out the details at this point.

Remark 4.5.1 If we allow the arrival rate to depend on the queue length then the proportionality property is lost. The steady-state distribution seen upon arrival, or at arbitrary time points, no longer equals the steady-state distribution of the time-changed process. Under exponentiality assumptions this is easily checked by comparing the balance equations.

Remark 4.5.2 A related result regarding the proportionality property was obtained in Cohen [19, Theorem 5.3]. The model studied there is a closed queueing model with L customers, who are served according to the processor-sharing discipline with queue-dependent service rates. After having completed his service, a customer waits for a generally distributed time, and then enters the system again with a new (independently drawn) service requirement. It is shown that if an exogenous customer with an amount of work τ is brought into the system in steady state, his mean sojourn time is proportional to τ. In this model, the arrival process is obviously queue-dependent, and hence the proportionality result seems to contradict Remark 4.5.1. However, the considered model in Cohen [19] is fundamentally different from the above models: The exogenous customer may cause the number of customers in service to become $L+1$. Moreover, the arrival process is still determined by the ordinary customers, so that the queue-dependent arrivals in the original process and the time-changed process "cancel out". Again, under exponentiality assumptions this is easily seen from the balance equations.

4.6 Computation and approximation

We return to the general queueing model of Figure 4.1. Let $V(\tau)$ be the sojourn time of a customer with an amount of work τ, arriving to the system in steady state. In this section we show how $\mathbf{E}[V(\tau)]$ can be computed accurately.

Remark 4.6.1 In our presentation we required $\lambda^{(L)}=0$ so that no customers are lost. In many applications the arrival process is a Poisson process, and customers arriving when there are L other customers present are lost. Then it must be explicitly stated that the sojourn time of a customer is conditional on this customer not being rejected. As said before, this conditioning is inherent to our formulation. Poisson arrivals are thus incorporated by defining $\lambda^{(L)}=0$ and $\lambda^{(k)}=\lambda$, $k=0,1,\ldots,L-1$.

For $(k,i)\in\mathbf{S}^*$, denote by $a_{k,i}$ the steady-state probability that the system is in state (k,i) immediately after the arrival of a customer. The $a_{k,i}$ are the steady-state probabilities of a discrete-time Markov chain with transition probability matrix:

$$\mathcal{A}:=\begin{bmatrix} T^{(1,1)} & T^{(1,0)} & 0 & \cdots & \cdots & 0 \\ T^{(2,2)} & T^{(2,1)} & T^{(2,0)} & 0 & \cdots & \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & & T^{(L-1,2)} & T^{(L-1,1)} & T^{(L-1,0)} \\ T^{(L,L)} & \cdots & & \cdots & T^{(L,2)} & T^{(L,1)} \end{bmatrix}.$$

Here, for $k=1,\ldots,L-1$,

$$T^{(k,0)}=\left[-Q_d^{(k)}\right]^{-1}\Lambda^{(k)},$$

and for $k=1,\ldots,L$; $n=1,\ldots,k$,

$$T^{(k,n)}=\prod_{m=0}^{n-1}\left(\left[-Q_d^{(k-m)}\right]^{-1}M^{(k-m)}\right)\left[-Q_d^{(k-n)}\right]^{-1}\Lambda^{(k-n)}.$$

We now show how $\mathbf{E}[V(\tau)]$ can be computed after having determined the steady-state probabilities immediately after the arrival of a customer. For an alternative probabilistic algorithm to compute the *distribution* of $V(\tau)$ we refer to De Souza e Silva and Gail [108].

We only show how the computations can be done for the case $\mathbf{S}_0^* = \emptyset$, see Remark 4.6.3 for the case $\mathbf{S}_0^* \neq \emptyset$. Our starting point is the set of differential equations and initial conditions given in (4.10) and (4.11). Obviously, for $n \geq 1$,

$$\frac{\mathrm{d}^n}{\mathrm{d}\tau^n} \left(\mathbf{E}\left[V_{k,i}(\tau)\right]\right)_{k,i} \; |_{\tau=0} = \left(\mathcal{R}^{-1}\mathcal{G}^*\right)^{n-1} \mathcal{R}^{-1}\overline{1}. \tag{4.18}$$

We use Jensen's method to uniformise the generator $\mathcal{R}^{-1}\mathcal{G}^*$, and define the probability matrix

$$\mathcal{P}^* := I + \frac{1}{\eta}\mathcal{R}^{-1}\mathcal{G}^*,$$

with the scalar $\eta > 0$ being equal to minus the entry with largest absolute value (along the diagonal) of $\mathcal{R}^{-1}\mathcal{G}^*$. Assuming the Taylor-series of $\mathbf{E}\left[V_{k,i}(\tau)\right]$ around $\tau = 0$ exists (at the end we verify the result), and using (4.18) we may find:

$$\left(\mathbf{E}\left[V_{k,i}(\tau)\right]\right)_{k,i} = \frac{1}{\eta}\sum_{l=0}^{\infty}\left(1 - \mathrm{e}^{-\eta\tau}\sum_{k=0}^{l}\frac{(\eta\tau)^k}{k!}\right)\left(\mathcal{P}^*\right)^l \mathcal{R}^{-1}\overline{1}. \tag{4.19}$$

Noting that $k! \geq (l+1)!(k-l-1)!$, when $0 < l+1 \leq k$, we have:

$$0 \leq 1 - \mathrm{e}^{-\eta\tau}\sum_{k=0}^{l}\frac{(\eta\tau)^k}{k!} = \frac{\sum_{k=l+1}^{\infty}\frac{(\eta\tau)^k}{k!}}{\mathrm{e}^{\eta\tau}} \leq \frac{(\eta\tau)^{l+1}}{(l+1)!},$$

and hence the infinite sum in (4.19) exists for every $\tau \geq 0$. Moreover, by substitution it may be seen that it satisfies the differential equations and initial conditions (4.10) and (4.11).

Expression (4.19) for the $\mathbf{E}\left[V_{k,i}(\tau)\right]$ provides a numerically stable algorithm, since it only involves multiplication and addition of positive terms. Within the summation one needs to evaluate the "coefficients" $\mathrm{e}^{-\eta\tau}\sum_{k=l+1}^{\infty}\frac{(\eta\tau)^k}{k!}$, which can be done accurately by proper scaling of the terms (to avoid problems when $\eta\tau$ is large).

Remark 4.6.2 Instead of starting from the differential equations (4.10), we may start from the final Expression (4.8) in Corollary 4.2.3. Again we may use Jensen's uniformisation method to derive:

$$\left(\mathbf{E}\left[V_{k,i}(\tau)\right]\right)_{k,i} = \frac{\tau}{c^* - \rho^*}\overline{1} + \overline{\gamma} - \mathrm{e}^{-\eta\tau}\exp\left\{\eta\tau\mathcal{P}^*\right\}\overline{\gamma}. \tag{4.20}$$

However, for this approach one first needs to compute the vector $\overline{\gamma}$. Moreover, the vector $\overline{\gamma}$ contains negative elements which may cause the evaluation of $\mathrm{e}^{-\eta\tau}\exp\left\{\eta\tau\mathcal{P}^*\right\}\overline{\gamma}$ to be numerically unstable. However, no problems were encountered in the numerical experiments of Núñez Queija et al. [85], where both methods were used to compute the exact value of $\mathbf{E}[V_{k,i}(\tau)]$. In all cases the relative difference between the outcomes was of the order 10^{-8} or smaller (with values of τ up to 10 times the mean $1/\mu$).

Remark 4.6.3 When $\mathbf{S}_0^* \neq \emptyset$ we may proceed in a similar way. The starting point is then the set of differential equations mentioned in the proof of Theorem 4.4.3. For $(k,i) \in \mathbf{S}_+^*$, the $\mathbf{E}[V_{k,i}(\tau)]$ are found as before. However, first the $\mathbf{E}[U_{l,j}]$, for $(l,j) \in \mathbf{S}_0^*$, and the probability matrix $\mathcal{U}(0)$, need to be computed. Using Lemma 4.4.2, from the $\mathbf{E}[V_{k,i}(\tau)]$, $(k,i) \in \mathbf{S}_+^*$, also the $\mathbf{E}[V_{l,j}(\tau)]$, for $(l,j) \in \mathbf{S}_0^*$ can be computed. Note that $\mathbf{E}[V_{l,j}(0+)] = \mathbf{E}[U_{l,j}] > 0$, for $(l,j) \in \mathbf{S}_0^*$. As a consequence, $\mathbf{E}[V(0+)] > 0$, unless $p_{ij}^{(l-1)} = 0$, $\forall (l,j) \in \mathbf{S}_0^*$, $i \in E^{(l-1)}$.

Although Expression (4.19) provides a numerically stable algorithm to compute the $\mathbf{E}[V_{k,i}(\tau)]$, in general this task requires considerable computation time and memory space. Therefore it would be convenient to have a good approximation which is less computationally demanding. From Corollary 4.2.4 we have for the mean of $V(\tau)$:

$$\lim_{\tau\to\infty} \mathbf{E}[V(\tau)] - \frac{\tau}{c^* - \rho^*} \;=\; \gamma := \sum_{(k,i)\in\mathbf{S}^*} a_{k,i}\gamma_{k,i}.$$

This asymptotic relation can be used for a first approximation, i.e., $\mathbf{E}[V(\tau)] \approx \frac{\tau}{c^*-\rho^*} + \gamma$. Indeed, when the number of states of the random environment is small ($N \leq 5$), the asymptotic result may serve as a useful approximation for $\mathbf{E}[V(\tau)]$. For this case, the exact value and the asymptote typically look as shown in Figure 4.3. However, we shall see in the next section that for larger values of N (≥ 30) the asymptote may give a very poor approximation, whereas for those cases the tangent in the origin is often an excellent approximation of $\mathbf{E}[V(\tau)]$, even for relatively large values of τ, see for instance Figure 4.7. The slope δ of the tangent line is equal to the initial expected "delay per unit of service" upon arrival of a customer in steady state:

$$\delta := \sum_{(k,i)\in\mathbf{S}^*} a_{k,i}\frac{k}{c_i^{(k)}}. \tag{4.21}$$

Remark 4.6.4 When the tangent line in the origin is close to the exact value, the mean of $V(\tau)$ is "almost" proportional to τ, the proportionality constant being given by the slope δ of the tangent line.

In practice it is not clear beforehand which of the two approximations (the asymptote or the tangent in the origin) is best, the more since the quality of both approximations also depends on the transition rates of the random environment. However, in Experiment 3 of the next section it is observed that both approximations are an *upper bound* for $\mathbf{E}[V(\tau)]$, and that for practical purposes the minimum of the two gives a useful approximation. Therefore

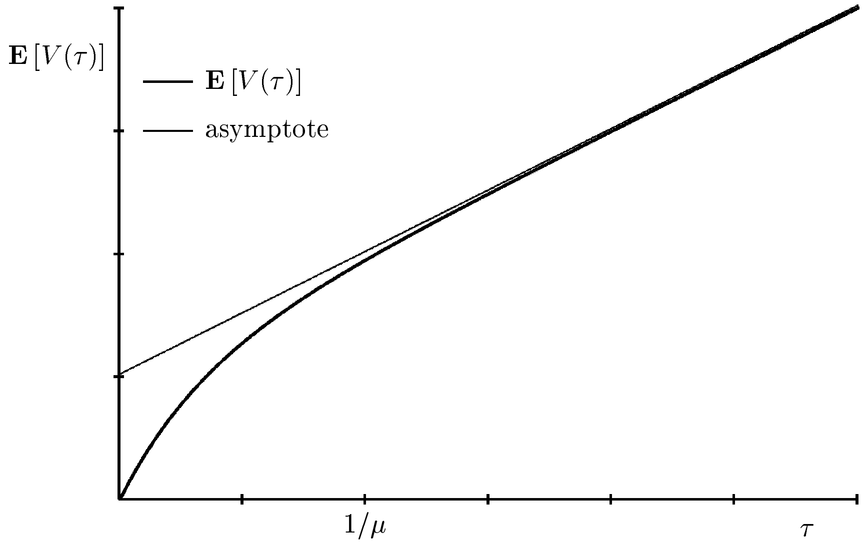

Figure 4.3: Example with $N = 5$

we propose to use the following refined approximation, by combining the two previously mentioned ones (for the case that $\mathbf{S}_0^* = \emptyset$).

$$\mathbf{E}[V(\tau)] \approx \min\left(\frac{\tau}{c^* - \rho^*} + \gamma, \delta\tau\right). \tag{4.22}$$

This approximation can be improved by computing more than the first coefficient (δ) of the Taylor-series. This may be done iteratively by using Expression (4.18), until two subsequent approximations are considered to be close enough. Note however that this procedure is not guaranteed to be numerically stable, since positive and negative numbers are added in each step. Therefore the roundoff errors may accumulate significantly in the iterative procedure.

Remark 4.6.5 The models evaluated in the next section form a special subclass of the general framework depicted in Figure 4.1. In particular, the capacity allocated to a single customer, $c_i^{(k)}/k$, is a non-increasing function of k (the total number of customers). For practical situations this seems to be a reasonable assumption.

4.7 Performance evaluation of a communication system

In this section we use the model and results of the previous sections of this chapter to evaluate the performance of a telecommunication system with elastic and stream traffic. The system under consideration may be modelled as described in Section 1.5. We briefly review the main features of that model. Two types of customers (traffic) are served by the system: elastic customers and stream customers. These customers arrive according to two independent Poisson processes with rates $\lambda^{(e)}$ and $\lambda^{(s)}$, respectively. Stream customers require a fixed amount of capacity $r^{(s)}$ during their holding times, which are assumed to be exponentially distributed with mean $h^{(s)}$. The work offered to the system by stream customers is denoted by $\rho^{(s)} := \lambda^{(s)}h^{(s)}$. The service requirements of elastic customers are exponentially distributed with mean $f^{(e)} = 1/\mu$, and the traffic load of elastic customers is given by $\rho^{(e)} := \rho = \lambda^{(e)}f^{(e)}$. At all times, the service rate of an individual elastic customer in the system must be between a minimum value $r_-^{(e)} \geq 0$ and a maximum value $r_+^{(e)} > 0$. The total service capacity of the system is constantly equal to C. By $X(t) = X^{(e)}(t)$ we denote the number of elastic customers in the system at time t. We note that these are served according to the processor-sharing discipline. The number of stream customers in the system at time t, $X^{(s)}(t)$, determines the state of the random environment of the processor-sharing queue: $Y(t) := X^{(s)}(t)+1 \in \{1, 2, \ldots, N\}$. Here it is more convenient to work with $X^{(s)}(t)$ instead of $Y(t)$. The state space of the process $(X^{(e)}(t), X^{(s)}(t))$ will be denoted by:

$$\mathbf{S}' = \left\{(k^{(e)}, k^{(s)}) \ : \ (k^{(e)}, k^{(s)} + 1) \in \mathbf{S}\right\}.$$

We denote the numbers of customers in steady state by $X^{(e)}$ and $X^{(s)}$, respectively. The capacity left over by stream customers, $C - r^{(s)}X^{(s)}(t)$, is potentially available to elastic customers. Thus if $X^{(e)}(t) = k^{(e)}$ and $X^{(s)}(t) = k^{(s)}$ then at time t the total capacity allocated to elastic customers $r_{k^{(s)}}^{(k^{(e)})} := c_{k^{(s)}+1}^{(k^{(e)})}$ is *at most*

$$\min\left\{k^{(e)}r_+^{(e)}, C - k^{(s)}r^{(s)}\right\},$$

and each of the $k^{(e)}$ elastic customers gets capacity $r_{k^{(s)}}^{(k^{(e)})}/k^{(e)}$. Occasionally customers must be rejected from the system upon arrival — we say that these customers are *blocked* — in order to ensure that at all times the capacity requirements are satisfied:

$$r_-^{(e)}X^{(e)}(t) + r^{(s)}X^{(s)}(t) \leq C,$$

for all $t \geq 0$. How the two different types of traffic are integrated in the system is now determined by the acceptance/rejection policy. Here, we consider policies that can be characterised by $\mathbf{S}'$, the state space of the process $(X^{(e)}(t), X^{(s)}(t))$, in the following way. Suppose $(X^{(e)}(t), X^{(s)}(t)) = (k^{(e)}, k^{(s)}) \in \mathbf{S}'$. Then, if a new elastic customer arrives at time t, it is accepted if $(k^{(e)} + 1, k^{(s)}) \in \mathbf{S}'$, and rejected otherwise. Similarly, a new stream customer is accepted if $(k^{(e)}, k^{(s)} + 1) \in \mathbf{S}'$. For this reason we henceforth call $\mathbf{S}'$ the *admissible region.*

Three "integration" strategies will be considered: complete segregation, full integration and a mixed strategy. Below, we describe these strategies and compare their respective efficiency gains. The QoS (Quality of Service) offered to stream customers is determined by the fraction of stream customers being blocked, which we denote by $p^{(s)}$. For elastic customers, there are two relevant performance measures: (i) the sojourn time (either V or $V(\tau)$, for given service requirement τ), and (ii) the blocking probability $p^{(e)}$. The steady-state probabilities $\mathbf{P}\left\{X^{(e)}=k^{(e)}, X^{(s)}=k^{(s)}\right\}$ determine the blocking probabilities and, using Little's formula, the (unconditional) mean sojourn time $\mathbf{E}[V]=\mathbf{E}\left[X^{(e)}\right] / \left((1-p^{(e)})\lambda^{(e)}\right)$.

Remark 4.7.1 In the model described in Section 4.1 there was no real blocking of (elastic) customers. If in state $(k^{(e)}, k^{(s)}) \in \mathbf{S}'$ a newly arriving elastic customer is blocked then the corresponding $\lambda_i^{(k)}=0$, where $k=k^{(e)}$ and $i=k^{(s)}+1$. The blocking probabilities of elastic and stream customers in the telecommunication model are obtained using the PASTA (Poisson Arrivals See Time Averages) property,

$$\begin{aligned} p^{(e)} &= \sum_{(k^{(e)},k^{(s)})\in\mathbf{B}^{(e)}} \mathbf{P}\left\{X^{(e)}=k^{(e)}, X^{(s)}=k^{(s)}\right\}, \\ p^{(s)} &= \sum_{(k^{(e)},k^{(s)})\in\mathbf{B}^{(s)}} \mathbf{P}\left\{X^{(e)}=k^{(e)}, X^{(s)}=k^{(s)}\right\}, \end{aligned}$$

with the "blocking regions" $\mathbf{B}^{(e)}$ and $\mathbf{B}^{(s)}$ defined by:

$$\begin{aligned} \mathbf{B}^{(e)} &:= \left\{(k^{(e)}, k^{(s)}) \in \mathbf{S}' \ : \ (k^{(e)}+1, k^{(s)}) \notin \mathbf{S}'\right\}, \\ \mathbf{B}^{(s)} &:= \left\{(k^{(e)}, k^{(s)}) \in \mathbf{S}' \ : \ (k^{(e)}, k^{(s)}+1) \notin \mathbf{S}'\right\}. \end{aligned}$$

4.7.1 Integration strategies

Complete segregation

In this first strategy there is no interaction between stream customers and elastic customers. The link capacity C is split into two parts: $C=C^{(e)}+C^{(s)}$. The capacity $C^{(e)}$ is permanently assigned to elastic customers, and $C^{(s)}$ to stream customers. Virtually there are two separate service systems. This strategy may be motivated by a need for low system complexity. However, it is to be expected that the efficiency of this strategy in terms of resource usage is not favourable. The admissible region in this case is given by:

$$\mathbf{S}'=\mathbf{S}^{(seg)} := \left\{(k^{(e)}, k^{(s)}) \in \mathbb{N}\times\mathbb{N} : k^{(e)} r_{-}^{(e)} \le C^{(e)}, k^{(s)} r^{(s)} \le C^{(s)}\right\}. \quad (4.23)$$

For stream customers this results in Erlang's loss model, see Tijms [112, Section 4.8.1]. In particular,

$$p^{(s)} = \frac{\left(\rho^{(s)}\right)^{L^{(s)}} / L^{(s)}!}{\sum_{k=0}^{L^{(s)}} \left(\rho^{(s)}\right)^{k} / k!}, \quad (4.24)$$

where $L^{(\mathrm{s})} := N - 1 = \lfloor C^{(\mathrm{s})}/r^{(\mathrm{s})} \rfloor$ is the maximum number of stream customers in the system. Recall that $\rho^{(\mathrm{s})} = \lambda^{(\mathrm{s})} h^{(\mathrm{s})}$ is the amount of work arriving per time unit due to stream customers.

For elastic customers, the resulting model is a processor-sharing queue with queue-length dependent service rates: for $0 \leq k^{(\mathrm{e})} \leq L^{(\mathrm{e})}$,

$$r^{(k^{(\mathrm{e})})} = \min\left\{k^{(\mathrm{e})} r_+^{(\mathrm{e})}, C^{(\mathrm{e})}\right\},$$

with $L^{(\mathrm{e})} := L = \left\lfloor C^{(\mathrm{e})}/r_-^{(\mathrm{e})} \right\rfloor$ the maximum number of elastic customers. For this queueing model, the performance measures of interest have been derived in closed form by Cohen [19]. Let,

$$\varphi_k := \left(\prod_{j=1}^{k} r^{(j)}\right)^{-1}, \quad k = 1, 2, ..., L^{(\mathrm{e})},$$

and $\varphi_0 := 1$. Then

$$p^{(\mathrm{e})} = \frac{\varphi_{L^{(\mathrm{e})}} \left(\rho^{(\mathrm{e})}\right)^{L^{(\mathrm{e})}}}{\sum_{j=0}^{L^{(\mathrm{e})}} \varphi_j \left(\rho^{(\mathrm{e})}\right)^j}, \tag{4.25}$$

$$\mathbf{E}\left[V(\tau)\right] = \frac{\tau \sum_{n=1}^{L^{(\mathrm{e})}} n\varphi_n \left(\rho^{(\mathrm{e})}\right)^{n-1}}{\sum_{j=0}^{L^{(\mathrm{e})}} \varphi_j \left(\rho^{(\mathrm{e})}\right)^j}. \tag{4.26}$$

As discussed in Section 4.5, $\mathbf{E}\left[V(\tau)\right]$ is proportional to the service requirement τ. Furthermore, the above results are valid for general service requirement distributions: $\mathbf{E}\left[V(\tau)\right]$ and $p^{(\mathrm{e})}$ depend only on the mean service requirement. If we take $r_-^{(\mathrm{e})} = 0$ (i.e., really "best-effort" customers) and $r_+^{(\mathrm{e})} \geq C^{(\mathrm{e})}$, the model for elastic customers reduces to the "standard" M/G/1 processor-sharing queue. Note that, in this case, $L^{(\mathrm{e})} = \infty$ and $\varphi_k = (1/C^{(\mathrm{e})})^k$, $k = 1, 2, \ldots$, hence, Expression (4.26) reads:

$$\mathbf{E}\left[V(\tau)\right] = \frac{\tau}{C^{(\mathrm{e})} - \rho^{(\mathrm{e})}}.$$

Full integration

Full integration is (in some sense) the "opposite" extreme strategy with respect to complete segregation: both types of customers completely share the capacity C. This strategy can potentially achieve a high system utilisation (compared to complete segregation). A new customer (of any type) is accepted if the guaranteed service rate is not violated for any customer in the system. Thus, the admissible region is given by:

$$\mathbf{S}' = \mathbf{S}^{(\mathrm{int})} := \left\{(k^{(\mathrm{e})}, k^{(\mathrm{s})}) \in \mathbb{N} \times \mathbb{N} : k^{(\mathrm{e})} r_-^{(\mathrm{e})} + k^{(\mathrm{s})} r^{(\mathrm{s})} \leq C\right\}.$$

In this case the capacity left over by stream customers is completely available to elastic customers:

$$r_{k^{(s)}}^{(k^{(e)})} = \min\left\{k^{(e)} r_+^{(e)}, C - k^{(s)} r^{(s)}\right\}. \tag{4.27}$$

Define the maximum number of stream customers when there are $k^{(e)}$ elastic customers, by

$$L_{k^{(e)}}^{(s)} := \left\lfloor \frac{C - k^{(e)} r_-^{(e)}}{r^{(s)}} \right\rfloor,$$

$k^{(e)} = 0, 1, \dots, L^{(e)}$. Obviously, $L_0^{(s)} = L^{(s)} = N - 1$.

The generator of the process $\left\{(X^{(e)}(t), X^{(s)}(t)), t \geq 0\right\}$ is given by $\mathcal{G}^{(\mathrm{int})} := \mathcal{G}$, using Definition (4.2) with $(L_k^{(s)} + 1)$-dimensional (square) matrices $Q_d^{(k)}$, $k = 0, 1, \dots, L$. Hence, $\Lambda^{(k-1)}$ is an $(L_{k-1}^{(s)} + 1) \times (L_k^{(s)} + 1)$ matrix and $M^{(k)}$ is an $(L_k^{(s)} + 1) \times (L_{k-1}^{(s)} + 1)$ matrix, for $k = 1, 2, \dots, L$. For appropriate indices, $\lambda_i^{(k)} = \lambda^{(e)}$, $p_{ij}^{(k)}$ and $m_{ij}^{(k)}$ are equal to 1 if $i = j$ and equal to 0 otherwise, and (except for the diagonal elements) the matrix $Q_d^{(k)}$ is equal to the generator of the queue-length process of the M/M/$L_k^{(s)}$/$L_k^{(s)}$ model. The diagonal elements are such that each row of $\mathcal{G}^{(\mathrm{int})}$ sums up to 0.

Mixed strategy

A possible draw-back of the full-integration strategy is that one type of traffic may be "blocked" from the system, when the other type (temporarily) generates a relatively large load. The idea behind the mixed strategy described below is to have the benefit of efficiency gain (as with full integration), but at the same time offering customers of both types a certain "protection" against the other.

As in the model with complete segregation, a fixed capacity $C^{(e)} > 0$ is exclusively reserved for elastic customers. The remaining capacity $C^{(s)} > 0$ is primarily dedicated to stream customers, but elastic customers may use the spare capacity (if any). However, this capacity is immediately re-allocated to stream customers, as soon as a new stream customer arrives. Therefore the capacity $C^{(e)}$ should always be sufficient to guarantee the minimum service rate $r_-^{(e)}$ to each elastic customer in the system. Hence, the admissible region is the same as for the complete-segregation model: $\mathbf{S}' = \mathbf{S}^{(\mathrm{mix})} := \mathbf{S}^{(\mathrm{seg})}$, see Definition (4.23). The service capacity allocated to the elastic customers is, as in the full-integration model, given by (4.27), with $C = C^{(e)} + C^{(s)}$.

Since elastic customers do not affect the acceptance/rejection nor the service of stream customers, $X^{(s)}(t)$ evolves as the queue-length process of the standard Erlang loss model, just as in the complete-segregation model. The process $(X^{(e)}(t), X^{(s)}(t))$ is again a finite inhomogeneous QBD process. Its generator $\mathcal{G}^{(\mathrm{mix})}$ is given by $\mathcal{G}$ from Definition (4.2) with $(L^{(s)}+1) \times (L^{(s)}+1)$ matrices $\Lambda^{(k)}$, $M^{(k)}$ and $D^{(k)}$. For appropriate indices, $\lambda_i^{(k)} = \lambda^{(e)}$, $p_{ij}^{(k)}$ and $m_{ij}^{(k)}$ are equal to 1 if $i = j$ and equal to 0 otherwise. Finally, for $k = 0, 1, \dots, L$, the matrix $Q_d^{(k)}$ is

Link	C	155	Mbit/sec.
	$C^{(e)}$	105–80–55–30–5	Mbit/sec.
Elastic customers	$f^{(e)}$	50	Mbit
	$r_-^{(e)}$	0	Mbit/sec.
	$r_+^{(e)}$	10–50–155	Mbit/sec.
Stream customers	$h^{(s)}$	10	sec.
	$r^{(s)}$	5	Mbit/sec.

Table 4.1: Reference parameter values

$C^{(s)}$	50	75	100	125	150
$\lambda^{(s)}$	0.446118	0.810804	1.203062	1.612456	2.033728

Table 4.2: Choice of $\lambda^{(s)}$ from Erlang's loss formula; $p^{(s)} = 0.01$

equal to the generator of the queue-length process of the M/M/$L^{(s)}$/$L^{(s)}$ model, except for the diagonal entries which are such that $\mathcal{G}^{(\mathrm{mix})}$ is a true generator.

4.7.2 Experiments

It should be emphasised that quite a number of parameters play a role in the model. This makes it impossible to draw general conclusions over the entire parameter space. Therefore, we fix a number of parameters at a realistic value. The reference parameter values that we used are listed in Table 4.1. We have chosen the parameters such that the means of the service requirements coincide for both customer types: $f^{(e)} = h^{(s)} r^{(s)}$. The "guaranteed rate" $r_-^{(e)}$ for elastic customers is taken equal to zero, i.e., elastic customers are best-effort customers. In the fourth experiment we also consider $r_-^{(e)} > 0$.

Experiment 1: Efficiency of the strategies

In our first experiment, we compare the efficiency of the three strategies (complete segregation, full integration and the mixed strategy). The efficiency is measured in terms of the maximum traffic load due to elastic customers (given by $\lambda^{(e)}$) under given performance restrictions. More precisely stated, for a given effective load of stream customers $\left(1 - p^{(s)}\right) \lambda^{(s)}$ we determine the maximum value for $\lambda^{(e)}$ such that the mean sojourn time $\mathbf{E}\,[V]$ for elastic customers does not exceed a pre-specified value. We choose the load of stream customers such that under complete segregation (or the mixed strategy) the blocking probability $p^{(s)}$ equals 0.01. The corresponding value of $\lambda^{(s)}$ for those two strategies is determined by Erlang's loss formula. In Table 4.2 the resulting values of $\lambda^{(s)}$ are given for several values of $C^{(s)}$. For example if $C^{(s)} = 50$ then $\left(1 - p^{(s)}\right) \lambda^{(s)} \approx 0.442$. Under the full-integration strategy, the acceptance of stream customers does de-

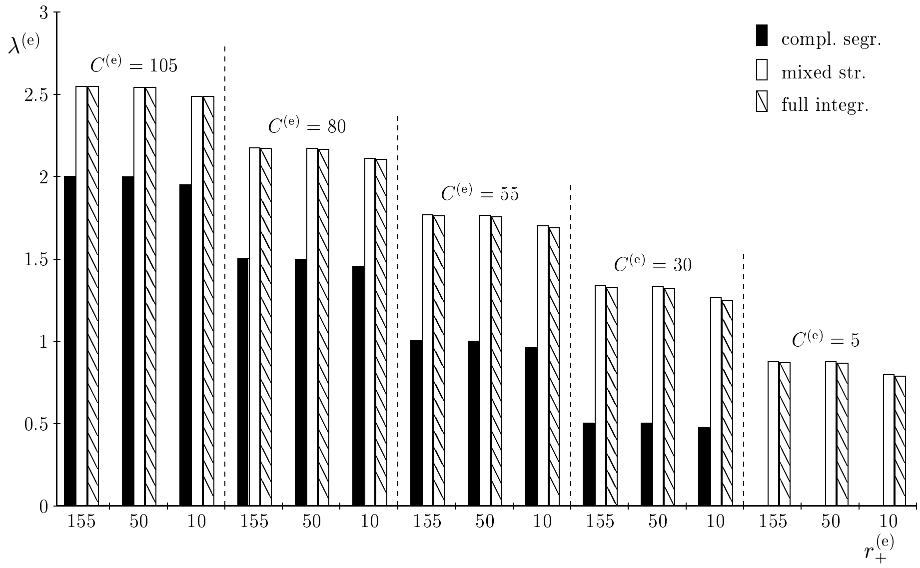

Figure 4.4: Efficiency of the three strategies (in terms of $\lambda^{(e)}$)

pend on the number of elastic customers. Hence, the load of stream customers can only be evaluated by computing the complete equilibrium distribution corresponding to $\mathcal{G}^{(\mathrm{int})}$. We need to do this repeatedly (for different values of $\lambda^{(s)}$ and $\lambda^{(e)}$) in order to achieve the same load of stream customers as under the other two strategies. Hence, for the full-integration strategy we simultaneously determine the appropriate $\lambda^{(s)}$ and the maximum $\lambda^{(e)}$. In all cases the blocking probability $p^{(s)}$ under full integration was smaller than 1%, hence the value of $\lambda^{(s)}$ chosen for the full-integration strategy was always between the chosen value from Table 4.2 and 0.99 times that value. Since these differences are only marginal, we do not report the values of $\lambda^{(s)}$ used for the full-integration strategy.

For the parameter values given in Table 4.1 we evaluated the efficiency of each of the three strategies. Note that we vary the parameters $C^{(s)} = C - C^{(e)}$ and $r_+^{(e)}$. The target value for the mean sojourn time of elastic customers is set at $\mathbf{E}[V] = h^{(s)} = 10$, i.e., on average both types of customers stay 10 seconds in the system. The results of this first experiment are shown in Figure 4.4. For $C^{(e)} = 5$ the allowed $\lambda^{(e)}$ (for the three values of $r_+^{(e)}$) in case of complete segregation is smaller than 10^{-5}. As expected, the mixed and full-integration strategies are considerably more efficient than the complete-segregation strategy: apparently, the elastic customers highly benefit from the fluctuating capacity that is left over by the stream customers. The differences between the mixed strategy and the full-integration strategy are very small. In all cases, the mixed strategy is at least as efficient as the full-integration strategy. Finally it is noted that the impact of $r_+^{(e)}$ on the efficiency of the strategies is very small. This is possibly due to the fact that the system is highly loaded: the number of elastic customers simultaneously present in the system is most of the time so large, that each of them receives less than 10 Mbit/sec. of the total available capacity (hence, it makes no difference whether $r_{+}^{(\mathrm{e})} = 10, 50$ or 155 Mbit/sec.).

Remark 4.7.2 As mentioned above, the above procedure to determine the appropriate $\lambda^{(\mathrm{s})}$ under full integration in all cases led to a marginally lower value of $\lambda^{(\mathrm{s})}$ than the value from Erlang's loss formula (the relative difference being less than 1%). Using the (higher) value of $\lambda^{(\mathrm{s})}$ from Erlang's loss formula under full integration leads to a lower maximum value of $\lambda^{(\mathrm{e})}$. However, the relative differences in the values of $\lambda^{(\mathrm{e})}$ never exceeded 1%.

Experiment 2: Time-scale differences

In the previous experiment, stream and elastic customers arrive/depart at more or less the same time scale. What if this is not the case, i.e., what if the number of stream customers fluctuates much faster or much slower than the number of elastic customers? To investigate this, we repeated Experiment 1 for the cases $h^{(\mathrm{s})} = 1$ (rapidly fluctuating stream traffic) and $h^{(\mathrm{s})} = 100$ (slowly fluctuating stream traffic). We refer to Section 2.8 for a similar experiment. All other parameters in Table 4.1 remain unchanged. The values of $\lambda^{(\mathrm{s})}$ in Table 4.2 are multiplied by a factor 10 in case $h^{(\mathrm{s})} = 1$, and by a factor 0.1 in case $h^{(\mathrm{s})} = 100$. This way $\rho^{(\mathrm{s})} = \lambda^{(\mathrm{s})}h^{(\mathrm{s})}$ is constant throughout the experiments and, hence, $p^{(\mathrm{s})}$ is always equal to 0.01 in the mixed and complete-segregation strategies. We observed that in all cases complete segregation is the least efficient, and that the mixed strategy outperforms the full-integration strategy (particularly when $h^{(\mathrm{s})} = 100$). For the mixed strategy, being the most efficient in all cases, the impact of the time-scale differences is illustrated in Figure 4.5. Based on the analysis in Section 2.6, one expects that when stream traffic fluctuates very fast ($h^{(\mathrm{s})} \downarrow 0$), the performance of elastic traffic is the same as for complete segregation with $C^{(\mathrm{e})}$ equal to the mean available capacity $C - (1 - p^{(\mathrm{s})})\rho^{(\mathrm{s})}$. In Figure 4.5, also the values of $\lambda^{(\mathrm{e})}$ are given for that case. The numerical results show that the maximum value of $\lambda^{(\mathrm{e})}$ — our measure for efficiency — increases when the number of stream customers fluctuates faster, that is when $h^{(\mathrm{s})}$ is smaller. Note that, as expected, the difference between the mixed strategy with $h^{(\mathrm{s})} = 1$ (i.e., the number of stream customers fluctuates relatively fast) and the complete-segregation strategy with $C^{(\mathrm{e})} = C - (1 - p^{(\mathrm{s})})\rho^{(\mathrm{s})}$ is negligible. We observe that, again, the impact of $r_{+}^{(\mathrm{e})}$ on the efficiency is very small.

Experiment 3: The conditional sojourn time

For the mixed strategy (the most efficient among the three), we consider the conditional mean sojourn time of elastic customers $\mathbf{E}[V(\tau)]$ as a function of the service requirement τ. In particular, we are interested in how fast $\mathbf{E}[V(\tau)]$ converges to its linear asymptote (as $\tau \to \infty$). The parameters $f^{(\mathrm{e})}$, $r_{-}^{(\mathrm{e})}$, $h^{(\mathrm{s})}$ and $r^{(\mathrm{s})}$ are fixed at their respective values given in Table 4.1, and $C^{(\mathrm{e})}$ is set equal to 80. The value of $\lambda^{(\mathrm{s})}$ (0.81) is again chosen such that $p^{(\mathrm{s})} = 0.01$, and $\lambda^{(\mathrm{e})}$

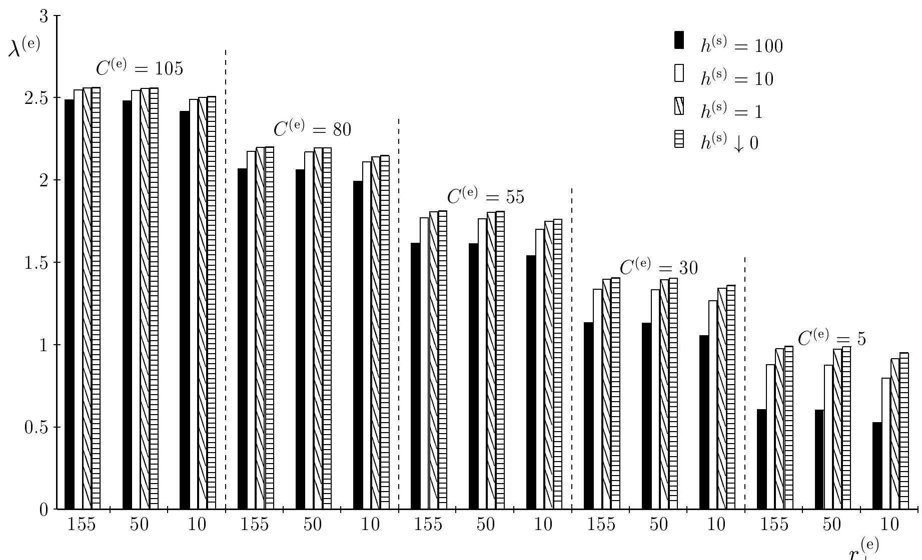

Figure 4.5: Impact of different time scales on the efficiency

is fixed at 2.17, which is the value computed in Experiment 1 with $r_+^{(e)} = \infty$. In Figure 4.6, $\mathbf{E}[V(\tau)]$ is given for the three values of $r_+^{(e)}$. We observe that $\mathbf{E}[V(\tau)]$ is considerably smaller for larger values of $r_+^{(e)}$. The exact curve of $\mathbf{E}[V(\tau)]$ and its asymptote for $r_+^{(e)} = 10$ are shown in Figure 4.7. For the other two values of $r_+^{(e)}$ the results are reported in Figure 4.8. For comparison, in the latter figure we have also plotted the results for $r_+^{(e)} = 10$. We observe that the distance between the actual curve and the asymptote increases with $r_+^{(e)}$.

Keeping $\lambda^{(e)}$ fixed, we repeated the above experiment for rapidly fluctuating stream traffic ($h^{(s)} = 1$) and for slowly varying stream traffic ($h^{(s)} = 100$). As in Experiment 2, the value of $\lambda^{(s)}$ when $h^{(s)} = 1$ and when $h^{(s)} = 100$ is found by multiplication by a factor 10 and by a factor 0.1, respectively, such that the load of stream customers (in terms of $\rho^{(s)}$) is the same in all cases. The outcomes can be found in Figures 4.9 and 4.10. For "fast" stream traffic we observed that the distance between $\mathbf{E}[V(\tau)]$ and its asymptote is considerably smaller. For "slow" stream traffic this distance is extremely large: the asymptotes lie outside the range of the vertical axis in Figure 4.10 (they intersect the vertical axis above the value 200).

The results show that in general the asymptote does not give a useful approximation for $\mathbf{E}[V(\tau)]$. An additional numerical study indicates that, as we mentioned in Section 4.6, a good approximation is often provided by the tangent of the curve in the origin. In Figure 4.7, for values of τ smaller than five times the mean service requirement $f^{(e)} = 50$ Mbit (with exponential services this is the case for 99.9% of the customers), the relative difference between $\mathbf{E}[V(\tau)]$ and the tangent in zero is less than 2.5%. Recall that the slope of this tangent

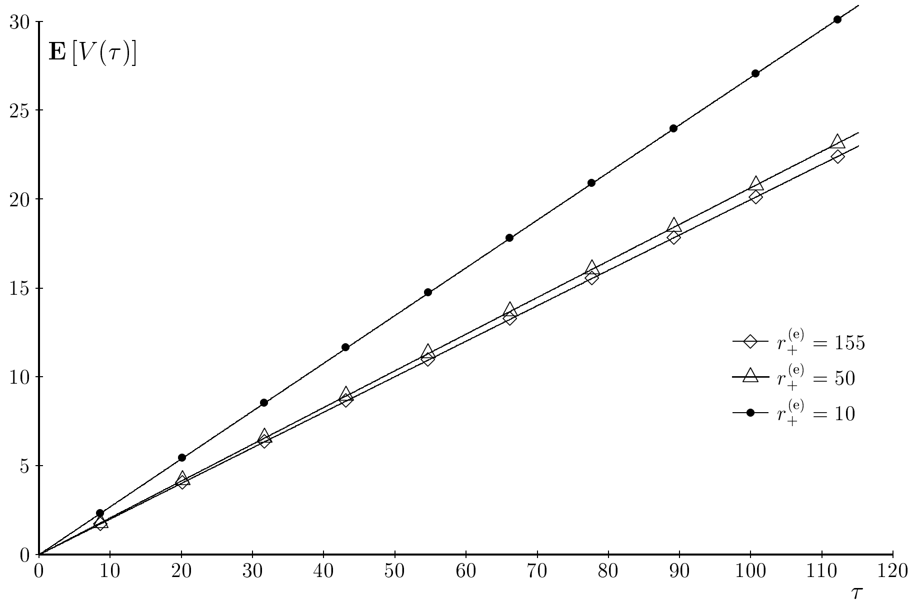

Figure 4.6: $\mathbf{E}[V(\tau)]$; for $h^{(s)} = 10$ and $r_+^{(e)} = 10, 50, 155$

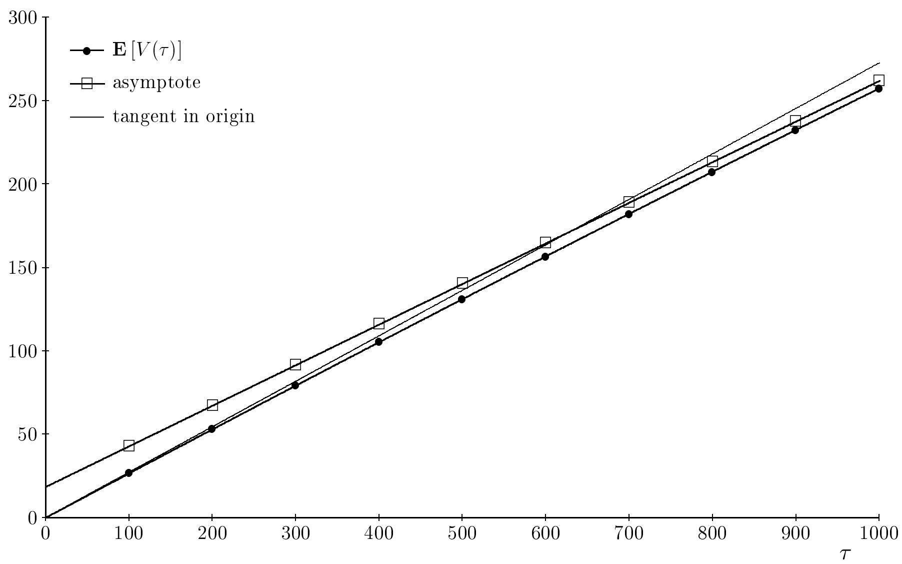

Figure 4.7: Asymptote and tangent line of $\mathbf{E}[V(\tau)]$; for $h^{(s)} = 10$ and $r_+^{(e)} = 10$

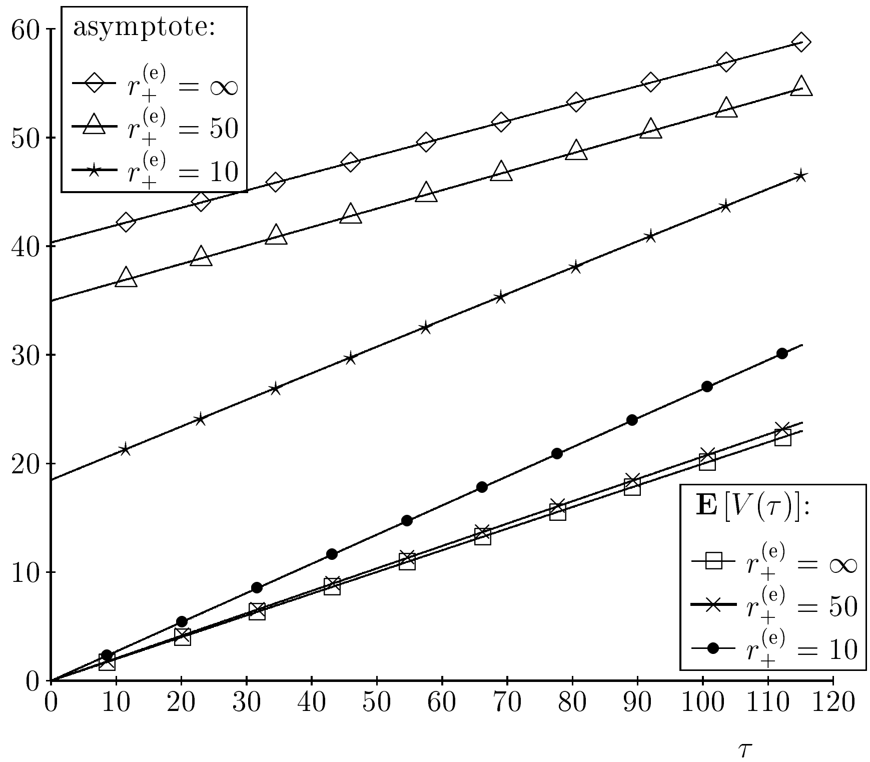

Figure 4.8: $\mathbf{E}[V(\tau)]$ and its asymptote; for $h^{(s)} = 10$

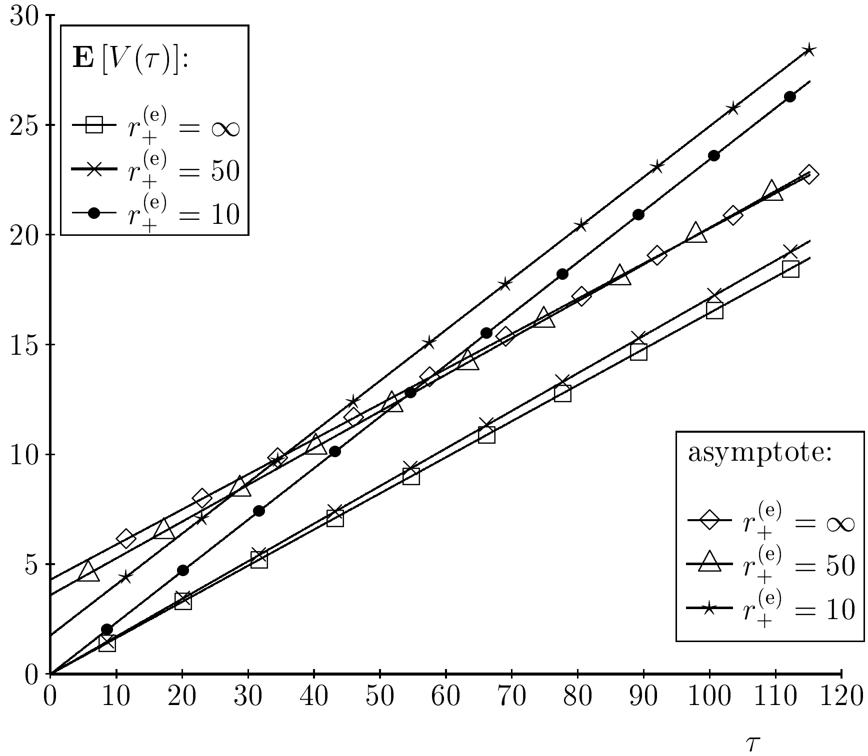

Figure 4.9: $\mathbf{E}[V(\tau)]$ and its asymptote; for $h^{(s)} = 1$

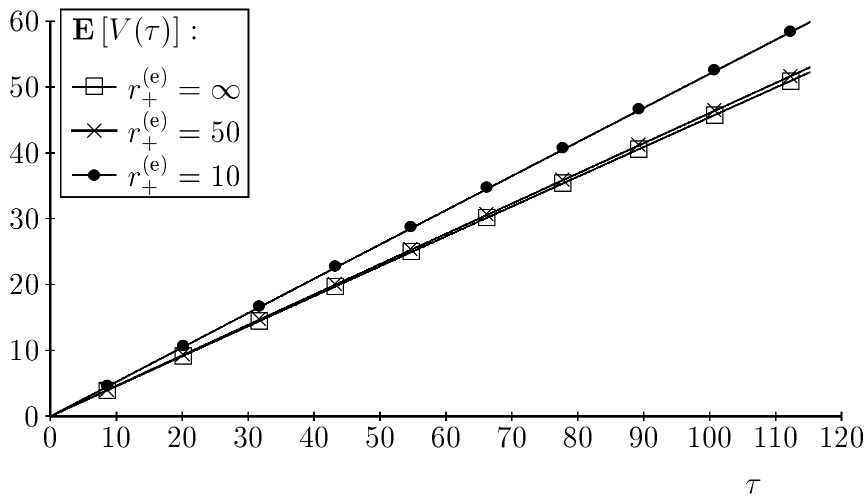

Figure 4.10: $\mathbf{E}[V(\tau)]$; for $h^{(s)} = 100$

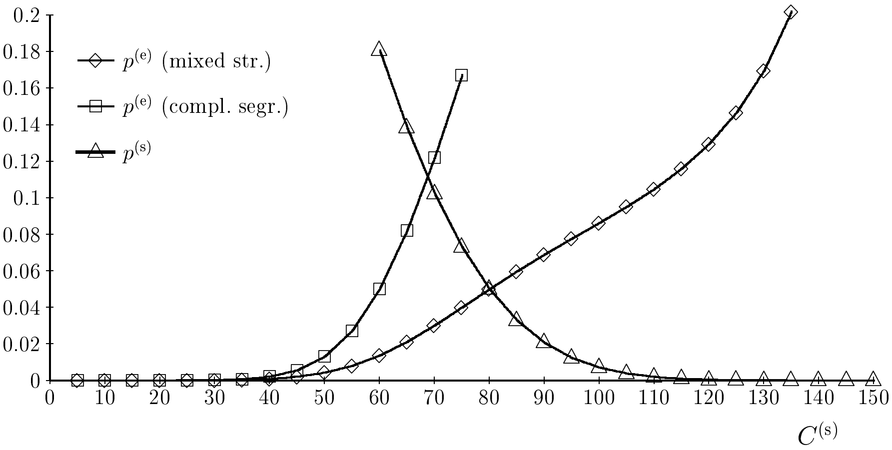

Figure 4.11: Blocking probabilities

line can be computed from the steady-state distribution, see Expression (4.21). The numerical results validate the proposed approximation of $\mathbf{E}[V(\tau)]$ given in Expression (4.22).

Experiment 4: Blocking probabilities

In our last experiment we consider the situation that the elastic customers are guaranteed a certain minimum capacity $r_-^{(e)}$. For the mixed strategy, we study the impact of $C^{(s)}$ on the blocking probabilities $p^{(e)}$ and $p^{(s)}$ of the elastic customers and the stream customers, respectively. As before, we choose $f^{(e)} = 50$ Mbit and $r^{(s)} = 5$ Mbit/sec. Furthermore, $h^{(s)} = 10$ seconds, $r_-^{(e)} = 5$ Mbit/sec. (i.e., the sojourn time of a customer with service requirement τ Mbit is bounded by $\tau/5$ seconds), and $r_+^{(e)} = 155$ Mbit/sec. We fix the customer arrival intensities at $\lambda^{(e)} = 1.90$ and $\lambda^{(s)} = 1.15$. These values are chosen such that $p^{(e)} = p^{(s)} = 0.05$ in the mixed strategy with $C^{(e)} = 75$ Mbit/sec. The results are shown in Figure 4.11. It is seen that the blocking probability for the stream customers decreases sharply when $C^{(s)}$ increases, while the blocking probability for the elastic customers grows only moderately. Note that, as $C^{(s)}$ increases, the amount of bandwidth $(C^{(e)} = C - C^{(s)})$ *reserved* for elastic customers decreases. A part of this re-assigned bandwidth is however not used by the stream customers. This amount of bandwidth, $C^{(s)} - (1-p^{(s)})\rho^{(s)}$, allocated to, but not used by the stream customers, is apparently very well exploited by the elastic customers. This is confirmed by the results for the blocking probability of elastic customers in the corresponding complete-segregation case, which are also shown in Figure 4.11.

4.7.3 Conclusions from the experiments

We used the model of this chapter to study the integration of stream traffic and elastic traffic in a multiservice communication system. Our model enables efficient computation of the relevant performance measures, in particular the blocking probabilities of both traffic types and the mean transfer (sojourn) times of elastic traffic (customers). Our numerical study validates the proposed approximation of the conditional sojourn time of elastic customers $\mathbf{E}[V(\tau)]$ given in Expression (4.22). This approximation only depends on steady-state characteristics and can therefore be efficiently computed.

The numerical output was used for assessing and comparing the efficiency gains achieved by the three integration strategies. We saw that "dynamic" integration (using full integration or the mixed strategy) of stream and elastic customers in a multiservice network is much more efficient with respect to the use of network resources, than having two dedicated networks for the two customer types (i.e., complete segregation). The so-called mixed strategy is slightly more efficient than full integration and has the additional advantage of offering both types of traffic a certain protection against the other, when the latter (temporarily) generates a relatively large load. Other integration schemes — like trunk reservation — also fall within the framework of our model. Comparison of the efficiency of such strategies with those considered here is an interesting topic for further research.

4.8 Generalisations

In Section 4.1 we made some assumptions which are not essential for the analysis, but facilitated the presentation and discussion. In this section we relax some of the assumptions and show how the resulting models either fit into the framework, or how they can be included in an analogous but generalised analysis.

4.8.1 Service requirements of phase-type

We may allow the service requirements of customers in the queueing system of Figure 4.1 to be of phase-type. By phase-type distributions we mean the class of distributions presented in Neuts [81, Chapter 2]. In order to preserve the Markovian description of our model, some additional state-descriptors must be added. For the analysis of the sojourn time conditioned on the amount of work, in the queueing model with one permanent customer, a state is determined by the number of customers (excluding the permanent one) in each service phase together with the state of the random environment. Thus if the service requirement distribution consists of P phases, then the state space is given by:

$$\mathbf{S}^* := \left\{ (k_1, k_2, \ldots, k_P; i) \;\middle|\; \begin{array}{l} 0 \leq k_1 + k_2 + \ldots + k_P \leq L-1, \\ k_j \in \{0, 1, 2, \ldots, L-1\}, \\ i \in \{1, 2, 3, \ldots, N\} \end{array} \right\}.$$

Note that we lose the QBD structure, which was convenient for computation of various entities (the new process could be called a multi-dimensional QBD process). Note also that, when studying the process with one permanent customer, the role of the random environment and the service phases of non-permanent customers is not fundamentally different. We may redefine the random environment such that it also contains the service phases of non-permanent customers, and then view the resulting model as a special case of the earlier model with $L = 1$. In particular we find that, also for phase-type services, the conditional mean sojourn time as a function of the amount of work τ, has an asymptote for $\tau \to \infty$.

For the representation of sojourn times (not conditioned on the amount of work) as absorption times in an appropriate Markov process, we need to add yet another descriptor to the state space, namely the phase of the tagged customer. Then Theorem 4.2.1 again applies.

4.8.2 Other service disciplines

Our model of Section 4.1 also includes other service disciplines. For instance discriminatory processor sharing (sometimes called weighted processor sharing) which contains (ordinary) processor sharing as a special case. Discriminatory processor sharing is of great interest for applications. For this service discipline several classes of customers are identified, numbered as $1, 2, \ldots, J$. With customer class j a weight $w_j > 0$ is associated. If there are k_j customers of class j, $j = 1, 2, \ldots, J$, then each of these gets a fraction $\frac{w_j}{k_1 w_1 + \ldots + k_J w_J}$ of the total (available) capacity. In our model this capacity may be a function of the state of a random environment and the numbers k_j, $j = 1, 2, \ldots, J$. If we are interested in the (conditional) sojourn time of customers of class 1, then we may view the model in the framework of Section 4.1 by extending the random environment with the tuples $(k_2, \ldots, k_J)$ containing the number of customers of all other classes.

As in Section 4.8.1, we may allow for phase-type distributions for each of the customer classes. In our state description we need to record the number of customers of any class in each particular service-phase. The number of (other) customers of the class under consideration in each possible service-phase also needs to be incorporated in the random environment.

Similarly, other service disciplines — including FCFS and LCFS (Last Come First Served) — may be incorporated by a proper definition of the random environment. Again we emphasise that with these generalisations, the computational complexity may be increased tremendously. These generalisations, however, retain the qualitative properties such as the existence of an asymptote for the conditional mean sojourn time.

4.8.3 Infinite state space

In Section 4.1 we assumed $L < \infty$ and $N < \infty$. Here we discuss the case where either of these, or both, are infinite. The results obtained in this chapter may be generalised to infinite state spaces, under recurrence conditions which are stronger than requiring ergodicity. For instance, the existence of the vector $\overline{\gamma}$ in Corollaries 4.2.3 and 4.4.4 is not ensured if we only assume ergodicity. This issue is related to convergence of the value iteration algorithm for Markov-Reward (decision) processes on countable state spaces, see for instance Sennott [106].

In applications, it is usually the case that the $c_i^{(k)}$ are uniformly bounded from above, so that the rewards in the Markov-Reward processes of the proofs of Corollaries 4.2.3 and 4.4.4 are uniformly bounded. In that case the vector $\overline{\gamma}$ exists under the assumption of ergodicity. To see this we may proceed as in Tijms [112, p. 188] to construct a relative reward vector which satisfies the conditions given for $\overline{\gamma}$ in Corollaries 4.2.3 and 4.4.4. In the same way, we may show that the mean of these constructed relative rewards exists and is finite, so that we may normalise as required in Corollaries 4.2.3 and 4.4.4.

Moreover, in applications when $L = \infty$ and $N < \infty$, it is often the case that the QBD process with generator $\mathcal{G}$ given by Definition (4.2) is homogeneous beyond some level, i.e. there is a positive integer K such that $M^{(k)} = M$, $Q_d^{(k)} = Q_d$, and $\Lambda^{(k)} = \Lambda$, for all $k \geq K$ (see for instance Núñez Queija et al. [85]). The ergodicity condition is then $\overline{p}\Lambda\overline{1} < \overline{p}M\overline{1}$, with $\overline{p}\,[M + Q_d + \Lambda] = \overline{0}$, where $\overline{0}$ is a vector of zeroes, see Neuts [81, Theorem 3.1.1].

We finally remark that for infinite generators, the exponential function as in (4.7) may be defined by its Taylor-series representation.

4.9 Concluding remarks

In this chapter we studied sojourn times of customers in a Markovian queueing system with processor sharing, in which arrival and service rates may depend on the number of customers already in the system *and* on the state of a random environment. The random environment itself may be dependent on the number of customers in the system. For this model we first represented the sojourn time as the absorption time in an appropriate Markov process. Particular attention was paid to sojourn times conditioned on the amount of work. For these, we found a closed-form solution for the LST, and in particular for its mean. We showed that as a function of the service requirement, the conditional mean sojourn time has a linear asymptote. By means of the method of random time change, the conditional sojourn times were represented by rewards in a particular Markov-Reward process. The latter was shown to be closely related to the branching process of the previous chapter, which previously has been used in the literature to study processor-sharing systems with constant (available) service capacity. For those systems it is known that the conditional mean sojourn time is proportional to the amount of work. This property (which does not hold for our model with fluctuating service capacity) was explained by comparing the steady-state distributions of the original queueing model and the model obtained by the random time change.

We discussed how the conditional mean of the sojourn times as a function of the service requirement may be computed. A numerically stable algorithm was developed, but the computational complexity calls for reliable and efficient approximations. We applied our results for the numerical evaluation of a particular telecommunication system with integration of stream and elastic traffic, and compared different integration strategies. The numerical results also motivated an approximation of the conditional mean sojourn time, which in all tested cases proved to be both conservative and useful for practical purposes.

The analysis was shown to include the case of service requirements with a phase-type distribution (leading to a "multi-dimensional" QBD process). We also saw that the more general discriminatory processor-sharing service discipline fits into our framework. We discussed extensions to infinite state spaces, and showed that for uniformly bounded service rates the analysis still applies. In particular we found that, for these generalisations, the conditional mean sojourn time as a function of the service requirement τ, has an asymptote for $\tau \to \infty$.

Chapter 5

Asymptotics for heavy-tailed sojourn time distributions

In Chapters 3 and 4 we studied sojourn times in processor-sharing queues with varying service capacity. The focus was on the distribution and the moments of the sojourn time of a customer conditional on the customer's service requirement. Different from the previous chapters, we now assume that the service requirements have a heavy-tailed distribution. We develop a new approach for analysing tail distributions of sojourn times under this assumption. It is based on an extension of our previous analysis of sojourn times conditional on the service requirements.

The analysis of queueing models with heavy-tailed service requirement distributions is an important issue in performance evaluation. The class of heavy-tailed distributions includes all distributions for which not all moments are finite. The exponential distribution is *not* heavy-tailed. An important subclass of heavy-tailed distributions consists of regularly varying distributions. A distribution function $H(x)$, $x \geq 0$, is said to have a regularly varying tail (at infinity) with index $-\zeta \in \mathbb{R}$ if, for arbitrary $t > 0$,

$$\lim_{x\to\infty} \frac{1 - H(tx)}{1 - H(x)} \quad = \quad t^{-\zeta}.$$

It was already shown by Cohen [18] that in the G/G/1 queue with the FCFS (First Come First Served) discipline, the waiting-time distribution is regularly varying of index $1 - \zeta$ if and only if the distribution of the service requirements is regularly varying of index $-\zeta$, where $\zeta > 1$ (to ensure a finite mean service requirement). Thus, the waiting-time distribution (and, hence, the sojourn time distribution) is as heavy as the integrated-tail distribution of the service requirement. Assuming Poisson arrivals and a regularly varying tail of the service requirement distribution with $1 < \zeta < 2$, Anantharam [6] has shown that the mean of the sojourn time is infinite for any *non-preemptive* service discipline. In contrast, Anantharam [6] also showed that, under the same assumptions on the arrival process and the service requirements, there exist preemptive service disciplines for which the mean sojourn time is finite. More specifically, Zwart and Boxma [128] proved that in the M/G/1 queue with *processor sharing* the tail of the sojourn time distribution is exactly as heavy as that of the service requirement distribution when the latter has a regularly varying tail. In this chapter we extend Zwart and Boxma's result to the model of Chapter 3 where the server is subject to interruptions. Our approach is based on Markov's inequality and it readily leads to a new and simpler proof of Zwart and Boxma's result (the original proof was based on transform techniques).

We also apply the method to the M/G/1 queue (with constant service capacity) for two other service disciplines: the FBPS (Foreground-Background Processor Sharing) and the SRPT (Shortest Remaining Processing Time) discipline. Under the SRPT discipline, at all times the total capacity is used to serve the customer in the system which has the smallest remaining service requirement, see Schrage and Miller [102]. By analogy, the FBPS discipline could alternatively be called the "least attained processing time" discipline: at all times the service capacity is used to serve the customer(s) which so far have received the least amount of service, see Kleinrock [55] or Yashkov [122]. For the FBPS and the SRPT disciplines the result (that the sojourn time distribution and the service distribution are equally heavy) is new. Moreover, it is unlikely that for these disciplines this result is easy to obtain using transform techniques, due to the complicated form of the LST (Laplace-Stieltjes Transform) of the sojourn time distributions.

The chapter is organised as follows. In Section 5.1 we derive conditions under which the tail of the sojourn time distribution is exactly as heavy as that of the service requirement distribution. The conditions involve the distribution of the service requirement and that of the sojourn time conditional on the service requirement. In Section 5.2 we show that these conditions are satisfied in the M/G/1 queue for three different service disciplines: processor sharing, FBPS and SRPT. The main goal of the chapter is to show that the conditions are also satisfied in the on/off processor-sharing model of Chapter 3 when the service requirements have a heavy-tailed distribution. In Section 5.3 we review the model and extend some of the results of Chapter 3 to the case of generally distributed service requirements. This is done, in particular, for the decomposition of the sojourn time into "fundamental random variables". In Section 5.4 we study the first and second moments of these random variables, specifically when the service requirement is large. Then, in Section 5.5, we study the system in steady state. Unfortunately, the distributions of the number of customers in the system and/or their respective remaining service requirement are not known. However, the distribution of the total amount of work in the system allows further analysis. The desired tail equivalence is then proved in Section 5.6. Finally, Section 5.7 concludes the chapter.

5.1 Sufficient conditions for tail equivalence

We state the main result in a general setting. Let B be a non-negative random variable with distribution function $B(x)$, $x \geq 0$. For every $\tau \geq 0$ let $V(\tau) \geq 0$ be a non-negative random variable independent of B. The random variable $V(B)$ is well defined and its distribution function is given by

$$\mathbf{P}\{V(B) \leq t\} \;=\; \int_{\tau=0}^{\infty} \mathbf{P}\{V(\tau) \leq t\} \,\mathrm{d}B(\tau).$$

Remark 5.1.1 The choice of notation for the random variables $V(\tau)$ and B is consistent with our previous notation. The results derived in this section will be applied in the next sections to service systems where the service requirements are distributed as B, and the sojourn times of customers with service requirement τ are distributed as $V(\tau)$. Furthermore, the unconditional sojourn time of an arbitrary customer is distributed as $V(B)$.

Assumption 5.1.1 *The tail of the service requirement distribution* $\overline{B}(x) := 1 - B(x)$ *is of intermediate regular variation at infinity, i.e.,*

$$\liminf_{\varepsilon \downarrow 0} \liminf_{x \to \infty} \frac{\overline{B}(x(1+\varepsilon))}{\overline{B}(x)} \;=\; 1.$$

When this assumption is satisfied, we write $\overline{B}(x) \in \mathcal{IRV}$. Observe that the above definition of intermediate regular variation is equivalent with

$$\limsup_{\varepsilon \downarrow 0} \limsup_{x \to \infty} \frac{\overline{B}(x(1-\varepsilon))}{\overline{B}(x)} \;=\; 1,$$

and that, in both characterisations, we may replace the first lim inf and the first lim sup, repectively, by the ordinary limit. In particular, all functions with a regularly varying tail are of intermediate regular variation, see Cline [16] for a discussion. Assumption 5.1.1 implies that there exist numbers $\zeta \in (0,\infty)$, $x_0 \in (0,\infty)$, and $\eta \in (0,1)$ such that, for all $x_2 \geq x_1 \geq x_0$,

$$\frac{\overline{B}(x_2)}{\overline{B}(x_1)} \;\geq\; \eta \left(\frac{x_2}{x_1}\right)^{-\zeta}, \tag{5.1}$$

see Appendix 5.A.

Assumption 5.1.2 *For some* $g^* > 0$,

$$\lim_{\tau \to \infty} \frac{\mathbf{E}[V(\tau)]}{\tau} \;=\; g^*, \tag{5.2}$$

and there exist $\kappa > \zeta$ *and* $\delta > 0$ *such that:*

$$\lim_{\tau \to \infty} \tau^{-\kappa+\delta} \mathbf{E}\left[\left| V(\tau) - \mathbf{E}[V(\tau)] \right|^{\kappa}\right] \;=\; 0, \tag{5.3}$$

i.e.,

$$\mathbf{E}\left[\left|\, V(\tau) - \mathbf{E}\left[V(\tau)\right]\,\right|^{\kappa}\right] \;=\; \mathrm{o}(\tau^{\kappa-\delta}), \quad \tau \to \infty,$$

where $\zeta \geq 0$ is as in Relation (5.1). Moreover, for all $t \geq 0$, the probability $\mathbf{P}\{V(\tau) > t\}$ is non-decreasing in $\tau \geq 0$. Hence, all moments $\mathbf{E}[V(\tau)^n]$, $n \in \mathbb{N}$, are non-decreasing in τ.

In this section we suppose that Assumptions 5.1.1 and 5.1.2 are satisfied, and g^*, ζ, κ and δ will be as described above. First we formulate the main result of this section in the next theorem. In the proof we use two lemmas, which we prove subsequently.

Theorem 5.1.1 *Suppose Assumptions 5.1.1 and 5.1.2 are satisfied. Then the tail distributions of the random variables B and $V(B)$ are equivalent in the sense that:*

$$\lim_{x\to\infty} \frac{\mathbf{P}\{V(B) > g^* x\}}{\mathbf{P}\{B > x\}} \;=\; 1.$$

Proof The proof is given in two parts. First we write, for $\varepsilon > 0$,

$$\begin{aligned}\mathbf{P}\{V(B) > g^* x\} \;\leq\; & \mathbf{P}\{V(B) > g^* x; B \leq x(1-\varepsilon)\} \\ & + \mathbf{P}\{B > x(1-\varepsilon)\}.\end{aligned}$$

By Lemma 5.1.2 below and the fact that $\overline{B}(x) \in \mathcal{IRV}$ we may neglect the first term in the right-hand side. Hence,

$$\limsup_{x\to\infty} \frac{\mathbf{P}\{V(B) > g^* x\}}{\mathbf{P}\{B > x\}} \;\leq\; \limsup_{x\to\infty} \frac{\mathbf{P}\{B > x(1-\varepsilon)\}}{\mathbf{P}\{B > x\}}.$$

Letting $\varepsilon \downarrow 0$, the right-hand side tends to 1.
For the second part of the proof we write, for $\varepsilon > 0$,

$$\mathbf{P}\{V(B) > g^* x\} \;\geq\; \mathbf{P}\{V(B) > g^* x; B > x(1+\varepsilon)\}.$$

Combining this with Lemma 5.1.3 below, we have

$$\liminf_{x\to\infty} \frac{\mathbf{P}\{V(B) > g^* x\}}{\mathbf{P}\{B > x\}} \;\geq\; \liminf_{x\to\infty} \frac{\mathbf{P}\{B > x(1+\varepsilon)\}}{\mathbf{P}\{B > x\}}.$$

By Assumption 5.1.1 the right-hand side tends to 1 as $\varepsilon \downarrow 0$. □

In the proof we used two lemmas which we prove next. The first one states that "when B is small, $V(B)$ can not be large". To prove this we use the following form of Markov's inequality (see Williams [118, Section 6.4]) for the tail distribution of $V(\tau)$:

$$\mathbf{P}\{V(\tau) - \mathbf{E}[V(\tau)] > t\} \leq \frac{\mathbf{E}\left[\left|\, V(\tau) - \mathbf{E}[V(\tau)]\,\right|^{\kappa}\right]}{t^{\kappa}}, \tag{5.4}$$

for all $\tau \geq 0$ and $t > 0$. When $\kappa = 2$, which is the case in most of the examples studied in the next sections, this reduces to Chebyshev's inequality:

$$\mathbf{P}\{V(\tau) - \mathbf{E}[V(\tau)] > t\} \leq \frac{\mathbf{Var}[V(\tau)]}{t^2}. \tag{5.5}$$

Lemma 5.1.2 *Suppose Assumptions 5.1.1 and 5.1.2 are satisfied. Then, for fixed* $\varepsilon \in (0,1)$,

$$\lim_{x\to\infty} \frac{\mathbf{P}\{V(B) > g^* x; B \leq x(1-\varepsilon)\}}{\mathbf{P}\{B > x(1-\varepsilon)\}} \quad = \quad 0.$$

Proof We prove the lemma using the following relations, which hold for x "large enough":

$$\begin{aligned}
&\mathbf{P}\{V(B) > g^* x; B \leq x(1-\varepsilon)\} \\
&= \int_{\tau=0}^{x(1-\varepsilon)} \mathbf{P}\{V(\tau) > g^* x\} \,\mathrm{d}B(\tau) \\
&\leq \int_{\tau=0}^{x(1-\varepsilon)} \mathbf{P}\{V(\tau) - \mathbf{E}[V(\tau)] > g^* x - \mathbf{E}[V(x(1-\varepsilon))]\} \,\mathrm{d}B(\tau) \\
&\leq \frac{\int_{\tau=0}^{x(1-\varepsilon)} \mathbf{E}\left[\left|\, V(\tau) - \mathbf{E}[V(\tau)]\, \right|^{\kappa}\right] \mathrm{d}B(\tau)}{(g^* x - \mathbf{E}[V(x(1-\varepsilon))])^{\kappa}}.
\end{aligned} \tag{5.6}$$

The first inequality is an immediate consequence of the monotonicity of $\mathbf{E}[V(\tau)]$ in τ, see Assumption 5.1.2. For the second inequality we use Relation (5.4). Note that, indeed, for x large enough it must be that $g^* x - \mathbf{E}[V(x(1-\varepsilon))]$ is positive, since by Assumption 5.1.2:

$$\frac{g^* x}{\mathbf{E}[V(x(1-\varepsilon))]} \longrightarrow \frac{1}{1-\varepsilon} \quad > \quad 1, \quad x \to \infty.$$

Hence, for large x, the denominator of the right-hand side of Relation (5.6) "behaves as" $(g^* x \varepsilon)^{\kappa}$.

Next we study the numerator. First note that, without loss of generality, we may assume that $\kappa - \delta > \zeta$, since if this is not the case, we can choose $\delta > 0$ smaller, and Assumption 5.1.2 will still be satisfied. Let x_0 be as in Relation (5.1), and $\tau_0 \geq x_0$ such that, for all $\tau \geq \tau_0$:

$$\mathbf{E}\left[\left|\, V(\tau) - \mathbf{E}[V(\tau)]\, \right|^{\kappa}\right] \quad \leq \quad \tau^{\kappa-\delta}.$$

Such a τ_0 exists by Assumption 5.1.2. If x is such that $x(1-\varepsilon) > \tau_0$ then Relation (5.6) leads to:

$$\int_{\tau=\tau_0}^{x(1-\varepsilon)} \mathbf{E}\left[\left|\, V(\tau) - \mathbf{E}[V(\tau)]\, \right|^{\kappa}\right] \mathrm{d}B(\tau)$$

$$
\begin{aligned}
&\leq -\int_{\tau=\tau_0}^{x(1-\varepsilon)} \tau^{\kappa-\delta} \mathrm{d}\overline{B}(\tau) \\
&\overset{\text{p.i.}}{=} \tau_0^{\kappa-\delta}\overline{B}(\tau_0) - (x(1-\varepsilon))^{\kappa-\delta}\,\overline{B}(x(1-\varepsilon)) \\
&\qquad +(\kappa-\delta)\int_{\tau=\tau_0}^{x(1-\varepsilon)} \tau^{\kappa-\delta-1}\overline{B}(\tau)\mathrm{d}\tau \\
&\leq \tau_0^{\kappa-\delta}\overline{B}(\tau_0) \\
&\qquad +(\kappa-\delta)\overline{B}(x(1-\varepsilon))\int_{\tau=\tau_0}^{x(1-\varepsilon)} \tau^{\kappa-\delta-1}\left(\frac{\tau}{x(1-\varepsilon)}\right)^{-\zeta}\mathrm{d}\tau \\
&\leq \tau_0^{\kappa-\delta}\overline{B}(\tau_0) + \frac{\kappa-\delta}{\kappa-\delta-\zeta}\overline{B}(x(1-\varepsilon))\,(x(1-\varepsilon))^{\kappa-\delta},
\end{aligned}
$$

where "p.i." indicates the use of partial integration. In the second inequality we used Relation (5.1) and the fact that

$$(x(1-\varepsilon))^{\kappa-\delta}\,\overline{B}(x(1-\varepsilon)) \;\geq\; 0.$$

Since

$$\int_{\tau=0}^{\tau_0} \mathbf{E}\left[\Big|\, V(\tau) - \mathbf{E}\left[V(\tau)\right] \,\Big|^{\kappa}\right] \mathrm{d}B(\tau)$$

is independent of x, and that $\kappa-\delta>\zeta$, the numerator of the right-hand side of Relation (5.6) is bounded from above by a function that tends to infinity as

$$\overline{B}(x(1-\varepsilon))\,(x(1-\varepsilon))^{\kappa-\delta}.$$

Recall that the denominator "behaves as" $(g^*x\varepsilon)^{\kappa}$. Therefore, dividing the right-hand side of Relation (5.6) by $\overline{B}(x(1-\varepsilon))$, and letting $x\to\infty$, proves the lemma. □

The following lemma complements the statements of Lemma 5.1.2 for the case that B is large.

Lemma 5.1.3 *If Assumption 5.1.2 is satisfied then, for all $\varepsilon>0$,*

$$\lim_{x\to\infty}\frac{\mathbf{P}\left\{V(B)>g^*x; B>x(1+\varepsilon)\right\}}{\mathbf{P}\left\{B>x(1+\varepsilon)\right\}} \;=\; 1.$$

Proof Clearly, the lim sup of the above expression can not be larger than 1. Therefore it suffices to show that the lim inf is at least 1. By Assumption 5.1.2 we have, for all $\tau\geq x(1+\varepsilon)$,

$$\mathbf{P}\left\{V(\tau)>g^*x\right\} \;\geq\; \mathbf{P}\left\{V(x(1+\varepsilon))>g^*x\right\}.$$

Hence,

$$\begin{aligned}\mathbf{P}\{V(B) > g^*x; B > x(1+\varepsilon)\} &= \int_{\tau=x(1+\varepsilon)}^{\infty} \mathbf{P}\{V(\tau) > g^*x\}\,\mathrm{d}B(\tau)\\ &\geq \mathbf{P}\{V(x(1+\varepsilon)) > g^*x\}\,\mathbf{P}\{B > x(1+\varepsilon)\}.\end{aligned}$$

From Relation (5.2) it follows that $\mathbf{E}[V(x(1+\varepsilon))] - g^*x > 0$, for x large enough. Hence, by Markov's inequality:

$$\begin{aligned}&\mathbf{P}\{V(x(1+\varepsilon)) \leq g^*x\}\\ &= \mathbf{P}\{\mathbf{E}[V(x(1+\varepsilon))] - V(x(1+\varepsilon)) \geq \mathbf{E}[V(x(1+\varepsilon))] - g^*x\}\\ &\leq \frac{\mathbf{E}\left[\left| V(x(1+\varepsilon)) - \mathbf{E}[V(x(1+\varepsilon))]\ \right|^{\kappa}\right]}{(\mathbf{E}[V(x(1+\varepsilon))] - g^*x)^{\kappa}},\end{aligned}$$

and, by Relations (5.2) and (5.3), this vanishes as $x \to \infty$. Therefore,

$$\lim_{x\to\infty} \mathbf{P}\{V(x(1+\varepsilon)) > g^*x\} = 1,$$

and the proof is completed. □

5.2 The M/G/1 queue for three service disciplines

In the remainder of the chapter we show that, under certain conditions, the on/off processor-sharing model of Chapter 3, with generally distributed service requirements, satisfies Assumptions 5.1.1 and 5.1.2 and thus exhibits the tail behaviour described in Theorem 5.1.1. As indicated in Remark 5.1.1, in the sequel the random variables B, $V(\tau)$ and $V(B)$ are distributed as the service requirement, the sojourn time conditional on the service requirement and the unconditional sojourn time, respectively. Before turning to the on/off model, however, we apply the theory of the previous section to the M/G/1 queue for three different standard service disciplines. First we show that Theorem 5.1.1 provides a new and simpler proof of the result of Zwart and Boxma [128], whose analysis relies on transform techniques. Then we establish the "tail equivalence" for two other service disciplines under the assumption that the second moment of the service requirement distribution is infinite. Besides the independent interest of these results, this section serves to illustrate the use of Theorem 5.1.1 and clarify which steps must be made in order to prove the result for the on/off model.

The remainder of this section is divided into five parts. In Section 5.2.0 we present preliminary results which facilitate the subsequent analysis. We prove the tail equivalence for the three above mentioned service disciplines in Sections 5.2.1 – 5.2.3. In Section 5.2.4 we discuss a common property of these three models and, anticipating on the results in subsequent sections, of the on/off processor-sharing model.

5.2.0 Preliminaries

First we review some common notation. In all three models of this section we consider an M/G/1 queue, the difference between the three cases being the service discipline, that is the way in which service is allocated to the customers in the system. Customers arrive according to a Poisson process with intensity λ and their service requirement distribution is $B(x)$ with mean $\beta_1 < \infty$ and k^{th} moment $\beta_k \leq \infty$, $k = 2, 3, \ldots$. Service is rendered at rate 1 whenever the system is not empty. The traffic load is denoted by $\rho = \lambda\beta_1$ and we assume that the system is stable: $\rho < 1$. When the second moment of the service requirement distribution (β_2) is infinite, we often need to impose the following conditions:

Assumption 5.2.1 $\mathbf{E}\left[B^{\alpha}\right] < \infty$ *for some* $\alpha \in (1, 2)$.

Assumption 5.2.2 $\mathbf{E}\left[B^{\zeta}\right] = \infty$ *for some* $\zeta \in (1, 2)$.

It is straightforward to see that when the second assumption is satisfied and $\overline{B}(x) \in \mathcal{IRV}$, then Relation (5.1) holds. Assumption 5.2.1 implies that the tail of the service requirement distribution is dominated by a Pareto tail. We formalise the latter statement in a more general context in the next lemma.

Lemma 5.2.1 *If* Z *is a non-negative random variable with* $\mathbf{E}\left[Z^{\theta}\right] < \infty$, *for some* $\theta \in \mathbb{R}$, *then*

$$\mathbf{P}\left\{Z > u\right\} \;=\; \mathrm{o}(u^{-\theta}),$$

for $u \to \infty$. *Hence, there exists a number* $u_0 > 0$ *such that* $\mathbf{P}\left\{Z > u\right\} \leq u^{-\theta}$, *for all* $u \geq u_0$.

The converse statement is not true, but if $\mathbf{P}\left\{Z > u\right\}$ *has the above asymptotic property then, for all* $\varepsilon \in (0, \theta)$, $\mathbf{E}\left[Z^{\theta-\varepsilon}\right] < \infty$.

Proof The first statement follows from the fact that

$$\begin{aligned}
&\lim_{u\to\infty} \left(u^{\theta}\mathbf{P}\left\{Z > u\right\} \right) \\
&= \lim_{u\to\infty} \left(\theta \int_{x=0}^{u} x^{\theta-1}\mathbf{P}\left\{Z > x\right\} \mathrm{d}x - \int_{x=0}^{u} x^{\theta} \mathrm{d}\mathbf{P}\left\{Z \leq x\right\} \right) \\
&= \mathbf{E}\left[Z^{\theta}\right] - \mathbf{E}\left[Z^{\theta}\right] \;=\; 0.
\end{aligned}$$

The existence of the number u_0 is trivial and the last statement follows from:

$$\mathbf{E}\left[Z^{\theta-\varepsilon}\right] \;=\; (\theta - \varepsilon) \int_{u=0}^{\infty} u^{\theta-\varepsilon-1}\mathbf{P}\left\{Z > u\right\} \mathrm{d}u \;<\; \infty.$$

$\square$

In the sequel, the random variable $W_{\lambda,B}$ is distributed as the (steady-state) waiting time in the M/G/1 FCFS queue with arrival rate λ and service time distribution $B(x)$, i.e.,

$$\mathbf{P}\{W_{\lambda,B} \leq t\} \quad = \quad (1-\rho)\sum_{n=0}^{\infty}\rho^n\left[\frac{1}{\beta_1}\int_{x=0}^{t}\mathbf{P}\{B > x\}\,\mathrm{d}x\right]^{n\star}, \qquad (5.7)$$

cf. Cohen [20, Part II, Expression (4.82)]. Here, the symbol * denotes the convolution operator for probability distributions, i.e., for a distribution function $H(x)$, $x \geq 0$, we define $H(x)^{0\star} := 1$, for all $x \geq 0$, and for $n \in \mathbb{N}_0$ and $x \geq 0$,

$$H(x)^{(n+1)\star} \quad := \quad \int_{u=0}^{x} H(x-u)^{n\star}\mathrm{d}H(u). \qquad (5.8)$$

In particular, $H(x)^{1\star} = H(x)$, $x \geq 0$.

The next lemma states a direct implication of Assumption 5.2.1 for the distribution of $W_{\lambda,B}$. This relation will be useful in the analysis of sojourn times in the case that $\beta_2 = \infty$.

Lemma 5.2.2 *If* $\mathbf{E}[B^\alpha] < \infty$ *then* $\mathbf{E}\left[(W_{\lambda,B})^{\alpha-1}\right] < \infty$.

Proof See Asmussen [7, Theorem VIII.2.1], or Appendix 5.B for an alternative proof. □

5.2.1 Processor sharing

In the M/G/1 processor-sharing queue, at any point in time all customers in the system share equally in the service capacity. For more on processor-sharing queues we refer to Section 1.6, where an overview of the literature on these models is provided. Here we are interested in the tail of the sojourn time distribution. Zwart and Boxma [128, Theorem 4.1] have shown that the tails of the service requirement distribution and the sojourn time distribution are equally heavy (in the sense of Theorem 5.1.1) when the service requirements are regularly varying. We now show that Theorem 5.1.1 provides a new proof of this fact (for intermediate regularly varying service requirement distributions).

Theorem 5.2.3 *Consider the M/G/1 queue with processor sharing. If* $\overline{B}(x) \in \mathcal{IRV}$*, and one of the following two conditions holds,*

(i) $\beta_2 < \infty$*, or,*

(ii) $\mathbf{E}[B^\alpha] < \infty$ *and* $\mathbf{E}\left[B^\zeta\right] = \infty$*, for some* $1 < \alpha < \zeta < 2$*,*

then

$$\lim_{x\to\infty}\frac{\mathbf{P}\left\{V(B) > \frac{x}{1-\rho}\right\}}{\mathbf{P}\{B > x\}} \quad = \quad 1.$$

Remark 5.2.1 Note that we exclude the case that $\beta_2 = \infty$ and $\mathbf{E}\left[B^{\zeta}\right] < \infty$ for all $\zeta \in (1,2)$. This case can be included by studying the fourth moment of the sojourn time.

Proof We show that Assumption 5.1.2 is satisfied. First we note that the monotonicity of $\mathbf{P}\{V(\tau) > t\}$ in τ, the last condition in Assumption 5.1.2, is easily seen using a sample-path argument: Comparing the sojourn times of two customers, for the same sequences of inter-arrival times and service requirements of other customers, it follows immediately that the one requiring the smaller amount of service leaves before the one with the larger service requirement.

We now focus on Relations (5.2) and (5.3). It is well known that the mean sojourn time is given by:

$$\mathbf{E}[V(\tau)] = \frac{\tau}{1-\rho},$$

see Sakata et al. [99, Formula (10)], Sakata et al. [100, Formula (49)], or Kleinrock [55, Formula (4.17)]. As a consequence, Relation (5.2) holds with $g^* = 1/(1-\rho)$.

To show that Relation (5.3) holds as well, we first consider the case that $\beta_2 < \infty$. We then have the following asymptotic result, with $k = 2, 3, \ldots$,

$$\mathbf{E}\left[V(\tau)^k\right] = \mathbf{E}[V(\tau)]^k + \frac{\beta_2}{2\beta_1}\frac{\rho}{1-\rho}\frac{k(k-1)}{(1-\rho)^k}\tau^{k-1} + \mathrm{o}(\tau^{k-1}), \quad \tau \to \infty,$$

cf. Zwart and Boxma [128, Remark 3.3]. This implies the following strong asymptotic result, for arbitrary $\varepsilon > 0$ and $k = 2, 3, \ldots$,

$$\mathbf{E}\left[(V(\tau) - \mathbf{E}[V(\tau)])^k\right] = \mathrm{o}(\tau^{k-1+\varepsilon}), \quad \tau \to \infty.$$

Thus, if $\overline{B}(x) \in \mathcal{IRV}$ and ζ is as in Relation (5.1), then let κ be an even integer which is larger than ζ. Then Assumption 5.1.2 is satisfied for any $\delta \in (0,1)$ with $g^* = 1/(1-\rho)$, hence, Theorem 5.1.1 can be applied.

In the case that $\mathbf{E}[B^{\alpha}] < \infty$ and $\mathbf{E}\left[B^{\zeta}\right] = \infty$, for some $1 < \alpha < \zeta < 2$, we note that $\overline{B}(x)$ satisfies Relation (5.1). We use Zwart and Boxma [128, Equations (3.5) and (3.10)] to write:

$$\mathbf{Var}[V(\tau)] = \frac{2}{(1-\rho)^2}\int_{u=0}^{\tau}(\tau - u)\mathbf{P}\{W_{\lambda,B} > u\}\,\mathrm{d}u,$$

where $W_{\lambda,B}$ is distributed as in Expression (5.7). Using Lemmas 5.2.1 and 5.2.2, we have $\mathbf{P}\{W_{\lambda,B} > u\} = \mathrm{o}(u^{1-\alpha})$, hence, $\mathbf{Var}[V(\tau)] = \mathrm{o}(u^{3-\alpha+\varepsilon})$ for all $\varepsilon > 0$. Thus, Assumption 5.1.2 is satisfied with $\kappa = 2$ and $0 < \delta < \alpha - 1$. Now apply Theorem 5.1.1. □

5.2.2 Foreground-background processor sharing

With the FBPS discipline, at all times, the service capacity is used to serve the customer(s) which so far have received the least amount of service, see Kleinrock [55] or Yashkov [122]. Note that more than one customer can have the (same) minimum amount of attained service. In that case the service capacity is shared equally among these customers, hence the term processor sharing.

Assuming $B(x)$ is absolutely continuous, the mean and variance of the sojourn time are given by:

$$\mathbf{E}[V(\tau)] = \frac{\tau}{1-\lambda h_1(\tau)} + \frac{\lambda h_2(\tau)}{2(1-\lambda h_1(\tau))^2}, \tag{5.9}$$

$$\mathbf{Var}[V(\tau)] = \frac{\lambda h_3(\tau)}{3(1-\lambda h_1(\tau))^3} + \frac{\lambda \tau h_2(\tau)}{(1-\lambda h_1(\tau))^3} + \frac{3(\lambda h_2(\tau))^2}{4(1-\lambda h_1(\tau))^4}, \tag{5.10}$$

cf. Yashkov [122, Formulas (6.2) and (6.3)]. The functions $h_j(\tau)$, $j = 1, 2, 3$, are given by

$$h_j(\tau) = j \int_{x=0}^{\tau} x^{j-1} \overline{B}(x) \mathrm{d}x. \tag{5.11}$$

Using these expressions we apply Theorem 5.1.1 to the case $\beta_2 = \infty$.

Theorem 5.2.4 *Consider the M/G/1 queue with the FBPS service discipline. If $\overline{B}(x) \in \mathcal{IRV}$, $\mathbf{E}[B^\alpha] < \infty$ and $\mathbf{E}\left[B^\zeta\right] = \infty$, for some $1 < \alpha < \zeta < 2$, then*

$$\lim_{x\to\infty} \frac{\mathbf{P}\left\{V(B) > \frac{x}{1-\rho}\right\}}{\mathbf{P}\{B > x\}} = 1.$$

Proof First we remark that, as in the proof of Theorem 5.2.3, the monotonicity of $\mathbf{P}\{V(\tau) > t\}$ in τ, follows from a sample-path argument. Hence, it remains to be shown that Relations (5.2) and (5.3) hold.

Note that the $h_j(\tau)$ defined in Equation (5.11) are non-decreasing, positive functions, and that

$$\lim_{\tau\to\infty} h_1(\tau) = \beta_1 < \infty.$$

By Lemma 5.2.1 there is a number $x_0 > 0$ such that $\overline{B}(x) \le x^{-\alpha}$, for all $x \ge x_0$. Using this in Equation (5.11) for $j = 2, 3$, we have, for arbitrary $\varepsilon > 0$,

$$h_j(\tau) = \mathrm{o}(\tau^{j-\alpha+\varepsilon}), \quad \tau \to \infty.$$

Hence, by Expressions (5.9) and (5.10),

$$\lim_{\tau\to\infty} \frac{\mathbf{E}[V(\tau)]}{\tau} = \frac{1}{1-\rho},$$

$$\lim_{\tau\to\infty} \frac{\mathbf{Var}[V(\tau)]}{\tau^{3-\alpha+\varepsilon}} = 0.$$

Taking ζ as specified above, Assumption 5.1.2 is satisfied with $\kappa = 2$ and $0 < \delta < \alpha - 1$. $\square$

5.2.3 Shortest remaining processing time first

Now we consider an M/G/1 queue in which the total service capacity is always allocated to the customer with the shortest remaining processing time. Note that the service of a customer is pre-empted when a new customer arrives with a service requirement smaller than the remaining service requirement of the customer being served. The service of the customer that is pre-empted is resumed as soon as there are no other customers with a smaller amount of work in the system. By sample-path arguments, it can be shown that, among all work-conserving service disciplines, the SRPT discipline minimises the number of customers in the system *at any point in time*, see Schrage and Miller [102]. For this model we show the tail equivalence of the service requirement distribution and the sojourn time distribution for the case that $\beta_2 = \infty$.

Remark 5.2.2 Note that if the service requirement distribution has discontinuity points then it may occur (with positive probability) that two customers have the same remaining service requirement, see Schrage and Miller [102]. Here we assume this is not the case, thus, $B(x)$ is a continuous function.

Following Schrage and Miller [102] we decompose the sojourn time into two different periods: The waiting time (the time until the customer is first served) and the residence time (the remainder of the sojourn time). For a customer with service requirement τ, we denote the waiting time by $W(\tau)$ and the residence time by $R(\tau)$. Thus, the sojourn time is given by $V(\tau) = W(\tau) + R(\tau)$. We emphasise that the residence time may contain service pre-emption periods caused by customers with a smaller service requirement. Schrage and Miller [102] obtained the LST of $W(\tau)$ and $R(\tau)$. For our purposes we only need the first two moments of these random variables. First we define $\rho(\tau)$ as the traffic load of customers with an amount of work less than or equal to τ,

$$\rho(\tau) := \lambda \int_{t=0}^{\tau} t \mathrm{d}B(t). \tag{5.12}$$

The first two moments of $W(\tau)$ are given by:

$$\mathbf{E}\left[W(\tau)\right] = \lambda \frac{\int_{t=0}^{\tau} t^2 \mathrm{d}B(t) + \tau^2 \overline{B}(\tau)}{2\left(1-\rho(\tau)\right)^2}, \tag{5.13}$$

$$\begin{aligned}\mathbf{E}\left[W(\tau)^2\right] &= \lambda \frac{\int_{t=0}^{\tau} t^3 \mathrm{d}B(t) + \tau^3 \overline{B}(\tau)}{3\left(1-\rho(\tau)\right)^3} \\ &\quad + \lambda^2 \int_{t=0}^{\tau} t^2 \mathrm{d}B(t) \frac{\int_{t=0}^{\tau} t^2 \mathrm{d}B(t) + \tau^2 \overline{B}(\tau)}{\left(1-\rho(\tau)\right)^4},\end{aligned} \tag{5.14}$$

and the mean and variance of $R(\tau)$ by

$$\mathbf{E}\left[R(\tau)\right] = \int_{t=0}^{\tau} \frac{1}{1-\rho(t)} \mathrm{d}t, \tag{5.15}$$

$$\mathbf{Var}\left[R(\tau)\right] = \lambda \int_{t=0}^{\tau} \frac{\int_{u=0}^{t} u^2 \mathrm{d}B(u)}{\left(1-\rho(t)\right)^3} \mathrm{d}t. \tag{5.16}$$

These expressions enable us to apply Theorem 5.1.1, thus showing the tail equivalence in the case that $\beta_2 = \infty$. This result is stated in the next theorem.

Theorem 5.2.5 *Consider the M/G/1 queue with the SRPT service discipline. If $\overline{B}(x) \in \mathcal{IRV}$, $\mathbf{E}[B^\alpha] < \infty$ and $\mathbf{E}\left[B^\zeta\right] = \infty$, for some $1 < \alpha < \zeta < 2$, then*

$$\lim_{x\to\infty} \frac{\mathbf{P}\left\{V(B) > \frac{x}{1-\rho}\right\}}{\mathbf{P}\{B > x\}} = 1.$$

Proof The proof proceeds along the same lines as those of Theorems 5.2.3 and 5.2.4. The monotonicity of $\mathbf{P}\{V(\tau) > t\}$ in τ follows from a sample-path argument. Furthermore, note that $\rho(\tau)$ defined by Equation (5.12) is a positive, non-decreasing function with $\rho(\tau) \longrightarrow \rho$, as $\tau \to \infty$. Using that the Césaro limit of a function is finite and equal to the ordinary limit when the latter exists, we have:

$$\lim_{\tau\to\infty} \frac{\mathbf{E}[R(\tau)]}{\tau} = \lim_{t\to\infty} \frac{1}{1-\rho(t)} = \frac{1}{1-\rho}.$$

Now consider Expression (5.16) and replace $\mathrm{d}B(u)$ by $-\mathrm{d}\overline{B}(u)$. By Lemma 5.2.1 there is a number $x_0 > 0$ such that $\overline{B}(x) \le x^{-\alpha}$, for all $x \ge x_0$. Using partial integration and the fact that $\rho(t) \le \rho$ for all $t \ge 0$, we have, for arbitrary $\varepsilon > 0$,

$$\begin{aligned}
\mathbf{Var}[R(\tau)] &= -\lambda \int_{t=0}^{\tau} \frac{\int_{u=0}^{t} u^2 \mathrm{d}\overline{B}(u)}{(1-\rho(t))^3} \mathrm{d}t \\
&\le \frac{-\lambda}{(1-\rho)^3} \int_{t=0}^{\tau} \left(t^2 \overline{B}(t) - 2\int_{u=0}^{t} u\overline{B}(u)\mathrm{d}u \right) \mathrm{d}t \\
&= \mathrm{o}(\tau^{3-\alpha+\varepsilon}), \quad \tau \to \infty.
\end{aligned}$$

In the same way, by partial integration we have for $\mathbf{E}[W(\tau)]$, using Formula (5.13),

$$\mathbf{E}[W(\tau)] = \lambda \frac{\int_{t=0}^{\tau} t\overline{B}(t)\mathrm{d}t}{(1-\rho(\tau))^2},$$

and similarly for $\mathbf{E}\left[W(\tau)^2\right]$. With the above bound for $\overline{B}(u)$, the following relations follow for all $\varepsilon > 0$:

$$\lim_{\tau\to\infty} \frac{\mathbf{E}[W(\tau)]}{\tau^{2-\alpha+\varepsilon}} = \lim_{\tau\to\infty} \frac{\mathbf{E}\left[W(\tau)^2\right]}{\tau^{3-\alpha+\varepsilon}} = 0,$$

hence, since $3 - \alpha > 2(2-\alpha)$,

$$\mathbf{Var}[W(\tau)] = \mathrm{o}(\tau^{3-\alpha+\varepsilon}), \quad \tau \to \infty.$$

Using the fact that the random variables $W(\tau)$ and $R(\tau)$ are independent for fixed $\tau > 0$, we have, for all $\varepsilon > 0$,

$$\begin{aligned} \lim_{\tau\to\infty} \frac{\mathbf{E}\left[V(\tau)\right]}{\tau} &= \frac{1}{1-\rho}, \\ \lim_{\tau\to\infty} \frac{\mathbf{Var}\left[V(\tau)\right]}{\tau^{3-\alpha+\varepsilon}} &= 0. \end{aligned}$$

Thus, Assumptions 5.1.1 and 5.1.2 are satisfied (for $\kappa = 2$ and $0 < \delta < \alpha - 1$) and we may apply Theorem 5.1.1. □

5.2.4 Intermediate discussion

We found in all three models of Sections 5.2.1 – 5.2.3 that, when applying Theorem 5.1.1, the factor g^* is equal to $1/(1-\rho)$. In the remainder of this chapter we shall see that this property is essentially shared by the unreliable processor-sharing model, the factor g^* being equal to $1/(c-\rho)$ where $c < 1$ is the *average* available service capacity. We now provide an intuitive interpretation of this finding. Theorems 5.2.3, 5.2.4 and 5.2.5, as well as Theorem 5.3.1 below, state that the probability that a customer's sojourn time exceeds the value $x/(c-\rho)$ is asymptotically (for $x \to \infty$) equal to the probability that a customer's service requirement exceeds a value x. This property can be understood partly by the same arguments used in Chapters 3 and 4 to explain the asymptotic linearity of $\mathbf{E}\left[V(\tau)\right]$, see Remarks 3.6.2 and 4.3.3. The above mentioned models share the property that if a customer with an infinite service requirement is placed in the queue, then the queue remains stable. Hence, after a very long time, the average capacity per unit of time devoted to the service of "non-permanent" customers is approximately ρ (all non-permanent customers eventually leave the system). Since the mean total service capacity rendered by the system per unit of time is c, the average service capacity received by the permanent customer is approximately $c - \rho$. Still, this is not sufficient for Theorem 5.1.1 to apply. For example, it is known for the M/M/1 processor-sharing queue that the result is not true. The reason for this is that with a "light-tailed" service requirement distribution, Lemma 5.1.2 does not hold: A large sojourn time is not necessarily caused by a large service requirement, but may be due to the fact that many other customers are requesting service. For heavy-tailed service requirement distributions, we showed by Lemma 5.1.2 that the probability of this happening is negligible compared to that of a large sojourn time and a large service requirement occurring simultaneously.

5.3 The on/off model with general service requirements

The remainder of this chapter is devoted to the M/G/1 processor-sharing queue with random service interruptions. Our goal is to extend Theorem 5.2.3 to this model for the case that $\beta_2 = \infty$. In order to apply Theorem 5.1.1 we need to study the first and second moment of the conditional sojourn time $V(\tau)$. Different from the presentation in the previous section, for the "on/off" model this turns out to be a complicated task. Our ultimate objective is stated in the next theorem. As before, the random variable $V(B)$ is distributed as the steady-state sojourn time of customers. The proof of the theorem will be provided by Theorem 5.6.2 which states that Assumption 5.1.2 is satisfied. Hence, we may apply Theorem 5.1.1. In the next theorem an additional condition (Assumption 5.5.1) is imposed. It postulates that if the random variable X is distributed as the number of customers in the system in steady state, then $\mathbf{E}[X^\gamma] < \infty$, for some $\gamma > 2$. At this point we do not go into the rationale for this assumption, but refer to Remark 5.5.3 below for a discussion.

Theorem 5.3.1 *Consider an M/G/1 processor-sharing queue with random service interruptions which satisfies Assumption 5.5.1 below. If* $\overline{B}(x) \in \mathcal{IRV}$, $\mathbf{E}[B^\alpha] < \infty$ *and* $\mathbf{E}\left[B^\zeta\right] = \infty$, *for some* $1 < \alpha < \zeta < 2$, *then*

$$\lim_{x\to\infty} \frac{\mathbf{P}\left\{V(B) > \frac{x}{c-\rho}\right\}}{\mathbf{P}\{B > x\}} \quad = \quad 1.$$

Here c is the average service capacity and ρ is the traffic load.

Proof Assumption 5.1.2 is satisfied because of Theorem 5.6.2 and Remark 5.6.1 below. Hence, the result follows from Theorem 5.1.1. □

The model under consideration was studied in Chapter 3 assuming exponentially distributed service requirements. Now the service requirements have a general (heavy-tailed) distribution. We first review the model in this section. Parts of the analysis of Chapter 3 are extended. This provides a basis for the analysis of the tail of the sojourn time distribution, ultimately leading to Theorem 5.6.2 and, hence, to Theorem 5.3.1 above. In the extensions of the results of Chapter 3 to the present model, the key ideas are essentially the same as with exponentially distributed service requirements. In the course of this section we point out where, and how, the analysis needs to be modified to include the more general case.

As before, customers arrive according to a Poisson process with rate λ. The service requirement distribution is $B(x)$, $x \geq 0$, with first and second moments $\beta_1 < \infty$ and $\beta_2 \leq \infty$, respectively. For $\mathrm{Re}(s) \geq 0$ we define the LST of $B(x)$ by:

$$\beta(s) := \mathbf{E}\left[\mathrm{e}^{-sB}\right] = \int_{x=0}^{\infty} \mathrm{e}^{-sx}\mathrm{d}B(x).$$

The service station alternates between "on-periods", which have an exponential distribution with mean $1/\nu$, and "off-periods" with a general distribution $F(t)$, $t \geq 0$, and LST $\phi(s)$, $\mathrm{Re}(s) \geq 0$. The duration of an off-period will again be generically denoted by the random variable T_{off}, its first three moments — m_1, m_2 and m_3, respectively — are assumed to be finite. During an on-period, service is rendered at a constant rate 1, and all customers in the system share equally in this capacity, according to the processor-sharing discipline. During off-periods there is no service. Define the traffic load by $\rho := \lambda\beta_1$ and the average service capacity by $c = 1/(1+\nu m_1)$. We assume that $\rho < c$, hence, the system is stable.

Remark 5.3.1 Our objective in this chapter is to apply Theorem 5.1.1. For this we need to verify that (under some conditions) Assumption 5.1.2 is satisfied. In particular we will need Condition (5.3) with $\kappa = 2$, and hence, we need $\mathbf{E}\left[V(\tau)^2\right] < \infty$. To ensure this, we assumed above that $m_3 < \infty$. In Chapter 3 we saw, for exponentially distributed service requirements, that if $m_2 = \infty$ then $\mathbf{E}\left[V(\tau)\right] = \infty$ for all $\tau > 0$, see Expression (3.41). Here $V(\tau)$ is distributed as the sojourn time (in steady state) of a customer with service requirement τ. More generally it is true that if the k^{th} moment of the off-periods is infinite then $\mathbf{E}\left[V(\tau)^{k-1}\right] = \infty$ for all $\tau > 0$, regardless of the distribution of the service requirements. To see this note that (assuming that the on/off process is in steady state) a customer arrives during an off-period with probability equal to $\nu m_1/(1+\nu m_1) = 1 - c > 0$. The remaining time until service is available again is distributed as the forward recurrence time of T_{off}. Note that during this period no service is rendered, hence the sojourn time of the customer is at least as large as this period, irrespective of the customer's service requirement. It is well known that the $k-1^{\text{st}}$ moment of the forward recurrence time is finite if and only if the k^{th} moment of T_{off} is finite.

As we shall see later (in Assumption 5.5.1 and Remark 5.5.3) we need to impose the condition that

$$\mathbf{E}\left[(T_{off})^{\gamma+1}\right] \;<\; \infty,$$

for some $\gamma > 2$, which is slightly more restrictive than requiring that $m_3 < \infty$.

In Remark 3.2.1 we outlined how the sojourn time, with generally distributed service requirements, can be decomposed into independent contributions of the customers in the system. We state this result in the next theorem. By $V_{n,1}(\tau; x_1, \ldots, x_n)$ we denote the conditional sojourn time of a customer with service requirement τ starting in an on-period with $n \in \{0, 1, 2 \ldots\}$ other customers in the system with service requirements $x_1, x_2, \ldots, x_n$. Similarly we use $V_{n,0}(\tau; x_1, \ldots, x_n)$ to represent the conditional sojourn time starting in an off-period. In this case we further denote the remaining duration of this off-period by D_0, with LST $\phi_0(s)$, $\mathrm{Re}(s) \geq 0$, and the number of arrivals during D_0 by A_0. The joint distribution of the pair (D_0, A_0) is given by:

$$\mathbf{E}\left[\mathrm{e}^{-sD_0} z^{A_0}\right] \;=\; \phi_0(s + \lambda(1-z)), \tag{5.17}$$

with $\mathrm{Re}(s) \geq 0$, and $|z| \leq 1$.

In the sequel we also use random variables $C_n(\tau; B_n)$, for $n \in \{-1, -2, \ldots\}$. These random variables are distributed as $C_1(\tau; B)$. The indices with negative values help to avoid confusion with different terms in the next theorem and in the sequel. The theorem is the analogue of Theorem 3.2.5 and Corollary 3.2.6, where we considered the case of exponentially distributed service requirements.

Theorem 5.3.2 *The conditional sojourn time $V_{n,1}(\tau; x_1, \ldots, x_n)$ can be decomposed as:*

$$V_{n,1}(\tau; x_1, \ldots, x_n) \stackrel{d}{=} C_0(\tau) + \sum_{i=1}^{n} C_i(\tau; x_i).$$

Similarly, for $V_{n,0}(\tau; x_1, \ldots, x_n)$:

$$V_{n,0}(\tau; x_1, \ldots, x_n) \stackrel{d}{=} D_0 + C_0(\tau) + \sum_{i=1}^{n} C_i(\tau; x_i) + \sum_{i=-1}^{-A_0} C_i(\tau; B_i).$$

Here $\stackrel{d}{=}$ denotes equality in distribution. The random variables B_i are distributed according to $B(\cdot)$. All random variables in the right-hand sides are independent except for the pair (D_0, A_0) whose joint distribution is given by Expression (5.17).

Proof Consider the population model described in Remark 3.2.1 with general life time distribution $B(x)$ and associated rewards. Similar to the exponential case in Section 3.2, we can show — by means of a random time change — the equivalence of the rewards in the population model, between times 0 and τ, and the sojourn time of a customer with service requirement τ in the on/off processor-sharing model. $C_i(\tau; x_i)$ is the reward for a family, between times 0 and τ, starting with one individual with a remaining life time x_i. $C_0(\tau)$ is the reward for the family of the permanent individual between times 0 and τ. The independence of the $C_0(\tau)$ and $C_i(\tau; x_i)$, $i = 1, 2, \ldots$, is a direct consequence of the construction of the population model. □

Similar to Chapter 3, we interpret $C_0(\tau)$ as the part of the sojourn time due to the service requirement τ of the customer under consideration itself, and $C_i(\tau; x_i)$ as the delay due to the service requirement x_i of a competing customer. As in the formulation of the previous theorem, the delay due to an "average" competing customer — that is a customer with service requirement B drawn from the distribution $B(x)$ — will be denoted by $C_1(\tau; B)$. For $t \geq 0$ its distribution is given by:

$$\mathbf{P}\left\{C_1(\tau; B) \leq t\right\} = \int_{x=0}^{\infty} \mathbf{P}\left\{C_1(\tau; x) \leq t\right\} \mathrm{d}B(x).$$

For $\mathrm{Re}(s) \geq 0$, we define the LSTs of $C_0(\tau)$ and $C_i(\tau; x)$, $i \in \{1, 2, \ldots\}$, by:

$$\begin{aligned} g_0(\tau; s) &:= \mathbf{E}\left[\mathrm{e}^{-sC_0(\tau)}\right], \\ g_1(\tau; x; s) &:= \mathbf{E}\left[\mathrm{e}^{-sC_i(\tau; x)}\right]. \end{aligned}$$

Obviously,

$$g_1(\tau;B;s) := \mathbf{E}\left[\mathrm{e}^{-sC_i(\tau)}\right] = \int_{x=0}^{\infty} g_1(\tau;x;s)\mathrm{d}B(x).$$

The following lemma identifies a useful relation between the distributions of the random variables $C_0(\tau)$ and $C_1(\tau;x)$. For the ordinary M/G/1 processor-sharing queue the result was obtained by Yashkov [120].

Lemma 5.3.3 *For all $\tau \geq x \geq 0$,*

$$C_0(\tau) \stackrel{d}{=} C_1(\tau;x) + C_0(\tau - x),$$

where the two random variables on the right-hand side are independent of each other. If $x \geq \tau \geq 0$ then $C_1(\tau;x) \stackrel{d}{=} C_0(\tau)$. Hence, for all $\mathrm{Re}(s) \geq 0$ *and $\tau \geq x \geq 0$,*

$$g_1(\tau;x;s) = \frac{g_0(\tau;s)}{g_0(\tau - x;s)},$$

and when $x \geq \tau \geq 0$, $g_1(\tau;x;s) = g_0(\tau;s)$.

Proof A formal technical proof in terms of LSTs can be given in the same way as in Yashkov [120]. Here we give a constructive proof. Consider the following accumulation of rewards in the population model. Start at time 0 with an individual with remaining life time $x \geq 0$. The reward earned until time $\tau \geq x$ by this individual and his descendants is distributed as $C_1(\tau;x)$. At time x the original individual dies. At this time *introduce* a permanent individual. The reward earned until time τ by this individual and his descendants is distributed as $C_0(\tau - x)$ and is independent of the reward for the other family. Using the fact that the "inter-birth" time intervals are exponentially distributed, i.e., they are memoryless, it is clear that the total reward of the two families is equally distributed as that earned by a single permanent individual over a time span of length τ. Thus, the individual with life time x and the permanent one introduced at time x can be replaced by a permanent individual starting at time 0, without affecting the total reward.

Finally, if an individual in the population model at time 0 has remaining life time $x \geq \tau$, then until time τ the individual acts as a permanent one. □

Similar to Lemma 3.3.1 (for exponentially distributed service requirements) we characterise the distribution of $C_0(\tau)$ — and, by the previous lemma, that of $C_1(\tau;x)$ — by means of a differential equation.

Lemma 5.3.4 *For* $\mathrm{Re}(s) \geq 0$ *and $\tau \geq 0$,*

$$\begin{aligned}\frac{\partial}{\partial \tau} g_0(\tau;s) = g_0(\tau;s)\{&-(s+\lambda+\nu) + \lambda g_1(\tau;B;s) \\ &+\nu\phi(s+\lambda(1-g_1(\tau;B;s)))\},\end{aligned}$$

and $g_0(0;s) = 1$, for all $\mathrm{Re}(s) \geq 0$.

Proof Using the same technique as in the proof of Lemma 3.3.1 we find, for $\Delta \downarrow 0$,

$$g_0(\tau+\Delta;s) = \mathrm{e}^{-s\Delta} g_0(\tau;s)\{1-\Delta(\lambda+\nu)+\Delta\lambda g_1(\tau;B;s) \\ +\Delta\nu\phi(s+\lambda(1-g_1(\tau;B;s)))\}+\mathrm{o}(\Delta).$$

After re-arranging terms, dividing by Δ and passing Δ to zero, we find the desired differential equations. The initial conditions are evident from the interpretation of sojourn times in the queueing model as rewards in the population model: $C_0(0) = 0$. □

5.4 Moments of the fundamental random variables

Our analysis of sojourn times builds upon the decompositions of Theorem 5.3.2. Our goal is to show that the variance of $V(\tau)$ satisfies Assumption 5.1.2 (with $\kappa = 2$) when $\beta_2 = \infty$. Therefore, we are particularly interested in the first and second moments of the (conditional) sojourn time. We analyse these via the first and second moments of the "fundamental" random variables $C_0(\tau)$, $C_1(\tau;x)$ and $C_1(\tau;B)$, which appear in the decompositions of Theorem 5.3.2. We also derive results for the case that $\beta_2 < \infty$, which serve as a guide for the analysis of the more complicated case that $\beta_2 = \infty$.

We study the moments of the random variables $C_0(\tau)$, $C_1(\tau;x)$ and $C_1(\tau;B)$, by means of LTs (Laplace Transforms). We define, for $s > 0$, $w > 0$, $x \geq 0$ and $k = 1, 2$,

$$\begin{aligned}
\widehat{f}_k(s) &:= \int_{\tau=0}^{\infty} \mathrm{e}^{-s\tau}\mathbf{E}\left[C_0(\tau)^k\right] \mathrm{d}\tau, \\
\widehat{f}_k(s;x) &:= \int_{\tau=0}^{\infty} \mathrm{e}^{-s\tau}\mathbf{E}\left[C_1(\tau;x)^k\right] \mathrm{d}\tau, \\
\widehat{f}_k(s;B) &:= \int_{\tau=0}^{\infty} \mathrm{e}^{-s\tau}\mathbf{E}\left[C_1(\tau;B)^k\right] \mathrm{d}\tau, \\
\widehat{f}_k(s;w) &:= \int_{\tau=0}^{\infty}\int_{x=0}^{\infty} \mathrm{e}^{-s\tau - wx}\mathbf{E}\left[C_1(\tau;x)^k\right] \mathrm{d}x\ \mathrm{d}\tau.
\end{aligned}$$

These functions are well defined if we allow them to take values in $[0,\infty)\cup\{\infty\}$. Note that the second argument of $\widehat{f}_k$ can have different meanings. This does not lead to confusion if we keep in mind that, as arguments of $\widehat{f}_k$, the variable x always stands for a remaining service requirement (or life time), B stands for a service requirement drawn from the distribution $B(x)$ and w is a transformation variable (in the frequency domain). Using Lemma 5.3.3 it can be verified — by substituting $\mathbf{E}[C_0(\tau)] - \mathbf{E}[C_0(\tau-x)]$ for $\mathbf{E}[C_1(\tau;x)]$ in the above definitions — that, for $s > 0$, $w > 0$ and $x \geq 0$,

$$\widehat{f}_1(s;x) = \left(1-\mathrm{e}^{-sx}\right)\widehat{f}_1(s), \tag{5.18}$$

$$\widehat{f}_1(s;B) = (1-\beta(s))\,\widehat{f}_1(s), \tag{5.19}$$
$$\widehat{f}_1(s;w) = \frac{s}{w(s+w)}\widehat{f}_1(s), \tag{5.20}$$

in the sense that in all three equalities both sides are infinite if one of them is.

As before, we denote the convolution operator for probability distributions by the symbol *, see Relation (5.8). In the next lemma we give a closed-form expression for $\mathbf{E}[C_0(\tau)]$. The proof is based on the derivation and solution of a differential equation that is satisfied by $\mathbf{E}[C_0(\tau)]$, similar to that given in Lemma 5.3.4. Combined with Lemma 5.3.3 we also have, as a corollary, closed-form expressions for $\mathbf{E}[C_1(\tau;x)]$ and $\mathbf{E}[C_1(\tau;B)]$. A prominent role is played by the (steady-state) waiting time distribution in the M/G/1 FCFS queue with arrival rate λ/c and service time distribution $B(x)$. Let the random variable $W_{\lambda/c,B}$ have this distribution:

$$\mathbf{P}\left\{W_{\lambda/c,B} \le t\right\} = \left(1-\frac{\rho}{c}\right)\sum_{n=0}^{\infty}\left(\frac{\rho}{c}\right)^n\left[\frac{1}{\beta_1}\int_{x=0}^{t}\mathbf{P}\{B>x\}\,\mathrm{d}x\right]^{n\star}, \tag{5.21}$$

see also Expression (5.7).

Lemma 5.4.1 *For $s>0$,*

$$\widehat{f}_1(s) = \frac{s^{-2}}{c-\rho\frac{1-\beta(s)}{s\beta_1}}.$$

Hence, for $\tau \ge 0$,

$$\begin{aligned}\mathbf{E}[C_0(\tau)] &= \frac{1}{c-\rho}\int_{t=0}^{\tau}\mathbf{P}\left\{W_{\lambda/c,B}\le t\right\}\mathrm{d}t\\ &= \frac{\tau}{c-\rho}-\frac{1}{c-\rho}\int_{t=0}^{\tau}\mathbf{P}\left\{W_{\lambda/c,B}>t\right\}\mathrm{d}t.\end{aligned}$$

Proof See Appendix 5.C. □

Corollary 5.4.2 *For $\tau\ge 0$ and $0\le x\le\tau$,*

$$\begin{aligned}\mathbf{E}[C_1(\tau;x)] &= \frac{x}{c-\rho}-\frac{1}{c-\rho}\int_{t=\tau-x}^{\tau}\mathbf{P}\left\{W_{\lambda/c,B}>t\right\}\mathrm{d}t,\\ \mathbf{E}[C_1(\tau;B)] &= \frac{1}{c-\rho}\mathbf{E}[\min\{\tau,B\}]\\ &\quad -\frac{1}{c-\rho}\int_{t=0}^{\tau}\mathbf{P}\left\{W_{\lambda/c,B}>t\right\}\mathbf{P}\{B>\tau-t\}\,\mathrm{d}t.\end{aligned}$$

Proof Directly from Lemmas 5.3.3 and 5.4.1. □

Recall that our aim is to provide conditions under which Assumption 5.1.2 is satisfied. The assumption is concerned with the asymptotic behaviour of the moments of $V(\tau)$, for $\tau \to \infty$. With this in mind, we establish a limiting result for $\mathbf{E}[C_0(\tau)]$ when $\tau \to \infty$. Here, a distinction must be made between the cases where β_2 is finite or infinite.

Lemma 5.4.3 *If $\beta_2 < \infty$ then*

$$\lim_{\tau\to\infty}\left(\mathbf{E}[C_0(\tau)] - \frac{\tau}{c-\rho}\right) = -\frac{1}{c-\rho}\mathbf{E}\left[W_{\lambda/c,B}\right] = -\frac{\lambda\beta_2}{2(c-\rho)^2},$$

and the limit equals $-\infty$ if $\beta_2 = \infty$. In the latter case, if $\mathbf{E}[B^\alpha] < \infty$, for some $\alpha \in (1,2)$, then, for all $\varepsilon > 0$,

$$\mathbf{E}[C_0(\tau)] - \frac{\tau}{c-\rho} = \mathrm{o}(\tau^{2-\alpha+\varepsilon}), \quad \tau \to \infty.$$

Proof For $\beta_2 < \infty$ the result is immediate from Lemma 5.4.1 and Expression (5.21) for the distribution of $W_{\lambda/c,B}$. When $\beta_2 = \infty$ and $\mathbf{E}[B^\alpha] < \infty$, for some $\alpha \in (1,2)$, we have by Lemmas 5.2.1 and 5.2.2 that, for t large enough, $\mathbf{P}\left\{W_{\lambda/c,B} > t\right\}$ is bounded from above by $t^{1-\alpha}$. Using this in the expression for $\mathbf{E}[C_0(\tau)]$ given in Lemma 5.4.1 we have, for arbitrary $\varepsilon > 0$,

$$\lim_{\tau\to\infty}\frac{\mathbf{E}[C_0(\tau)] - \frac{\tau}{c-\rho}}{\tau^{2-\alpha+\varepsilon}} = 0,$$

which was to be proved. □

The next lemma is concerned with the asymptotic behaviour of $\mathbf{E}[C_1(\tau;x)]$ and $\mathbf{E}[C_1(\tau;B)]$. It states that these expectations have a finite limit as $\tau \to \infty$, irrespective of β_2 being finite or infinite. However, later in our analysis of sojourn times we need more refined asymptotic results for $\mathbf{E}[C_1(\tau;x)]$ and $\mathbf{E}[C_1(\tau;B)]$. To be more precise, we need to know the rate at which these quantities converge to their respective limits. For this second-order analysis we again need to make a distinction between the cases $\beta_2 < \infty$ and $\beta_2 = \infty$. The results are combined in the following lemma.

Lemma 5.4.4 *If $\beta_2 < \infty$ then*

$$\mathbf{E}[C_1(\tau;x)] - \frac{x}{c-\rho} = \mathrm{o}(\tau^{-1}), \quad \tau \to \infty,$$
$$\mathbf{E}[C_1(\tau;B)] - \frac{\beta_1}{c-\rho} = \mathrm{o}(\tau^{-1}), \quad \tau \to \infty.$$

If $\mathbf{E}[B^\alpha] < \infty$, for some $\alpha \in (1,2)$, then

$$\mathbf{E}[C_1(\tau;x)] - \frac{x}{c-\rho} = \mathrm{o}(\tau^{1-\alpha}), \quad \tau \to \infty,$$
$$\mathbf{E}[C_1(\tau;B)] - \frac{\beta_1}{c-\rho} = \mathrm{o}(\tau^{1-\alpha}), \quad \tau \to \infty.$$

Proof We give the proof for the case $\beta_2 = \infty$, which is the more difficult one. When $\beta_2 < \infty$ we follow essentially the same steps, replacing $\alpha - 1$ by 1. So we assume $\beta_2 = \infty$ and $\mathbf{E}[B^\alpha] < \infty$. To prove the result for $\mathbf{E}[C_1(\tau;x)]$ we use the expression given in Corollary 5.4.2 together with the inequality

$$\int_{t=\tau-x}^{\tau} \mathbf{P}\left\{W_{\lambda/c,B} > t\right\} \mathrm{d}t \quad \leq \quad x\mathbf{P}\left\{W_{\lambda/c,B} > \tau - x\right\}.$$

By Lemmas 5.2.1 and 5.2.2 we have

$$\lim_{t\to\infty} t^{\alpha-1}\mathbf{P}\left\{W_{\lambda/c,B} > t\right\} \quad = \quad 0,$$

which leads to the required result for $\mathbf{E}[C_1(\tau;x)]$.

The derivation of the result for $\mathbf{E}[C_1(\tau;B)]$ is somewhat more delicate. Using the expression for $\mathbf{E}[C_1(\tau;B)]$ given in Corollary 5.4.2, we write:

$$\begin{aligned}
&\lim_{\tau\to\infty} \tau^{\alpha-1} \int_{t=0}^{\tau} \mathbf{P}\left\{W_{\lambda/c,B} > t\right\} \mathbf{P}\left\{B > \tau - t\right\} \mathrm{d}t \\
&= \lim_{\tau\to\infty} \tau^{\alpha-1} \int_{t=0}^{\infty} \mathbf{1}_{\{t\leq\tau\}} \mathbf{P}\left\{W_{\lambda/c,B} > \tau - t\right\} \mathbf{P}\left\{B > t\right\} \mathrm{d}t \\
&= \int_{t=0}^{\infty} \lim_{\tau\to\infty} \mathbf{1}_{\{t\leq\tau\}} \tau^{\alpha-1} \mathbf{P}\left\{W_{\lambda/c,B} > \tau - t\right\} \mathbf{P}\left\{B > t\right\} \mathrm{d}t \\
&= 0,
\end{aligned}$$

where $\mathbf{1}_{\{\cdot\}}$ is the indicator function. The interchange of limit and integral is justified by the Dominated Convergence Theorem, because

$$\begin{aligned}
&\sup_{\tau\geq t} \left\{\mathbf{1}_{\{t\leq\tau\}} \tau^{\alpha-1} \mathbf{P}\left\{W_{\lambda/c,B} > \tau - t\right\}\right\} \\
&\leq \sup_{\tau\geq t} \left\{\left(\frac{\tau}{1+\tau-t}\right)^{\alpha-1}\right\} \sup_{\tau\geq t} \left\{(1+\tau-t)^{\alpha-1} \mathbf{P}\left\{W_{\lambda/c,B} > \tau - t\right\}\right\} \\
&= (\max\{1,t\})^{\alpha-1} \sup_{\tau\geq t} \left\{(1+\tau-t)^{\alpha-1} \mathbf{P}\left\{W_{\lambda/c,B} > \tau - t\right\}\right\}.
\end{aligned}$$

Now use that $(1+\tau-t)^{\alpha-1}\mathbf{P}\left\{W_{\lambda/c,B} > \tau - t\right\}$ is bounded from above by a constant (we use $1+\tau-t$ instead of $\tau - t$ to avoid difficulties with $\tau - t = 0$) and that

$$\int_{t=0}^{\infty} (\max\{1,t\})^{\alpha-1} \mathbf{P}\left\{B > t\right\} \mathrm{d}t \; < \; \infty,$$

which is true by assumption. □

This concludes the analysis of the expectations of the fundamental random variables. We now turn to their second moments. A key result is stated in the next lemma where $\mathbf{E}\left[C_0(\tau)^2\right]$ is expressed in terms of the function $R_0(\tau)$ defined by:

$$\begin{aligned} R_0(\tau) \quad := \quad & \nu m_2 \Big(1 + \lambda \mathbf{E}\left[C_1(\tau;B)\right] \Big)^2 \\ & +2(1+\nu m_1)\mathbf{E}\left[C_0(\tau)\right] \Big(1 + \lambda \mathbf{E}\left[C_1(\tau;B)\right] \Big) \\ & -2\lambda\left(1+\nu m_1\right) \int_{x=0}^{\tau} \mathbf{E}\left[C_0(\tau - x)\right] \mathbf{E}\left[C_1(\tau;x)\right] \mathrm{d}B(x). \quad (5.22) \end{aligned}$$

Note that $R_0(\tau)$ is completely determined by expectations of the fundamental random variables, which we studied previously.

Lemma 5.4.5 *For* $\tau \geq 0$,

$$\mathbf{E}\left[C_0(\tau)^2\right] \quad = \quad \frac{c}{c-\rho} \int_{u=0}^{\tau} R_0(\tau - u) \mathbf{P}\left\{W_{\lambda/c,B} \leq u\right\} \mathrm{d}u. \qquad (5.23)$$

Proof As in the proof of Lemma 5.4.1, let $\overline{C_0}(\tau)$ be the reward in the time interval $[0, \tau]$ of a family of permanent individuals, starting with one individual. We can show (see the proof of Lemma 5.4.1 for more details) that

$$\int_{\tau=0}^{\infty} \mathrm{e}^{-s\tau} \mathbf{E}\left[\overline{C_0}(\tau)^2\right] \mathrm{d}\tau$$

is finite for $s > 2\lambda/c$ and, hence, so is $\widehat{f_2}(s)$. For $\mathbf{E}\left[C_0(\tau)^2\right]$ the following differential equation may be derived:

$$\begin{aligned} \frac{\partial}{\partial \tau} \mathbf{E}\left[C_0(\tau)^2\right] \quad = \quad & \lambda\left(1+\nu m_1\right) \mathbf{E}\left[C_1(\tau;B)^2\right] + \nu m_2 \Big(1 + \lambda \mathbf{E}\left[C_1(\tau;B)\right] \Big)^2 \\ & +2(1+\nu m_1)\mathbf{E}\left[C_0(\tau)\right] \Big(1 + \lambda \mathbf{E}\left[C_1(\tau;B)\right] \Big). \end{aligned}$$

By Lemma 5.3.3 we may write, for $0 \leq x \leq \tau$,

$$\begin{aligned} \mathbf{E}\left[C_1(\tau;x)^2\right] \quad = \quad & \mathbf{E}\left[C_0(\tau)^2\right] - \mathbf{E}\left[C_0(\tau - x)^2\right] \\ & -2\mathbf{E}\left[C_1(\tau;x)\right] \mathbf{E}\left[C_0(\tau - x)\right], \qquad (5.24) \end{aligned}$$

and, $\mathbf{E}\left[C_1(\tau;x)^2\right] = \mathbf{E}\left[C_0(\tau)^2\right]$ when $x \geq \tau \geq 0$. Hence,

$$\begin{aligned} \mathbf{E}\left[C_1(\tau;B)^2\right] \quad = \quad & \mathbf{E}\left[C_0(\tau)^2\right] - \int_{x=0}^{\tau} \mathbf{E}\left[C_0(\tau - x)^2\right] \mathrm{d}B(x) \\ & -2 \int_{x=0}^{\tau} \mathbf{E}\left[C_0(\tau - x)\right] \mathbf{E}\left[C_1(\tau;x)\right] \mathrm{d}B(x). \end{aligned}$$

The differential equation for $\mathbf{E}\left[C_0(\tau)^2\right]$ reduces to

$$\begin{aligned} \frac{\partial}{\partial \tau} \mathbf{E}\left[C_0(\tau)^2\right] \quad = \quad & \lambda\left(1+\nu m_1\right) \left(\mathbf{E}\left[C_0(\tau)^2\right] - \int_{x=0}^{\tau} \mathbf{E}\left[C_0(\tau - x)^2\right] \mathrm{d}B(x)\right) \\ & +R_0(\tau), \end{aligned}$$

where $R_0(\tau)$ is given by Expression (5.22). Taking LTs, we have, for $s > 2\lambda/c$,

$$s\widehat{f}_2(s) \quad = \quad \lambda\,(1+\nu m_1)\,\widehat{f}_2(s)\,(1-\beta(s))+\widehat{r}_0(s),$$

where $\widehat{r}_0(s)$ is the LT of $R_0(\tau)$. Hence,

$$\widehat{f}_2(s) \quad = \quad \frac{c\widehat{r}_0(s)/s}{c-\rho\frac{1-\beta(s)}{s\beta_1}}.$$

By inverting this transform (as in the proof of Lemma 5.4.1) the proof is completed. □

As with the expectations of the fundamental random variables, we need to study the second moments for $\tau \to \infty$. The next lemma provides the basis for that.

Lemma 5.4.6 *If $\beta_2 < \infty$ then*

$$\lim_{\tau\to\infty}\left(R_0(\tau)-\frac{2\tau}{c(c-\rho)}\right) \quad = \quad \nu m_2\left(\frac{c}{c-\rho}\right)^2+\frac{\lambda\beta_2}{c(c-\rho)^2}.$$

When $\beta_2 = \infty$ and $\mathbf{E}\left[B^\alpha\right] < \infty$, for some $\alpha \in (1,2)$, then, for any $\varepsilon > 0$,

$$R_0(\tau)-\frac{2\tau}{c(c-\rho)} \quad = \quad \mathrm{o}(\tau^{2-\alpha+\varepsilon}), \quad \tau\to\infty.$$

Proof See Appendix 5.D. □

For the asymptotic analysis of the second moments of the fundamental random variables we consider the cases $\beta_2 < \infty$ and $\beta_2 = \infty$ separately. In the case $\beta_2 < \infty$ we are able to get much sharper results. The condition $\beta_2 < \infty$ is inherited from Lemma 5.4.6, but surprisingly enough, in the final result for $\mathbf{E}\left[C_0(\tau)^2\right]$, β_2 does not show up.

Lemma 5.4.7 *If $\beta_2 < \infty$ then*

$$\mathbf{E}\left[C_0(\tau)^2\right] \quad = \quad \left(\frac{\tau}{c-\rho}\right)^2+\tau\nu m_2\left(\frac{c}{c-\rho}\right)^3+h(\tau), \tag{5.25}$$

where the function $h(\tau)$ is such that, for all $x \geq 0$,

$$\lim_{\tau\to\infty}\Big(h(\tau)-h(\tau-x)\Big) \quad = \quad 0.$$

In particular,

$$\lim_{\tau\to\infty}\frac{\mathbf{E}\left[C_0(\tau)^2\right]-\left(\frac{\tau}{c-\rho}\right)^2}{\tau} \quad = \quad \nu m_2\left(\frac{c}{c-\rho}\right)^3.$$

Proof See Appendix 5.E. □

Corollary 5.4.8 *If* $\beta_2 < \infty$ *then, for* $x \geq 0$,

$$\begin{aligned}\lim_{\tau\to\infty} \mathbf{E}\left[C_1(\tau;x)^2\right] &= \frac{x^2}{(c-\rho)^2} + \frac{x\left(\nu m_2 c^3 + \lambda\beta_2\right)}{(c-\rho)^3},\\ \lim_{\tau\to\infty} \mathbf{E}\left[C_1(\tau;B)^2\right] &= \frac{c\beta_2 + c^3\beta_1\nu m_2}{(c-\rho)^3}.\end{aligned}$$

Proof Using Expression (5.24) and Lemmas 5.4.3 and 5.4.4 we have that

$$\begin{aligned}&\lim_{\tau\to\infty}\left(\mathbf{E}\left[C_1(\tau;x)\right]\mathbf{E}\left[C_0(\tau-x)\right] - \frac{x\tau}{(c-\rho)^2}\right)\\ &= \lim_{\tau\to\infty}\left(\left(\mathbf{E}\left[C_1(\tau;x)\right] - \frac{x}{c-\rho}\right)\mathbf{E}\left[C_0(\tau-x)\right]\right.\\ &\qquad\left. + \frac{x}{c-\rho}\left(\mathbf{E}\left[C_0(\tau-x)\right] - \frac{\tau-x+x}{c-\rho}\right)\right)\\ &= -\frac{x}{c-\rho}\left(\frac{x}{c-\rho} + \frac{\lambda\beta_2}{2(c-\rho)^2}\right). \end{aligned} \tag{5.26}$$

Now use Relations (5.25) and (5.26) in Expression (5.24) and the result for $\mathbf{E}\left[C_1(\tau;x)^2\right]$ is proved. Then write,

$$\lim_{\tau\to\infty} \mathbf{E}\left[C_1(\tau;B)^2\right] = \lim_{\tau\to\infty}\int_{x=0}^{\infty} \mathbf{E}\left[C_1(\tau;x)^2\right] \mathrm{d}B(x).$$

Since $\mathbf{E}\left[C_1(\tau;x)^2\right]$ is non-decreasing in τ we may interchange the limit and integral. □

The previous lemma and corollary characterised the asymptotic behaviour of $\mathbf{E}\left[C_0(\tau)^2\right]$, $\mathbf{E}\left[C_1(\tau;x)^2\right]$ and $\mathbf{E}\left[C_1(\tau;B)^2\right]$ for large τ, in the case that $\beta_2 < \infty$. We now turn to the case $\beta_2 = \infty$. In the result of Lemma 5.4.7, β_2 does not show up. However, the condition $\beta_2 < \infty$ is crucial for the proof of Lemma 5.4.7 to apply. Nevertheless, for the case that $\beta_2 = \infty$ the somewhat weaker result stated in the next lemma suffices for our purposes. The asymptotic properties of $\mathbf{E}\left[C_0(\tau)^2\right]$ reveal the asymptotic behaviour of $\mathbf{E}\left[C_1(\tau;x)^2\right]$ and $\mathbf{E}\left[C_1(\tau;B)^2\right]$.

Lemma 5.4.9 *If* $\mathbf{E}\left[B^\alpha\right] < \infty$, *for some* $\alpha \in (1,2)$, *then, for all* $\varepsilon > 0$,

$$\mathbf{E}\left[C_0(\tau)^2\right] - \left(\frac{\tau}{c-\rho}\right)^2 = \mathrm{o}(\tau^{3-\alpha+\varepsilon}), \quad \tau\to\infty.$$

Proof We treat the two terms in the right-hand side of Expression (5.40) in the proof of Lemma 5.4.7 (see Appendix 5.E) separately. For the first we have

$$
\begin{aligned}
&\lim_{\tau\to\infty} \frac{1}{\tau(1+\tau)^{2-\alpha+\varepsilon}} \left| \frac{c}{c-\rho}\int_{u=0}^{\tau} R_0(\tau-u)\mathrm{d}u - \left(\frac{\tau}{c-\rho}\right)^2 \right| \\
&\leq \lim_{\tau\to\infty} \frac{1}{\tau}\int_{u=0}^{\tau} \frac{c}{c-\rho}(1+u)^{\alpha-2-\varepsilon} \left| R_0(u) - \frac{2u}{c(c-\rho)} \right| \mathrm{d}u \\
&= \lim_{u\to\infty} \frac{c}{c-\rho}(1+u)^{\alpha-2-\varepsilon} \left| R_0(u) - \frac{2u}{c(c-\rho)} \right| \;=\; 0.
\end{aligned}
$$

Here we used Lemma 5.4.6 and the fact that if the limit of a function is finite, then so is the Césaro limit and the two limits are equal. Note that we used a factor $(\tau+1)^{2-\alpha+\varepsilon}$ instead of $\tau^{2-\alpha+\varepsilon}$ to avoid problems with $u=0$ inside the integral. For the second part we use the fact that

$$
\left| \frac{R_0(\tau-u)}{1+\tau-u} \right| \;\leq\; K,
$$

for some constant $K>0$ (this follows from Lemma 5.4.6). Hence,

$$
\begin{aligned}
&\lim_{\tau\to\infty} \frac{1}{(1+\tau)^{3-\alpha+\varepsilon}} \left| \int_{u=0}^{\tau} R_0(\tau-u)\mathbf{P}\left\{W_{\lambda/c,B}>u\right\}\mathrm{d}u \right| \\
&\leq \lim_{\tau\to\infty} K(1+\tau)^{-\varepsilon}\int_{u=0}^{\tau} (1+u)^{\alpha-2}\mathbf{P}\left\{W_{\lambda/c,B}>u\right\}\mathrm{d}u \\
&= 0,
\end{aligned}
$$

where we used that, by Lemmas 5.2.1 and 5.2.2,

$$
\mathbf{P}\left\{W_{\lambda/c,B}>u\right\} \;<\; u^{1-\alpha},
$$

for u large enough. □

Corollary 5.4.10 *If* $\mathbf{E}\left[B^{\alpha}\right]<\infty$, *for some* $\alpha\in(1,2)$, *then, for all* $\varepsilon>0$ *there exists a constant* $K_\varepsilon>0$ *such that, for all* $0\leq x\leq\tau$,

$$
\mathbf{E}\left[C_1(\tau;x)^2\right] \;\leq\; \frac{2x\tau}{(c-\rho)^2} + K_\varepsilon(1+\tau)^{3-\alpha+\varepsilon}.
$$

Hence, for all $\varepsilon>0$,

$$
\mathbf{E}\left[C_1(\tau;B)^2\right] \;=\; \mathrm{o}(\tau^{3-\alpha+\varepsilon}), \quad \tau\to\infty.
$$

Proof By Lemma 5.4.9 there exists a constant $K'_\varepsilon>0$ such that, for all $\tau\geq 0$,

$$
\left| \mathbf{E}\left[C_0(\tau)^2\right] - \frac{\tau^2}{(c-\rho)^2} \right| \;\leq\; K'_\varepsilon(1+\tau)^{3-\alpha+\varepsilon}.
$$

Using this result and Expression (5.24) we now write:

$$\begin{aligned}\mathbf{E}\left[C_1(\tau;x)^2\right] &\le \mathbf{E}\left[C_0(\tau)^2\right]-\mathbf{E}\left[C_0(\tau-x)^2\right]\\ &= \frac{\tau^2}{(c-\rho)^2}-\frac{(\tau-x)^2}{(c-\rho)^2}+\left(\mathbf{E}\left[C_0(\tau)^2\right]-\frac{\tau^2}{(c-\rho)^2}\right)\\ &\quad -\left(\mathbf{E}\left[C_0(\tau-x)^2\right]-\frac{(\tau-x)^2}{(c-\rho)^2}\right)\\ &\le \frac{2\tau x}{(c-\rho)^2}-\frac{x^2}{(c-\rho)^2}+K'_\varepsilon(1+\tau)^{3-\alpha+\varepsilon}\\ &\quad +K'_\varepsilon(1+\tau-x)^{3-\alpha+\varepsilon}\\ &\le \frac{2\tau x}{(c-\rho)^2}+2K'_\varepsilon(1+\tau)^{3-\alpha+\varepsilon}.\end{aligned}$$

Set $K_\varepsilon := \frac{1}{2}K'_\varepsilon$ and the result for $\mathbf{E}\left[C_1(\tau;x)^2\right]$ is proved. Using the definition of $\mathbf{E}\left[C_1(\tau;B)^2\right]$ we write, for arbitrary $\varepsilon>0$,

$$\begin{aligned}\mathbf{E}\left[C_1(\tau;B)^2\right] &= \int_{x=0}^{\tau}\mathbf{E}\left[C_1(\tau;x)^2\right]\mathrm{d}B(x)+\mathbf{E}\left[C_0(\tau)^2\right]\overline{B}(\tau)\\ &\le \frac{2\tau}{(c-\rho)^2}\int_{x=0}^{\tau}x\mathrm{d}B(x)+K_{\frac{1}{2}\varepsilon}(1+\tau)^{3-\alpha+\frac{1}{2}\varepsilon}B(\tau)\\ &\quad +\mathbf{E}\left[C_0(\tau)^2\right]\overline{B}(\tau)\\ &\le \frac{2\beta_1\tau}{(c-\rho)^2}+K_{\frac{1}{2}\varepsilon}(1+\tau)^{3-\alpha+\frac{1}{2}\varepsilon}+\mathbf{E}\left[C_0(\tau)^2\right]\overline{B}(\tau)\\ &= \mathrm{o}(\tau^{3-\alpha+\varepsilon}).\end{aligned}$$

In the last step we use Lemmas 5.2.1 and 5.4.9 and the fact that $3-\alpha>1$. □

5.5 Work load and queue length in steady state

The analysis of sojourn times in steady state will be based on the decompositions of the sojourn time given in Theorem 5.3.2. The fundamental random variables $C_0(\tau)$, $C_1(\tau;x)$ and $C_1(\tau;B)$ were studied in detail in the previous section. The yet unknown element in the above mentioned decomposition is the state of the system upon arrival, that is the number of customers in the system, their individual remaining service requirements and whether or not service is available. In this section we give a (partial) characterisation of their steady-state distributions.

As in the case with exponentially distributed service requirements, we denote the number of customers at time $t\ge 0$ by $X(t)$ and the state of the server by $Y(t)$, i.e., $Y(t)$ equals 1 if the server is available and 0 otherwise. If $X(t)>0$ then $W_n(t)$ is the remaining service requirement of the n^{th} customer, $n=1,2,\dots,X(t)$. If $Y(t)=0$ then $D_0(t)$ is the remaining duration of the unavailability period and, similarly, $E_0(t)$ equals the length of the elapsed part of the unavailability period. When $Y(t)=1$ then by definition $D_0(t)=E_0(t)=0$. We assume that the joint probability distribution and the moments of the above random variables converge as $t\to\infty$, and denote their steady-state equivalents by X, Y, W_n, D_0 and E_0, respectively, i.e., as $t\to\infty$,

$$\begin{aligned}\mathbf{P}\Big\{X(t)=n;W_1(t)\le x_1,\ldots,W_n(t)\le x_n;Y(t)=i;D_0(t)\le u_1;E_0(t)\le u_2\Big\}\\ \longrightarrow\ \mathbf{P}\Big\{X=n;W_1\le x_1,\ldots,W_n\le x_n;Y=i;D_0\le u_1;E_0\le u_2\Big\},\end{aligned}$$

where $n\in\{1,2,\ldots\}$, $x_1,\ldots,x_n\ge 0$, $i\in\{0,1\}$, and $u_1,u_2\ge 0$. Similarly,

$$\begin{aligned}\mathbf{P}\Big\{X(t)=0;Y(t)=i;D_0(t)\le u_1;E_0(t)\le u_2\Big\}\\ \longrightarrow\ \mathbf{P}\Big\{X=0;Y=i;D_0\le u_1;E_0\le u_2\Big\}.\end{aligned}$$

Remark 5.5.1 In the ordinary M/G/1 processor-sharing queue (with no service interruptions), it is known that

$$\mathbf{P}\Big\{X=n;W_1\le x_1,\ldots,W_n\le x_n\Big\}\ =\ (1-\rho)\rho^n\prod_{k=1}^{n}\int_{u=0}^{x_k}\frac{\overline{B}(u)}{\beta_1}\mathrm{d}u,$$

see for instance Cohen [19, Theorem 3.1] or Yashkov [120, Theorem 3]. Thus, the number of customers in the system has a geometric distribution (which only depends on the service requirement distribution through its mean β_1) and each of the customers in the system has a residual service requirement which is distributed as the forward recurrence time of the services, independent of the number of other customers in the system and of their individual service requirements. When service is subject to interruptions the situation changes completely, and even the mean number of customers in the system is hard to obtain. This is even true for the model in this chapter with hyper-exponentially distributed service requirements and exponentially distributed off-periods. We will not go into details here. For exponentially distributed service requirements the steady-state distributions of interest were obtained in Section 3.1.

As before, let $V(\tau)$ be the sojourn time of a customer with service requirement τ, arriving to the system in steady state. From Theorem 5.3.2 we know that, for $\tau\ge 0$,

$$V(\tau)\ \stackrel{d}{=}\ C_0(\tau)+\sum_{n=1}^{X}C_n(\tau;W_n)+D_0+\sum_{n=-1}^{-A_0}C_n(\tau;B_n), \tag{5.27}$$

with the random variables in the right-hand side being distributed as above. Recall that A_0 is the number of arrivals during D_0 (when Y=0). Their joint distribution is given in Equation (5.17) with $\phi_0(s)$ being the LST of the forward recurrence time of the off-periods:

$$\phi_0(s)\ =\ \frac{1-\phi(s)}{m_1 s}.$$

When $Y = 1$ and (by definition) $D_0 = 0$ we also set A_0 equal to 0. Given that $Y = 0$, the marginal distribution of E_0 is the same as that of D_0, namely that of the backward (and forward) recurrence time of the off-periods with LST $\phi_0(s)$ defined above. We emphasise that (if we condition on $Y = 0$) the random variables D_0 and E_0 — and, hence, X and D_0 — are *not* independent. We come back to this later in this section.

The analysis of sojourn times is complicated by the fact that the (joint) distribution of the number of customers and their service requirements, as needed in Relation (5.27), is unknown. However, using that the total amount of work in the system is the same for all work-conserving service disciplines, we do know the joint distribution of Y and the total amount of work in the system, which we denote by:

$$W := \sum_{n=1}^{X} W_n, \tag{5.28}$$

(the empty sum being equal to 0 by definition). A work-conserving service discipline is one under which the server works at rate 1 whenever the system is not empty *and* the server is available. We know that $\mathbf{P}\{Y = 1\} = c$, and (see Remark 5.5.2 below), for $\mathrm{Re}(s) \geq 0$,

$$\mathbf{E}\left[\mathrm{e}^{-sW} \,|Y = 1\right] = \frac{1 - \rho/c}{1 - (\lambda + \nu)\frac{1 - \widehat{\beta}(s)}{s}}, \tag{5.29}$$

$$\mathbf{E}\left[\mathrm{e}^{-sW} \,|Y = 0\right] = \frac{1 - \phi(\lambda(1 - \beta(s)))}{\lambda m_1 (1 - \beta(s))} \mathbf{E}\left[\mathrm{e}^{-sW} \,|Y = 1\right], \tag{5.30}$$

where, for $\mathrm{Re}(s) \geq 0$,

$$\widehat{\beta}(s) = \frac{\lambda}{\lambda + \nu}\beta(s) + \frac{\nu}{\lambda + \nu}\phi(\lambda(1 - \beta(s))). \tag{5.31}$$

In particular, we find from Expressions (5.29) and (5.30) that, if $\beta_2 < \infty$,

$$\mathbf{E}[W|Y = 1] = \frac{\rho}{c - \rho}\left(\frac{\beta_2}{2\beta_1} + (1 - c)\rho\frac{m_2}{2m_1}\right),$$

$$\mathbf{E}[W|Y = 0] = \mathbf{E}[W|Y = 1] + \rho\frac{m_2}{2m_1},$$

hence, using that $\mathbf{P}\{Y = 1\} = c$,

$$\mathbf{E}[W] = \frac{\rho}{c - \rho}\left(\frac{\beta_2}{2\beta_1} + c(1 - c)\frac{m_2}{2m_1}\right). \tag{5.32}$$

If $\beta_2 = \infty$ then the above expressions for the (conditional) expectation of W are also equal to $+\infty$.

Remark 5.5.2 Expressions (5.29) and (5.30) may be found using the same approach by which we derived the steady-state distribution of the number of customers for the case of exponentially distributed service requirements, see Section 3.1. The key idea is to consider the system only during availability periods. All customers that arrive during one particular off-period in the original model are seen as one "large" customer whose service requirement is distributed as the total amount of work that arrives during one off-period. The LST of this distribution is given by $\phi(\lambda(1-\beta(s)))$. In the new model the arrival rate is $\lambda+\nu$, the LST of the service requirement distribution of an arbitrary customer is $\widehat{\beta}(s)$ and the mean service requirement is

$$\frac{\lambda}{\lambda+\nu}\beta_1 + \frac{\nu}{\lambda+\nu}m_1\lambda\beta_1,$$

hence, the traffic load is $\lambda\beta_1 + \nu m_1\lambda\beta_1 = \rho/c$. Formula (5.29) now follows from the Pollaczek-Khintchine formula for the new model, see also Expression (5.7). Since the on-periods have an exponential distribution, the state of the system when an off-period starts is distributed the same as the state at any arbitrary availability epoch. Formula (5.30) is then found by multiplying the LST of the amount of work at the time that the off-period started with the LST of the amount of work that arrives during the backward recurrence time of the off-periods (E_0). For a discussion on the amount of work at arbitrary time instants, see for instance Gaver [37, Section 8] or Fuhrmann and Cooper [32, Proposition 4]. A transient analysis by means of LTs is provided by Li et al. [66].

Using Expressions (5.29) and (5.30) we are able to prove the next lemma, which is an analogue of Lemma 5.2.2.

Lemma 5.5.1 *If* $\mathbf{E}[B^\alpha] < \infty$, *for some* $\alpha \in (1,2)$, *then* $\mathbf{E}\left[W^{\alpha-1}\right] < \infty$. *Hence,*

$$\mathbf{P}\{W > x\} = \mathrm{o}(x^{1-\alpha}),$$

for $x \to \infty$.

Proof See Appendix 5.F. □

The steady-state distribution of the amount of work in the system is determined by Expressions (5.29) and (5.30). Still, nothing is said about the distribution of X, the number of customers in the system. In the analysis of sojourn times in steady state presented in the next section we will need to make the following assumption. In Remark 5.5.3 below we give arguments to motivate this assumption.

Assumption 5.5.1 *There exists a* $\gamma > 2$ *for which* $\mathbf{E}[X^\gamma] < \infty$.

This assumption has the following implication for the joint distribution of the pair (X, W) which will prove to be useful in the next section.

Lemma 5.5.2 *If* $\mathbf{E}[B^{\alpha}] < \infty$ *and* $\mathbf{E}[X^{\gamma}] < \infty$, *for some* $\alpha \in (1,2)$ *and* $\gamma > 2$, *then, for* $\delta < (\alpha - 1)\left(1 - \frac{2}{\gamma}\right)$,

$$\mathbf{E}\left[X^2 \mathbf{1}_{\{W>u\}}\right] \;=\; \mathrm{o}(u^{-\delta}),$$

for $u \to \infty$.

Proof Note that from Jensen's inequality we have, since $\gamma > 2$,

$$\mathbf{E}\left[X^2 \,|W > u\right] \;\leq\; \left(\mathbf{E}\left[X^{\gamma} \,|W > u\right]\right)^{\frac{2}{\gamma}}.$$

After multiplying both sides of the inequality by $\mathbf{P}\{W > u\}$ we have by Lemma 5.5.1 that, for arbitrary $\varepsilon > 0$,

$$\begin{aligned}
\mathbf{E}\left[X^2 \mathbf{1}_{\{W>u\}}\right] &\leq \left(\mathbf{E}\left[X^{\gamma} \,|W > u\right]\right)^{\frac{2}{\gamma}} \mathbf{P}\{W > u\} \\
&= \left(\mathbf{E}\left[X^{\gamma} \mathbf{1}_{\{W>u\}}\right]\right)^{\frac{2}{\gamma}} \mathbf{P}\{W > u\}^{1-\frac{2}{\gamma}} \\
&\leq \left(\mathbf{E}\left[X^{\gamma}\right]\right)^{\frac{2}{\gamma}} \mathbf{P}\{W > u\}^{1-\frac{2}{\gamma}} \\
&= \mathrm{o}(u^{(1-\alpha+\varepsilon)\left(1-\frac{2}{\gamma}\right)}), \quad u \to \infty,
\end{aligned}$$

because $\mathbf{E}[X^{\gamma}] < \infty$. □

Above we mentioned that if we condition on $Y = 0$ then the random variables X and D_0 are *not* independent. The reason for this is that, during an off-period, the number of customers in the system (X) depends on E_0, the elapsed part of the current off-period. It is well known that in general D_0 and E_0 are dependent, see Expression (5.35) below. By similar arguments as used in the derivation of Expression (5.30) above, we find, for $|z| \leq 1$,

$$\mathbf{E}\left[z^X \,|Y = 0, E_0 = y\right] \;=\; \mathbf{E}\left[z^X \,|Y = 1\right] \mathrm{e}^{-\lambda y(1-z)}, \tag{5.33}$$

and, hence,

$$\mathbf{E}\left[z^X \,|Y = 0\right] \;=\; \mathbf{E}\left[z^X \,|Y = 1\right] \phi_0(\lambda(1-z)). \tag{5.34}$$

The latter formula is an analogue of Expression (5.30). For $x \geq 0$ and $y \geq 0$, the joint distribution of D_0 and E_0 is given by:

$$\mathbf{P}\{D_0 \geq x, E_0 \geq y \,|Y = 0\} \;=\; \frac{1}{m_1} \int_{u=x+y}^{\infty} (1 - F(u))\,\mathrm{d}u, \tag{5.35}$$

cf. Cohen [20, Expression (I.6.23)]. Recall that $F(u)$ is the distribution function of the off-periods. From the above or, more easily, from the second moment of the distribution of $D_0 + E_0$, we have

$$\mathbf{E}\left[D_0 E_0 \,|Y = 0\right] \;=\; \frac{m_3}{6m_1} \;<\; \infty,$$

hence, using Expression (5.33) above,

$$\mathbf{E}[D_0 X \mid Y = 0] \;=\; \frac{m_2}{2m_1}\mathbf{E}[X \mid Y = 1] + \lambda\frac{m_3}{6m_1} < \infty. \tag{5.36}$$

Here we also used that D_0 is independent of the number of customers present at the beginning of the off-period, and that $\mathbf{E}[X] < \infty$ and, hence, $\mathbf{E}[X \mid Y = 1] < \infty$. We need Relation (5.36) in the next section.

Remark 5.5.3 In Remark 5.3.1 we motivated the assumption that

$$m_3 \;=\; \mathbf{E}\left[(T_{off})^3\right] \;<\; \infty.$$

Now we use similar arguments to motivate Assumption 5.5.1. If $\mathbf{E}\left[X^2\right] = \infty$ then, from the decomposition in Relation (5.27), $\mathbf{E}\left[V(\tau)^2\right] = \infty$ for all $\tau > 0$. Since our interest is in the analysis of the first and second moment of $V(\tau)$, it is reasonable to assume that these moments are finite, and hence that $\mathbf{E}\left[X^2\right] < \infty$. Assumption 5.5.1 is only slightly more restrictive. We note that the assumption can not be satisfied unless

$$\mathbf{E}\left[(T_{off})^{\gamma+1}\right] \;<\; \infty. \tag{5.37}$$

This follows from Expression (5.34). Conversely, if Relation (5.37) holds, then Assumption 5.5.1 is satisfied. We give a sketch of the proof of this fact for the case that γ is integer. First note (using sample-path arguments) that the number of customers in the system with one permanent customer is stochastically larger than the number of customers in the system without permanent customers. The finiteness of "integer" moments of the queue-length distribution of the system with one permanent customer can be shown by considering the time-changed process. For the latter we can derive a differential equation for the (moments of the) distribution of the population size, as we did for the cumulative rewards in Lemma 5.3.4. Then let the time parameter (τ) tend to infinity and use that the $k-1^{\text{st}}$ moment of the steady-state queue-length distribution in the system with one permanent customer is finite if and only if the k^{th} moment of the limiting population-size distribution is finite.

5.6 Sojourn times in steady state

In this section we show by means of the auxiliary Theorem 5.6.2 below, that if Assumptions 5.2.1 and 5.5.1 are satisfied then the sojourn times of customers in the on/off processor-sharing model satisfy Assumption 5.1.2 and, hence, Theorem 5.3.1 applies. We start, however, with the following theorem for the case that $\beta_2 < \infty$. In that case, the analysis is quite straightforward and, as before, it serves as a guide for the analysis of the case that $\beta_2 = \infty$, which is our ultimate goal. The theorem states that the conditional mean sojourn time $\mathbf{E}[V(\tau)]$ converges to a linear function for $\tau \to \infty$, just as in the case with exponentially distributed service requirements (see Chapter 3). Remarkably enough, β_2 does not appear in the final result, still the assumption is needed in the proof.

Theorem 5.6.1 *If $\beta_2 < \infty$ then the steady-state conditional mean sojourn time satisfies*

$$\lim_{\tau\to\infty}\left(\mathbf{E}\left[V(\tau)\right] - \frac{\tau}{c-\rho}\right) = (1-c)\frac{m_2}{2m_1}\left(\frac{c}{c-\rho}\right)^2.$$

Proof We use Relation (5.27) term by term. From Lemma 5.4.3 we have

$$\lim_{\tau\to\infty}\left(\mathbf{E}\left[C_0(\tau)\right] - \frac{\tau}{c-\rho}\right) = -\frac{\lambda\beta_2}{2(c-\rho)^2}.$$

Obviously,

$$\mathbf{E}\left[D_0\,|Y=0\right] = \frac{m_2}{2m_1},$$

and using Wald's identity (see Feller [30, Relation (XII.2.7)]), we have

$$\mathbf{E}\left[\sum_{n=-1}^{-A_0} C_1(\tau;B_n)\,|Y=0\right] = \mathbf{E}\left[A_0\right]\mathbf{E}\left[C_n(\tau;B)\right] = \lambda\frac{m_2}{2m_1}\mathbf{E}\left[C_1(\tau;B)\right].$$

By Lemma 5.4.4 we then have:

$$\lim_{\tau\to\infty}\mathbf{E}\left[\sum_{n=-1}^{-A_0} C_n(\tau;B_n)\,|Y=0\right] = \frac{\rho m_2}{2m_1\,(c-\rho)}.$$

For the remaining term we can not use Wald's identity, since it is not clear whether X is a stopping time. However, after conditioning on $X = k, W_1 = x_1, \ldots, W_k = x_k$, we have

$$\begin{aligned}
&\mathbf{E}\left[\sum_{n=1}^{X} C_n(\tau;W_n)\,\Big|X=k;W_1=x_1,\ldots,W_k=x_k\right] \\
&= \sum_{n=1}^{k}\mathbf{E}\left[C_n(\tau;x_n)\right] \longrightarrow \sum_{n=1}^{k}\frac{x_n}{c-\rho},
\end{aligned}$$

as $\tau\to\infty$. For fixed x, $\mathbf{E}\left[C_1(\tau;x)\right]$ is obviously non-decreasing in τ (this is also apparent from Corollary 5.4.2). Then using the Monotone Convergence Theorem for the interchange of limit and integral we have:

$$\lim_{\tau\to\infty}\mathbf{E}\left[\sum_{n=1}^{X} C_n(\tau;W_n)\right] = \mathbf{E}\left[\frac{1}{c-\rho}\sum_{n=1}^{X} W_n\right] = \frac{1}{c-\rho}\mathbf{E}\left[W\right].$$

Now use Expression (5.32) for $\mathbf{E}\left[W\right]$. □

In the proof of the theorem the two terms containing β_2 cancel out. Therefore, when $\beta_2 = \infty$ we can not use the same arguments. However, under the Assumptions 5.2.1 and 5.5.1 we are able to prove weaker results which are sufficient for $V(\tau)$ to satisfy Assumption 5.1.2. The condition in Assumption 5.2.1 is inherited from the analysis of the moments of the fundamental random variables in Section 5.4 for the case that $\beta_2 = \infty$. We need Assumption 5.5.1 in order to apply Lemma 5.5.2. Before proving the next theorem, we explain (in Remark 5.6.1 below) how it leads to the result aimed for in Theorem 5.3.1.

Theorem 5.6.2 *If* $\mathbf{E}[B^\alpha] < \infty$ *and* $\mathbf{E}[X^\gamma] < \infty$, *for some* $\alpha \in (1,2)$ *and* $\gamma > 2$, *then:*

$$(i) \quad \lim_{\tau\to\infty} \frac{\mathbf{E}[V(\tau)] - \frac{\tau}{c-\rho}}{\tau^{1-\delta}} = 0,$$

$$(ii) \quad \lim_{\tau\to\infty} \frac{\mathbf{E}\left[V(\tau)^2\right] - \left(\frac{\tau}{c-\rho}\right)^2}{\tau^{2-\delta}} = 0,$$

for all $\delta < (\alpha-1)\left(1-\frac{2}{\gamma}\right)$.

Remark 5.6.1 The previous theorem ensures that Relations (5.2) and (5.3) in Assumption 5.1.2 are satisfied. To prove Theorem 5.3.1, it remains to be shown that the last condition in Assumption 5.1.2 is satisfied as well, that is the monotonicity of $\mathbf{P}\{V(\tau) > t\}$ in τ. As in the proofs of Theorems 5.2.3, 5.2.4 and 5.2.5 this can be done using a sample-path argument. However, here the monotonicity of $\mathbf{P}\{V(\tau) > t\}$ follows directly from the decomposition of the sojourn times given in Relation (5.27), and the interpretation of the components as rewards in the time-changed population model (see the proof of Theorem 5.3.2).

Proof of Theorem 5.6.2

Part (i). As in the proof of Theorem 5.6.1 we use Relation (5.27), together with

$$\begin{aligned}
\mathbf{E}[D_0 \,|Y=0] &= \frac{m_2}{2m_1},\\
\mathbf{E}\left[\sum_{n=-1}^{-A_0} C_n(\tau;B_n)\,|Y=0\right] &= \lambda\frac{m_2}{2m_1}\mathbf{E}[C_1(\tau;B)],
\end{aligned}$$

and Lemma 5.4.4. The above terms vanish after dividing by $\tau^{1-\delta}$ and letting $\tau\to\infty$, because $\delta < \alpha-1 < 1$. For the term $\mathbf{E}[C_0(\tau)]$ we use Lemma 5.4.3 to conclude:

$$\lim_{\tau\to\infty} \frac{\mathbf{E}[C_0(\tau)] - \frac{\tau}{c-\rho}}{\tau^{1-\delta}} = 0,$$

because $1-\delta > 2-\alpha$.

For the remaining term we condition on $X = k$ and $W_n = x_n$, $n = 1, 2, \ldots, k$, and write for $\tau \geq 0$,

$$\begin{aligned}
&\mathbf{E}\left[\sum_{n=1}^{X} C_n(\tau; W_n) \,\Big|\, X = k, W_1 = x_1, \ldots, W_k = x_k\right] \\
&= \sum_{n=1}^{k} \mathbf{E}\left[C_n(\tau; x_n)\right] \;\leq\; \frac{1}{c-\rho} \sum_{n=1}^{k} \min\{\tau, x_n\},
\end{aligned}$$

where we used Lemma 5.4.1 and Corollary 5.4.2 for the inequalities:

$$\begin{aligned}
\mathbf{E}\left[C_1(\tau; x)\right] &= \mathbf{E}\left[C_0(\tau)\right] \;\leq\; \frac{\tau}{c-\rho}, \quad x \geq \tau \geq 0, \\
\mathbf{E}\left[C_1(\tau; x)\right] &\leq \frac{x}{c-\rho}, \quad 0 \leq x \leq \tau.
\end{aligned}$$

From the above we may conclude that

$$\mathbf{E}\left[\sum_{n=1}^{X} C_n(\tau; W_n)\right] \;\leq\; \frac{1}{c-\rho} \mathbf{E}\left[\sum_{n=1}^{X} \min\{\tau, W_n\}\right].$$

We now treat the cases $W = W_1 + \ldots + W_X \leq \tau$ and $W > \tau$ separately. For the first we write

$$\mathbf{E}\left[\sum_{n=1}^{X} \min\{\tau, W_n\} \,|W \leq \tau\right] \;\leq\; \mathbf{E}\left[W \,|W \leq \tau\right],$$

and use that, because of Lemma 5.5.1, for all $\varepsilon > 0$ there exists a k_ε such that, for all $u \geq 0$,

$$\mathbf{P}\{W > u\} \;\leq\; k_\varepsilon (1+u)^{1-\alpha+\varepsilon}.$$

Now take $\varepsilon \in (0, \alpha - 1 - \delta)$ and write:

$$\begin{aligned}
&\limsup_{\tau\to\infty} \tau^{-1+\delta} \mathbf{E}\left[W \,|W \leq \tau\right] \mathbf{P}\{W \leq \tau\} \\
&= \limsup_{\tau\to\infty} \tau^{-1+\delta} \left(\int_{u=0}^{\tau} \mathbf{P}\{W > u\}\,\mathrm{d}u - \tau \mathbf{P}\{W > \tau\}\right) \\
&\leq \limsup_{\tau\to\infty} \tau^{-1+\delta} \int_{u=0}^{\tau} k_\varepsilon (1+u)^{1-\alpha+\varepsilon}\mathrm{d}u \\
&= 0.
\end{aligned}$$

For the second case we write

$$\mathbf{E}\left[\sum_{n=1}^{X} \min\{\tau, W_n\} \,|W > \tau\right] \;\leq\; \tau \mathbf{E}\left[X \,|W > \tau\right],$$

and use that, by Lemma 5.5.2,

$$\begin{aligned}
&\limsup_{\tau\to\infty} \tau^{\delta}\mathbf{E}\left[X \,|W > \tau\right]\mathbf{P}\left\{W > \tau\right\} \\
&\leq \limsup_{\tau\to\infty} \tau^{\delta}\mathbf{E}\left[X^2 \,|W > \tau\right]\mathbf{P}\left\{W > \tau\right\} \;=\; 0,
\end{aligned}$$

which completes the proof of Part (i).

Part (ii). Using Relation (5.27) for the second moment of $V(\tau)$, we may write:

$$\begin{aligned}
\mathbf{E}\left[V(\tau)^2\right] \;=\; & \mathbf{E}\left[C_0(\tau)^2\right] + \mathbf{E}\left[\left(\sum_{n=1}^{X} C_n(\tau;W_n)\right)^2\right] \\
& +2\mathbf{E}\left[C_0(\tau)\right]\mathbf{E}\left[\sum_{n=1}^{X} C_n(\tau;W_n) + D_0 + \sum_{n=-1}^{-A_0} C_n(\tau;B_n)\right] \\
& +2\mathbf{E}\left[\left(\sum_{n=1}^{X} C_n(\tau;W_n)\right)\left(D_0 + \sum_{n=-1}^{-A_0} C_n(\tau;B_n)\right)\right] \\
& +\mathbf{E}\left[\left(D_0 + \sum_{n=-1}^{-A_0} C_n(\tau;B_n)\right)^2\right]. \qquad (5.38)
\end{aligned}$$

We study each of the terms on the right-hand side separately. For the first term, $\mathbf{E}\left[C_0(\tau)^2\right]$, we use Lemma 5.4.9:

$$\lim_{\tau\to\infty} \tau^{\delta-2}\left(\mathbf{E}\left[C_0(\tau)^2\right] - \left(\frac{\tau}{c-\rho}\right)^2\right) \;=\; 0,$$

because $\delta - 2 < \alpha - 3$. We now study the second term,

$$\mathbf{E}\left[\left(\sum_{n=1}^{X} C_n(\tau;W_n)\right)^2\right],$$

for two cases, namely, when $W = W_1 + \ldots + W_X \leq \tau$ and when $W > \tau$. First condition on $X = k$ and $W_n = x_n$, $n = 1,\ldots,k$, with $x_1 + \ldots + x_k \leq \tau$. Recall that, after conditioning, the random variables $C_n(\tau;x_n)$ are independent of each other. Choose an arbitrary $\varepsilon > 0$ and let $K_{\frac{1}{2}\varepsilon}$ be as in Corollary 5.4.10. Then, using Corollaries 5.4.2 and 5.4.10,

$$\begin{aligned}
&\mathbf{E}\left[\left(\sum_{n=1}^{X} C_n(\tau;W_n)\right)^2 \,\middle|\, X = k; W_1 = x_1,\ldots,W_k = x_k\right] \\
&= \mathbf{E}\left[\left(\sum_{n=1}^{k} C_n(\tau;x_n)\right)^2\right]
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{n=1}^{k} \mathbf{E}\left[C_n(\tau;x_n)^2\right] + 2\sum_{n_1=1}^{k-1}\sum_{n_2=n_1+1}^{k} \mathbf{E}\left[C_{n_1}(\tau;x_{n_1})\right]\mathbf{E}\left[C_{n_2}(\tau;x_{n_2})\right] \\
&\leq \sum_{n=1}^{k} \frac{2\tau x_n}{(c-\rho)^2} + 2\sum_{n_1=1}^{k-1}\sum_{n_2=n_1+1}^{k} \frac{x_{n_1}}{c-\rho}\times\frac{x_{n_2}}{c-\rho} + kK_{\frac{1}{2}\varepsilon}(1+\tau)^{3-\alpha+\frac{1}{2}\varepsilon} \\
&\leq \frac{2\tau}{(c-\rho)^2}\sum_{n=1}^{k} x_n + \frac{1}{(c-\rho)^2}\left(\sum_{n=1}^{k} x_n\right)^2 + kK_{\frac{1}{2}\varepsilon}(1+\tau)^{3-\alpha+\frac{1}{2}\varepsilon} \\
&= \frac{2\tau}{(c-\rho)^2}\mathbf{E}\left[\sum_{n=1}^{X} W_n \;\middle|\; X=k; W_1=x_1,\ldots,W_k=x_k\right] \\
&\quad + \frac{1}{(c-\rho)^2}\mathbf{E}\left[\left(\sum_{n=1}^{X} W_n\right)^2 \;\middle|\; X=k; W_1=x_1,\ldots,W_k=x_k\right] \\
&\quad + K_{\frac{1}{2}\varepsilon}(1+\tau)^{3-\alpha+\frac{1}{2}\varepsilon}\mathbf{E}\left[X \;\middle|\; X=k; W_1=x_1,\ldots,W_k=x_k\right].
\end{aligned}
$$

Hence, summing over all k and integrating over all x_n, $n=1,\ldots,k$ for which $x_1+\ldots+x_k\leq\tau$, we have

$$
\begin{aligned}
&\mathbf{E}\left[\left(\sum_{n=1}^{X} C_n(\tau;W_n)\right)^2 \;\middle|\; W\leq\tau\right] \\
&\leq \frac{2\tau}{(c-\rho)^2}\mathbf{E}\left[W \;\middle|\; W\leq\tau\right] + \frac{1}{(c-\rho)^2}\mathbf{E}\left[W^2 \;\middle|\; W\leq\tau\right] \\
&\quad + K_{\frac{1}{2}\varepsilon}(1+\tau)^{3-\alpha+\frac{1}{2}\varepsilon}\mathbf{E}\left[X \;\middle|\; W\leq\tau\right].
\end{aligned}
$$

Note that

$$
\mathbf{E}\left[X \;\middle|\; W\leq\tau\right]\mathbf{P}\left\{W\leq\tau\right\} \;\leq\; \mathbf{E}\left[X\right] \;<\; \infty.
$$

Now use that, for all $\varepsilon>0$, $\mathbf{P}\left\{W>\tau\right\}=\mathrm{o}(\tau^{1-\alpha+\varepsilon})$, when $\tau\to\infty$, to conclude (using partial integration) that, for any $\varepsilon>0$,

$$
\begin{aligned}
\mathbf{E}\left[W \;\middle|\; W\leq\tau\right]\mathbf{P}\left\{W\leq\tau\right\} &= \int_{x=0}^{\tau} x\mathrm{d}\mathbf{P}\left\{W\leq x\right\} \\
&= \mathrm{o}(\tau^{2-\alpha+\varepsilon}), \\
\mathbf{E}\left[W^2 \;\middle|\; W\leq\tau\right]\mathbf{P}\left\{W\leq\tau\right\} &= \int_{x=0}^{\tau} x^2\mathrm{d}\mathbf{P}\left\{W\leq x\right\} \\
&= \mathrm{o}(\tau^{3-\alpha+\varepsilon}),
\end{aligned}
$$

for $\tau\to\infty$. Therefore,

$$
\lim_{\tau\to\infty}\tau^{\delta-2}\mathbf{E}\left[\left(\sum_{n=1}^{X} C_n(\tau;W_n)\right)^2 \;\middle|\; W\leq\tau\right]\mathbf{P}\left\{W\leq\tau\right\} \;=\; 0.
$$

This settles the case $W \leq \tau$ for the term under consideration. Next condition on $X = k$ and $W_n = x_n$, $n = 1, \ldots, k$, with $x_1 + \ldots + x_k > \tau$. For this case

$$\begin{aligned}
&\mathbf{E}\left[\left(\sum_{n=1}^{X} C_n(\tau; W_n)\right)^2 \,\middle|\, X = k; W_1 = x_1, \ldots, W_k = x_k\right] \\
&= \sum_{n=1}^{k} \mathbf{E}\left[C_n(\tau; x_n)^2\right] + 2\sum_{n_1=1}^{k-1} \sum_{n_2=n_1+1}^{k} \mathbf{E}\left[C_{n_1}(\tau; x_{n_1})\right] \mathbf{E}\left[C_{n_2}(\tau; x_{n_2})\right] \\
&\leq \sum_{n=1}^{k} \mathbf{E}\left[C_n(\tau; \tau)^2\right] + 2\sum_{n_1=1}^{k-1} \sum_{n_2=n_1+1}^{k} \mathbf{E}\left[C_{n_1}(\tau; \tau)\right] \mathbf{E}\left[C_{n_2}(\tau; \tau)\right] \\
&= k\mathbf{E}\left[C_0(\tau)^2\right] + k(k-1)\left(\mathbf{E}\left[C_0(\tau)\right]\right)^2.
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbf{E}\left[\left(\sum_{n=1}^{X} C_n(\tau; W_n)\right)^2 \,\middle|\, W > \tau\right] \quad &\leq \quad \mathbf{E}\left[C_0(\tau)^2\right] \mathbf{E}\left[X \,\middle|\, W > \tau\right] \\
&\quad + \left(\mathbf{E}\left[C_0(\tau)\right]\right)^2 \mathbf{E}\left[X(X-1) \,\middle|\, W > \tau\right].
\end{aligned}$$

Now use that, from Lemmas 5.4.3 and 5.4.9,

$$\lim_{\tau\to\infty} \frac{\mathbf{E}\left[C_0(\tau)^2\right]}{\tau^2} \quad = \quad \lim_{\tau\to\infty} \frac{\left(\mathbf{E}\left[C_0(\tau)\right]\right)^2}{\tau^2} \quad = \quad \frac{1}{(c-\rho)^2},$$

and that, by Lemma 5.5.2,

$$\mathbf{E}\left[X \,\middle|\, W > \tau\right] \mathbf{P}\left\{W > \tau\right\} \;\leq\; \mathbf{E}\left[X^2 \,\middle|\, W > \tau\right] \mathbf{P}\left\{W > \tau\right\} \;=\; \mathrm{o}(\tau^{-\delta}),$$

as $\tau \to \infty$. Hence,

$$\lim_{\tau\to\infty} \tau^{\delta-2}\mathbf{E}\left[\left(\sum_{n=1}^{X} C_n(\tau; W_n)\right)^2 \,\middle|\, W > \tau\right] \mathbf{P}\left\{W > \tau\right\} \quad = \quad 0.$$

For the third term in Expression (5.38) we use that, by Lemma 5.4.3,

$$\lim_{\tau\to\infty} \tau^{-1}\mathbf{E}\left[C_0(\tau)\right] \quad = \quad \frac{1}{c-\rho},$$

and that, as shown in the proof of Part (i),

$$\lim_{\tau\to\infty} \tau^{\delta-1}\mathbf{E}\left[\sum_{n=1}^{X} C_n(\tau; W_n) + D_0 + \sum_{n=-1}^{-A_0} C_n(\tau; B_n)\right] \quad = \quad 0.$$

For the fourth term in Expression (5.38), condition on $Y = 0$. Although the two factors inside the expectation are dependent, both can be handled similarly as before:

$$\begin{aligned}
&\mathbf{E}\left[\left(\sum_{n=1}^{X} C_n(\tau; W_n)\right)\left(D_0 + \sum_{n=-1}^{-A_0} C_n(\tau; B_n)\right) |Y = 0\right] \\
\le\ &\mathbf{E}\left[\frac{1}{c-\rho} X \left(D_0 + \frac{\beta_1}{c-\rho} A_0\right) |Y = 0\right] \\
=\ &\mathbf{E}\left[\frac{1}{c-\rho} X \left(D_0 + \frac{\rho}{c-\rho} D_0\right) |Y = 0\right] \\
<\ &\infty,
\end{aligned}$$

where the finiteness follows from Relation (5.36). Finally, the last term in Expression (5.38) can be written out, and then use Lemma 5.4.4 and Corollary 5.4.10, and the fact that $\mathbf{E}\left[(D_0)^2\right] < \infty$, because $m_3 < \infty$. □

5.7 Concluding remarks

We presented a new approach for the analysis of the tail of the sojourn time distribution when the service requirement distribution has a heavy tail. The approach relies on the analysis of the moments of the distribution of the sojourn time *conditional* on the service requirement. We provided a new proof for the result of Zwart and Boxma [128], which states that, in the M/G/1 processor-sharing queue, the sojourn time distribution is exactly as heavy as the service requirement distribution, when the latter is regularly varying. Our method allows the extension of this result to distributions with an intermediate regularly varying tail. We also established the above tail equivalence in the M/G/1 queue with (i) the FBPS service discipline, (ii) the SRPT discipline, and (iii) processor sharing and random service interruptions. Different from the ordinary M/G/1 processor-sharing queue, we assumed an infinite second moment of the service requirement distribution in the three latter models.

In all four models, Theorem 5.1.1 holds with the same factor $g^* = 1/(c-\rho)$, where c is the average service capacity and ρ is the traffic load. Thus, the probability of a customer's sojourn time exceeding a value $x/(c-\rho)$ is asymptotically, for large x, equal to the probability that a customer's service requirement exceeds the value x. This property can be explained as follows. We showed in Lemmas 5.1.2 and 5.1.3 that, with a heavy-tailed service requirement distribution, a large service requirement leads to a large sojourn time, and, conversely, a large sojourn time must be caused by a large service requirement. The above mentioned models share the property that if a permanent customer (that is a customer with an infinite service requirement) is placed in the queue, then the queue is still stable. Hence, after a very long period, say t time units with $t \to \infty$, the average capacity per unit of time devoted to the service of non-permanent customers is approximately equal to the average traffic load ρ. This is true because the system is stable and, hence, all non-permanent customers eventually leave the system (see also Remarks 3.6.2 and 4.3.3). The average total service capacity rendered by the system (per unit of time) is approximately c. Thus, the average service capacity devoted to the permanent customer is approximately $c-\rho$. If the amount of service received by the permanent customer at time t is denoted by $S(t)$, we have that

$$\frac{t}{S(t)} \approx \frac{1}{c-\rho},$$

hence, the factor g^* above.

The above reasoning for the ratio $t/S(t)$ also holds when the service requirement distribution is not heavy tailed. In the latter case, however, Lemma 5.1.2 — which we need in the proof of Theorem 5.1.1 — does not hold, since a large sojourn time is not necessarily caused by a large service requirement. A large sojourn time may then also be caused by the fact that many other customers are requesting service. With heavy-tailed service requirements, the probability of this happening is negligible compared to the probability that a large sojourn time and a large service requirement occur simultaneously.

A large part of this chapter was devoted to the analysis of the M/G/1 processor-sharing queue with random service interruptions. Extension of the tail equivalence result from the ordinary M/G/1 processor-sharing queue to this on/off model is of interest for the performance evaluation of multi-service telecommunication networks. The result indicates that the processor-sharing service discipline preserves this desirable property, even when the service capacity varies over time.

Appendix

5.A Proof of Relation (5.1)

First we repeat the relation in the next lemma.

Lemma *Let $\overline{B}(x) \in \mathcal{IRV}$. Then there exist numbers $\zeta \in (0,\infty)$, $x_0 \in (0,\infty)$, and $\eta \in (0,1)$ such that, for all $x_2 \geq x_1 \geq x_0$,*

$$\frac{\overline{B}(x_2)}{\overline{B}(x_1)} \geq \eta \left(\frac{x_2}{x_1}\right)^{-\zeta}.$$

Proof Let $\varepsilon > 0$. Because $\overline{B}(x) \in \mathcal{IRV}$, there exists a $K = K(\varepsilon) \in (0,1)$ and an $x_0 = x_0(\varepsilon, K)$ such that, for all $x \geq x_0$,

$$\frac{\overline{B}(x(1+\varepsilon))}{\overline{B}(x)} \geq K.$$

Let x_1 and x_2 be such that $x_2 \geq x_1 \geq x_0$, and let

$$n := \left\lceil \frac{\ln(x_2) - \ln(x_1)}{\ln(1+\varepsilon)} \right\rceil,$$

where $\lceil y \rceil$ is the smallest integer which is larger than or equal to $y \in \mathbb{R}$. Obviously, $n > 0$ and $x_2 \leq x_1(1+\varepsilon)^n$. We may write:

$$\begin{aligned}\overline{B}(x_1) \;&\leq\; K^{-1}\overline{B}(x_1(1+\varepsilon)) \;\leq\; \ldots \\ &\leq\; K^{-n}\overline{B}(x_1(1+\varepsilon)^n) \;\leq\; K^{-n}\overline{B}(x_2).\end{aligned}$$

Now the lemma is proved by setting

$$\zeta \;=\; \frac{-\ln(K)}{\ln(1+\varepsilon)} \;>\; 0,$$

and $\eta = (1+\varepsilon)^{-\zeta}$. □

5.B Proof of Lemma 5.2.2

Lemma *If* $\mathbf{E}\left[B^{\alpha}\right] < \infty$ *then* $\mathbf{E}\left[(W_{\lambda,B})^{\alpha-1}\right] < \infty$.

Proof We define the probability distribution of the backward recurrence time of the service requirement by

$$H(t) \;:=\; \frac{1}{\beta_1}\int_{x=0}^{t} \overline{B}(x)\mathrm{d}x.$$

From this, it is straightforward to see that if the random variable H has distribution $H(t)$ then

$$\mathbf{E}\left[H^{\alpha-1}\right] \;=\; \frac{1}{\alpha\beta_1}\mathbf{E}\left[X^{\alpha}\right] \;<\; \infty.$$

Let N be a discrete random variable with $\mathbf{P}\{N=n\} = (1-\rho)\rho^n$, for $n \in \{0,1,2,\ldots\}$, and let $H_1, H_2, \ldots$, be a sequence of i.i.d. random variables (independent of N) with distribution function $H(t)$. From the Pollaczek-Khintchine formula for $W_{\lambda,B}$, given by Expression (5.7), we know that $W_{\lambda,B}$ is distributed as $H_1 + H_2 + \ldots + H_N$, where the empty sum is set equal to 0. Therefore,

$$\begin{aligned}\mathbf{E}\left[(W_{\lambda,B})^{\alpha-1}\right] \;&=\; \mathbf{E}\left[\left(\sum_{n=1}^{N} H_n\right)^{\alpha-1}\right] \\ \leq\; \mathbf{E}\left[N^{\max\{0,\alpha-2\}}\sum_{n=1}^{N}(H_n)^{\alpha-1}\right] \;&=\; \mathbf{E}\left[N^{\max\{1,\alpha-1\}}\right]\mathbf{E}\left[(H_n)^{\alpha-1}\right] \\ <\; \infty.\end{aligned}$$

In the first inequality we use the fact that, for any numbers $k \in \mathbb{N}$ and $x_1 \geq 0, \ldots, x_k \geq 0$,

$$\begin{aligned}\left(\textstyle\sum_{j=1}^{k} x_j\right)^{\xi} \;&\leq\; \textstyle\sum_{j=1}^{k}(x_j)^{\xi}, \quad \text{if } \xi \leq 1, \\ \left(\tfrac{1}{k}\textstyle\sum_{j=1}^{k} x_j\right)^{\xi} \;&\leq\; \tfrac{1}{k}\textstyle\sum_{j=1}^{k}(x_j)^{\xi}, \quad \text{if } \xi \geq 1.\end{aligned} \tag{5.39}$$

The latter is essentially Jensen's inequality. □

5.C Proof of Lemma 5.4.1

Lemma *For $s > 0$,*

$$\widehat{f}_1(s) \quad = \quad \frac{s^{-2}}{c - \rho\frac{1-\beta(s)}{s\beta_1}}.$$

Hence, for $\tau \geq 0$,

$$\begin{aligned}\mathbf{E}\left[C_0(\tau)\right] &= \frac{1}{c-\rho}\int_{t=0}^{\tau}\mathbf{P}\left\{W_{\lambda/c,B} \leq t\right\}\mathrm{d}t \\ &= \frac{\tau}{c-\rho} - \frac{1}{c-\rho}\int_{t=0}^{\tau}\mathbf{P}\left\{W_{\lambda/c,B} > t\right\}\mathrm{d}t.\end{aligned}$$

Proof First we prove that for $s > \lambda/c$ the function $\widehat{f}_1(s)$ is finite. For this purpose we construct a new random variable $\overline{C_0}(\tau)$ which is (stochastically) larger than $C_0(\tau)$, see the related approach in the proof of Theorem 3.4.1 given in Appendix 3.B. In the population model let $\overline{C_0}(\tau)$ be the reward in the time interval $[0,\tau]$ of a family of permanent individuals, starting with one individual. Thus births occur as before, but all new individuals are permanent themselves. Since no individual dies, the number of individuals is at all times (stochastically) larger than if the new individuals had life times distributed as B. Permanent individuals generate children and rewards exactly as living non-permanent ones, hence, the stochastic inequality also holds for the total rewards until time τ.

For $\mathbf{E}\left[\overline{C_0}(\tau)\right]$ we derive the following differential equation, in the same way as we did for the LST of $C_0(\tau)$ in Lemma 5.3.4,

$$\frac{\partial}{\partial\tau}\mathbf{E}\left[\overline{C_0}(\tau)\right] \quad = \quad (1+\nu m_1)\left(1+\lambda\mathbf{E}\left[\overline{C_0}(\tau)\right]\right),$$

and $\mathbf{E}\left[\overline{C_0}(0)\right] = 0$. This gives:

$$\mathbf{E}\left[\overline{C_0}(\tau)\right] \quad = \quad \frac{\mathrm{e}^{\frac{\lambda}{c}\tau}-1}{\lambda}.$$

Hence, for $s \geq \lambda/c$,

$$\widehat{f}_1(s) \quad \leq \quad \int_{\tau=0}^{\infty}\mathrm{e}^{-s\tau}\mathbf{E}\left[\overline{C_0}(\tau)\right]\mathrm{d}\tau \;=\; \frac{1}{s(cs-\lambda)} \;<\; \infty.$$

Now, $\mathbf{E}\left[C_0(\tau)\right]$ satisfies the following differential equation (which can be derived as above),

$$\begin{aligned}\frac{\partial}{\partial\tau}\mathbf{E}\left[C_0(\tau)\right] &= (1+\nu m_1)\left(1+\lambda\mathbf{E}\left[C_1(\tau;B)\right]\right) \\ &= (1+\nu m_1)\left(1+\lambda\mathbf{E}\left[C_0(\tau)\right]-\lambda\int_{x=0}^{\tau}\mathbf{E}\left[C_0(\tau-x)\right]\mathrm{d}B(x)\right),\end{aligned}$$

and $\mathbf{E}[C_0(0)] = 0$. For $s > \lambda/c$ we may “take Laplace transforms” on both sides leading to:

$$s\widehat{f}_1(s) \quad = \quad (1+\nu m_1)\left(\frac{1}{s} + \lambda\widehat{f}_1(s) - \lambda\widehat{f}_1(s)\beta(s)\right),$$

which is equivalent to the expression for $\widehat{f}_1(s)$ given in the lemma. Then, using that the LST of $W_{\lambda/c,B}$ is given by the (distributional version of the) Pollaczek-Khintchine formula (cf. Cohen [20, Part II, Expression (4.81)]),

$$\mathbf{E}\left[\mathrm{e}^{-sW_{\lambda/c,B}}\right] \quad = \quad \frac{1-\rho/c}{1-(\rho/c)\frac{1-\beta(s)}{\beta_1 s}}.$$

We can derive the second part of the lemma by inverting this transform and then integrating twice with respect to τ. □

5.D Proof of Lemma 5.4.6

Lemma *If $\beta_2 < \infty$ then*

$$\lim_{\tau\to\infty}\left(R_0(\tau) - \frac{2\tau}{c(c-\rho)}\right) \quad = \quad \nu m_2\left(\frac{c}{c-\rho}\right)^2 + \frac{\lambda\beta_2}{c(c-\rho)^2}.$$

When $\beta_2 = \infty$ and $\mathbf{E}[B^\alpha] < \infty$, for some $\alpha \in (1,2)$, then, for any $\varepsilon > 0$,

$$R_0(\tau) - \frac{2\tau}{c(c-\rho)} \quad = \quad \mathrm{o}(\tau^{2-\alpha+\varepsilon}), \quad \tau\to\infty.$$

Proof The proofs for the cases $\beta_2 < \infty$ and $\beta_2 = \infty$ proceed along the same lines. When $\beta_2 = \infty$, the constants α and ϵ are as stated above. We study the three terms in the right-hand side of Expression (5.22) separately. By Lemma 5.4.4, the first term converges to a constant as $\tau \to \infty$. For the second term we have, by Lemmas 5.4.3 and 5.4.4,

$$\begin{aligned}
&\mathbf{E}[C_0(\tau)]\Big(1+\lambda\mathbf{E}[C_1(\tau;B)]\Big) - \frac{c\tau}{(c-\rho)^2} \\
&= \left(\mathbf{E}[C_0(\tau)] - \frac{\tau}{c-\rho}\right)\left(1+\frac{\rho}{c-\rho}+\lambda\left(\mathbf{E}[C_1(\tau;B)] - \frac{\beta_1}{c-\rho}\right)\right) \\
&\qquad + \frac{\lambda\tau}{c-\rho}\left(\mathbf{E}[C_1(\tau;B)] - \frac{\beta_1}{c-\rho}\right) \\
&\begin{cases} \longrightarrow -\dfrac{\lambda c\beta_2}{2(c-\rho)^3}, & \tau\to\infty, \text{ when } \beta_2<\infty, \\ = \mathrm{o}(\tau^{2-\alpha+\varepsilon}), & \tau\to\infty, \text{ when } \beta_2=\infty. \end{cases}
\end{aligned}$$

We now turn to the third and last term of Expression (5.22). First note that

$$\tau\int_{x=0}^{\tau}\mathbf{E}[C_1(\tau;x)]\,\mathrm{d}B(x) - \frac{\beta_1\tau}{c-\rho}$$

$$
\begin{aligned}
&= \tau\left(\mathbf{E}\left[C_1(\tau;B)\right]-\mathbf{E}\left[C_0(\tau)\right]\overline{B}(\tau)\right)-\frac{\beta_1\tau}{c-\rho}\\
&= \tau\left(\mathbf{E}\left[C_1(\tau;B)\right]-\frac{\beta_1}{c-\rho}\right)-\tau\left(\mathbf{E}\left[C_0(\tau)\right]-\frac{\tau}{c-\rho}\right)\overline{B}(\tau)\\
&\quad -\frac{\tau^2}{c-\rho}\overline{B}(\tau)\\
&\begin{cases} = \mathrm{o}(1), & \tau\to\infty, \text{ when } \beta_2<\infty,\\ = \mathrm{o}(\tau^{2-\alpha}), & \tau\to\infty, \text{ when } \beta_2=\infty.\end{cases}
\end{aligned}
$$

To obtain the limiting behaviour we combined Lemmas 5.4.3 and 5.4.4. For the case $\beta_2<\infty$ we further used the fact that $\tau^2\overline{B}(\tau)$ vanishes as τ tends to infinity, see also Lemma 5.2.1. We now may write

$$
\begin{aligned}
&\int_{x=0}^{\tau}\mathbf{E}\left[C_0(\tau-x)\right]\mathbf{E}\left[C_1(\tau;x)\right]\mathrm{d}B(x)-\frac{\beta_1\tau}{(c-\rho)^2}\\
&= \int_{x=0}^{\tau}\left(\mathbf{E}\left[C_0(\tau-x)\right]-\frac{\tau-x}{c-\rho}\right)\mathbf{E}\left[C_1(\tau;x)\right]\mathrm{d}B(x)\\
&\qquad -\int_{x=0}^{\tau}\frac{x}{c-\rho}\mathbf{E}\left[C_1(\tau;x)\right]\mathrm{d}B(x)\\
&\qquad +\frac{\tau}{c-\rho}\left(\int_{x=0}^{\tau}\mathbf{E}\left[C_1(\tau;x)\right]\mathrm{d}B(x)-\frac{\beta_1}{c-\rho}\right)\\
&\begin{cases} \longrightarrow -\dfrac{\beta_2}{(c-\rho)^2}\left(\dfrac{\rho}{2(c-\rho)}+1\right), & \tau\to\infty, \text{ when } \beta_2<\infty,\\ = \mathrm{o}(\tau^{2-\alpha+\varepsilon}), & \tau\to\infty, \text{ when } \beta_2=\infty.\end{cases}
\end{aligned}
$$

Here we used that the order of limit and the first two integrals may be interchanged. For the case $\beta_2<\infty$ this is justified by the Monotone Convergence Theorem, the monotonicity being clear from Lemmas 5.4.3 and 5.4.4. For the case $\beta_2=\infty$ we need the Dominated Convergence Theorem, using (for the first integral) that

$$
\begin{aligned}
\mathbf{1}_{\{0\le x\le\tau\}}\frac{\mathbf{E}\left[C_0(\tau-x)\right]-\frac{\tau-x}{c-\rho}}{(1+\tau-x)^{2-\alpha+\varepsilon}}\left(\frac{1+\tau-x}{1+\tau}\right)^{2-\alpha+\varepsilon} &\le K,\\
\mathbf{E}\left[C_1(\tau;x)\right] &\le \frac{x}{c-\rho},
\end{aligned}
$$

and that $\beta_1<\infty$. The existence of the constant K follows from Lemma 5.4.3. Furthermore, (for the second integral) we need that

$$
\int_{x=0}^{\tau}x^2\mathrm{d}B(x) = \mathrm{o}(\tau^{2-\alpha+\varepsilon}),
$$

which follows easily by partial integration. Combining the above results in Expression (5.22) we have, for $\beta_2<\infty$,

$$
\lim_{\tau\to\infty}\left(R_0(\tau)-2(1+\nu m_1)\frac{c\tau}{(c-\rho)^2}+2\lambda(1+\nu m_1)\frac{\beta_1\tau}{(c-\rho)^2}\right)
$$

$$
\begin{aligned}
&= \nu m_2 \left(\frac{c}{c-\rho}\right)^2 - 2(1+\nu m_1)\frac{\lambda c \beta_2}{2(c-\rho)^3} \\
&\quad +2\lambda(1+\nu m_1)\frac{\beta_2}{(c-\rho)^2}\left(\frac{\rho}{2\,(c-\rho)}+1\right),
\end{aligned}
$$

and, for $\beta_2 = \infty$,

$$
R_0(\tau) - 2(1+\nu m_1)\frac{c\tau}{(c-\rho)^2} + 2\lambda(1+\nu m_1)\frac{\beta_1 \tau}{(c-\rho)^2} \;=\; \mathrm{o}(\tau^{2-\alpha+\varepsilon}),
$$

which completes the proof. □

5.E Proof of Lemma 5.4.7

Lemma *If $\beta_2 < \infty$ then*

$$
\mathbf{E}\left[C_0(\tau)^2\right] \;=\; \left(\frac{\tau}{c-\rho}\right)^2 + \tau\nu m_2\left(\frac{c}{c-\rho}\right)^3 + h(\tau),
$$

where the function $h(\tau)$ is such that, for all $x \geq 0$,

$$
\lim_{\tau\to\infty}\Big(h(\tau) - h(\tau - x)\Big) \;=\; 0.
$$

In particular,

$$
\lim_{\tau\to\infty}\frac{\mathbf{E}\left[C_0(\tau)^2\right] - \left(\frac{\tau}{c-\rho}\right)^2}{\tau} \;=\; \nu m_2\left(\frac{c}{c-\rho}\right)^3.
$$

Proof First we rewrite Expression (5.23) as

$$
\begin{aligned}
\mathbf{E}\left[C_0(\tau)^2\right] \;&=\; \frac{c}{c-\rho}\int_{u=0}^{\tau} R_0(\tau-u)\mathrm{d}u \\
&\quad -\frac{c}{c-\rho}\int_{u=0}^{\tau} R_0(\tau-u)\mathbf{P}\left\{W_{\lambda/c,B} > u\right\}\mathrm{d}u. \qquad (5.40)
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&\mathbf{E}\left[C_0(\tau)^2\right] - \mathbf{E}\left[C_0(\tau-x)^2\right] \\
&= \frac{c}{c-\rho}\Bigg(\int_{u=0}^{\tau} R_0(\tau-u)\mathrm{d}u - \int_{u=0}^{\tau-x} R_0(\tau-x-u)\mathrm{d}u \\
&\qquad -\int_{u=0}^{\tau-x}\Big(R_0(\tau-u) - R_0(\tau-x-u)\Big)\mathbf{P}\left\{W_{\lambda/c,B} > u\right\}\mathrm{d}u \\
&\qquad -\int_{u=\tau-x}^{\tau} R_0(\tau-u)\mathbf{P}\left\{W_{\lambda/c,B} > u\right\}\mathrm{d}u\Bigg). \qquad (5.41)
\end{aligned}
$$

Taking the first two integrals together, we have

$$
\begin{aligned}
&\frac{c}{c-\rho}\left(\int_{u=0}^{\tau} R_0(\tau-u)\mathrm{d}u - \int_{u=0}^{\tau-x} R_0(\tau-x-u)\mathrm{d}u\right)\\
&= \frac{c}{c-\rho}\int_{u=\tau-x}^{\tau} R_0(u)\mathrm{d}u\\
&= \frac{2x\tau - x^2}{(c-\rho)^2} + \frac{c}{c-\rho}\int_{u=\tau-x}^{\tau}\left(R_0(u) - \frac{2u}{c(c-\rho)}\right)\mathrm{d}u,
\end{aligned}
$$

and by Lemma 5.4.6 we have

$$
\lim_{\tau\to\infty}\int_{u=\tau-x}^{\tau}\left(R_0(u) - \frac{2u}{c(c-\rho)}\right)\mathrm{d}u \quad = \quad \frac{x\left(\nu m_2 c^3 + \lambda\beta_2\right)}{c(c-\rho)^2}.
$$

Now focus on the third integral in the right-hand side of Expression (5.41). From Lemma 5.4.6 we can derive that, for fixed $x \geq 0$,

$$
\lim_{\tau\to\infty}\Big(R_0(\tau) - R_0(\tau-x)\Big) \quad = \quad \frac{2x}{c(c-\rho)},
$$

and hence,

$$
\mathbf{1}_{\{0\leq u\leq \tau-x\}}\Big|\, R_0(\tau-u) - R_0(\tau-x-u)\,\Big| \quad \leq \quad K(x),
$$

for some constant $K(x)$ that is independent of u and τ. Since $\beta_2 < \infty$ and, hence, $\mathbf{E}\left[W_{\lambda/c,B}\right] < \infty$, we may interchange the order of limit and integral in:

$$
\begin{aligned}
&\lim_{\tau\to\infty}\int_{u=0}^{\infty}\mathbf{1}_{\{0\leq u\leq\tau-x\}}\Big(R_0(\tau-u) - R_0(\tau-x-u)\Big)\mathbf{P}\left\{W_{\lambda/c,B} > u\right\}\mathrm{d}u\\
&= \frac{2x}{c(c-\rho)}\mathbf{E}\left[W_{\lambda/c,B}\right] \quad = \quad \frac{x\lambda\beta_2}{c(c-\rho)^2}.
\end{aligned}
$$

For the last integral in the right-hand side of Expression (5.41) we write

$$
\begin{aligned}
&\left|\int_{u=\tau-x}^{\tau} R_0(\tau-u)\mathbf{P}\left\{W_{\lambda/c,B} > u\right\}\mathrm{d}u\right|\\
&\leq \int_{u=\tau-x}^{\tau}\Big|\,R_0(\tau-u)\,\Big|\,\mathbf{P}\left\{W_{\lambda/c,B} > u\right\}\mathrm{d}u\\
&\leq \mathbf{P}\left\{W_{\lambda/c,B} > \tau-x\right\}\int_{u=\tau-x}^{\tau}\Big|\,R_0(\tau-u)\,\Big|\,\mathrm{d}u\\
&\leq \mathbf{P}\left\{W_{\lambda/c,B} > \tau-x\right\}\int_{u=0}^{x}\Big|\,R_0(u)\,\Big|\,\mathrm{d}u\\
&\longrightarrow 0, \quad \tau\to\infty.
\end{aligned}
$$

Combining the above results for the integrals in the right-hand side of Expression (5.41), we have:

$$
\lim_{\tau\to\infty}\left(\mathbf{E}\left[C_0(\tau)^2\right] - \mathbf{E}\left[C_0(\tau-x)^2\right] - \frac{2x\tau}{(c-\rho)^2}\right)
$$

$$= x\nu m_2 \left(\frac{c}{c-\rho}\right)^3 - \frac{x^2}{(c-\rho)^2}.$$

Hence, the function

$$h(\tau) \quad := \quad \mathbf{E}\left[C_0(\tau)^2\right] - \left(\frac{\tau}{c-\rho}\right)^2 - \tau\nu m_2 \left(\frac{c}{c-\rho}\right)^3,$$

satisfies, for all $x \geq 0$,

$$\lim_{\tau\to\infty} \Big(h(\tau) - h(\tau - x) \Big) \quad = \quad 0.$$

The first statement is now proved. To prove the second part, note that for all $x \geq 0$ and $\varepsilon > 0$ there exists a $\tau_{x,\varepsilon}$ such that, for all $\tau \geq \tau_{x,\varepsilon}$,

$$\Big| \, h(\tau) - h(\tau - x) \, \Big| \leq \varepsilon,$$

which implies that

$$\lim_{\tau\to\infty} \left| \frac{h(\tau)}{\tau} \right| \quad \leq \quad \frac{\varepsilon}{x}.$$

Now let $\varepsilon \downarrow 0$. □

5.F Proof of Lemma 5.5.1

Lemma *If $\mathbf{E}\left[B^\alpha\right] < \infty$, for some $\alpha \in (1, 2)$, then $\mathbf{E}\left[W^{\alpha-1}\right] < \infty$. Hence,*

$$\mathbf{P}\{W > x\} \quad = \quad \mathrm{o}(x^{1-\alpha}),$$

for $x \to \infty$.

Proof Let the random variable $\widetilde{B}$ be distributed as the amount of work that arrives during an off-period:

$$\mathbf{E}\left[\mathrm{e}^{-s\widetilde{B}}\right] \quad = \quad \phi(\lambda(1 - \beta(s))).$$

Obviously,

$$\widetilde{B} \quad \stackrel{d}{=} \quad \sum_{n=1}^{A} B_n,$$

where the B_n form an i.i.d. sequence with distribution $B(x)$, and A is independent from the B_n and distributed as the number of customers that arrive during an off-period:

$$\mathbf{E}\left[z^A\right] \quad = \quad \phi(\lambda(1 - z)).$$

Note that, because $\alpha > 1$,

$$\mathbf{E}\left[\left(\sum_{n=1}^{A} B_n\right)^{\alpha}\right] \le \mathbf{E}\left[A^{\alpha-1}\sum_{n=1}^{A}(B_n)^{\alpha}\right] = \mathbf{E}\left[A^{\alpha}\right]\mathbf{E}\left[(B_n)^{\alpha}\right] < \infty.$$

The first inequality is a consequence of Relations (5.39) with $\xi = \alpha > 1$. For last inequality we used that $m_3 < \infty$ and, hence, $\mathbf{E}\left[A^3 < \infty\right]$. Thus, we may conclude that $\mathbf{E}\left[\widetilde{B}^{\alpha}\right] < \infty$. Clearly, if the random variable $\widehat{B}$ has LST $\widehat{\beta}(s)$, given by Expression (5.31), then $\mathbf{E}\left[\widehat{B}^{\alpha}\right] < \infty$, because the distribution of $\widehat{B}$ is a weighted combination of those of B and $\widetilde{B}$. Note that W *given* that $Y = 1$ is distributed as $W_{\lambda/c,\widehat{B}}$, the waiting time in the M/G/1 FCFS queue with arrival rate λ/c and service requirements distributed as $\widehat{B}$. Hence, by Lemma 5.2.2, $\mathbf{E}\left[W^{\alpha-1}\,|Y = 1\right] < \infty$.

By similar arguments we may conclude that if the random variable $\widetilde{B'}$ is distributed as the amount of work that arrives during E_0, the backward recurrence time of an off-period, then $\mathbf{E}\left[\widetilde{B'}^{\alpha}\right] < \infty$. Note that it suffices that the *second* moment of the distribution of E_0 is finite to use the above arguments. Then use that W given that $Y = 0$ is distributed as the sum of $\widetilde{B'}$ and W given that $Y = 1$, to conclude that also $\mathbf{E}\left[W^{\alpha-1}\,|Y = 0\right] < \infty$. Hence, $\mathbf{E}\left[W^{\alpha-1}\right] < \infty$ and the final statement follows by Lemma 5.2.1. $\square$

References

[1] ATM Forum Technical Committee. Traffic Management Specification. Version 4.0, April 1996.

[2] J. Abate, W. Whitt. The Fourier-series method for inverting transforms of probability distributions. *Queueing Systems* 10 (1992), 5–87.

[3] B. Almási. A queuing model for a processor-shared multi-terminal system subject to breakdowns. *Acta Cybernetica* 10 (1992), 273–282.

[4] B. Almási, J. Sztrik. A queueing model for a non-homogeneous terminal system subject to breakdowns. *Computers and Mathematics with Applications* 25 (1993), 105–111.

[5] E. Altman, D. Artiges, K. Traore. On the integration of best-effort and guaranteed performance services. INRIA Research Report 3222 (1997).

[6] V. Anantharam. Scheduling strategies and long-range dependence. Report, Department of Electrical Engineering & Computer Sciences, University of California at Berkeley (1997).

[7] S. Asmussen. *Applied Probability and Queues.* Wiley, Chichester (1987).

[8] F. Baskett, K.M. Chandy, R.R. Muntz, F.G. Palacios. Open, closed and mixed networks of queues with different classes of customers. *Journal of the Association for Computing Machinery* 22 (1975), 248–260.

[9] J.L. van den Berg. *Sojourn times in feedback and processor-sharing queues.* Ph.D. thesis, Rijksuniversiteit Utrecht (1990).

[10] J.L. van den Berg, O.J. Boxma. The M/G/1 queue with processor sharing and its relation to a feedback queue. *Queueing Systems* 9 (1991), 365–401.

[11] A. Berger, Y. Kogan. Multi-class elastic data traffic: Bandwidth engineering via asymptotic approximations. *In: Teletraffic Engineering in a Competitive World — Proceedings of ITC 16, Edinburgh.* Eds. D. Smith and P. Key. Elsevier, Amsterdam (1999), 77–86.

[12] D.P. Bertsekas, R. Gallager. *Data Networks.* Prentice-Hall, Englewood Cliffs (1987).

[13] N.H. Bingham, C.M. Goldie, J.L. Teugels. *Regular Variation.* Cambridge University Press, Cambridge (1987).

[14] S. Blaabjerg, G. Fodor, A.T. Andersen, M. Telek. A partial blocking-queueing system with CBR/VBR and ABR/UBR arrival streams. *In: Proceedings of the 5th International Conference on Telecommunication Systems, Nashville* (1997).

[15] A. Brandt, M. Brandt. On the sojourn times in many-queue head-of-the-line processor-sharing systems with permanent customers. *Mathematical Methods of Operations Research* 47 (1998), 181–220.

[16] D.B.H. Cline. Intermediate regular and Π variation. *Proceedings of the London Mathematical Society (3^{rd} series)* 68 (1994), 594–616.

[17] E.G. Coffman, R.R. Muntz, H. Trotter. Waiting-time distributions for processor-sharing systems. *Journal of the Association for Computing Machinery* 17 (1970), 123–130.

[18] J.W. Cohen. Some results on regular variation in queueing and fluctuation theory. *Journal of Applied Probability* 10 (1973), 343–353.

[19] J.W. Cohen. The multiple phase service network with generalised processor sharing. *Acta Informatica* 12 (1979), 245–284.

[20] J.W. Cohen. *The Single Server Queue.* (2nd ed.) North-Holland, Amsterdam (1982).

[21] J.W. Cohen, O.J. Boxma. A survey of the evolution of queueing theory. *Statistica Neerlandica* 39 (1985), 143–158.

[22] P.J. Courtois. *Decomposability — Queueing and Computer System Applications.* Academic Press, New York (1977).

[23] J.N. Daigle, D.M. Lucantoni. Queueing systems having phase-dependent arrival and service rates. *In: Numerical Solution of Markov Chains.* Ed. W.J. Stewart. Marcel Dekker, New York (1991), 161–202.

[24] V. De Nitto Personè, V. Grassi. Solution of finite QBD processes. *Journal of Applied Probability* 33 (1996), 1003–1010.

[25] A. Elwalid, D. Mitra. Analysis, approximations and admission control of a multi-service multiplexing system with priorities. *In: Proceedings of IEEE INFOCOM '95* (1995), 463–472.

[26] G. Falin, Z. Khalil, D.A. Stanford. Performance analysis of a hybrid switching system where voice messages can be queued. *Queueing Systems* 16 (1994), 51–65.

[27] A. Federgruen, L. Green. Queueing systems with service interruptions. *Operations Research* 34 (1986), 752–768.

[28] A. Federgruen, L. Green. Queueing systems with service interruptions II. *Naval Research Logistics* 35 (1988), 345–358.

[29] W. Feller. *An Introduction to Probability Theory and its Applications — Volume I.* (3rd ed.) Wiley, New York (1968).

[30] W. Feller. *An Introduction to Probability Theory and its Applications — Volume II.* Wiley, New York (1966).

[31] R.D. Foley, G.-A. Klutke. Stationary increments in the accumulated work process in processor-sharing queues. *Journal of Applied Probability* 26 (1989), 671–677.

[32] S.W. Fuhrmann, R.B. Cooper. Stochastic decompositions in the M/G/1 queue with generalised vacations. *Operations Research* 33 (1985), 1117–1129.

[33] H.R. Gail, S.L. Hantler, A.G. Konheim, B.A. Taylor. An analysis of a class of telecommunications models. *Performance Evaluation* 21 (1994), 151–161.

[34] H.R. Gail, S.L. Hantler, B.A. Taylor. Analysis of a non-preemptive priority multi-server queue. *Advances in Applied Probability* 20 (1988), 852–879.

[35] H.R. Gail, S.L. Hantler, B.A. Taylor. On a preemptive Markovian queue with multiple servers and two priority classes. *Mathematics of Operations Research* 17 (1992), 365–391.

[36] H.R. Gail, S.L. Hantler, B.A. Taylor. Spectral analysis of $M/G/1$ and $G/M/1$ type Markov chains. *Advances in Applied Probability* 28 (1996), 114–165.

[37] D.P. Gaver, Jr. A waiting line with interrupted service, including priorities. *Journal of the Royal Statistical Society* 24 (1962), 73–90.

[38] I. Gohberg, P. Lancaster, L. Rodman. *Matrix Polynomials.* Academic Press, New York (1982).

[39] S.A. Grishechkin. Crump-Mode-Jagers branching processes as a method of investigating M/G/1 systems with processor sharing. *Theory of Probability and its Applications* 36 (1991), 19–35; translated from *Teoriya Veroyatnostei i ee Primeneniya* 36 (1991), 16–33 (in Russian).

[40] S.A. Grishechkin. On a relationship between processor sharing queues and Crump-Mode-Jagers branching processes. *Advances in Applied Probability* 24 (1992), 653–698.

[41] B. Hajek. Birth-and-death processes on the integers with phases and general boundaries. *Journal of Applied Probability* 19 (1982), 488–499.

[42] B.R. Haverkort, A. Ost. Steady-state analysis of infinite stochastic Petri nets: Comparing the spectral expansion and the matrix-geometric method. *In: Proceedings of the 7th International Workshop on Petri Nets and Performance Models.* IEEE Computer Society Press (1997), 36–45.

[43] D.P. Heyman, T.V. Lakshman, A.L. Neidhardt. A new method for analysing feedback-based protocols with applications to engineering Web traffic over the Internet. *In: Proceedings of ACM SIGMETRICS 97 — Performance Evaluation Review* 25 (1997), 24–38.

[44] R.A. Horn and C.R. Johnson. *Matrix Analysis.* Cambridge University Press, Cambridge (1987).

[45] R.A. Howard. *Dynamic Programming and Markov Processes.* M.I.T. Press, New York (1960).

[46] V. Jacobson. Congestion avoidance and control. *Proceedings of ACM SIGCOMM '88,* 314–329.

[47] F.P. Kelly. Networks of queues. *Advances in Applied Probability* 8 (1976), 416–432.

[48] F.P. Kelly. *Reversibility and Stochastic Networks.* Wiley, Chichester (1979).

[49] D.G. Kendall. Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *Annals of Mathematical Statistics* 24 (1953), 338–354.

[50] G. Kesidis. *ATM Network Performance.* Kluwer, Boston (1996).

[51] M.Yu. Kitayev, S.F. Yashkov. Analysis of a single-channel queueing system with the discipline of uniform sharing of a device. *Engineering Cybernetics* 17 (1979), 42–49; translated from *Izvestiya Akademii Nauk SSSR Tekhnicheskaya Kibernetika* (1979), 64–71 (in Russian).

[52] L. Kleinrock. Analysis of a time-shared processor. *Naval Research Logistics Quarterly* 11 (1964), 59–73.

[53] L. Kleinrock. Time-shared systems: A theoretical treatment. *Journal of the Association for Computing Machinery* 14 (1967), 242–261.

[54] L. Kleinrock. *Queueing Systems, Vol. I: Theory.* Wiley, New York (1975).

[55] L. Kleinrock. *Queueing Systems, Vol. II: Computer Applications.* Wiley, New York (1976).

[56] U.R. KRIEGER, V. NAOUMOV, D. WAGNER. Analysis of a versatile multiclass delay-loss system with a superimposed Markovian arrival process. *European Journal of Operations Research* 108 (1998), 425–437.

[57] L.A. KULKARNI, S.-Q. LI. Performance analysis of rate based feedback control for ATM networks. *In: Proceedings of IEEE INFOCOM '97, Kobe* (1997), 795–804.

[58] A. KUMAR, K.V.S. HARI, R. SHOBHANJALI, S. SHARMA. Long-range dependence in the aggregate flow of TCP-controlled elastic sessions: An investigation via the processor-sharing model. *In: Proceedings of the National Conference on Communications 2000, New Delhi.*

[59] V.P. KUMAR, T.V. LAKSHMAN, D. STILIADIS. Beyond best effort: Architectures for the differentiated services of tomorrow's Internet. *IEEE Communications Magazine* 36 (1998), 152–164.

[60] T.V. LAKSHMAN, V.P. KUMAR. Guest editorial. *IEEE Communications Magazine* 36 (1998), p. 127.

[61] G. LATOUCHE, C.E.M. PEARCE, P.G. TAYLOR. Invariant measures for quasi birth and death processes. *Communications in Statistics — Stochastic Models* 14 (1998), 443–460.

[62] G. LATOUCHE, V. RAMASWAMI. A logarithmic reduction algorithm for quasi-birth-death processes. *Journal of Applied Probability* 30 (1993), 650–674.

[63] D.-S. LEE. Analysis of a single server queue with semi-Markovian service interruption. *Queueing Systems* 27 (1997), 153–178.

[64] J.P. LEHOCZKY, D.P. GAVER. Diffusion approximation for the cooperative service of voice and data messages. *Journal of Applied Probability* 18 (1981), 660–671.

[65] S.-Q. LI, H.-D. SHENG. Generalised folding-algorithm for sojourn time analysis of finite QBD processes and its queueing applications. *Communications in Statistics — Stochastic Models* 12 (1996), 507–522.

[66] W. LI, D. SHI, X. CHAO. Reliability analysis of M/G/1 queueing systems with server breakdowns and vacations. *Journal of Applied Probability* 34 (1997), 546–555.

[67] D.V. LINDLEY. The theory of queues with a single server. *Proceedings of the Cambridge Philosophical Society* 48 (1952), 277–289.

[68] R.M. LOYNES. The stability of a queue with non-independent interarrival and service times. *Proceedings of the Cambridge Philosophical Society* 58 (1962), 497–520.

[69] M. Marcus, H. Minc. *A Survey of Matrix Theory and Matrix Inequalities.* Allyn and Bacon, Inc., Boston (1964).

[70] L. Massoulié, J.W. Roberts. Arguments in favour of admission control for TCP flows. *In: Teletraffic Engineering in a Competitive World — Proceedings of ITC 16, Edinburgh.* Eds. D. Smith and P. Key. Elsevier, Amsterdam (1999), 33–44.

[71] L. Massoulié, J.W. Roberts. Bandwidth sharing: Objectives and algorithms. *In: Proceedings of IEEE INFOCOM '99, New York* (1999), 1395–1403.

[72] A. de Meyer, J.L. Teugels. On the asymptotic behaviour of the distributions of the busy period and service time in M/G/1. *Journal of Applied Probability* 17 (1980), 802–813.

[73] I. Mitrani. The spectral-expansion solution method for Markov processes on lattice strips. *In: Advances in Queueing — Theory, Methods, and Open Problems.* Ed. J.H. Dshalalow. CRC Press, Boca Raton (1995), 337–352.

[74] I. Mitrani, B. Avi-Itzhak. A many server queue with service interruptions. *Operations Research* 16 (1968), 628–638.

[75] I. Mitrani, R. Chakka. Spectral expansion solution for a class of Markov models: Application and comparison with the matrix-geometric method. *Performance Evaluation* 23 (1995), 241–260.

[76] I. Mitrani, P.J.B. King. Multiprocessor systems with preemptive priorities. *Performance Evaluation* 1 (1981), 118–125.

[77] I. Mitrani, D. Mitra. A spectral expansion method for random walks on semi-infinite strips. *In: Iterative Methods in Linear Algebra — Proceedings of the IMACS International Symposium, Brussels.* Eds. R. Beauwens and P. de Groen, Elsevier, Amsterdam (1991), 141–149.

[78] J.A. Morrison. Response-time distribution for a processor sharing system. *SIAM Journal on Applied Mathematics* 45 (1985), 152–167.

[79] V.A. Naoumov. Matrix-multiplicative approach to Quasi-Birth-and-Death processes analysis. *In: Matrix-Analytic Methods in Stochastic Models — Lecture Notes in Pure and Applied Mathematics.* Eds. A.S. Alfa and S.R. Chakravarthy. Marcel Dekker, New York (1996), 87–106.

[80] R. Nelson. *Probability, Stochastic Processes, and Queueing Theory — The Mathematics of Computer Performance Modeling.* Springer, New York (1995).

[81] M.F. Neuts. *Matrix-Geometric Solutions in Stochastic Models — An Algorithmic Approach.* Johns Hopkins, Baltimore (1981).

[82] R. Núñez Queija. A queueing model with varying service rate for ABR. *In: Computer Performance Evaluation — Modelling Techniques and Tools — Proceedings of TOOLS '98, Mallorca.* Eds. R. Puigjaner, N.N. Savino, and B. Serra. Springer Verlag, Berlin (1998), 93–104.

[83] R. Núñez Queija. Sojourn times in a processor-sharing queue with service interruptions. Accepted for publication in *Queueing Systems.* Preprint: CWI Report PNA-R9807 (1998).

[84] R. Núñez Queija. Sojourn times in non-homogeneous QBD processes with processor sharing. CWI Report PNA-R9901 (1999). Submitted for publication.

[85] R. Núñez Queija, J.L. van den Berg, M.R.H. Mandjes. Performance evaluation of strategies for integration of elastic and stream traffic. *In: Teletraffic Engineering in a Competitive World — Proceedings of ITC 16, Edinburgh.* Eds. D. Smith and P. Key. Elsevier, Amsterdam (1999), 1039–1050.

[86] R. Núñez Queija, O.J. Boxma. Analysis of a multi-server queueing model of ABR. *Journal of Applied Mathematics and Stochastic Analysis* 11 (1998), 339–354.

[87] T.J. Ott. The sojourn time distributions in the $M/G/1$ queue with processor sharing. *Journal of Applied Probability* 21 (1984), 360–378.

[88] B.N. Parlett. *The Symmetric Eigenvalue Problem.* Prentice-Hall, Englewood Cliffs (1980).

[89] N.U. Prabhu. A bibliography of books and survey papers on queueing systems: Theory and applications. *Queueing Systems* 2 (1987), 393–398.

[90] M.L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* Wiley, New York (1994).

[91] K.M. Rege, B. Sengupta. A decomposition theorem and related results for the discriminatory processor sharing queue. *Queueing Systems* 18 (1994), 333–351.

[92] M.I. Reiman, J.A. Schmitt. Performance models of multirate traffic in various network implementations. *In: The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks — Proceedings of ITC 14, Antibes Juan-les-Pins.* Eds. J. Labetoulle and J.W. Roberts. Elsevier, Amsterdam (1994), 1217–1228.

[93] Y. Rekhter, B. Davie, E. Rosen, G. Swallow, D. Farinacci, D. Katz. Tag Switching architecture overview. *Proceedings of the IEEE* 85 (1997), 1973–1983.

[94] J.W. Roberts. Realising quality of service guarantees in multiservice networks. *In: Proceedings of the IFIP Seminar PMCCN '97, Tsukuba.* Eds. T. Hasegawa, H. Takagi, Y. Takahashi. Chapman and Hall, London (1998), 277–293.

[95] J.W. Roberts. Quality of service guarantees and charging in multiservice networks. *IEICE Transactions on Communications* 81-B (1998), 824–831.

[96] J.W. Roberts. Engineering for quality of service. *In: Self-similar Network Traffic and Performance Evaluation.* Eds. K. Park and W. Willinger. Wiley, New York (2000).

[97] J.W. Roberts, L. Massoulié. Bandwidth sharing and admission control for elastic traffic. *In: Proceedings of the ITC Specialist Seminar, Yokohama,* (1998).

[98] S.M. Ross. *Applied Probability Models with Optimization Applications.* Holden-Day, San Francisco (1970).

[99] M. Sakata, S. Noguchi, J. Oizumi. Analysis of a processor-shared queueing model for time-sharing systems. *In: Proceedings of the 2nd Hawaii International Conference on System Sciences* (1969), 625–628.

[100] M. Sakata, S. Noguchi, J. Oizumi. An analysis of the M/G/1 queue under round-robin scheduling. *Operations Research* 19 (1971), 371–385.

[101] R. Schassberger. A new approach to the $M/G/1$ processor sharing queue. *Advances in Applied Probability* 16 (1984), 202–213.

[102] L.E. Schrage, L.W. Miller. The queue M/G/1 with the shortest remaining processing time discipline. *Operations Research* 14 (1966), 670–684.

[103] B. Sengupta. A queue with service interruptions in an alternating random environment. *Operations Research* 38 (1990), 308–318.

[104] B. Sengupta. An approximation for the sojourn-time distribution for the $GI/G/1$ processor-sharing queue. *Communications in Statistics — Stochastic Models* 8 (1992), 35–57.

[105] B. Sengupta, D.L. Jagerman. A conditional response time of the M/M/1 processor-sharing queue. *AT&T Technical Journal* 64 (1985), 409–421.

[106] L.I. Sennott. Value iteration in countable state average cost Markov decision processes with unbounded costs. *Annals of Operations Research* 28 (1991), 261–271.

[107] H.A. Simon, A. Ando. Aggregation of variables in dynamic systems. *Econometrica* 29 (1961), 111–138.

[108] E. DE SOUZA E SILVA, H.R. GAIL. An algorithm to calculate transient distributions of cumulative rate and impulse based reward. *Communications in Statistics — Stochastic Models* 14 (1998), 509–536.

[109] A.J. STAM. Regular variation of the tail of a subordinated probability distribution. *Advances in Applied Probability* 5 (1973), 308–327.

[110] H. TAKAGI. *Queueing Analysis — A Foundation of Performance Evaluation.* (3 volumes.) Elsevier, Amsterdam (1991–1993).

[111] T. TAKINE, B. SENGUPTA. A single server queue with server interruptions. *Queueing Systems* 26 (1997), 285–300.

[112] H.C. TIJMS. *Stochastic Models — An Algorithmic Approach.* Wiley, Chichester (1994).

[113] A. VISWANATHAN, N. FELDMAN, Z. WANG, R. CALLON. Evolution of multiprotocol label switching. *IEEE Communications Magazine* 36 (1998), 165–173.

[114] K. VAN DER WAL, M. MANDJES, H. BASTIAANSEN. Delay performance analysis of the new Internet services with guaranteed QoS. *Proceedings of the IEEE* 85 (1997), 1947–1957.

[115] P.D. WELCH. On a generalised M/G/1 queueing process in which the first customer of each busy period receives exceptional service. *Operations Research* 12 (1964), 736–752.

[116] H. WHITE, L.S. CHRISTIE. Queuing with preemptive priorities or with breakdown. *Operations Research* 6 (1958), 79–95.

[117] P.P. WHITE, J. CROWCROFT. The integrated services in the Internet: State of the art. *Proceedings of the IEEE* 85 (1997), 1934–1946.

[118] D. WILLIAMS. *Probability with Martingales.* Cambridge University Press, Cambridge (1991).

[119] R.W. WOLFF. Poisson arrivals see time averages. *Operations Research* 30 (1982), 223–231.

[120] S.F. YASHKOV. A derivation of response time distribution for a $M/G/1$ processor-sharing queue. *Problems of Control and Information Theory* 12 (1983), 133–148.

[121] S.F. YASHKOV. New applications of random time change to the analysis of processor-sharing queues. *In: Proceedings of the 4th International Vilnius Conference on Probability Theory and Mathematical Statistics* (1985), 343–345.

[122] S.F. YASHKOV. Processor-sharing queues: Some progress in analysis. *Queueing Systems* 2 (1987), 1–17.

[123] S.F. Yashkov. Mathematical problems in the theory of processor-sharing queueing systems. *Journal of Soviet Mathematics* 58 (1992), 101–147.

[124] S.F. Yashkov. On a heavy traffic limit theorem for the $M/G/1$ processor-sharing queue. *Communications in Statistics — Stochastic Models* 9 (1993), 467–471.

[125] U. Yechiali. A queueing-type birth and death process defined on a continuous-time Markov chain. *Operations Research* 21 (1973), 604–609.

[126] Y.Q. Zhao, W. Li, W.J. Braun. Infinite block-structured transition matrices and their properties. *Advances in Applied Probability* 30 (1998), 365–384.

[127] W.H.M. Zijm. Exponential convergence in undiscounted continuous-time Markov decision chains. *Mathematics of Operations Research* 12 (1987), 700–717.

[128] A.P. Zwart, O.J. Boxma. Sojourn time asymptotics in the M/G/1 processor sharing queue. Accepted for publication in *Queueing Systems*. Pre-print: CWI Report PNA-R9802 (1998).

Summary

In this thesis we study *queueing* models which can be used in the performance analysis of *integrated-services* telecommunication networks. Chapter 1 gives an overview of the evolution of these networks and describes the most relevant features. Modern telecommunication systems offer a wide range of services (data, voice, video) which are carried simultaneously in the network on an integrated basis. We can roughly divide the traffic into two broad classes: *stream* traffic and *elastic* traffic. Stream traffic mainly consists of "real-time" connections (such as telephony and interactive video applications) which are extremely sensitive to transmission delays. Stream connections therefore require a certain guaranteed capacity. Elastic traffic (data transmission, e-mail) on the other hand allows for fluctuations in the transmission rate, as long as the total delay is "acceptable". The transmission capacity available to elastic traffic varies as stream traffic connections are set up or terminated. Each elastic traffic connection gets an equal share of the capacity left over by stream traffic. In the thesis we focus on the performance analysis of elastic traffic, using so-called processor-sharing models with varying service capacity. An elastic traffic connection is represented by a customer in a queueing model. Hence, the service requirement of a customer in the model corresponds to the size of, for instance, a data file. The service capacity, which fluctuates according to some stochastic process, is shared among the customers in the queue according to the processor-sharing discipline, i.e., each customer gets an equal share. Processor-sharing models with *constant* service capacity are well-studied in the literature. Fluctuations in the service capacity, however, turn out to make the analysis considerably more complicated. This thesis presents the first analytic results concerning *sojourn* times in processor-sharing queues with varying service capacity (which correspond to the transmission times of elastic services).

In Chapter 2 we first derive the queue-length distribution in an M/M/1 (processor-sharing) queue of which the service capacity (and arrival intensity) varies depending on the state of a birth-death process. The queue-length distribution is obtained by combining the theory of matrix-geometric solutions with the method of spectral expansion. The theory of matrix-geometric solutions enables a transparent analysis using probabilistic arguments, while the spectral expansion allows for a more detailed analysis. We also show how the alternative method of generating functions can be applied, and we discuss the intimate relation between the three approaches. Special attention is devoted to the in-

fluence of the (capacity) fluctuations when these occur either very fast or very slow (relative to the service times of customers). We show that approximating the system by one with constant service capacity, equal to the average service capacity in the model with fluctuations, is only justified when the fluctuations occur very fast (so that they average out). The formal analysis is illustrated by numerical experiments for a specific telecommunication system.

In the remainder of the thesis we concentrate on the sojourn times of customers, in particular conditional on the service requirement. In Chapter 3 we study a processor-sharing model of which the service capacity is constant during so-called *on-periods* and no service is rendered during *off-periods*. We again assume that the service requirements have an exponential distribution. The sojourn time distribution is given in terms of its LST (Laplace-Stieltjes Transform). The analysis is based on a random time-scale transformation, via which sojourn times in the original model are represented by transient rewards in a branching process with a specific reward structure. We further show that the decomposition of the sojourn time into *independent* components, which is known for processor-sharing models with constant service capacity, also applies to the on/off model. Another well-known property of standard processor-sharing models is that the expected conditional sojourn time is a linear function of the service requirement. In the on/off model it turns out that this is only true asymptotically, that is, for large service requirements.

In Chapter 4 we study sojourn times in the case that the service capacity depends on the state of a general Markov process. In contrast to the on/off model, service can be rendered at *different* positive rates. This generalisation prohibits an analysis as detailed as the one presented for the on/off model. In particular, the above mentioned decomposition of sojourn times no longer applies. However, the asymptotic linearity of the expected conditional sojourn time as a function of the service requirement is preserved. This is shown using the LST of the conditional sojourn time, which is again derived using the method of time-scale transformation. We also discuss *why* the above mentioned linearity is lost when the service capacity fluctuates. The results of the analysis are then used in numerical experiments for the performance evaluation of a communication system. The analytic and numerical results lead to a good and simple approximation of the expected conditional sojourn time. The analysis can be extended to the case that the service requirements have a phase-type distribution. Furthermore, the analysis also applies to the more general service discipline *discriminatory* processor sharing. Both generalisations, however, are at the expense of a higher computational complexity.

In Chapter 5 we study the tail of the sojourn time distribution in the case that the service requirement distribution has a so-called *heavy tail*. It is well-known that when the latter is the case and customers are served in the order of arrival (the so-called *First Come First Served* discipline), then the tail of the sojourn time distribution is "one degree" heavier: it is as heavy as the *integrated* tail of the service requirement distribution. As a consequence, the mean sojourn time is infinite when the variance of the service requirements is infinite. It is also known that with the processor-sharing discipline (and constant service capacity) the tails of the sojourn time and the service requirement distributions are exactly as heavy. This is generally seen as a desirable property. We generalise this result to the on/off model assuming a heavy-tailed service requirement distribution (this was not the case in Chapter 3). We do so by generalising the decomposition property of the sojourn times in the on/off model to the case of generally distributed service requirements. The approach also leads to a new and simpler proof of the result in the standard processor-sharing model. Furthermore, we establish the "tail equivalence" of the sojourn time and service requirement distributions for two other disciplines: *foreground-background processor sharing* (only the customers that have received the least amount of service are served in processor-sharing fashion), and *shortest remaining processing time first* (in which the customers with the smallest remaining service requirement are served).

Samenvatting

In dit proefschrift worden *wachtrijmodellen* bestudeerd, die gebruikt kunnen worden in de prestatie-analyse van telecommunicatiesystemen met *geïntegreerde* diensten. Hoofdstuk 1 geeft achtergrondinformatie over de historische ontwikkeling van telecommunicatiesystemen met geïntegreerde diensten en beschrijft de meest relevante eigenschappen. Moderne communicatiesystemen bieden de mogelijkheid om simultaan zeer verschillende typen verkeer (data, geluid, video) in geïntegreerde vorm over hetzelfde netwerk te versturen. We kunnen de verschillende soorten verkeer ruwweg indelen in twee klassen: *stroom* verkeer en *elastisch* verkeer. Stroom verkeer bestaat voornamelijk uit "real-time" verbindingen (o.a. telefonie en interactieve video-applicaties) die nauwelijks vertragingen in de transmissie tolereren; derhalve is voor die diensten een zekere capaciteitsgarantie vereist. Elastisch verkeer (o.a. datatransmissie, e-mail) daarentegen laat fluctuaties in de transmissiesnelheid toe, zolang de *totale transmissieduur* "acceptabel" is. Als gevolg van het opzetten en afbreken van connecties van stroom verkeer, varieert de transmissiecapaciteit die beschikbaar is voor elastisch verkeer. Elke connectie van elastisch verkeer deelt in gelijke mate in de capaciteit die overgelaten wordt door het stroom verkeer. In het proefschrift gaat de aandacht uit naar de prestatie-analyse van elastisch verkeer door middel van zogenaamde *processor-sharing* modellen met variërende capaciteit. Een connectie van een elastische dienst wordt gerepresenteerd door een klant in een wachtrijmodel. De bedieningsvraag van de klant in het wachtrijmodel correspondeert dus bijvoorbeeld met de omvang van een data bestand in het oorspronkelijke communicatiesysteem. De bedieningscapaciteit, die fluctueert volgens een stochastisch proces, wordt op elk moment gelijk verdeeld ("processor sharing") onder de aanwezige klanten. Processor-sharing modellen met *constante* bedieningscapaciteit zijn reeds uitvoerig bestudeerd. De fluctuerende capaciteit blijkt echter een belangrijke complicerende factor te zijn in de analyse. In dit proefschrift worden voor het eerst analytische resultaten verkregen voor de verdeling van *verblijftijden* van klanten in zo'n systeem.

In Hoofdstuk 2 wordt allereerst de verdeling van het aantal klanten in een M/M/1 (processor-sharing) wachtrij bepaald, waarbij de bedieningscapaciteit (*en* de aankomstintensiteit) varieert volgens een geboorte-sterfte proces. De rijlengte verdeling wordt verkregen door middel van resultaten van de theorie van matrix-geometrische oplossingen in combinatie met de techniek van spectrale ontwikkeling. De theorie van matrix-geometrische oplossingen maakt de afleiding inzichtelijk door het gebruik van probabilistische argumenten, terwijl de spectrale ontwikkeling een gedetailleerdere analyse mogelijk maakt. Tevens wordt aangegeven hoe, als alternatief, de methode van genererende functies gebruikt kan worden en wordt de relatie tussen de drie verschillende methoden besproken. Speciale aandacht wordt geschonken aan het effect van de fluctuaties wanneer deze zeer snel of juist zeer traag plaats vinden (ten opzichte van de verblijftijd van klanten). Aangetoond wordt dat de benadering door middel van een systeem met *constante* bedieningscapaciteit en aankomstintensiteit gelijk aan de overeenkomstige *gemiddelde* waarden in het model met fluctuaties, alleen gerechtvaardigd is wanneer de fluctuaties zeer snel plaats vinden (waardoor uitmiddeling optreedt). De formele analyse wordt geïllustreerd met behulp van numerieke experimenten voor een specifiek telecommunicatiesysteem.

In de rest van het proefschrift concentreren we ons op de verblijftijd van klanten (dit correspondeert met de transmissieduur van elastische diensten), in het bijzonder geconditioneerd op de bedieningsvraag. In Hoofdstuk 3 bestuderen we een processor-sharing model waarbij de bedieningscapaciteit constant is gedurende zogenaamde *aan-periodes* van de bediende en er geen bediening is gedurende *uit-periodes*. We nemen wederom aan dat de bedieningsvraag exponentieel verdeeld is. De kansverdeling van de conditionele verblijftijd wordt gegeven in termen van de LST (Laplace-Stieltjes Transformatie). Hiervoor wordt, door middel van een (stochastische) tijdschaal-transformatie, het verblijftijden-probleem geformuleerd in termen van een vertakkingsproces met een specifieke opbrengsten-structuur. We tonen verder aan dat de — voor processor-sharing modellen met constante capaciteit — bekende decompositie van de verblijftijd in *onafhankelijke* componenten, behouden blijft in het aan/uit-model. Een andere eigenschap van standaard processor-sharing modellen (met constante capaciteit) is dat de verwachte conditionele verblijftijd een lineaire functie is van de bedieningsvraag. Voor het aan/uit-model blijkt deze eigenschap echter alleen asymptotisch (voor grote bedieningsvraag) te gelden.

In Hoofdstuk 4 bestuderen we de verblijftijden in een model waarbij de bedieningscapaciteit afhangt van de toestand van een algemeen Markov proces. Anders dan in het aan/uit-model kan de bedieningscapaciteit *verschillende* positieve waarden aannemen. Deze generalisatie staat een gedetailleerde analyse zoals in het aan/uit-model in de weg. In het bijzonder blijkt de bovengenoemde decompositie van de verblijftijd in onafhankelijke componenten niet langer te gelden. De asymptotische lineariteit van de verwachte conditionele verblijftijd blijft echter wel behouden. Dit wordt aangetoond met behulp van de LST van de conditionele verblijftijd-verdeling, die wederom gevonden wordt door middel van een tijdschaal-transformatie. Ook wordt verklaard *waarom* de lineariteit verstoord wordt, wanneer de bedieningscapaciteit fluctueert. Door middel van numerieke experimenten worden de verkregen resultaten toegepast in de prestatie-analyse van een specifiek communicatiesysteem met stroom en elastisch verkeer. De analytische en numerieke resultaten leiden tot een goede en eenvoudige benadering van de verwachte conditionele verblijftijd. De analyse kan gegeneraliseerd worden naar het geval dat de bedieningsvraag een *fase-type* verdeling heeft. Ook geldt de analyse voor hetzelfde model met de algemenere bedieningsdiscipline *discriminatory* processor-sharing. Beide generalisaties brengen echter een hogere numerieke complexiteit met zich mee.

In Hoofdstuk 5 bestuderen we de staart van de verblijftijd-verdeling in het geval dat de bedieningsvraag-verdeling een zogenaamde *zware staart* heeft. Het is bekend dat als dit laatste het geval is en klanten in volgorde van aankomst bediend worden (de zogenaamde *First Come First Served* discipline), dan is de staart van de verblijftijd-verdeling "één graad" zwaarder dan die van de bedieningsvraag-verdeling (namelijk even zwaar als de geïntegreerde staart van de laatst genoemde). Hierdoor is bijvoorbeeld de verwachte verblijftijd oneindig wanneer de variantie van de bedieningsvraag oneindig is. Ook is bekend dat onder de processor-sharing discipline geldt dat de staarten *precies even zwaar* zijn, wat in het algemeen gezien wordt als een wenselijke eigenschap. Dit laatste resultaat generaliseren we voor het aan/uit-model waarbij de bedieningsvraag-verdeling een zware staart heeft (in Hoofdstuk 3 was dat niet het geval). Hiervoor generaliseren we onder meer de decompositie eigenschap van de verblijftijd voor het geval dat de bedieningsvraag in het aan/uit-model een algemene verdeling heeft. De gekozen aanpak leidt tevens tot een eenvoudiger bewijs van het reeds bekende resultaat in het gewone processor-sharing model (met constante capaciteit). Met behulp van dezelfde bewijstechniek wordt de eigenschap ook bewezen voor twee andere bedieningsdisciplines: *foreground-background processor sharing* (waarbij de klanten met de minste reeds verkregen bediening volgens processor sharing worden bediend) en *shortest remaining processing time first* (waarbij de klanten met de minste resterende hoeveelheid werk eerst worden bediend).

About the author/Over de auteur

Sindo (Rudesindo) Núñez Queija was born in Heemskerk (The Netherlands) on May 10, 1972. He graduated from Grammar School (Augustinus College, Beverwijk) on June 13, 1990. For two years he was a student assistant in Operations Research at the Econometrics Department of the Vrije Universiteit in Amsterdam. He also participated in the European project *Human Capital and Mobility* at the Universitat Politècnica de Catalunya (Barcelona, Spain). He received his master's degree in econometrics (cum laude) from the Vrije Universiteit (Amsterdam) on June 29, 1995, after which he became research assistant at CWI (Centre for Mathematics and Computer Science, Amsterdam). Since August 1, 1999, he is a post-doc at CWI. He defends this PhD thesis at the Technische Universiteit Eindhoven on January 20, 2000.

Sindo (voluit: Rudesindo) Núñez Queija werd geboren op 10 mei 1972 in Heemskerk. Op 13 juni 1990 behaalde hij het VWO diploma aan het Augustinus College te Beverwijk, waarna hij econometrie ging studeren aan de Vrije Universiteit van Amsterdam. Daar was hij twee jaar lang student-assistent in de Operations Research (vakgroep Econometrie). Verder werkte hij een half jaar in het Europese project *Human Capital and Mobility* aan de Universitat Politècnica de Catalunya (Barcelona, Spanje). Op 29 juni 1995 behaalde hij het doctoraal diploma Econometrie (cum laude). Hierna werd hij onderzoeker-in-opleiding aan het Centrum voor Wiskunde en Informatica in Amsterdam. Sinds 1 augustus 1999 is hij daar werkzaam als post-doc. Op 20 januari 2000 verdedigt hij dit proefschrift aan de Technische Universiteit Eindhoven.