Queues with state-dependent rates

© Bekker, R.
Queues with state-dependent rates / by René Bekker

Eindhoven : Technische Universiteit Eindhoven, 2005
Proefschrift. ISBN 90-386-0734-2
NUR 919
Subject headings: queueing theory, production systems, communication systems
2000 Mathematics Subject Classification: 60K25, 90B22, 60K30, 90B05, 90B18

Cover design by Loekie Karstens
Printed by Dékavé, Alkmaar

Queues with state-dependent rates

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
Rector Magnificus, prof.dr.ir. C.J. van Duijn, voor een
commissie aangewezen door het College voor
Promoties in het openbaar te verdedigen
op maandag 12 december 2005 om 16.00 uur

door

René Bekker

geboren te Leiderdorp

Dit proefschrift is goedgekeurd door de promotoren:

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

Dankwoord (Acknowledgments)

Dit proefschrift markeert het eindpunt van een fijne en bijzondere periode in mijn leven. Ik maak dan ook met alle plezier gebruik van de mogelijkheid om enkele mensen te bedanken die een belangrijke bijdrage hebben geleverd aan de totstandkoming hiervan.

Allereerst wil ik graag mijn promotoren Onno Boxma en Sem Borst van harte bedanken. Zij hebben mij niet alleen op een voortreffelijke manier weten te begeleiden, maar hebben me eveneens altijd enthousiast weten te maken voor het doen van wetenschappelijk onderzoek. Ik realiseer me nu hoe belangrijk het is om met veel plezier naar je "werk" te gaan. Daarnaast bezit ik heel fijne herinneringen aan onze persoonlijke gesprekken en contacten.

Mijn eerste stappen op wetenschappelijk terrein heb ik pas gezet nadat Rein Nobel en professor Tijms mij wezen op een vrije AiO-plaats bij een zekere professor Boxma. Ik wil hen dan ook graag bedanken voor het openen van deze deur. Tijdens mijn AiO-tijd heb ik met veel plezier samengewerkt met Sindo Núñez Queija en Bert Zwart. De hoofdstukken 7 en 4, respectievelijk, berusten op gezamenlijk werk. Verder kon ik altijd bij ze aankloppen, zowel voor wiskundige vragen als voor persoonlijke adviezen. Offer Kella ben ik dankbaar voor de aangename samenwerking aan hoofdstuk 2. Professor J.W.M. Bertrand en professor Ton de Kok wil ik bedanken voor het stellen van vragen die hebben geleid tot het onderzoek in de hoofdstukken 5 en 6. Philips Research ben ik erkentelijk voor de financiële ondersteuning van mijn promotieproject. Daarnaast ben ik Dee Denteneer van Philips Research dank verschuldigd voor diverse nuttige discussies en plezierige contacten. Michel Mandjes bedank ik graag voor nuttige opmerkingen op een eerdere versie van dit proefschrift, alsmede voor de goede gesprekken op weg naar en tijdens het verblijf in Ottawa.

Van alle kamergenoten wil ik Ton Dieker in het bijzonder bedanken. Hij is de enige met wie ik vrijwel de gehele periode een kamer heb mogen delen. Ik ben blij met zijn plezierige gezelschap, onze wachtrijtheoretische discussies en de wandelingen naar de snoepautomaat. Dennis van Ooteghem bedank ik graag voor de prettige samenwerking tijdens de LNMB en Beta cursussen. Uiteraard wil ik ook graag iedereen van PNA 2 (CWI) en B&S (TU/e) bedanken voor de uitermate goede sfeer en gezellige lunches.

Tenslotte wil ik graag het woord richten tot de mensen die op persoonlijk vlak van groot belang zijn geweest, te beginnen met mijn vader. Koos, ik ben je zeer dankbaar voor de steun die jij me hebt gegeven om dit te kunnen bereiken.

Tijdens mijn tweedaagse verblijf in Eindhoven heb ik altijd met veel plezier in Nuenen overnacht. Jos en Francien, bedankt voor de enorme gastvrijheid de afgelopen vier jaar; Michelle en Ronald, bedankt voor het altijd klaar staan als zij er niet waren. De Loek-en, jullie wil ik graag bedanken voor de creativiteit bij het ontwerp van de omslag en de uitvoering van het proefschrift. Ik ben onder de indruk van het mooie resultaat. Jacob en Judith, bedankt voor de hulp met de Stellingen en het gevoel om thuis te zijn. Saar, als laatste wil ik jou bedanken voor je steun, motivatie en onvoorwaardelijke liefde. Jouw aanwezigheid heeft deze periode tot een fijne tijd gemaakt en is steeds van essentieel belang geweest.

Contents

Chapter 1

Introduction

In many sectors of today's society there is an increasing demand for resources that are scarce in, for instance, an economical sense. Such scarcity of resources often gives rise to congestion phenomena. The study of congestion and the optimization of resource performance form the realm of queueing theory and performance analysis. A prominent example is the tremendous growth of the Internet where huge capacity improvements are required to keep up with the rapidly growing demand.

In the study of queueing systems, the service rate is generally assumed to be constant. There are however various application areas where this assumption may not be valid, e.g., water dams, communication networks and production systems. In addition to the service rate, the arrival rate of new demands may also be affected by the level of congestion. Queueing models with such congestion-dependent rates form the main subject of this monograph.

The organization of this introductory chapter is as follows: In Section 1.1, we describe the classical single-server queue. Various examples of queues with congestion-dependent rates are highlighted in Section 1.2. We discuss three examples providing the main motivation for this thesis in more detail in Sections 1.3-1.5; production systems are considered in Section 1.3, we address the mathematical study of water dams in Section 1.4, and applications to communication networks can be found in Section 1.5. Moreover, a considerable amount of literature on queueing models with congestion-dependent rates, including dams, fluid queues, and fluid networks, is presented in Section 1.3. In Section 1.6, we give a flavor of the techniques used in this monograph by applying them to the single-server M/G/1 queue. In addition, we rederive several known results which may serve as reference framework for results obtained later in this thesis. We conclude with an outline of the thesis in Section 1.7.

1.1 Classical queueing models

The discipline of queueing (or congestion) theory originates from the study of teletraffic issues in the beginning of the 20th century. At that time, A.K. Erlang initiated a systematic study of the dimensioning of telephone switches using principles of queues and waiting lines. Ever since, communication engineering and queueing theory have heavily influenced each others development. Motivated by the development of computer systems, the area of queueing networks flourished during the 1960's and 1970's, see for instance the famous papers of Baskett *et al.* [18] and Kelly [103]. In telecommunications, the huge expansion of the Internet and the introduction of the Transmission Control Protocol (TCP) to regulate the transmission of traffic flows gave rise to a broad range of new queueing models.

Besides the applications in computer and telecommunication systems, the theory of queues may also be applied in many other situations, like in production systems and road traffic. For the evolution of queueing theory until the 1980's, we refer to [57]. An extensive introduction on queueing theory may be found in, e.g., Asmussen [10], Cohen [56], Kleinrock [107], and Tijms [163].

In general, congestion occurs when there is a demand for a scarce resource. The demands are typically generated by customers arriving at the service station, where the capacity of the server represents the limited resource. A classical and intuitively appealing example is the waiting line at the post office or supermarket, where customers arrive and (possibly) wait in line until they receive their service. Other examples are jobs in a production system, or data packets in communication networks (in which case the server represents the communication link). Usually, a system is designed such that all arriving customers can eventually be served. Congestion typically arises due to "temporary overload" caused by the randomness in the arrival process and in the service requests of customers.

1.1.1 The single-server queue

The classical and most elementary queueing model is the single-server queue where the server works at constant speed. In such a model, customers arrive at the service station according to a stochastic process: A renewal process. This means that customers arrive one at a time and the times between two consecutive arrivals, referred to as *interarrival times*, are i.i.d. (independent identically distributed). The *service requirements* (or requested service times if customers are served at unit speed) also form a sequence of i.i.d. random variables, which are also independent of the sequence of interarrival times. Customers that have received their full service request leave the system.

The above model is often denoted as the G/G/1 queue, sometimes followed by the service discipline (default is FCFS, see below). This notational convention is due to Kendall [104]. The first G reflects that the interarrival times are generally distributed, while the second G refers to the general distribution of the service requirements. The notation GI is sometimes used if the independence assumption needs to be stressed. Commonly used distributions are the exponential (denoted with M, the acronym for memoryless) and deterministic (denoted with D). The 1 in Kendall's notation indicates that there is a single server.

There are many variations of this basic queueing model. One main stream of variations includes queues with finite capacity, or restricted accessibility. The finite capacity may be due to the fact that there is only a finite number N of waiting places (the G/G/1/$N+1$ queue in Kendall's notation, where the latter 1 stems from the position at the server). When all places are occupied arriving customers are blocked and typically assumed to be lost. The workload-equivalent of this system is the finite dam: The total amount of work is bounded by a finite number K and excess work is lost. The finite capacity may also be interpreted as a bounded waiting time. Customers that are waiting too long may get impatient and decide to leave. This phenomenon is usually called customer impatience or reneging. For details and other systems with restricted accessibility, we refer to Chapter 3.

To describe the number of customers in the queue and the waiting times, we also need to specify the *service discipline*. At a post office, the First-Come First-Served (FCFS) discipline where customers are served in the order of arrival, is the most natural one. However, in other systems different service disciplines may be more appropriate, such as Last-Come First-Served (with or without preemption) and the Processor Sharing (PS) discipline. PS is an idealized version of round-robin; the service capacity is equally divided among all customers present, who all simultaneously receive service.

1.1.2 Performance measures

The performance of queueing systems may be expressed in terms of one or more *performance measures*. The performance measures of interest strongly depend on system-specific objectives and model-specific assumptions. The most commonly used are *queue lengths*, *waiting times*, *sojourn times*, and *workloads*. The waiting time is the time spent by customers in the queue, while the sojourn time represents the total time in the system (including the service time). Other possible performance measures are the duration of a busy period or the maximum workload during such a period. In case of finite buffers, the fraction of customers lost or lost amount of work are further natural performance measures.

The main performance measure considered in this thesis is the *workload* or *virtual waiting time*, which is defined as the sum of the remaining service requirements or the unfinished amount of work present. The term virtual waiting time stems from the observation that the workload at time t equals the waiting time under the FCFS discipline of a fictitious customer arriving at time t assuming that the server works at unit speed. A useful property is that the workload process is independent of the service discipline provided that the discipline is work-conserving. The service discipline is work-conserving when the total service capacity is used as long as there is any work present and work is neither created nor destroyed during the service process.

A typical sample path of the workload process in the ordinary single-server queue is depicted in Figure 1.1. The jumps correspond to customer arrivals, where the size of the jump represents the service requirement. The relation between the waiting time in the FCFS single-server queue and the workload is

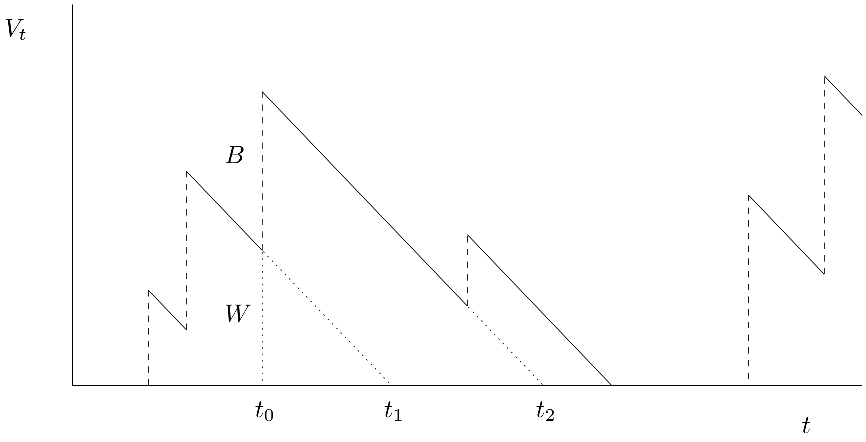

Figure 1.1: A typical sample path of the workload process $\{V_t, t \geq 0\}$ in the single-server queue.

now immediate: The customer arriving at time t_0 is taken into service at time t_1 and leaves the system at time t_2. Since the server works at unit speed, it holds that $W = t_1 - t_0$ and $W + B = t_2 - t_0$. Hence, the waiting time just equals the workload embedded at epochs just before arrival instants and the sojourn time is given by the workload right after arrival epochs.

1.2 Queues with workload-dependent rates

In the queueing model of Section 1.1 the server is usually assumed to work at constant speed as long as there is any work present. However, this assumption may not always be appropriate as the state of the system may affect the server productivity in some practical situations. Below, we give several examples where such behavior may occur. We specifically focus on queueing situations where the service speed depends on the workload.

In addition, the arrival rate of new customers may also be influenced by the amount of work present. In traditional queueing theory, finite buffers partly capture this issue. However, in queueing systems with finite buffers, the arrival rate immediately reduces to zero when the workload exceeds a certain threshold, which does not allow for a smoother change of arrival rates. This limitation is especially pertinent in the field of communication systems where the transmission rate is gradually adapted based on the buffer content.

Below, we sketch some practical scenarios leading to models with state-dependent arrival or service rates. Examples 1.2.1-1.2.3 constitute the main motivation for this thesis and are addressed in more detail in Sections 1.3-1.5, respectively.

Example 1.2.1 In systems where the server represents a human being, the perception of the workload may directly influence the human's productivity. For instance, in the post office example, it is plausible that the server gets lazy when there is hardly any work, but gets stressed when the office is crowded. A more detailed discussion of this effect is given in Section 1.3, where we particularly focus on production systems in which the shop floor personnel is being affected by the level of work-in-process.

Example 1.2.2 Dams form a historically important area of systems with state-dependent rates. In dams, inflowing water, caused by large rainfalls, is temporarily stored and released according to a content-dependent release rule. Since the 1950's a rich body of literature has emerged on the mathematical analysis of dams and storage systems, see Section 1.4 for an overview. In that respect, this monograph builds on the mathematical foundation of dam studies.

Example 1.2.3 Due to the immense growth of the Internet, a considerable amount of queueing literature has focused on congestion control mechanisms to regulate transmission rates in packet-switched communication networks. These studies are very diverse, focusing on different applications and various time scales. An important characteristic of best-effort applications in packet-switched systems is that the transmission rates are dynamically adapted based on implicit information about the buffer content. A particular example is TCP which is the dominant protocol used to regulate the transmission rate of Internet flows. We refer to Section 1.5 for a more elaborate discussion.

Example 1.2.4 The study of road traffic has also been very much inspired by queueing theory. Queueing at traffic lights is probably the most famous example, but also the study of traffic flows on the highway has received some attention. On the highway, each driver has its own preferred speed when the driver would be alone, which may be characterized by the free-speed distribution. However, the driving speed may be influenced by interactions with other cars. Let a flow ("server") be measured as the passed numbers of cars per hour and the concentration ("workload") as the number of cars per mile (or kilometer). Typically, a flow will first increase as the concentration of cars increases up to some optimal level. After this optimum, traffic jams arise and the flow will gradually (but usually quite fast) slow down. For details, we refer to [171].

Example 1.2.5 A phenomenon often arising in physics is a *shot noise* process. A typical example is the analysis of the shot effect in vacuum tubes, see for instance [144] and [76], p. 178. Suppose that shocks occur according to a Poisson process. The value of the ith shock is being represented by a random variable X_i. The values of the shocks are assumed to be additive and to decrease with an exponential rate α over time. Then, the total shock value at time t is given by, see [149, Subsection 8.7],

$$X(t) = \sum_{i=1}^{N(t)} X_i e^{-\alpha(t-T_i)},$$

where $N(t)$ is the number of shocks during $[0, t]$ and T_i is the time of occurrence of the ith shock. In fact, the shock-value process is identical to the workload process in a queue with service speed αx when the workload equals x. See for instance [76, 95, 144, 149] for further details of shot noise models.

In Subsection 1.1.2, we observed that the workload just before an arrival epoch constitutes the waiting time in the corresponding FCFS queue. As illustrated in Figure 1.1, the argument crucially relies on the independence between workload and service speed. In queues with workload-dependent service rates, the waiting time also depends on customers arriving at later instants as they influence the current service speed. Hence, the nice relation between waiting times and workloads in ordinary single-server queues is lost.

Remark 1.2.1 Above we have focused on queueing situations in which the arrival and service rate depend on the workload. Another important class of queueing models with state-dependent rates is the class of general birth-and-death processes. In a birth-and-death process the state-space is assumed to be discrete (or denumerable), typically consisting of the set of non-negative integers. In addition, the only transitions possible are the transitions to neighbouring states. In queueing theory, birth-and-death processes are frequently used to model the number of customers in various M/M/1-type models. Then, the arrival (birth) rate is λ_n and the service (death) rate is μ_n when the number of customers in the system equals n. Theory on birth-and-death processes can be found in many textbooks on applied probability, see for instance [10, 56, 107]. $\diamond$

Remark 1.2.2 A different class of models with workload-dependent rates is formed by diffusion processes on $[0, \infty)$ with drift parameter $\mu(x)$ and variance parameter $\sigma^2(x)$ when the state equals x. Such diffusions are studied in the literature on risk processes; see e.g. [9], p. 205 and 303, for expressions of ruin probabilities. However, due to the equivalence between ruin probabilities and workload distributions, this also yields the steady-state workload distribution in a queue driven by a diffusion with infinitesimal drift and variance parameters $-\mu(x)$ and $\sigma^2(x)$, respectively. We refer to [9] and [92], p. 191–195, for details. Note that the case of constant parameters just reduces to a reflected Brownian motion. $\diamond$

1.3 Production systems

In classical queueing systems, the server is assumed to work at a constant speed whenever there is any work in the system. As we briefly discussed in Example 1.2.1, this supposition may not be valid when the server is a human being. The perception of the unfinished amount of work has a strong impact on the information processing functions of the human brain and hence on the human-server performance.

In [172, Chapter 12], the authors describe the psychological effects of *arousal* on human performance. Arousal is caused by *stressors*, such as sleep loss, anxiety, or incentives, but may also involve phenomena as time pressure. On the one hand, an increased level of arousal is typically associated with the psychological effect of “trying harder” (see [172]) or “increase the efforts”. On the other hand, stress effects also have major drawbacks in the area of information processing. One of the most important aspects is *attentional narrowing*, or *tunneling*; under high stress a human provides all attention to tasks of (its subjective) highest priority. This may have some undesirable side-effects, especially when the task is complex and involves many channels of information. A second consequence of stress is the loss in working-memory, which directly affects the human performance.

The pattern of arousal and its impact on human performance is characterized in psychology by the Yerkes Dodson law, see Figure 1.2. The figure suggests that at low levels of stress an increased level of arousal mediates productivity, while at high levels problems arise with attentional narrowing and working-memory loss.

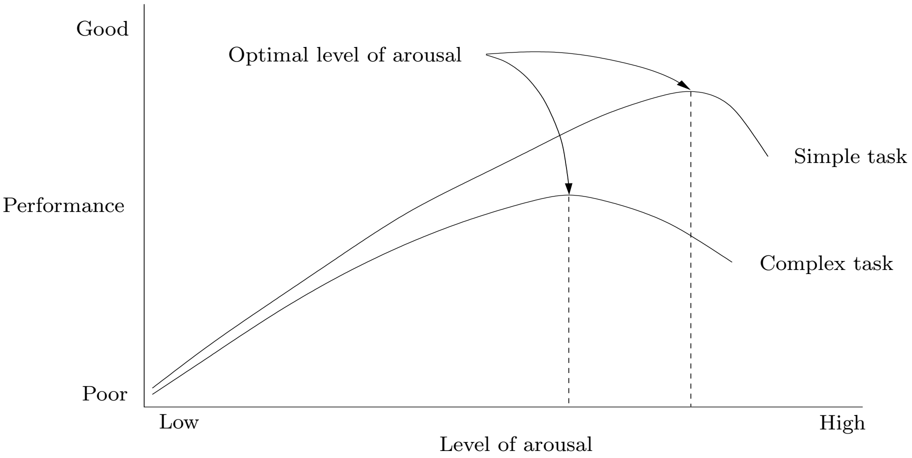

Figure 1.2: Yerkes Dodson law (taken from [172]); the relation between the level of arousal and human performance.

In production environments there is also evidence that productivity and the amount of work-in-process are related. In job shops, work-in-process is measured as the number of work orders on the shop floor. The studies [30, 154] and references therein indicate that decreasing the production lead time raises the output (performance) per employee. Here, production lead times correspond to the differences between order arrival times and their release. Because high lead times are common in job shops and go along with high levels of work-in-process (or workloads), these studies support the psychological arguments described above: At the right side of the curve the performance increases as the workload and level of arousal decrease. We refer to [30] for details.

The manager of a job shop can usually not directly influence the productivity of the shop floor personnel. However, a manager naturally strives for a high utilization of resources, or in other words, an efficient use of personnel. The arrival of jobs typically occurs on a much faster time scale than changes in the service capacity, i.e., the number of employees that can be utilized. As a consequence, over short to intermediate time periods, a shop manager can only influence the workload by controlling the arrival rate of new jobs. A way to accomplish this is, for instance, by rejecting new orders or by a different price setting or changing the terms of customer contracts. The effect of arrival control on job shops has been investigated in [169]. We refer to Chapter 5 for arrival rate control in queueing systems with such workload-dependent service rates.

1.4 Dams and storage processes

In this section, we consider the evolution of the mathematical theory of dams and its relation to queueing. Our aim is threefold; we first give a brief historical overview on the development in the study of dams. Second, in Subsection 1.4.2 we provide a formal mathematical description of a dam with a state-dependent release rule. This is of particular interest since the basic model in our thesis shows strong resemblance with such a dam. Third, we give several references on storage models with content-dependent release rates; literature on dams can be found in Subsection 1.4.2, while literature on the related field of fluid queues and fluid networks is reviewed in Subsection 1.4.3.

From a practical perspective the dam is designed for a temporary storage of water, which may have several purposes. In [128], Moran mentions three of them: (i) to provide a head of water for hydroelectric power, (ii) to prevent floods in times of exceptional rainfall, and (iii) to create a buffer of water during the wet season that can be used for irrigation during the dry season. The input of the dam consists of water flowing in from rivers and creeks caused by (large) rainfalls, and typically is of a random nature. The output is simply generated by the release of water from the reservoir.

1.4.1 Dams with constant release

The mathematical study of the operation of dams became popular in the 1950's. In 1954, Moran [126] was the first to formulate a stochastic model for this type of storage. He described a discrete-time model and considered a dam with finite buffer capacity and constant release. More precisely, the model was formulated as follows (see also [10] and [139, Chapter 6]): Let Z_n be the content of the dam at the beginning of year n and denote by X_n the amount of water that flows into the dam during year n. The dam has capacity k and at the end of the year an amount m of water is released (or the amount available if the content is less than m). These dynamics give rise to the following recursion:

$$Z_{n+1} = \max(\min(Z_n + X_n, k) - m, 0).$$

We refer to [139, Chapter 6] for a further discussion.

A natural extension of Moran's model is the dam in continuous time. In a first attempt, Moran [127] considered the discrete-time model with X_n, $n =$

$1, 2, \ldots$, geometrically distributed and then applied an appropriate limiting procedure. The resulting system is a dam with continuous inflow at unit rate and a Poisson process of (deterministic) release. In fact, the residual capacity $k - Z(t)$ may be identified with the workload in an M/D/1 queue with finite capacity k. At the same time, Gani [78] considered a dam with constant release at unit rate and Poisson input (which is equivalent to a finite-buffer M/D/1 queueing system). Both authors were however not aware of those similarities. It was Smith [157] who already pointed out the mathematical equivalence between waiting times in queues and content levels in storage systems in 1953 (he didn't specifically focus on dams though, and considered a storage system with unlimited capacity and with random instead of deterministic release). We also refer to the discussion in [79, 105] of Lindley for a clear comparison between the two systems at embedded epochs.

More generally, the discrete-time model of Moran can be extended to a continuous-time system by considering an input process $\{X(t), t \geq 0\}$ with stationary and independent increments. Additionally, we require that $X(t)$ is non-decreasing. In case the jump rate is finite, the input of the dam simply reduces to a compound Poisson process and the dam is identical to an M/G/1 queueing system (see the discussion above). Other studies considered the case of an infinite jump rate. In that case, infinitely many jumps occur in any finite interval with probability one. In the 1950's, the main focus was on the so-called *gamma input*, see [79, 105] for details. We refer to [139, Chapter 7] for an overview of dams in continuous time.

1.4.2 Dams with state-dependent release

Before we present a brief historical overview on dams with a state-dependent release rule, we first describe such a storage system in more detail. The water content $Z(t)$ increases with jumps that have a generally distributed size. The time intervals between jumps are exponentially distributed, so that $Z(t)$ constitutes a Markov process. In between jumps the content of the dam drains at a deterministic rate $r(x)$ when the content equals x. Thus, if the content at time 0 is w and no jumps occur in $(0, t)$, the content process during the interval $(0, t)$ behaves as a deterministic process: $Z(s) = w - \int_0^s r(Z(u))\mathrm{d}u$. An important quantity is

$$R(x) := \int_0^x \frac{1}{r(y)}\mathrm{d}y,$$

representing the time required to drain the dam in the absence of any jumps. In particular, a necessary and sufficient condition for a return to zero to be possible is $R(x) < \infty$, see e.g. [10, 45, 80, 83] and later chapters of this thesis. Note that $r(x) = 1$ yields the classical M/G/1 queue or dam. We also refer to [10, Chapter 14] for an overview of dams with state-dependent release rates.

Dams where the release of water depends on the content in the reservoir started to appear in the literature in the 1960's. For convenience, we continue to refer to these models as dams, although the focus increasingly shifted to storage systems in general. It seems that Gaver and Miller [80] were the first to study storage problems with content-dependent release rates. Using similar arguments as Takács [159], they constructed the Kolmogorov forward equations for several models. One of their models concerns a two-stage release rule; if the content of the dam exceeds a fixed value $R > 0$, then water is released at rate r_2, while water is released at rate r_1 in case $0 < Z(t) < R$. Using a clever idea for the inversion of the product of two Laplace transforms, they described a procedure for computing the content distribution in steady state for generally distributed inputs at Poisson instants. Later, a similar model was considered by Cohen [53]; see [56], p. 557, for an extension to an m-step release rule.

Gaver and Miller [80] were not yet able to solve the Kolmogorov equations for the system with a general release rule. As a special case, they did determine the Laplace-Stieltjes Transform (LST) of the steady-state buffer content in the case $r(x) = x$. This model is also well-known as *shot noise* and had already been studied by, e.g., Keilson and Mermin [95]. Observe that $r(x) = x$ corresponds to systems with proportional release.

In 1969, Moran [129] obtained the dam as a limit of a discrete-time Markov chain under some conditions on the release-rate function. He also showed that sample paths of $Z(t)$ satisfy

$$Z(t) = Z(0) + A(0,t) - \int_0^t r(Z(s))\mathrm{d}s, \tag{1.1}$$

where $A(0,t)$ denotes the amount of input during $(0,t)$. This equation is well-known as the *storage equation.* In 1958, Reich [141] obtained a similar equation in case of constant release (in that case $r(x) = 1$ for $x > 0$ and $r(x) = 0$ otherwise).

In 1971, Çinlar and Pinsky [47] studied the storage equation (1.1) and showed that, under some conditions, the sample paths of Z are uniquely defined by Equation (1.1). The conditions involved a finite jump rate and a continuous and non-decreasing release rate function. Moreover, they showed the intuitively appealing result that a limiting distribution of $Z(t)$ exists if $\sup r(x) > \mathbb{E}A(0,1)$, where $\mathbb{E}A(0,1)$ is the mean input rate. Motivated by some studies in the 1950's with *gamma input* (see e.g. [105]), the same authors extended their results to the case of an infinite jump rate [48], where $r(\cdot)$ is assumed to be non-decreasing.

Harrison and Resnick [83] continued the study of the storage equation (1.1), and succeeded in removing some undesirable assumptions on the release rate. They did, however, assume that the jump rate is finite and that the content process has an atom at state zero (see the discussion above for implications on $r(\cdot)$). In [83] the authors gave necessary and sufficient conditions for the existence of a stationary distribution of the dam content (which also is unique). The steady-state density of the buffer content is in terms of an infinite sum of iterates, which reduces to nice analytical expressions in some special cases. The paper of Brockwell *et al.* [45] is much in the same spirit. They extended the results of [47, 48, 83] by considering both finite and infinite jump rates and imposing only very mild conditions on the release rate function, including the case in which the content process does not have an atom at zero.

During the last two decades, storage systems with state-dependent release received some renewed attention. In 1992, Doshi [68] wrote a survey on level crossings including systems with *workload*-dependent release. We refer to Miyazawa [125] for a description, based on so-called rate conservation laws, of the time-dependent behavior of storage problems with state-dependent release rates (and a stationary marked point process as input). Furthermore, a whole body of literature appeared on various model extensions and dualities between storage processes. Without aiming for completion, we just mention [93, 94, 138]. Other interesting related papers are [11, 46]. In [46], Boxma *et al.* derive sufficient conditions for stability in case the input rate also depends on the buffer content. Asmussen and Kella [11] consider a release rate that depends both on the content and some underlying modulating process. They exploit the duality relation between risk and storage processes and specifically focus on an extension of the shot noise model.

1.4.3 Related results: Fluid queues and fluid networks

Closely related to traditional queueing systems are the so-called *fluid queues*. In such models, water (or fluid) gradually enters the system over time rather than in jumps. In this subsection, we first give a brief overview of the literature on fluid models, mainly with content-dependent rates. Second, we give some references on extensions to networks.

Fluid queues

The study of fluid models has mainly been triggered by the applications in packet-switched communication systems. In such systems, traffic is divided into small entities (packets) which are sent over the network, see also Section 1.5. Focusing on somewhat larger time scales, traffic is usually modeled as a continuous flow, thereby neglecting the discrete nature of the relatively small packets. More generally, fluid models may be valuable when a separation of time scales applies. In particular, the fluctuations around a certain drift on a shorter time scale may sometimes be neglected (i.e., approximated as a fluid) on a longer time scale.

Queues with gradual input were already introduced in the paper of Gaver and Miller [80]. However, the study of fluid queues mainly started in the 1970's and 1980's, see for instance [6, 109, 110]. We refer to [112] for a survey and to [43] for a detailed discussion of the history of fluid queues.

In communication systems, sources are generally assumed to transmit packets according to an *On-Off* process. An On-Off source sends at a constant rate when the source is On and at rate zero when the source is Off. Of particular interest is the case when several On-Off sources are multiplexed. A related branch of fluid models assumes that the buffer content varies linearly over time depending on some underlying semi-Markov process. In fact, the case where there is only one state of the background process in which the buffer content decreases has a strong relationship with the ordinary G/G/1 queue: Deleting all parts of the content process with an upward slope and glueing together the remaining parts (i.e., the parts with upward slope correspond to jumps in the workload process) directly provides the workload process in the corresponding G/G/1 queue. For a rigorous proof of this relationship in a general setting we refer to [51, 100].

Elwalid and Mitra [72] seem to be the first to study fluid models with content-dependent drifts. They considered the so-called Markov-modulated case meaning that the drifts are governed by some underlying continuous time Markov process. Moreover, they studied a fluid queue with finite buffer and a (piecewise linear) drift that is constant between certain threshold values of the buffer content, see also Subsection 1.5.2. Note that the drift corresponds to input (output) when the drift is positive (negative). Kella and Stadje [98] also considered the Markov-modulated fluid model with finite capacity, but they allowed a continuous content-dependent drift function. In [41], the authors analyze a fluid queue in which during Off times the buffer increases at piecewise linear rates, depending on some semi-Markov process, and during exponentially distributed On times the buffer decreases with a content-dependent rate. They find a decomposition result and give the complete steady-state content distribution for the case of constant and linear output rates.

In addition, the generator of the background process may depend on the buffer content. Models incorporating such dependencies are often called *feedback fluid queues*, reflecting the fact that feedback information of the buffer state governs the background process, see for instance [2] and [151, 167]. In [120], feedback fluid queues are used to model the access regulation in communication networks, see also Subsection 1.5.2. A feedback fluid queue with N background states, finite buffer, and content-dependent rates is considered in [152]. In [40], the authors study a similar model, however with $N = 2$ and infinite buffer. In addition to an explicit expression for the stationary distribution, which can also be found in [152] for the finite-buffer case, they obtain conditions for the existence of a stationary distribution.

Fluid networks

A natural extension of the single dam, with either pure jump or gradual input, is a system with several reservoirs in tandem. In contrast to Jackson networks where customers traverse the network as discrete entities, in tandem fluid queues the output from station i is directly fed into station $i+1$ (as a fluid). Nearly all studies on fluid networks focus on the case that the server at station i works at a fixed rate r_i as long as buffer i is not empty. Using martingale arguments, the authors of [101] find the LST of the joint workload distribution if there is only compound Poisson input into the first station. In [96], Kella considers several stations in parallel with dependent non-decreasing Lévy input and shows that such a parallel system may be considered as a generalization of the tandem model. The authors of [63] obtain the distribution of the second queue in a two-station tandem network where the first station is fed by (general) Lévy input.

More recently, fluid networks with gradual input have been analyzed as well. Kroese and Scheinhardt [111] analyzed several systems of fluid queues where the underlying Markov process has two states, see also [151]. Kella [97] extended the transform version of their result to Markov-modulated feedforward networks. Other extensions of [111] are [1, 153]. In [1], a tandem fluid network fed by multiple On-Off sources with generally distributed On times is considered. In [153], the LST of the joint buffer contents distribution is found in a two-station tandem queue driven by an On-Off source with a general input process during On times. Moreover, in case of finite buffers, the authors of [153] obtained a proportionality relation similar to the proportionality between single finite and infinite-buffer queues.

Exact results for fluid networks with content-dependent service rates are hardly known. Exceptions are [99, 102], where an extension of the shot noise model is studied. More specifically, in both papers a network of stations is considered where each station has a service rate proportional to the workload in that station (i.e., $r_i(x) = r_i x$). In fact, the authors of [102] first consider the general case in which the proportion of input served in a given time interval is governed by a distribution function. The internal flows are routed according to a substochastic transition matrix. In [102], structural results are given for the case where the external input to each station is completely general, which provides explicit expressions in case the external input is a multivariate non-decreasing Lévy process. These so-called linear stochastic fluid networks are motivated as the limit of a network of infinite-server queues with batch arrivals. In [99], both the external multivariate non-decreasing Lévy input and the transition matrix are modulated by a background process. The authors identify conditions for stability and present a functional equation for the LST of the joint buffer contents distribution.

1.5 Communication networks

As mentioned in Section 1.1, communication systems and queueing theory have evolved in close connection since the beginning of the 20th century. Originally, the study of queues was mainly motivated by applications in telephony. However, due to the immense growth of the Internet, the main research interests in communications have gradually shifted to the analysis of voice (also video-conferencing) and data (e.g., files, stored video) traffic in *packet-switched* networks. In such networks traffic is digitized and divided into small entities (*packets*), which are treated as independent entities as they traverse the network. At their destination, packets are reassembled to recover the original information flow.

The remainder of this section is organized as follows. In Subsection 1.5.1 we describe the mechanics of TCP, which is the dominant protocol for the transfer of packets on the Internet. Traffic models of packet-switched communication systems at various time scales are treated in Subsection 1.5.2. In Subsection 1.5.3 we discuss the integration of different traffic types (streaming and elastic) on a common infrastructure.

1.5.1 TCP

The introduction of the Transmission Control Protocol (TCP) has played a critical role in the huge expansion of the Internet since the 1980's. TCP turned out to be a highly effective protocol for the transfer of packets and continues to be the dominant transport protocol used in the current Internet. In this subsection we give a brief description of TCP, referring to, e.g., [87, 113] for further details.

As stated in [113], the two main functions of TCP are congestion control and error control. In order to guarantee error-free communication, the receiver (destination) sends an acknowledgment (ack) to the source after each (group of) correctly received packet(s). The sender maintains a timer to limit the period of time during which no ack is returned. If the sender (source) receives no ack before a time-out occurs, or the sender receives a duplicate ack indicating that a packet in the sequence is missing, the packet is considered to be lost and the source retransmits the lost packet. Retransmission also occurs when a negative ack is received, indicating that the packet contains errors.

The implementation of TCP's congestion control is accomplished using a *congestion window*. The window specifies the maximum number of outstanding packets, i.e., the number of packets sent by the source without having received an ack. TCP does not directly observe the level of congestion in the network itself, but infers the information from returned ack's. If packets are lost, TCP concludes that the level of congestion is high and reduces the window size (typically by a factor 2). Transmissions without packet losses are interpreted as an indication of a lightly loaded network, and the window size is consequently increased until some maximum window size is reached (the receiver window). In the *Slow Start* phase the window size increases at an exponential rate over time, but this phase is often neglected because it only occurs at the beginning of a transfer or after a time-out. In the *Congestion Avoidance* phase the window size increases linearly at rate $1/RTT$ (where RTT stands for roundtrip time) for every correctly received ack. This is effectively done by increasing the window size W by $1/W$ for every ack'ed packet. In fact, TCP in the Congestion Avoidance phase may be viewed as a special case of the family of Additive-Increase-Multiplicative-Decrease (AIMD) congestion control mechanisms. In AIMD the congestion window increases linearly when no losses are detected and the window size is reduced by a multiplicative factor when a loss event occurs, see e.g. [124] for details.

1.5.2 Packet-, burst-, and flow-level models

In packet-switched networks, it is common to distinguish between three different time scales. At the lowest level, the *packet level*, the primary interest is in individual packets traveling across the network. At the highest level, the *flow* or *connection level*, we abstract from packet-level details and consider all packets from the beginning of the transfer until the end as a single flow. The chief interest on the latter time scale concerns issues of fairness, bandwidth utilization, and performance as perceived by the end-user. During a connection, inactive periods, in which no packets are sent, usually alternate with intervals in which bursts of packets are transmitted. This gives rise to an intermediate time scale, the *burst* level.

Packet-level models
The connection between queueing theory and packet-level dynamics of transmission protocols (such as TCP and AIMD) becomes especially apparent in a series of studies starting with [124]. In those studies the increase in the window size is supposed to be continuous rather than in discrete increments. Thus, if $W(t)$ denotes the window size at time t, then $\mathrm{d}W(t)/\mathrm{d}t = 1/RTT$. Moreover, the authors in [124] use a network-centric loss model meaning that loss events are generated by the network. More specifically, they assume that losses occur according to a Poisson process. Let M denote the maximum window size, which may be due to either a capacity limitation or a peak rate limitation, see [4]. Then, $V(t) := M - W(t)$ corresponds to the workload in an M/D/1 queue with state-dependent service requirements. (To see this, flip the window process along a horizontal axis and interpret the loss events as customer arrivals and the sliding window size as the server working at speed $1/RTT$). We note that a slightly different transformation was proposed in [5] to remove the window dependency of the service requirements: The process $\hat{V}(t) := \ln M - \ln W(t)$ corresponds to an M/D/1 queue with workload-dependent service and arrival rates.

Queueing systems with workload-dependent rates may also be applied to model general Adaptive Window Protocols (AWPs) and loss rates that depend on the window size. A characteristic feature of AWP is that both the increase and decrease profiles of the congestion window are general. Results of Chapter 2 are used in [5] to obtain relationships between the steady-state window sizes of two AWPs with related loss rates and increase and decrease profiles.

Burst- and flow-level models
The study of communication systems at the burst and flow levels leaves out all packet-level details concerning the transmission mechanism, and focuses on somewhat larger time scales. Below, we first address traffic models on the burst level and then briefly discuss some properties of traffic at the flow level.

As mentioned in Subsection 1.4.3, fluid queues have proven to be appropriate to model traffic flows at the burst level. For instance, Markov-modulated queueing systems or fluid queues fed by (multiple) On-Off sources are often used to model bursty traffic that alternates between periods of activity and periods of inactivity. Some examples of fluid queues applied to packet-switched communications networks at the burst level, where the transmission rate is also dynamically adapted based on feedback information, are [71, 73, 120]. In [71, 73], bursty sources with loss priorities are multiplexed. Packets of the lowest priority are marked in the access regulator and are the first to be dropped if certain levels (typically thresholds in terms of the buffer contents) of congestion occur. In [120], feedback information on the buffer content also determines the transmission rate. Each source has a guaranteed minimum transmission rate, raising the need for admission control in case the system is heavily loaded. In case the system is relatively lightly loaded, the active sources send at their peak rate, while the sources share the link capacity in the intermediate region.

When identical TCP-controlled flows compete for bandwidth, each flow receives about an equal share of the bandwidth at the flow level. A common (and idealized) way to model this bandwidth sharing is by means of the PS discipline (see also Subsection 1.1.1) which equally divides the link capacity among all users present. There exists a rich body of literature on PS models, see e.g. [29, 108, 130] and [173, 174] for an overview. A limitation of the PS model is the assumption that TCP always efficiently uses the link capacity and that it reacts instantaneously to changes in the number of flows present.

Quite often, the distributions of On periods and service requirements in the above-described burst- and flow-level models are assumed to be heavy-tailed. This is motivated by extensive measurement studies which showed that file sizes and activity periods in the Internet commonly exhibit extreme variability, see for instance [60]. Because traditional Markovian models with phase-type distributions are unable to capture such features, these findings triggered a renewed interest in queueing models with heavy-tailed characteristics. Since an exact analysis is often intractable, most literature on queueing systems with heavy tails analyzes the asymptotic behavior as the buffer content or the delay gets large. Further asymptotic approximations are in regimes where the number of multiplexed sources grows large or the traffic load converges to the link capacity (heavy traffic).

An advantage of asymptotic approximations is that the results often provide useful qualitative insights. Typically, rare events in queueing systems with heavy tails occur as the consequence of a fairly simple most likely scenario. For instance, the most probable way for the workload in a queue fed by a single On-Off source to build up is due to one exceedingly long On-period (see [89]). In case several On-Off sources are multiplexed, the most likely scenario consists of a minimum dominant set of sources having extremely long On-periods while the other sources show average behavior (see, e.g., [37, 69, 89, 146, 179]). Similar observations apply to the G/G/1 queue, see Subsection 1.6.4. We refer to, e.g., [82, 90, 132, 180] for PS models with heavy-tailed traffic characteristics. Moreover, the fairly simple heuristic arguments usually give guidance for complicated proofs. See for instance [34, 170, 177] for a general framework to convert heuristics into rigorous proofs and more references on queueing systems with heavy tails.

1.5.3 Traffic integration

Based on their traffic characteristics and Quality-of-Service (QoS) requirements, we may roughly divide the traffic flows in packet-switched communication systems into two categories: *Elastic* and *streaming* traffic. The Internet was originally designed to support best-effort elastic applications, such as file transfers. Characteristic of these best-effort applications is the absence of any stringent bandwidth requirements. Instead, the performance of elastic traffic depends on "acceptable" total transmission times and error-free communication. This directly explains the success of TCP, which is especially suitable for best-effort traffic. In contrast, streaming traffic typically involves real-time applications, such as telephony, real-time video, or video conferencing. Streaming applications are extremely sensitive to packet transmission delays and rate variations.

TCP and its various extensions (including AIMD) are typically highly reactive transmission protocols. To avoid the wild oscillations in transmission rates, the User Data Protocol (UDP) could be used as an alternative. Because UDP does not respond to congestion however, this gives rise to unfairness in the competition for bandwidth with TCP-controlled flows. As an intermediate option, *TCP-friendly* rate control protocols were proposed [77, 134, 142]. These protocols estimate the throughput that a long-lived TCP flow would receive to determine the transmission rate for a subsequent period of time. Since TCP-friendly protocols are "fair" to TCP-controlled flows, the latter is especially suitable for integrated systems as considered in, e.g., Chapter 7.

In particular, in Chapter 7 we consider asymptotics in a model with integrated elastic and streaming traffic flows. The elastic flows are TCP-controlled, while the transmission rates of the streaming applications are governed by a TCP-friendly rate control protocol. Under the assumption that the file sizes (i.e., elastic flows) are heavy-tailed we obtain asymptotic results for the workload of the streaming traffic.

In fact, the model of Chapter 7 may also be interpreted as a dam where the service speed is determined by the state of a random environment. More specifically, the service rate is equal to the bandwidth share of a permanent customer in the G/G/1 queue under the PS discipline. In the first part of Chapter 7 we assume that the input of the dam is a fluid flow with fixed rate. In Section 7.8 we allow more general input processes with renewal processes and On-Off sources as special cases.

1.6 Methods and results for the M/G/1 queue

The aim of this section is to give a flavor of the methods used in this thesis by applying (some of) them to the standard M/G/1 queue. In addition, the results may serve as a reference source for results presented in later chapters.

In Subsection 1.6.1, we study the steady-state workload (and waiting-time) distribution using the classical Kolmogorov forward equations and the method of successive substitution. Stochastic recursions are introduced in Subsection 1.6.2 and are used as a starting point for the derivation of the famous Pollaczek-Khinchine (PK) formula. In Subsection 1.6.3 sample-path constructions are applied to find the workload distribution in the ordinary finite dam. Sample-path arguments are also applied in Subsection 1.6.4, however, in a different fashion; we obtain asymptotic results for the workload and cycle maximum in case the service requirements exhibit heavy-tailed characteristics. We also state two results of Asmussen [8] providing asymptotics for the M/G/1 queue with workload-dependent service speed (similar models are studied in Chapters 2–5).

1.6.1 Level crossings and successive substitutions

In this subsection, we apply level crossing arguments and the method of successive substitutions to obtain the well-known steady-state workload distribution in the M/G/1 queue. Similar techniques are at the core of (parts of) Chapters 2 and 3. However, we first introduce some notation that will be used throughout this monograph.

Let λ be the arrival rate of customers. Denote by B a generic service requirement with distribution function $B(\cdot)$, LST $\beta(\cdot)$ and mean β. We assume that the traffic load $\rho := \lambda\beta < 1$. Define V_t as the workload at time t, V as the steady-state amount of work, and let $V(\cdot)$ be the distribution of V, with corresponding density $v(\cdot)$, assuming that it exists.

A classical starting point for Markov-type models (in particular fluid models) is the construction of the Kolmogorov forward equations. The derivation below is similar to [159] and [56], p. 263. First, note that in the interval $(t, t+\Delta t)$ a new customer may arrive with probability $\lambda\Delta t + o(\Delta t)$. Then, for $x, t > 0$, and some finite constant $\theta \in (0, 1]$,

$$V_{t+\Delta t}(x) \quad = \quad (1-\lambda\Delta t)V_t(x+\Delta t) + \lambda\Delta t\int_{0^-}^{x} B(x-y)\mathrm{d}_y V_t(y+\theta\Delta t) + o(\Delta t).$$

Now, write for $\Delta t \to 0$,

$$\begin{aligned} V_{t+\Delta t}(x) &= V_t(x) + \Delta t\frac{\partial}{\partial t}V_t(x) + o(\Delta t), \\ V_t(x+\Delta t) &= V_t(x) + \Delta t\frac{\partial}{\partial x}V_t(x) + o(\Delta t). \end{aligned}$$

Using similar arguments as in, e.g., [56, 159] we find that the Kolmogorov forward equation of the process is

$$\frac{\partial}{\partial t}V_t(x) = \frac{\partial}{\partial x}V_t(x) - \lambda\int_{0^-}^{x}(1-B(x-y))\mathrm{d}_y V_t(y).$$

This relation is commonly known as the *integro-differential equation of Takács*. Letting $t \to \infty$ and using the fact that $v(\cdot)$ denotes a density, we have for $x > 0$,

$$v(x) = \lambda V(0)(1-B(x)) + \lambda\int_0^x v(y)(1-B(x-y))\mathrm{d}y. \tag{1.2}$$

The above equation is also well-known as the level crossing equation, see for instance [68]. It reflects the fact that the rate of crossing level x from above should equal, in steady-state, the rate of crossing level x from below. Note that (1.2) is a Volterra integral equation of the second kind, see, e.g., [165] for details. In fact, (1.2) is also a renewal equation.

Define

$$H(x) := \beta^{-1}\int_0^x (1-B(y))\mathrm{d}y$$

as the distribution of the residual service requirement with density $h(\cdot)$. Note that the level crossing equation (1.2) may be rewritten as

$$v(x) = \rho V(0)h(x) + \rho \int_0^x v(y)h(x-y)\mathrm{d}y. \tag{1.3}$$

There are different ways to obtain $v(\cdot)$ from (1.3), for instance, by observing that it is a renewal equation. However, in view of Chapters 2 and 3, we apply the method of successive substitutions: The $v(y)$ term on the right-hand side (rhs) of Equation (1.3) is substituted by the same expression given by Equation (1.3). Defining $h_n(\cdot)$, $H_n(\cdot)$, as the n-fold convolutions of $h(\cdot)$, $H(\cdot)$, with itself, respectively, we have for $x > 0$,

$$\begin{aligned} v(x) &= \rho V(0)h(x) + \rho \int_0^x \left[\rho V(0)h(y) + \rho \int_0^y v(z)h(y-z)\mathrm{d}z\right] h(x-y)\mathrm{d}y \\ &= \rho V(0)h(x) + \rho^2 V(0)h_2(x) + \rho^2 \int_0^x v(z)h_2(x-z)\mathrm{d}z, \end{aligned}$$

where the second equality follows from changing the order of integration in the third term. Iterating this argument leads to

$$v(x) = \sum_{n=1}^{\infty} \rho^n V(0)h_n(x).$$

Observe that the infinite sum is well-defined if and only if $\rho < 1$. Using normalization, it is easily checked that the well-known relation $V(0) = 1 - \rho$ indeed holds. The steady-state workload distribution $V(x)$ is now directly obtained from the atom at 0 and integrating the density $v(\cdot)$ (and rearranging integral and sum), which results in the following theorem:

Theorem 1.6.1 *For the ordinary M/G/1 queue, with $\rho < 1$, we have*

$$V(x) = (1-\rho)\sum_{n=0}^{\infty} \rho^n H_n(x). \tag{1.4}$$

Although the result (1.4) is surprisingly simple no direct interpretation from the FCFS perspective exists. However, the authors of [59] exploited the fact that the workload process is identical under any work-conserving discipline and considered the LCFS-PR (Last-Come First-Served Preemptive-Resume) discipline to explain the elegant form of (1.4). Under LCFS-PR, the customer in service is interrupted (preempted) when a new customer arrives and its service is resumed once this newly arrived customer and all subsequently arriving customers have completed service.

The LCFS-PR discipline has several useful properties. One of them concerns the geometric distribution of the steady-state number of customers in the system (denoted as N): $\mathbb{P}(N = n) = (1-\rho)\rho^n$. Also, the remaining service requirements of the customers present are independent and identically distributed as a residual service requirement with distribution $H(\cdot)$. Observe that the workload just consists of the sum of residual service requests of customers present. Thus,

$$\mathbb{P}(V \leq x | N = n) = H_n(x).$$

Now, conditioning on N,

$$V(x) = \sum_{n=0}^{\infty} \mathbb{P}(N = n)\mathbb{P}(V \leq x | N = n),$$

and combining the above relations yields (1.4).

The structure of the (tail of the) steady-state waiting time in the G/G/1 queue is roughly the same as (1.4). The number of customers present is still geometrically distributed but with a different parameter that can not be explicitly computed in general (see [131] for an intuitively appealing proof). Furthermore, the excess distribution of the waiting time involves a sum of "ascending ladder heights" instead of residual service requirements. We refer to [10] and [38, Theorem 3.2] for details.

1.6.2 Stochastic recursions

Stochastic recursions serve to describe the evolution of stochastic processes in discrete time. A well-known example in queueing theory is Lindley's equation giving the relationship between the waiting times of two successive customers. In this subsection, we first present Lindley's recursion and mention some monotonicity properties of this recursion. These properties form a key element in Chapter 4, where we exploit a machinery of monotone stochastic recursions developed in [15] to obtain a relationship with a dual system. Then, starting with Lindley's equation, we derive the LST of the steady-state waiting time in the ordinary M/G/1 queue. A similar derivation forms the first step of Chapter 6, where we determine the sojourn time in queues with feedback information on the workload only available at embedded epochs.

Denote by A_n, $n = 0, 1, \ldots$, the interarrival time between customer n and $n + 1$ and by B_n the service requirement of customer n. Also, let W_n be the waiting time of customer n and let W be a generic waiting time. The famous recursion equation of Lindley reads

$$W_{n+1} = (W_n + B_n - A_n)^+, \tag{1.5}$$

where $x^+ = \max(x, 0)$. In the context of random walks, a more familiar notation is $X_n := B_n - A_n$ such that W_n is an ordinary random walk with reflection at zero.

Because W_{n+1} is monotone in W_n, B_n, and A_n, the recursion (1.5) is also called a *monotone stochastic recursion*. Asmussen and Sigman [15] developed a theory for the equivalence between such recursions and their dual processes. The easiest application, and also the most common one, concerns the duality between the workload distribution in the G/G/1 queue and the ruin probability in a corresponding risk process. In Chapter 4 we apply this duality theory to obtain an equivalence between the loss probability in the finite dam and the cycle maximum in its infinite-buffer counterpart.

In Chapter 6, where the service speed is only adapted at embedded epochs, we assume that arrivals occur according to a Poisson process and we use a stochastic recursion to derive an equation for the steady-state sojourn time distribution. A similar analysis may be applied to recover the PK formula for the steady-state waiting time in M/G/1 queues, i.e., the LST of (1.4).

As in [56], p. 253, we first obtain after some calculations, for $s \neq \lambda$

$$\mathbb{E}[e^{-s(W_n+B_n-A_n)^+}|W_n+B_n=x] = \frac{s}{s-\lambda}e^{-\lambda x} - \frac{\lambda}{s-\lambda}e^{-sx}. \tag{1.6}$$

Then, applying (1.5), conditioning on $W_n + B_n$, and using (1.6) yields

$$\begin{aligned}\mathbb{E}[e^{-sW_{n+1}}] &= \int_0^\infty \mathbb{E}[e^{-s(W_n+B_n-A_n)^+}|W_n+B_n=x]\mathrm{d}\mathbb{P}(W_n+B_n<x)\\ &= \frac{s}{s-\lambda}\mathbb{P}(W_{n+1}=0) - \frac{\lambda}{s-\lambda}\mathbb{E}[e^{-s(W_n+B_n)}].\end{aligned}$$

Letting $n \to \infty$ and using the fact that $\mathbb{P}(W=0) = 1-\rho$, we get

$$(s-\lambda)\mathbb{E}[e^{-sW}] = s(1-\rho) - \lambda\mathbb{E}[e^{-sW}]\beta(s).$$

Some rewriting then directly yields the famous PK formula:

Theorem 1.6.2 *For the ordinary M/G/1 queue, with $\rho < 1$, we have*

$$\mathbb{E}[e^{-sW}] = \frac{s(1-\rho)}{s-\lambda(1-\beta(s))}.$$

1.6.3 Sample paths

Sample-path arguments are commonly used in queueing theory. In general, sample paths describe the evolution of the system under a particular realization of the underlying stochastic processes. This may be especially convenient in comparing different model variants with a common generator, such as the finite and infinite-buffer M/G/1 queue. A sample-path comparison may also provide certain bounds on the performance measure(s) of interest.

In this subsection, we first apply a sample-path construction for the workload in the M/G/1 queue to obtain a proportionality relation between the steady-state workload distribution in the finite-buffer queue and its infinite-buffer counterpart, see also [85]. Similar constructions are used in Chapters 3 and 5. Secondly, we briefly introduce some notions related to *cycle maxima* and provide an elegant proof of Takács' formula for the cycle maximum in the M/G/1 queue. The result serves as reference for Chapters 3 and 4 where cycle maxima are considered. Finally, sample-path techniques for queueing systems with heavy tails are discussed in Subsection 1.6.4.

Let V_t^K be the workload at time t in the M/G/1 queue with finite buffer K. In the present context, the finite buffer gives an upper bound for the amount of work in the system while the excess work is lost (see also the finite dam in Section 1.4). Denote by $V^K(\cdot)$ the steady-state distribution of V_t^K. Below, we show that $V(x)$ and $V^K(x)$ are proportional by applying the sample-path approach of [85].

Consider a realization of the workload process $\{V_t, t \geq 0\}$ in the infinite-buffer M/G/1 queue. Now, delete the parts from each upcrossing of level K until the first subsequent downcrossing of level K and paste together the remaining parts, see Figure 1.3 for an illustration. Observe that, at a downcrossing of K, the time until the next customer arrival is still exponential. Hence, the constructed sample path may be considered as a typical realization of the workload process in the M/G/1 queue with finite buffer K.

To formalize the proportionality relation between $V(\cdot)$ and $V^K(\cdot)$, we consider an arbitrary busy cycle of the infinite-buffer queue. Assume that at time 0 a customer arrives in an empty system and let $\tau_0 := \inf\{t > 0 : V_t = 0\}$ denote the length of the cycle. Then, applying the theory of regenerative processes, see e.g. [52], we have for $x \in [0, K]$,

$$V(x) = \frac{1}{\mathbb{E}\tau_0}\mathbb{E}\left[\int_0^{\tau_0} I(V_t \leq x)\mathrm{d}t\right].$$

By the "cut and paste" construction described above, the expected value of the integral is the same for both the original and the constructed process. Thus, $V(x)/V^K(x) = \mathbb{E}\tau_0^K/\mathbb{E}\tau_0$, $x \in [0, K]$, with τ_0^K the length of the busy cycle in the finite dam. The results are summarized in the following theorem:

Theorem 1.6.3 *(Proportionality). For the M/G/1 queue with a finite buffer of size K, we have, for $\rho < 1$ and $x \in [0, K]$,*

$$V^K(x) = \frac{V(x)}{V(K)}. \tag{1.7}$$

Remark 1.6.1 A similar relation holds when $\rho \geq 1$. In that case, the steady-state distribution of V does not exist, but (1.7) holds when $\rho H(x)$ in (1.4) is replaced by $L(x) = \int_0^x e^{-\delta u}\mathrm{d}\rho H(u)$, with δ the unique positive zero of $\int_0^\infty e^{-xu}\mathrm{d}\rho H(u) - 1$. We refer to [53] for further details. ◇

Similar "cut and paste" constructions are applied in Chapters 3 and 5. In fact, the proportionality is generalized along the same lines to queues with workload-dependent arrival and service rates in Chapter 3.

In Chapters 3 and 4, we also consider the notion of cycle maxima. The cycle maximum $C_{\max}$ is the maximum workload during a busy period, i.e., $C_{\max} := \max\{0 \leq t \leq \tau_0 : V_t\}$. A well-known formula for the cycle maximum in the M/G/1 queue, commonly referred to as Takács' formula, is given in the following theorem:

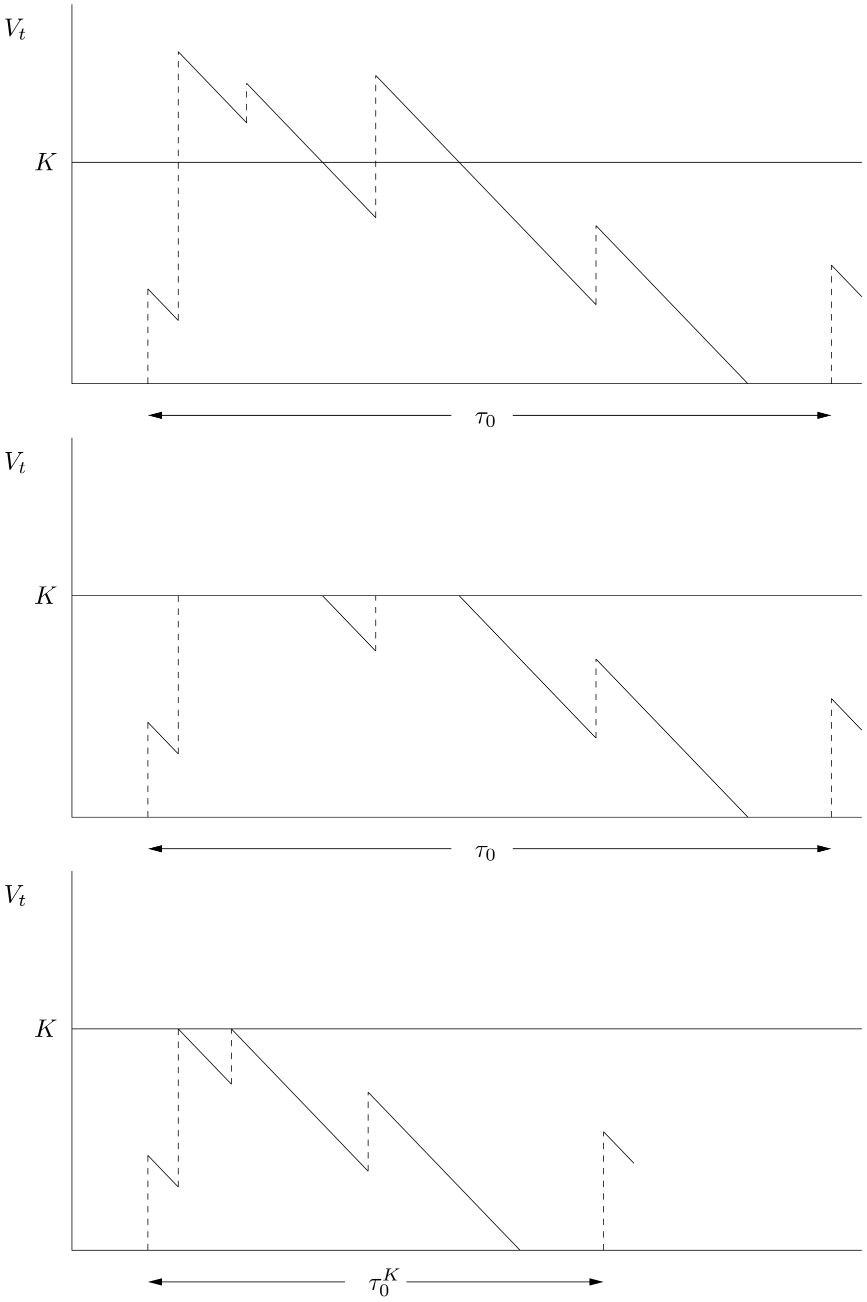

Figure 1.3: A sample-path construction of the workload process in an M/G/1 queue (based on [85]).

Theorem 1.6.4 *The distribution of the cycle maximum in the ordinary M/G/1 queue with $\rho < 1$, satisfies*

$$\mathbb{P}(C_{\max} \leq x) = \frac{\mathbb{P}(V + B \leq x)}{\mathbb{P}(V \leq x)}.$$

Next, we provide an elegant proof of Theorem 1.6.4, taken from [12], which is based on sample-path arguments. Crucial for the proof is the well-known workload representation in queues with unit service rate and time-reversible input processes (and work-conserving service disciplines), introduced by Reich [141],

$$V = \sup_{t \geq 0} \{A(0,t) - t\}, \tag{1.8}$$

with $A(0,t) = \sum_{n=1}^{N(t)} B_n$ the amount of input during $(0,t)$ and $N(t)$ the number of arrivals in an interval of length t. This workload representation is used in many cases. However, the representation is no longer valid when the service speed depends on the amount of work present. Therefore, the proof is not only an interesting exposition of sample-path arguments, but also indicates difficulties arising in proof techniques for queues with workload-dependent rates.

Proof of Theorem 1.6.4 Let $S(t) = A(0,t) - t$ denote the generator of the workload process, where $S(0) = 0$. Because a busy period starts with a jump in an empty system, $C_{\max}$ is also given by $\sup_{0 \leq t \leq \tau_0} \{S(t) + B\}$. (For notational convenience we suppress the dependence between τ_0 and B). Then, using (1.8) and splitting the overall maximum into the maximum during the first busy cycle and the maximum in the remainder of the process, we obtain

$$\begin{aligned} \mathbb{P}(V + B \leq x) &= \mathbb{P}(\sup_{t \geq 0} \{S(t) + B\} \leq x) \\ &= \mathbb{P}(C_{\max} \leq x) \mathbb{P}(\sup_{t \geq \tau_0} \{S(t) + B\} \leq x) \\ &= \mathbb{P}(C_{\max} \leq x) \mathbb{P}(V \leq x), \end{aligned}$$

where the final step follows from the fact that $S(\tau_0) + B = 0$, the stationarity of $S(t)$ and the representation (1.8). □

1.6.4 Sample paths: Heavy tails and asymptotics

Queueing systems with heavy tails have received much attention in the area of communication systems, see Subsection 1.5.2. In this subsection, we first consider the asymptotic behavior of the steady-state workload distribution and the cycle maximum in the standard M/G/1 queue. In particular, we sketch a heuristic derivation of the workload asymptotics using sample-path arguments. Similar arguments are applied in Chapter 7 to obtain the asymptotic behavior for a queue in a random environment (with applications to communication systems). Second, we present results of Asmussen [8] for the asymptotics in queues where the service rate depends on the amount of work present.

A general outline of methods for obtaining tail asymptotics in queueing systems with heavy-tailed characteristics is presented in, e.g., [34, 177]. We specifically focus on derivations based on sample-path techniques. A sample-path approach is especially suitable if the model is relatively complex and no LST of the performance measure of interest is available. Furthermore, sample-path arguments commonly provide qualitative insight into the occurrence of rare events. More specifically, in queueing models with heavy tails a rare event tends to occur as the consequence of a fairly simple most likely scenario. Below, we sketch this scenario for the asymptotic tail probability of the workload in the ordinary M/G/1 queue.

Here and in Chapter 7, we assume that the service requirement distribution function is *regularly varying*. Regularly varying functions are defined as follows.

Definition 1.6.1 *The function $f(\cdot)$ is regularly varying of index $\alpha \in \mathbb{R}$, denoted as $f \in \mathcal{R}_\alpha$, if for all $y > 0$,*

$$\lim_{x\to\infty} \frac{f(yx)}{f(x)} = y^\alpha.$$

Now, consider the workload at time 0 in the standard M/G/1 queue with a regularly varying service requirement distribution of index $-\nu < -1$. The premise is that the most likely scenario for a large workload at time 0 to occur is the arrival of a "tagged" customer with a large service requirement B_{tag} at some time $-y$, while the system shows average behavior otherwise. In the time interval $(-y, 0]$ the amount of work thus roughly decreases at linear rate $1-\rho$. The rare event $V_0 > x$ then implies that the service requirement B_{tag} must exceed $x + (1-\rho)y$. So, integrating with respect to y and neglecting the asymptotically small probability of two or more customer arrivals in $(-y, 0]$ with large service requirements, we obtain

$$\begin{aligned} \mathbb{P}(V > x) &= \int_0^\infty \lambda \mathbb{P}(B_{\text{tag}} > x + (1-\rho)y)\mathrm{d}y \\ &= \frac{\rho}{1-\rho}\mathbb{P}(B^r > x), \end{aligned}$$

where B^r represents a generic residual service requirement (which has distribution function $H(\cdot)$, see Subsection 1.6.1).

The above result also holds for the G/G/1 queue, see [50], and is presented in the following well-known theorem (with the convention that $f(x) \sim g(x)$ indicates $f(x)/g(x) \to 1$ as $x \to \infty$).

Theorem 1.6.5 *Assume that $\rho < 1$. Then, $B(\cdot) \in \mathcal{R}_{-\nu}$ iff $V(\cdot) \in \mathcal{R}_{1-\nu}$, and then*

$$\mathbb{P}(V > x) \sim \frac{\rho}{1-\rho}\mathbb{P}(B^r > x). \tag{1.9}$$

A general outline of converting the heuristic arguments into a rigorous proof for the regularly varying case can be found in [34, 177]. The heuristics and structure of the proof in Chapter 7 are along similar lines as the discussion of the M/G/1 queue given above. In case the arrival process is compound Poisson, a direct and short derivation of (1.9) using Theorem 1.6.1 is provided in [34, 177].

In fact, Pakes [135] extended the asymptotic tail equivalence to subexponential residual service requirements. Subexponential distribution functions include regularly varying distributions and are defined as

Definition 1.6.2 *The distribution function $F(x) := \mathbb{P}(X_i \leq x)$, $i = 1, 2, \ldots$, is subexponential if, for any $n \geq 2$,*

$$\mathbb{P}(X_1 + \ldots + X_n > x) \sim n\mathbb{P}(X_1 > x).$$

Similar intuitive arguments as presented for Theorem 1.6.5 hold for the excess probability of the cycle maximum in the G/G/1 queue, see [8, 64], leading to the following result: Let N denote the number of customers in a cycle.

Theorem 1.6.6 *For the cycle maximum in the G/G/1 queue with $\rho < 1$,*

$$\mathbb{P}(C_{\max} > x) \sim \mathbb{E}N\mathbb{P}(B > x).$$

Finally, we consider the M/G/1 queue with service speed function $r(x)$ when the workload equals x as described in Subsection 1.4.2. Related models are considered in Chapters 2–5. For the case of subexponential service requirements, Asmussen [8] obtained asymptotics for both the steady-state workload density and the cycle maximum. To complete the study of queues with workload-dependent rates, we here present Theorems 3.1 and 3.2 of [8].

Theorem 1.6.7 *(i) Assume that $r(x) \to r_\infty$ as $x \to \infty$, where $\lambda\beta < r_\infty < \infty$. Then,*

$$v(x) \sim \frac{\lambda}{r_\infty - \lambda\beta}\mathbb{P}(B > x).$$

(ii) Assume that $r(x) \to \infty$ as $x \to \infty$. Then,

$$v(x) \sim \frac{\lambda\mathbb{P}(B > x)}{r(x)}.$$

Theorem 1.6.8 *For the M/G/1 queue with service speed function $r(\cdot)$,*

$$\mathbb{P}(C_{\max} > x) \sim \lambda\mathbb{E}\tau_0\mathbb{P}(B > x).$$

Note that, for the standard M/G/1 queue, it holds that $\mathbb{E}N = \lambda\mathbb{E}\tau_0$. In that case, Theorem 1.6.8 reduces to the M/G/1 case of Theorem 1.6.6.

1.7 Overview of the thesis

In this first chapter we discussed various practical scenarios where queues with state-dependent rates may occur. In the remainder of the thesis we focus on the analysis of such queueing systems, where the methods and results of Section 1.6 may serve as a reference framework.

In Chapter 2 we consider two types of queues with workload-dependent service and arrival rates and an infinite buffer. First, in the M/G/1 case, we compare the steady-state distributions of the workload (both at arbitrary epochs and arrival instants) in two models, in which the ratios of arrival and service rates are equal, and show that the steady-state distributions are proportional. Second, for a G/G/1 queue with workload-dependent interarrival times and service rates, we generalize several well-known relations for the workload in the ordinary G/G/1 queue at various epochs. The results of this chapter appeared in [22] and are used in [5].

In Chapter 3, we extend the M/G/1 model of Chapter 2 to queues with restricted accessibility. The proportionality relation of the steady-state workload distributions between two queues with identical ratios of arrival and service rates is generalized to queues with general workload-dependent rejection rules. In addition, we obtain a formal solution for the steady-state workload density and extend the proportionality relation between finite and infinite-buffer queues. We also give an explicit expression for the cycle maximum in the M/G/1 queue with workload-dependent service and arrival rates. The content in this chapter is a combination of [19] and [20].

In Chapter 4 we analyze the G/G/1 queue with finite buffer and workload-dependent service speed. We first show that in the ordinary G/G/1 queue, with the server working at unit speed, the loss probability of a customer may be identified with the tail distribution of the cycle maximum in the associated infinite-buffer queue. A slight modification of this equivalence is required in case the service rate depends on the amount of work present. The results of this chapter are published in [28].

In Chapter 5 we also study an M/G/1 queue with workload-dependent service rate. We specifically assume that the service rate is first increasing and then decreasing as a function of the amount of work. The admission of work into the system is controlled by a policy for accepting or rejecting jobs. We seek an admission control policy that maximizes the long-run throughput. Under certain conditions, we show that a threshold policy is optimal, and derive a criterion for the optimal threshold value. This chapter is based on [21].

In the models of Chapters 2–5 the service rate may be continuously adapted based on the amount of work present. In Chapter 6 the service speed is only determined at epochs right after an arrival depending on the workload and is constant in intervals between customer arrivals. For the two-step service rule we present a procedure to obtain the steady-state workload distribution at various epochs, which provides quite explicit results in case of exponential service requirements. We also consider the generalization to the N-step service-rate function. This chapter is based on [26], while a short version focusing on the exponential case has appeared in [27].

In Chapter 7 we consider a queue where the service rate is determined by the state of a random environment. In particular, the service rate is equal to the bandwidth share of a permanent customer in the G/G/1 queue under the PS discipline. We focus on the case that the traffic process of the G/G/1 queue has heavy-tailed features. In the first part of Chapter 7, we assume that the input is a fluid with constant rate. In the second part we allow a fairly general non-decreasing input process, with renewal input and On-Off sources as special cases. The main result of this chapter is the exact asymptotic behavior of the workload in case the queue is critically loaded. We note that Chapter 7 is presented in the specific context of the integration of streaming and elastic traffic. In that case, the random environment consists of elastic users and the workload as performance measure for the streaming applications is especially relevant as it may be interpreted as the deficit in service compared to a nominal service target. This chapter is based on [24]. A shortened version has appeared in [25] and an extended abstract may be found in [23].

Chapter 2

Queues with workload-dependent service and arrival rates

2.1 Introduction

In practical queueing scenarios, the speed of the server often depends on the amount of work present. Several situations where this phenomenon may occur are extensively described in Chapter 1. One example is a production system where the server is not a machine but a human being. For instance, Bertrand and Van Ooijen [30, 169] describe a production system where the productivity of the shop personnel, that is, the speed of the server, is relatively low when there is much work (stress) or when there is very little work (laziness), see also Section 1.3. In addition, the rate at which jobs arrive at the service system may also depend on the amount of work present. In the human-server example, we may try to control the arrival of jobs to optimize server performance. Another application area of queues with workload-dependent rates are packet-switched communication systems where the transmission rate of data flows may be dynamically adapted based on the buffer content, see for instance [71, 73, 119, 140] or Section 1.5. In particular, implicit feedback information on the buffer state provides the basis for TCP to regulate the transmission rate of Internet flows. We refer to Subsection 1.5.2 for literature on the application of queueing models with workload-dependent rates in communication systems.

The above considerations lead us to study single-server queues with state-dependent interarrival times and general (state-dependent) service speed, which forms the basic model of the thesis. In the first part of this chapter, we analyze an extension of the dam discussed in Subsection 1.4.2. Customers arrive at the queueing system according to a Poisson process, where the arrival rate depends on the workload. The service requirement of a customer is generally distributed, and work is served according to a general release-rate function that also depends on the workload. In the second part, the Markovian case is extended to the regenerative case: We consider a similar model, however with general interarrival times, which may depend on the amount of work present.

In ordinary queueing systems, the speed of the server and the arrival rate of customers are usually assumed to be constant over time. In such systems, the Markovian case amounts to an ordinary M/G/1 queue, whereas the regenerative case represents the classical G/G/1 queue. As noted in Section 1.4, the workload process in a queueing model with general release rule constitutes a dam process with state-dependent release. For references on dams with compound Poisson input, we refer to Subsection 1.4.2. Literature on dams with more general input processes can be found in Cohen and Rubinovitch [58], Kaspi *et al.* [93], and references therein.

The two main goals of this chapter are the following. (i) To establish relationships between two queueing models with arrival rates $\lambda_i(x)$ and release rates $r_i(x)$, $i = 1, 2$, for which $\frac{\lambda_1(x)}{r_1(x)} = \frac{\lambda_2(x)}{r_2(x)}$, $\forall x > 0$. Such relationships will allow us to obtain results for a whole class of models from the analysis of one particular model. (ii) To extend relations between the steady-state workload and the workload at arrival times (waiting times) for the G/G/1 queue to queues with workload-dependent arrival rates and service speeds. We now discuss these two aspects in slightly greater detail.

Ad (i). We consider two related dams, or M/G/1 queues, with general (state-dependent) arrival rate and service speed. We show that the workload distributions are proportional and observe that the difference between the two models is just a rescaling of time. A similar result holds for the workload just before arrival instants — a quantity that does not necessarily equal the waiting time when the service speed is workload-dependent. The derivation of the proportionality relations is partly based on level crossing arguments that lead to a Volterra integral equation of the second kind, see the methodology described in Subsection 1.6.1. These insights also provide an important tool in determining the steady-state workload distribution in an individual model. It turns out to make a crucial difference whether or not the release-rate function $r(\cdot)$ allows the possibility of an empty system.

Ad (ii). The G/G/1 queue with state-dependent release rate requires a different method. Using a Palm-theoretic approach, we establish some general relations between the workload just before arrival instants and the workload at arbitrary time epochs. In the case of Poisson arrivals, we generalize the PASTA property for a continuous-state Markov process. Moreover, various well-known relations for ordinary G/G/1-type queues are extended to queues with general release rates.

This chapter is organized as follows. In Section 2.2, we introduce the M/G/1-type model with state-dependent arrival rate and service speed and consider the level crossing equations. In Section 2.3, we present the proportionality relations with respect to the workload process when we consider two related M/G/1-type queues. The steady-state densities in some special cases are determined explicitly in Section 2.4. In Section 2.5, we present several relations between the steady-state workload and the workload just before arrival instants for the G/G/1 queue with state-dependent release. Finally, Section 2.6 contains conclusions and suggestions for further research.

2.2 Model description and preliminaries

In this section, we introduce some notation for the M/G/1-type queue with general state-dependent arrival rate and service speed. The level crossing equation is stated for this particular model and some special attention is paid to the case that the workload process has no atom at state 0.

Model description
Consider a Markovian workload process with the following dynamics. Between arrivals, the server serves according to some workload-dependent service rate function $r(x)$. Arrivals are governed by a workload-dependent rate function $\lambda(x)$. More precisely, let V_t be the workload at time t and W_n the workload immediately before the n-th arrival epoch. Given that the workload at time t_0 is w and the next arrival is at time $t_1 > t_0$, the workload process during the interval (t_0, t_1) behaves as $V_{t_0+t} = w - \int_{t_0}^{t_0+t} r(V_s)\mathrm{d}s$ (a deterministic process). If A is distributed as the time until the next arrival (starting from initial workload w), then $\mathbb{P}_w(A > t) = e^{-\int_0^t \lambda(V_s)\mathrm{d}s}$, meaning that the hazard rate function of A (its density divided by the tail) at t is given by $\lambda(V_t)$. We assume that $\lambda(\cdot)$ is nonnegative, left-continuous, and has a right limit on $[0,\infty)$. Also, we assume that $r(0) = 0$ and that $r(\cdot)$ is strictly positive, left-continuous, and has a strictly positive right limit on $(0,\infty)$. Each arrival increases the workload by some positive amount (job size), where these amounts form a sequence of i.i.d. random variables $B_1, B_2, \dots$, which are also independent of the interarrival intervals. The random variables B_i have distribution function $B(\cdot)$, with mean β, and LST $\beta(\cdot)$. See Figure 2.1 for a typical realization of the workload process.

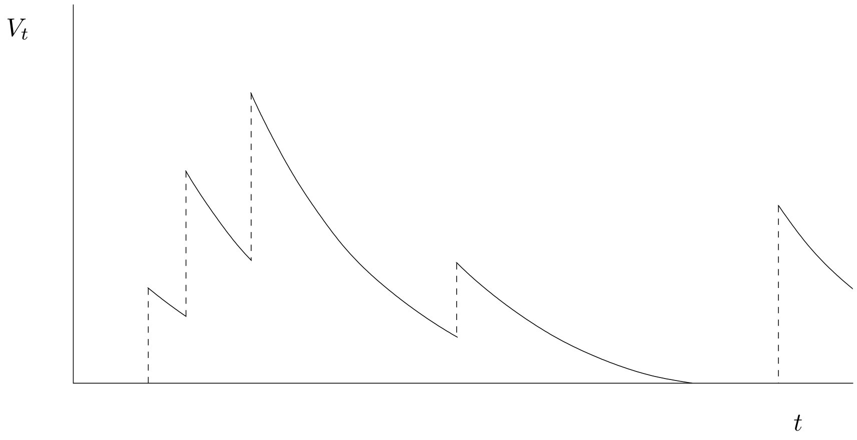

Figure 2.1: A typical sample path of the workload process $\{V_t, t \geq 0\}$ in a queue with strictly increasing service-rate function.

Throughout, we assume that the workload process is ergodic and has a stationary distribution. In order to prevent a general drift to infinity, the rate functions must satisfy $\limsup_{x\to\infty} \beta \frac{\lambda(x)}{r(x)} < 1$ (see also Cohen [54] and Gaver and Miller [80]). We refer to Browne and Sigman [46] for a more detailed discussion on stability issues in case $\lambda(\cdot)$ is non-increasing. Next, let the steady-state random variables corresponding to V_t and W_n be denoted by V and W, and let $v(\cdot)$, $w(\cdot)$ denote their densities.

Define

$$R(x,z) := \int_z^x \frac{1}{r(y)} \mathrm{d}y, \qquad 0 \le z < x < \infty, \tag{2.1}$$

representing the time required to move from state x down to state z in the absence of any arrivals. Of particular interest is $R(x) := R(x,0)$ representing the time required for a workload x to drain, again in the absence of arrivals. A related quantity is

$$\Lambda(x,z) := \int_z^x \frac{\lambda(y)}{r(y)} \mathrm{d}y, \qquad 0 \le z < x < \infty.$$

In particular, $\Lambda(x) := \Lambda(x,0)$ determines whether or not the workload process has an atom at 0 (see also Asmussen [10], p. 381, in case $\lambda(\cdot)$ is fixed). The case $\Lambda(x) < \infty$, for all $0 < x < \infty$, represents the situation that the workload process has an atom at state 0, whereas $\Lambda(x) = \infty$, for some $0 < x < \infty$ (and then for all) corresponds to the case that state 0 cannot be reached by the workload process. We assume that $\int_0^x \lambda(y)\mathrm{d}y$ and $\int_0^x r(y)^{-1}\mathrm{d}y$ cannot be both infinite.

Level crossings

Taking $r(x) \equiv 1$ and $\lambda(x) \equiv \lambda$ results in the ordinary M/G/1 queue. The level crossing identity for the workload is well-known in this case, see e.g. Cohen [52, 56]. In M/G/1-type queues with time-varying arrival rate, the workload level crossing identity has been obtained by Takács [159], while Hasofer [84] shows some additional properties. The proof proposed by Takács may be extended in a rather straightforward way to queues with workload-dependent service and arrival rates, see for instance Section 3.2. This results in the following theorem:

Theorem 2.2.1 *The workload density $v(x)$ exists and satisfies the equation*

$$r(x)v(x) = \lambda(0)V(0)(1 - B(x)) + \int_{y=0^+}^x (1 - B(x-y))\lambda(y)v(y)\mathrm{d}y, \qquad x > 0. \tag{2.2}$$

This integro-differential equation has the following interpretation. The left-hand side of the equation corresponds to the downcrossing rate through level x, while the right-hand side represents the long-run average number of upcrossings through x from the workload level 0 and workload levels between 0 and x respectively. If the workload process has an atom at state 0, it is obvious that $\{V_t, t \ge 0\}$ is a regenerative process, with arrivals of customers in an empty system as regeneration points. Under the assumption of an ergodic process, the expected cycle length is finite and it follows from level crossing theory that the workload density is well-defined. With some minor modifications, the result can be extended to workload processes that do not reach state 0 (see e.g. [54] for details). We also refer to [45, 83] for formal proofs of a stationary density in case $\lambda(\cdot)$ is constant. Note that if $\Lambda(x) = \infty$, then $V(0) = 0$. However, the level crossing equation still holds, and just the first term on the right-hand side of (2.2) disappears, see [45] for details.

2.3 Relations between two M/G/1 queues

In this section we consider two isolated M/G/1 queues with arrival rates $\lambda_i(\cdot)$, release rates $r_i(\cdot)$ and service requirements B_n^i for the n-th customers ($i = 1, 2$). Let $B_1^i, B_2^i, \ldots$ be i.i.d. random variables with distribution function $B(\cdot)$, and let $r_i(\cdot), \lambda_i(\cdot)$ have the same analytical properties as $r(\cdot), \lambda(\cdot)$ specified in Section 2.2. Furthermore, define $\Lambda_i(\cdot,\cdot), \Lambda_i(\cdot), V_i(\cdot), v_i(\cdot)$, and $w_i(\cdot)$ in a similar way as we defined $\Lambda(\cdot,\cdot), \Lambda(\cdot), V(\cdot), v(\cdot)$, and $w(\cdot)$ in Section 2.2. We assume that the two queueing models, to be denoted as Models 1 and 2, are related in the following way:

$$\frac{\lambda_1(x)}{r_1(x)} = \frac{\lambda_2(x)}{r_2(x)}, \qquad \forall x > 0. \tag{2.3}$$

Note that $\Lambda_1(x)$ thus equals $\Lambda_2(x)$. As a consequence, the workload process in both models either has an atom at state 0, or does not hit state 0 at all.

We now state the three theorems of this section.

Theorem 2.3.1 *For all $x > 0$,*

$$\frac{v_1(x)}{v_2(x)} = C\frac{r_2(x)}{r_1(x)},$$

with $C = \frac{\lambda_1(0)V_1(0)}{\lambda_2(0)V_2(0)}$ if $\Lambda_i(x) < \infty$ for all $0 < x < \infty$, and $C = 1$ if $\Lambda_i(x) = \infty$ for some $0 < x < \infty$.

We now turn to the density $w(\cdot)$. Without proof, we claim that it exists just like $v(\cdot)$ (see Theorem 2.2.1 and the end of this section).

Theorem 2.3.2 *$W_i(0) = \lambda_i(0)V_i(0)/\bar{\lambda}_i, i = 1, 2$, with $\bar{\lambda}_i := \int_{0^+}^{\infty} \lambda_i(x)v_i(x)\mathrm{d}x + \lambda_i(0)V_i(0)$, and for all $x > 0$,*

$$w_i(x) = \frac{1}{\bar{\lambda}_i}\lambda_i(x)v_i(x), \qquad i = 1, 2.$$

Theorem 2.3.3 *$W_1(0) = W_2(0)$, and for all $x > 0$,*

$$w_1(x) = w_2(x).$$

Remark 2.3.1 In principle Theorem 2.3.3 (Theorem 2.3.1) can be derived from Theorems 2.3.1 and 2.3.2 (2.3.3 and 2.3.2). To give more insight into the underlying similarities between the two models, we prove each of the three theorems separately. ◇

Remark 2.3.2 Note that $\lambda_i(x) \equiv \lambda$ would yield the PASTA property that the workload at an arbitrary time and the workload at an arrival epoch have the same distribution. Theorem 2.3.2 may thus be viewed as a generalization of the PASTA property yet under the assumption of a continuous-state stationary workload process. $\diamond$

Proof of Theorem 2.3.1 We apply the level crossing identity to Model i and define $z_i(x) := r_i(x)v_i(x)$, $i = 1, 2$. Then (2.2) reduces to

$$z_i(x) = \lambda_i(0)V_i(0)(1-B(x)) + \int_{y=0^+}^{x} (1-B(x-y))\frac{\lambda_i(y)}{r_i(y)} z_i(y)\mathrm{d}y, \qquad x > 0. \tag{2.4}$$

If $\Lambda_i(x) = \infty$ for some $0 < x < \infty$, then $V_i(0) = 0$ and the result follows easily. So assume that $\Lambda_i(x) < \infty$ for all $0 < x < \infty$.

Observe that (2.4) is a Volterra integral equation of the second kind. Let its kernel be $K^{(i)}(x, y) := (1 - B(x - y))\frac{\lambda_i(y)}{r_i(y)}$ for $0 < y < x < \infty$ and $K_*^{(i)}(x, 0) := 1 - B(x)$ for $0 < x < \infty$. Notice that due to (2.3) the kernels of both models are the same and we may drop the index i from our notation. Now define recursively

$$K_n(x, y) := \int_y^x K(x, z)K_{n-1}(z, y)\mathrm{d}z, \qquad 0 < y < x < \infty,\ n = 2, 3, \ldots,$$

and

$$K_{n*}(x, 0) := \int_{0^+}^x K_n(x, y)K_*(y, 0)\mathrm{d}y, \qquad 0 < x < \infty,\ n = 1, 2, \ldots,$$

where $K_1(x, y) := K(x, y)$ and $K_{0*}(x, 0) := K_*(x, 0)$. So the classical successive-substitution method for Volterra integral equations gives:

$$\begin{aligned} z_i(x) &= \lambda_i(0)V_i(0)K_*(x, 0) + \int_{0^+}^x K(x, y)K_*(y, 0)\lambda_i(0)V_i(0)\mathrm{d}y + \ldots \\ &= K_{0*}(x, 0)\lambda_i(0)V_i(0) + K_{1*}(x, 0)\lambda_i(0)V_i(0) + \ldots \\ &= \lambda_i(0)V_i(0)\sum_{n=0}^{\infty} K_{n*}(x, 0). \end{aligned} \tag{2.5}$$

Dividing $z_1(x)$ by $z_2(x)$ and substituting $z_i(x) = r_i(x)v_i(x), i = 1, 2$, yields

$$\frac{v_1(x)}{v_2(x)} = \frac{r_2(x)}{r_1(x)}\frac{\lambda_1(0)V_1(0)}{\lambda_2(0)V_2(0)},$$

and we have shown the result. $\square$

The Volterra approach provides a useful tool for determining the workload densities. For instance, Harrison and Resnick [83] solved the densities in case $\lambda(\cdot)$ is fixed and $\Lambda_i(x) < \infty$ for all $0 < x < \infty$. Perry and Asmussen [137] used the same approach to find workload densities for models with workload-dependent arrival rates, but fixed $r(\cdot)$. We adopt the analysis carried out in

[83] and use the bound $K(x,y) \leq \frac{\lambda(y)}{r(y)}$ to show inductively that $K_{(n+1)*}(x,y) \leq \frac{(\Lambda(x,y))^n}{n!}\frac{\lambda(y)}{r(y)}$. Note that the sum in (2.5) is well-defined (in case $\Lambda_i(x) < \infty$ for all $0 < x < \infty$) and we thus have a closed-form expression for $z_i(x)$ and, hence, for $v_i(x)$. An explicit formula is presented in Chapter 3, Section 3.3.

If $\Lambda_i(x) = \infty$ for some $0 < x < \infty$, then the workload process approaches the state 0, but never reaches it. In this case the integrated kernel, $\int_0^x K(y,0)\mathrm{d}y$, is unbounded and equation (2.4) is often referred to as a singular integral equation. Brockwell *et al.* [45, Theorem 5] obtained an explicit expression for the stationary workload distribution in case $\lambda(\cdot)$ is fixed. For references on the solution of singular integral equations in general, we refer to Linz [116], Ch. 1 and 3.5, Mikhlin [123] Ch. 1 and 3, and Zabreyko *et al.* [175], Ch. 1, 6, and 9. In Section 2.4 we give the steady-state workload distribution for some special cases.

Let us now give an intuitive explanation of Theorem 2.3.2, based on a Bayesian argument. In Section 2.5 we derive a more general result, from which the theorem follows as a special case.

Consider either of the two models and drop the index i from the notation. The probability of having two or more arrivals in a small time interval $(t, t+\Delta)$ is of order $o(\Delta)$. Then, by simple conditioning arguments we have $\mathbb{P}(\text{arrival in } (t,t+\Delta)) = \int_{y=0^+}^{\infty} \lambda(y)\Delta v(y)\mathrm{d}y + \lambda(0)\Delta V(0) + o(\Delta)$ and $\mathbb{P}(\text{arrival in } (t,t+\Delta)|V_t > x)\mathbb{P}(V_t > x) = \int_{y=x}^{\infty}\lambda(y)\Delta v(y)\mathrm{d}y + o(\Delta)$. Let us consider the tail probability of the workload at jump epochs,

$$\begin{aligned}
\mathbb{P}(W > x) &= \lim_{\Delta\to 0} \mathbb{P}(V_t > x|\text{arrival in } (t,t+\Delta)) \\
&= \lim_{\Delta\to 0} \frac{\mathbb{P}(\text{arrival in } (t,t+\Delta)|V_t > x)\mathbb{P}(V_t > x)}{\mathbb{P}(\text{arrival in } (t,t+\Delta))} \\
&= \frac{1}{\bar{\lambda}}\int_{y=x}^{\infty}\lambda(y)v(y)\mathrm{d}y, \qquad x \geq 0,
\end{aligned}$$

where $\bar{\lambda} = \int_{y=0^+}^{\infty}\lambda(y)v(y)\mathrm{d}y + \lambda(0)V(0)$. The intuitive explanation of the theorem is completed by differentiation:

$$w(x) = \frac{\mathrm{d}}{\mathrm{d}x}(1 - \mathbb{P}(W > x)) = \frac{1}{\bar{\lambda}}v(x)\lambda(x).$$

We now turn to Theorem 2.3.3. Let us consider either of the two models and define

$A^w_{n,x} =$ workload decrement between arrivals n and $n+1$, when workload immediately *after* n-th arrival epoch is x, $\quad n \in \mathbb{N}, x > 0$.

Observe that $A^w_{n,x}$ may be interpreted as some kind of interarrival time between the n-th and $(n+1)$-th customer. While the interarrival time is usually expressed in terms of time, $A^w_{n,x}$ represents the workload decrement between two successive arrivals. Remember that a similar argument holds for the service requirement, which in general does not equal the service time, and the workload at jump epochs, which in general differs from the waiting time. This demonstrates that the following well-known recursion, usually interpreted in terms of time, holds again in terms of workload:

$$W_{n+1} = (W_n + B_n - A^w_{n,W_n+B_n})^+, \qquad n = 1, 2, \ldots. \tag{2.6}$$

If we omitted the times between successive arrivals, we would have a system of only upward (arrival of a customer) and downward jumps (workload decrement during an interarrival interval). The distribution of the workload at arrival epochs only depends on the sizes of these jumps, as can be concluded from (2.6). Hence, the workload at jump epochs only depends indirectly on the time between two successive arrivals. The distribution of the service requirements (upward jumps) is by assumption identical for Models 1 and 2. In order to prove Theorem 2.3.3, it suffices to show that the sizes of the downward jumps, i.e., the workload decrements during an interarrival interval, are identically distributed for Models 1 and 2.

Thus let us consider the interarrival time and the workload decrement during an interarrival interval of either of the two models. Assume that at time 0 a customer arrives and the workload just after the arrival is $W + B$, with realization $W + B = y$. Denote by A^t_y the conditional interarrival time and by A^w_y the conditional workload decrement during the interarrival time, i.e., the event $\{A^w_y > v\}$ represents the situation that when the next customer arrives the workload is smaller than $y - v$. If we let t_v be the conditional time required for a workload decrease of v, then the events $\{A^w_y > v\}$ and $\{A^t_y > t_v\}$ are identical.

Next, use an alternative characterization of a Poisson arrival process with rate $\lambda(x)$ when the workload equals x, to determine the excess distribution of the conditional interarrival time,

$$\mathbb{P}(A^t_y > t_v) = e^{-\int_{t=0}^{t_v} \lambda(V_t)\mathrm{d}t}, \qquad y > v. \tag{2.7}$$

Recall that $r(x)$ is the depletion rate at time t if the workload V_t equals x. Hence, between successive arrivals the workload process satisfies (see, e.g., [10, 54, 80, 83])

$$\frac{\mathrm{d}V_t}{\mathrm{d}t} = -r(V_t). \tag{2.8}$$

Since the amount of work at $t = 0$ equals y, and t_v is defined such that $\int_{y-v}^{y} \frac{1}{r(x)}\mathrm{d}x = t_v$, the next proposition follows easily from (2.7) and (2.8).

Proposition 2.3.1 *Assume that the workload just after an arrival is y ($W + B = y$), then, for $y > v$:*

$$\mathbb{P}(A^w_y > v) = e^{-\int_{u=y-v}^{y} \frac{\lambda(u)}{r(u)}\mathrm{d}u}. \tag{2.9}$$

Differentiation of (2.9) yields the conditional density:

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{P}(A^w_{W+B} \leq x | W+B=y) = \frac{\lambda(x)}{r(x)} e^{-\int_{z=x}^{y} \frac{\lambda(z)}{r(z)}\mathrm{d}z}, \qquad 0<x<y. \qquad (2.10)$$

Proof of Theorem 2.3.3 Since $A^w_{n,x}$ depends on $W_n + B_n$, $n = 1, 2, \ldots$, (2.6) leads to:

$$\mathbb{P}(W_{n+1} > z) = \int_{u=z}^{\infty} \mathbb{P}(A^w_{n,u} < u - z \mid W_n + B_n = u)\mathrm{d}\mathbb{P}(W_n + B_n \leq u).$$

Notice that the distribution of $A^w_{n,x}$ (cf. (2.9)) depends only on the ratio $\frac{\lambda(\cdot)}{r(\cdot)}$ and hence is the same in both models. Since the distribution of the service requirements is the same by assumption, we can use a stochastic coupling argument to complete the proof. □

From Theorem 2.3.1 and the discussion of Theorem 2.3.3 we can conclude that changing between Model 1 and Model 2 is just a rescaling of time. If we consider the workload process of Model i, with the speed of time equal to $1/r_i(x)$ when the workload is x, then Models 1 and 2 are equivalent. The special case with $r_1(x) \equiv r_1$ and $r_2(x) \equiv r_2$ can thus be interpreted as observing the workload at two different (but constant) time scales, which clearly does not affect the workload distribution. See also [137] for the same time-transformation argument in the extreme case of $r_1(x) = 1/\lambda_2(x)$ and $r_2(x) = \lambda_1(x) \equiv 1$. It immediately follows from the rescaling arguments and the existence of a workload density at arrival instants in the ordinary M/G/1 queue, that the density $w(\cdot)$ in the more general model is also well-defined.

2.4 Special cases

The main results of Section 2.3 provide us with a tool to translate known results for a particular model to a whole class of related models. In this section we consider several examples. Throughout the section we use the notation $f(x) \propto g(x)$ if $f(x) = cg(x)$ for all $x > 0$ and some constant c.

Case (i): Arrival control follows service rate.
We start with an M/G/1 queue with $\lambda(x) = Cr(x)$. Note that this special case might be applicable to queues where the arrival control must follow the service rate, see for instance [30]:

- high workload: panicking server, reduce the arrival rate
- medium workload: fast server, send more work
- small workload: lazy server, send less work

Note that the third case does not seem desirable. However, in practical situations the main focus will be on a server with a relatively high workload, and regimes with a small workload are usually of limited interest.

It is obvious that $\frac{\lambda(x)}{r(x)} = \frac{C}{1}$. Apply Theorems 2.3.1 and 2.3.3 to see that $r(x)v(x) \propto v_{M/G/1}(x)$ and $w(x) = w_{M/G/1}(x)$, where $v_{M/G/1}(\cdot)$ and $w_{M/G/1}(\cdot)$ are the workload densities of the ordinary M/G/1 queue with arrival rate C and service speed 1, at arbitrary and arrival epochs, respectively. Hence the workload process in the M/G/1 queue with general service speed $r(x)$ and arrival rate $\lambda(x) = Cr(x)$ can be analyzed in detail.

Case (ii): Exponential service times.
Consider the M/M/1 queue with general $\lambda(\cdot)$ and $r(\cdot)$ functions. Substituting $B(x) = 1 - e^{-\mu x}$ in (2.2) gives

$$r(x)v(x) = \lambda(0)V(0)e^{-\mu x} + \int_{y=0^+}^{x} \lambda(y)v(y)e^{-\mu(x-y)}\mathrm{d}y, \qquad x > 0. \quad (2.11)$$

Multiply by $e^{\mu x}$, define $f(x) := e^{\mu x}r(x)v(x)$ and differentiate to obtain

$$\frac{\mathrm{d}}{\mathrm{d}x}f(x) = f(x)\frac{\lambda(x)}{r(x)}.$$

The solution of this differential equation is unique up to a constant and may be written in the form

$$f(x) = C\exp\left\{\Lambda(x,1)\right\}, \qquad x > 0. \quad (2.12)$$

Using a straightforward extension of Asmussen [10], p. 388, it follows that we have positive recurrence (and hence, C can be determined by normalization) if and only if

$$\alpha := \int_0^\infty \frac{1}{r(x)}\exp\left\{\Lambda(x,1) - \mu x\right\}\mathrm{d}x < \infty.$$

If $V(0) = 0$, then $C = \alpha^{-1}$. However, if $V(0) > 0$, observe from (2.11) that $\lim_{x\downarrow 0} r(x)v(x) = \lambda(0)V(0)$ and use the straightforward extension $\Lambda(x,y) = -\Lambda(y,x)$ for $0 < x < y < \infty$ to see that $C = \lambda(0)V(0)\exp\{\Lambda(1,0)\}$. Hence, if $V(0) > 0$, then

$$v(x) = \frac{\lambda(0)V(0)}{r(x)}\exp\left\{\int_0^x \left(\frac{\lambda(y)}{r(y)} - \mu\right)\mathrm{d}y\right\}.$$

From this solution it becomes immediately clear that for two models with $\lambda_i(\cdot)$ and $r_i(\cdot)$ $(i = 1,2)$ satisfying (2.3), Theorem 2.3.1 holds in the M/M/1 case.

Case (iii): Shot noise.
Another well-known example is the shot noise model, i.e., $\lambda(x) \equiv \lambda$ and $r(x) = rx$, cf. [95], [149], p. 393, [56], p. 558, or Example 1.2.5. First notice that for the

shot noise model $\Lambda(x) = \infty$ for $x > 0$ and the Volterra successive-substitution approach cannot (at least not directly) be applied. However, it is still possible to analyze this model due to the special form of $\lambda(\cdot)$ and $r(\cdot)$. First consider the level crossing identity (2.2) for this case,

$$rxv(x) = \lambda \int_{y=0^+}^{x} (1 - B(x-y))v(y)\mathrm{d}y, \qquad x > 0.$$

Let $\phi(s) := \int_0^\infty e^{-sx}v(x)\mathrm{d}x$ be the Laplace Transform (LT) of the workload density, then $-\frac{\mathrm{d}}{\mathrm{d}s}\phi(s) = \int_0^\infty e^{-sx}xv(x)\mathrm{d}x$ and we have

$$-\frac{\mathrm{d}}{\mathrm{d}s}\phi(s) = \frac{\lambda}{r}\frac{1-\beta(s)}{s}\phi(s).$$

Solving this differential equation yields (cf. [95], [149], p. 393, or [56], p. 558)

$$\phi(s) = \exp\left\{-\frac{\lambda}{r}\int_0^s \frac{1-\beta(u)}{u}\mathrm{d}u\right\}. \tag{2.13}$$

Because of the PASTA property, the LT of the workload density at arrival epochs also equals $\phi(\cdot)$.

Since the shot noise model is thus solved, we immediately obtain the LT of the workload density at arbitrary and at arrival epochs of models with $\lambda(x) = f(x)x^a$ and $r(x) = f(x)x^{a+1}$, where $\int_0^x f(y)y^a\mathrm{d}y$ and $\int_0^x f(y)^{-1}y^{-(a+1)}\mathrm{d}y$ are not both infinite (use Theorems 2.3.1-2.3.3).

Furthermore, let us consider some special properties of the shot noise model. Take $\lambda \equiv r$, and use (2.9) to observe that, given $W_n + B_n$, W_{n+1} is uniformly distributed on the stochastic interval $(0, W_n + B_n)$. This means that the steady-state distribution of W must satisfy

$$W =^d U(0, W + B), \tag{2.14}$$

with $U(x, y)$ denoting the uniform distribution on the interval (x, y) and $=^d$ indicating equality in distribution. But if W is $U(0, W + B)$ distributed, then B must be $U(0, W + B)$ distributed as well. Hence, in a model with exponentially distributed service requirements, W must have the same exponential distribution as the service requirement B.

Remark 2.4.1 Using LT's it can also be readily shown that Equation (2.14), with B being $\exp(\mu)$ distributed, implies that W is $\exp(\mu)$ distributed. Indeed, (2.14) implies that the LT $\psi(s)$ of W satisfies

$$\psi(s) = \frac{1}{s}\int_{\sigma=0}^{s} \psi(\sigma)\frac{\mu}{\mu+\sigma}\mathrm{d}\sigma,$$

with boundary condition $\psi(0) = 1$, which after differentiation yields $\psi(s) = \frac{\mu}{\mu+s}$. ◇

Another special case arises when the service requirements are exponential random variables with rate μ, while λ, r are not necessarily identical. Substituting $\beta(s) = \frac{\mu}{\mu+s}$ in (2.13) yields

$$\phi(s) = \left(\frac{\mu}{\mu+s}\right)^{\lambda/r}.$$

This is the LT of the Gamma distribution, i.e., $v(\cdot) \propto \text{Gamma}(\frac{\lambda}{r}, \mu)$ and due to the PASTA property also $w(\cdot) \propto \text{Gamma}(\frac{\lambda}{r}, \mu)$. Furthermore, if we consider the model with $\lambda(x) = \frac{\lambda}{x}$ and $r(x) \equiv r$, it follows that $w(\cdot) \propto \text{Gamma}(\frac{\lambda}{r}, \mu)$ and $v(x)$ is proportional to $xw(x)$, hence, $v(\cdot) \propto \text{Gamma}(\frac{\lambda}{r}+1, \mu)$. Note that taking $\lambda \equiv r$ indeed gives that $w(\cdot)$ is exponentially distributed with parameter μ.

Remark 2.4.2 Related to shot noise is the model with $\lambda(x) \equiv \lambda$ and $r(x) = p+rx$. Note that the workload process has an atom at 0 and the workload density can thus be formally written as an infinite sum of Volterra kernels. Paulsen and Gjessing [136] obtained some more explicit formulas for this case in the setting of a risk process. In particular, the workload excess probability may be written in terms of Bessel functions for Erlang$(2, \mu)$ service requirements, and in terms of confluent hypergeometric functions for hyperexponential service requirements consisting of two exponentials. ◇

Case (iv): Some other models.
Consider the M/G/1 queue with $\lambda(x) \equiv \lambda$ and $r(x) = x^2$. Using the level crossing identity (2.2) and noting that $\int_0^\infty x^2 v(x) e^{-sx} \mathrm{d}x = \phi^{''}(s)$ yields

$$\phi^{''}(s) = \lambda \frac{1-\beta(s)}{s} \phi(s). \tag{2.15}$$

Denoting $g(s) := \lambda \frac{1-\beta(s)}{s}$ and using the transformations $\phi(s) = e^{f(s)}$ and $h(s) = f^{'}(s)$ gives

$$h^{'}(s) + \big[h(s)\big]^2 = g(s).$$

This non-linear first-order differential equation is in general very difficult to solve. Note that $\Lambda(x) = \infty$, for all $x > 0$, and the workload process can thus not reach state 0. However, for some specific choices of the service requirement distribution we obtain an expression for the LT. For instance, the LT of $v(x)$ in the M/M/1 case is given by (use (2.12))

$$\phi(s) = G \int_0^\infty \frac{e^{-(s+\mu)x}}{x^2} e^{-\frac{\lambda}{x}} \mathrm{d}x,$$

with G some normalizing constant. Integrating $\phi^{''}(s)$ by parts shows that (2.15) is satisfied for the case of exponential service requirements.

If the service requirements are Erlang$(2,\mu)$ distributed, then the LT may be written as a weighted sum of Bessel functions. Alternatively, substituting $\lambda(x) \equiv \lambda, r(x) = x^2$ into (2.2), defining $z(x) := e^{\mu x}v(x)$, and differentiating twice, yields

$$x^2 z^{''}(x) + (4x - \lambda)z^{'}(x) + (2 - \lambda\mu)z(x) = 0. \tag{2.16}$$

Applying Maple to solve this second-order differential equation, it is seen that the steady-state density $v(\cdot)$ itself is also a weighted sum of Bessel functions (just like the LT).

We conclude with the observation that if we can calculate $v(\cdot)$ and/or $w(\cdot)$ this immediately gives results for M/G/1 models with $\lambda(x) = f(x)x^a$ and $r(x) = f(x)x^{a+2}$, where $\int_0^x f(y)y^a \mathrm{d}y$ and $\int_0^x f(y)^{-1}y^{-(a+2)}\mathrm{d}y$ are not both infinite (again use Theorems 2.3.1-2.3.3).

Remark 2.4.3 If $\lambda(x) \equiv \lambda$ and $r(x) = e^{ax}$, then $R(x) = \int_0^x e^{-ay}\mathrm{d}y < \infty$ for $0 < x < \infty, |a| < \infty$. In this case, we can follow [83], using the bound $K(x,y) \leq \frac{\lambda}{r(y)}$ and show inductively that, for $0 < y < x < \infty$,

$$K_{n+1}(x,y) \leq \frac{\lambda^{n+1}(R(x) - R(y))^n}{r(y)n!} = \frac{\lambda^{n+1}e^{-ay}(e^{-ay} - e^{-ax})^n}{a^n n!}.$$

Thus the sum $\sum_{n=1}^{\infty} K_n(x,y)$ is well-defined, and hence, $\sum_{n=0}^{\infty} K_{n*}(x,y)$ is well-defined as well. So the steady-state workload density is given by (2.5):

$$v(x) = \lambda V(0)e^{-ax}\sum_{n=0}^{\infty} K_{n*}(x,0), \qquad x > 0,$$

with $V(0)$ determined via normalization:

$$V(0) = \left[1 + \int_{0^+}^{\infty} \lambda e^{-ax}\sum_{n=0}^{\infty} K_{n*}(x,0)\mathrm{d}x\right]^{-1}.$$

Of course, this can be extended to models with $\lambda(x) = f(x)e^{bx}$ and $r(x) = f(x)e^{ax}$, where $\int_0^x f(y)e^{by}\mathrm{d}y$ and $\int_0^x f(y)^{-1}e^{-ay}\mathrm{d}y$ are not both infinite. ◇

2.5 Palm-theoretic approach

So far, we considered M/G/1-type queues with general arrival and service rate, $\lambda(\cdot)$, $r(\cdot)$, depending on the amount of work in the system. Recall that B_n denotes the service requirement of the n-th customer and A_n denotes the inter-arrival time between the n-th and $(n+1)$-th customer, $n = 0, 1, \ldots$, where B and A are their steady-state random variables. Furthermore, let V_t again denote the workload at time t and let W_n again be the workload immediately before the n-th jump epoch, with steady-state random variables V and W, respectively.

In this section, we first continue the study of this Markovian case. Using Palm-theoretic principles, we establish a general relation between V and W, or rather $f(V)$ and $f(W)$. Some specific well-chosen $f(\cdot)$-functions yield convenient relations for, e.g., the tail probability, the expectation, or the LST of the considered random variables. In addition, Theorem 2.3.2 follows easily as a corollary. We proceed by allowing general renewal arrival processes and again establish a general relation between V and W. Some examples show that the relations may be viewed as extensions of some well-known relations for ordinary G/G/1 queues. Furthermore, in case of Poisson arrivals, the level crossing equations are derived in an alternative way. We conclude with an extension of the dynamics driving the workload process in the ordinary G/G/1 queue to similar dynamics in G/G/1-type queues with general release rate.

In Sections 2.3 and 2.4 the arrivals followed a Poisson process. We applied the level crossing equations (2.2) to determine the limiting distribution and to show some equivalence properties. In order to handle the general renewal nature of the input process, we adopt a totally different approach based on Palm-theoretic principles, see for instance [16, Section 1.3]. Specifically, we express $\mathbb{E}[f(V)]$ as a stochastic mean value over one arbitrary interarrival interval. Let $W+B$ and W_A denote the workload at the beginning and end of the (arbitrary) interarrival interval A, respectively. If we assume that the function $f(\cdot)$ is such that the considered expectations exist and are finite, then

$$\mathbb{E}[f(V)] = \frac{1}{\mathbb{E}A}\mathbb{E}\left[\int_{t=0}^{A} f(V_t)\mathrm{d}t\right]. \tag{2.17}$$

Remark 2.5.1 In fact, this stochastic mean-value formula is an application of Campbell's theorem, see e.g. [16, Section 1.2] or [156, Section 5.4]. Campbell's theorem establishes a link between functions of time-stationary and event-stationary marked point processes. A special case is the so-called (first) inversion formula, which relates stationary probabilities to Palm probabilities. Combined with the definition of Palm probabilities this directly provides Equation (2.17). $\diamond$

First, let us consider the Markovian case.

Theorem 2.5.1 *Let $f(\cdot)$ be such that $\mathbb{E}[f(V)]$ exists and is finite, then*

$$\mathbb{E}[f(V)] = \mathbb{E}\left[\frac{f(W)}{\lambda(W)}\right]\frac{1}{\mathbb{E}\left[\frac{1}{\lambda(W)}\right]}. \tag{2.18}$$

Proof Starting with the stochastic mean-value result (2.17) and introducing the indicator function $I(\cdot)$, we have

$$\begin{aligned}\mathbb{E}[f(V)] &= \frac{1}{\mathbb{E}A}\mathbb{E}\left[\int_{t=0}^{A} f(V_t)\mathrm{d}t\right]\\ &= \frac{1}{\mathbb{E}A}\mathbb{E}\left[\int_{t=0}^{\infty} f(V_t)I(A>t)\mathrm{d}t\right].\end{aligned}$$

Note that $\mathbb{E}[I(A > t)|W + B] = \mathbb{P}(A > t|W + B) = \mathbb{P}(W_A < V_t|W + B)$, and use (2.8) to see that

$$\begin{aligned}
\mathbb{E}[f(V)] &= \frac{1}{\mathbb{E}A}\mathbb{E}\Bigg[\int_{x=W+B}^{0} f(x)\mathbb{P}(W_A < x|W_0 = W + B)\frac{\mathrm{d}x}{-r(x)}\Bigg] \\
&= \frac{1}{\mathbb{E}A}\mathbb{E}\Bigg[\int_{x=0}^{W+B} \frac{f(x)}{\lambda(x)}\mathrm{d}\mathbb{P}(W_A \le x|W_0 = W + B)\Bigg] \\
&= \frac{1}{\mathbb{E}A}\mathbb{E}\Bigg[\frac{f(W_A)}{\lambda(W_A)}\Bigg]. \qquad (2.19)
\end{aligned}$$

The second equality sign follows by combining (2.9) and (2.10). Notice that W_A and W have the same distribution. Furthermore, taking $f(x) \equiv 1$ yields

$$\mathbb{E}A = \mathbb{E}\left[\frac{1}{\lambda(W)}\right],$$

which completes the proof. □

Let us briefly consider some special cases of $f(\cdot)$. Taking $f(x) = x$, respectively $f(x) = e^{-sx}$, gives a relation between $\mathbb{E}V$ and $\mathbb{E}W$, respectively a relation between the LST's of V and W. Furthermore, taking $f(x) = I(x > v)$ expresses the steady-state excess distribution of the workload in terms of $\lambda(\cdot)$ and the steady-state workload density at arrival epochs. Taking $f(x) = \lambda(x)g(x)$ yields

$$\mathbb{E}[\lambda(V)g(V)] = \frac{\mathbb{E}[g(W)]}{\mathbb{E}\left[\frac{1}{\lambda(W)}\right]} = \mathbb{E}[\lambda(V)]\mathbb{E}[g(W)], \qquad (2.20)$$

or equivalently

$$\mathbb{E}[g(W)] = \frac{\mathbb{E}[\lambda(V)g(V)]}{\mathbb{E}[\lambda(V)]},$$

where the second equality sign in (2.20) follows from taking $f(x) = \lambda(x)$ in (2.18). Now taking $g(x) = e^{-sx}$ yields

$$\mathbb{E}\left[e^{-sW}\right] = \frac{\mathbb{E}\left[\lambda(V)e^{-sV}\right]}{\mathbb{E}[\lambda(V)]}. \qquad (2.21)$$

Because of the one-to-one correspondence between an LST and its inverse, (2.21) implies that the steady-state workload density at arrival epochs $w(x)$ is proportional to the product of $\lambda(x)$ and the steady-state workload density $v(x)$. Note that we have just proven Theorem 2.3.2.

Now let us consider a generalization of the above-described M/G/1-type model, by allowing *generally* distributed interarrival times, which may depend on the workload W found upon arrival according to some distribution $\mathbb{P}(A < x|W = w)$. We again derive a relation between V and W by starting from the stochastic mean-value result (2.17). Let B^r denote the residual service requirement with density $h(\cdot) = \frac{1-B(\cdot)}{\mathbb{E}B}$.

Theorem 2.5.2 *Let $f(\cdot)$ be such that $\mathbb{E}[f(V)]$ exists and is finite, then*

$$\mathbb{E}[f(V)|V>0] = \mathbb{E}[r(V)|V>0]\mathbb{E}\left[\frac{f(W+B^r)}{r(W+B^r)}\right]. \tag{2.22}$$

Proof First define $g(w,z) := \int_0^z f(w+u)\mathrm{d}u$ and consider

$$\begin{aligned}\mathbb{E}\left[\int_0^B f(w+x)\mathrm{d}x\right] &= \mathbb{E}[g(w,B)] \\ &= \int_0^\infty g'(w,x)\mathbb{P}(B>x)\mathrm{d}x + g(w,0) \\ &= \int_0^\infty f(w+x)\mathbb{P}(B>x)\mathrm{d}x \\ &= \mathbb{E}B\mathbb{E}[f(w+B^r)]. \end{aligned} \tag{2.23}$$

Starting with the stochastic mean-value result (2.17), making the substitution $u = V_t$ and using (2.8) yields

$$\mathbb{E}[f(V)] = \frac{1}{\mathbb{E}A}\left(\mathbb{E}\left[\int_{u=W+B}^{W} f(u)\frac{\mathrm{d}u}{-r(u)}\right] + f(0)\mathbb{E}(A-\tau)^+\right), \tag{2.24}$$

where $x^+ = \max(0,x)$ and $\tau := \inf\{t>0 : V_t = 0\}$. As $V=0$ might be a special workload level, we focus on $V>0$, resulting in:

$$\mathbb{E}[f(V)|V>0]\mathbb{P}(V>0) = \frac{1}{\mathbb{E}A}\mathbb{E}\left[\int_{u=W}^{W+B}\frac{f(u)}{r(u)}\mathrm{d}u\right]. \tag{2.25}$$

Since W_{n+1} depends on $W_n + B_n$, the boundaries in $\int_W^{W+B}$ really are dependent, as they represent the workloads at two successive arrival epochs. But we can rewrite $\mathbb{E}[\int_W^{W+B}] = \mathbb{E}[\int_0^{W+B} - \int_0^W]$ and observe that both W_n and W_{n+1} have the same steady-state distribution as W. Thus we can rewrite (2.25) into

$$\begin{aligned}\mathbb{E}[f(V)|V>0]\mathbb{P}(V>0) &= \frac{1}{\mathbb{E}A}\mathbb{E}_B\left[\int_{w=0}^\infty \mathrm{d}\mathbb{P}(W\le w)\int_{u=w}^{w+B}\frac{f(u)}{r(u)}\mathrm{d}u\right] \\ &= \frac{1}{\mathbb{E}A}\int_{w=0}^\infty \mathrm{d}\mathbb{P}(W\le w)\mathbb{E}B\mathbb{E}\left[\frac{f(w+B^r)}{r(w+B^r)}\right] \\ &= \frac{\mathbb{E}B}{\mathbb{E}A}\mathbb{E}\left[\frac{f(W+B^r)}{r(W+B^r)}\right], \end{aligned} \tag{2.26}$$

where we have used (2.23) in the second equality. The theorem follows by taking $f(x) = r(x)$, leading to $\mathbb{E}[r(V)|V>0]\mathbb{P}(V>0) = \frac{\mathbb{E}B}{\mathbb{E}A}$. □

Again, taking respectively $f(x) = x$, $f(x) = e^{-sx}$, and $f(x) = I(x>v)$ gives a relation between the workload at arbitrary epochs V and the workload at arrival epochs W, for respectively the expectation, the LST, and the tail probabilities.

Taking $f(x) = r(x)g(x)$ yields

$$\mathbb{E}[r(V)g(V)|V > 0] = \mathbb{E}[r(V)|V > 0]\mathbb{E}[g(W + B^r)],$$

and in particular

$$\mathbb{E}\left[r(V)e^{-sV}|V > 0\right] = \mathbb{E}\left[r(V)|V > 0\right]\mathbb{E}\left[e^{-s(W+B^r)}\right]. \tag{2.27}$$

The latter relation implies that the steady-state density of $W + B^r$ is proportional to the product of $r(\cdot)$ and the conditional steady-state density of V.

Using (2.27) we obtain an alternative proof of the level crossing identity (2.2). To show this, we let the arrival process be Poisson with intensity $\lambda(x)$ when the workload equals x. Note that the interarrival time and workload at arrival epochs are dependent; however, we can still apply Theorem 2.5.2 and (2.24). Furthermore, observe that for the choice of $f(x) = r(x)e^{-sx}$, we have that $f(0) = 0$, since we assumed that $r(0) = 0$. Then, by conditioning, it follows directly from (2.26) that $\mathbb{E}[r(V)] = \frac{\mathbb{E}B}{\mathbb{E}A}$, and similarly, we can rewrite (2.27) into

$$\mathbb{E}\left[r(V)e^{-sV}\right] = \frac{\mathbb{E}B}{\mathbb{E}A}\mathbb{E}\left[e^{-s(W+B^r)}\right].$$

Using the one-to-one correspondence between an LST and its inverse again, yields

$$r(x)v(x) = \frac{\mathbb{E}B}{\mathbb{E}A}\int_{y=0^-}^{x} \frac{1 - B(x-y)}{\mathbb{E}B} w(y)\mathrm{d}y,$$

where the 0^- in the integral denotes the inclusion of the (possibly exceptional) point 0. Furthermore, we had proven that $w(y) = \lambda(y)v(y)/\bar{\lambda}$, with $\bar{\lambda} = \int_{0^-}^{\infty} \lambda(y)v(y)\mathrm{d}y$ (see for instance Theorem 2.3.2). Take $f(x) = \lambda(x)$ in (2.19) to see that $1/\bar{\lambda} = \mathbb{E}A$ and the constants cancel,

$$r(x)v(x) = \int_{y=0^-}^{x} (1 - B(x-y))\lambda(y)v(y)\mathrm{d}y, \qquad x > 0.$$

Hence, we have shown the level crossing identity (2.2) in an alternative way.

Remark 2.5.2 Formula (2.25) is also valid when the n-th service requirement B_n is dependent on the workload W_n at its arrival. ◇

Remark 2.5.3 It is worth noting that taking $r(x) \equiv 1$ in (2.27) results in a well-known result for the GI/G/1 queue:

$$V|V > 0 =^d W + B^r,$$

see also [56], p. 296, or [10], p. 274. ◇

Another interesting relation between V and W in the ordinary GI/G/1 queue is presented in Asmussen [10], p. 274:

$$V =^d (W + B - A^r)^+, \tag{2.28}$$

where A^r denotes a residual interarrival time. We now generalize (2.28) to a G/G/1 queue with general service rate $r(x)$ when the workload equals x. By the stochastic mean-value result (2.17) and some similar manipulations as we did proving Theorems 2.5.1 and 2.5.2, one can find the following relation:

$$\begin{aligned}
\mathbb{E}[f(V)] &= \frac{1}{\mathbb{E}A}\mathbb{E}_{A,V_t}\left[\int_{t=0}^{\infty} f(V_t)I(A > t)\mathrm{d}t\right] \\
&= \mathbb{E}_{V_t}\left[\int_{t=0}^{\infty} f(V_t)\frac{\mathbb{P}(A > t)}{\mathbb{E}A}\mathrm{d}t\right] \\
&= \mathbb{E}[f(V_{A^r})].
\end{aligned}$$

Take $f(x) = I(x > v)$, where $I(\cdot)$ is the indicator function, then it follows that $\mathbb{P}(V > v) = \mathbb{P}(V_{A^r} > v)$. The latter probability equals the probability that A^r is less than the time required for a process, which decreases according to the function $r(\cdot)$, to go from $W + B$ (workload at $t = 0$) to v. Recall that $R(x)$ (see (2.1)) represents the time required for a workload x to drain in the absence of any arrivals. Moreover, the time required to go from $W + B$ to v according to the described process equals $R(W + B) - R(v) = R(W + B, v)$. Hence, we obtain

$$\mathbb{P}(V > v) = \mathbb{P}(R(W + B) - R(v) > A^r), \qquad v \geq 0.$$

When $r(x) \equiv 1$, this indeed yields (2.28).

In the remainder of this section we assume that the workload process $\{V_t, t \geq 0\}$ has an atom at zero, or equivalently, that $R(x) < \infty$. Our goal is to study the process $\{R(V_t), t \geq 0\}$. Note that $R(x)$ (like $R(w + x, w)$) is strictly increasing in x so we can speak unambiguously of $R^{-1}(t)$. We are interested in the service (release) process, and assume for the moment that the arrival process is shut off. Besides the time required for a workload u to go down to $x, 0 \leq x \leq u$, in the absence of any arrivals (which equals $R(u) - R(x)$), we are also interested in, for instance, the workload level at time $t > 0$ when $V_0 = u$. It is well-known that the latter expression equals $R^{-1}(R(u) - t)$ [83]. So, in principle it is possible to switch from the workload interpretation to the time interpretation and vice versa. However, there does not seem to be much hope for convenient expressions.

Using the definition of $R(\cdot)$ and the transformation property (2.8) between workload and time, it is easy to see that $R(V)$ transforms the workload V into the time required to finish the work in the system when no arrivals occur. This means that as long as there are no jumps, the process $R(V_t)$ decreases linearly with slope -1 until $R(V_t) = 0$ and then remains 0 until the next arrival. To get some feeling for the $R(V_t)$ process, it is easiest to think of its graphical representation: we have rescaled the workload axis such that in each time interval of length Δx where no arrival occurs, the decrement of the function $R(\cdot)$ is Δx. This means that every very small workload interval $(x, x + r(x)\Delta x)$ of the V_t process is compressed (or expanded) to an interval $(x, x + \Delta x)$. Since $r(\cdot)$ is left-continuous and has right-hand limits, it is bounded on closed intervals and hence the time required to move down from level $x + r(x)\Delta x$ to level x is indeed $r(x)\frac{\Delta x}{r(x)} = \Delta x$ for Δx small enough.

The jump sizes of the $R(V_t)$ process consist of the differences of the time required to empty the system just before, and just after, the arrival epoch. Hence, in steady state this service requirement equals $R(W + B) - R(W)$, or alternatively $R(W + B, W)$. Thus the $R(V_t)$ process behaves like an ordinary G/G/1 queue with workload-dependent service requirements, and follows the same sample path as V_t if we transform the jump size distribution according to the above integral. From the arguments above we can observe that for the G/G/1 queue with service rate $r(x)$ when the workload equals x, we have the following relations between V and W:

Theorem 2.5.3 *If $R(x) < \infty$ for all $0 < x < \infty$, then*

$$R(V) \;=^d\; \big(R(W + B) - A^r\big)^+, \qquad (2.29)$$

$$R(W) \;=^d\; \big(R(W + B) - A\big)^+. \qquad (2.30)$$

Furthermore, if we were able to solve the stationary distribution (density) of the $R(V_t)$ process, denoted by $V^R(\cdot)$ $(v^R(\cdot))$, we would have the stationary distribution (density) of V_t, since $v(x) = \frac{v^R(x)}{r(x)}$.

Remark 2.5.4 Taking $r(x) \equiv 1$ in (2.29) and (2.30) results respectively in (2.28) and the well-known relation for the G/G/1 queue (see for instance [56], p. 167). $\diamond$

Remark 2.5.5 In a similar way like (2.22) we can derive that the expected jump size of the $R(V_t)$ process equals $\mathbb{E}B\mathbb{E}\left[\frac{1}{r(W+B^r)}\right]$, where B^r denotes a residual service time. $\diamond$

Hence, the transformation from a G/G/1 queue with general service rate function $r(\cdot)$ to an ordinary G/G/1 queue (with a server working at unit speed) can be interpreted as a rescaling of the service requirement. In the transformed $R(V_t)$ model, the amount of work a customer brings upon arrival includes the time required to finish this additional workload. If, for instance, $r(x) \equiv r$, we rescale the service requirement by the factor r^{-1} to take into account that the server would have been working at speed r in the V_t process. Note that, in the absence of any arrivals, the time required to finish a workload B at speed r indeed equals the time required to finish a workload $r^{-1}B$ at unit speed.

2.6 Conclusions and topics for further research

We studied single-server queues with state-dependent interarrival times and service speed. The two main contributions of this chapter may be summarized as follows.

Firstly, in the case of Poisson arrivals, we derived proportionality relations between the workload distributions of two queues that have the same ratio of arrival rate and service speed. Such relationships allow us to obtain results for a whole class of models from the analysis of one particular model. Secondly, we analyzed G/G/1-type queues with workload-dependent service speed and interarrival times. Using a Palm-theoretic approach, several well-known relations for the workload at various epochs in the ordinary G/G/1 queue were generalized. Moreover, an extension of the PASTA result to M/G/1 queues with state-dependent arrival rate followed as a by-product.

Finally, we like to mention a topic for further research. In production systems, for example, workload management may be realized by controlling the arrival rate of new jobs, or by regulating the speed of the server. Arrival control to optimize throughput is discussed, under some assumptions on the service speed function, in Chapter 5. Another important issue is the design of the system such that a target steady-state behavior of the workload is achieved. This so-called *reverse engineering* (cf. [70]) is left for a further investigation.

Chapter 3

Finite-buffer queues

3.1 Introduction

In the previous chapter we analyzed the single-server queue with workload-dependent service and arrival rates and an infinite buffer. In particular, we distinguished between general and exponential interarrival times. In this chapter, we extend the latter (Markovian) model to a model with finite buffer. The renewal case is further addressed in Chapter 4, where we consider the relation between the loss probability and the distribution of the cycle maximum in its infinite-buffer counterpart. We refer to Chapter 1 for an exposition on the applicability of queueing systems with state-dependent service and arrival rates.

Various queueing models with restricted accessibility have been considered in the literature. In this chapter, the admission of customers typically depends on the amount of work upon arrival in addition to their own service requirements. In such systems, we may distinguish three main admission rules: (i) the finite dam, governed by the partial-rejection rule (scenario f), (ii) systems with impatience of customers depending on the amount of work found upon arrival (scenario i), and (iii) queues regulated by the complete-rejection discipline (scenario c).

The three main goals of the present chapter are the following. First, we establish relationships between two queueing models with arrival rates $\lambda_i(x)$ and service speeds $r_i(x), i = 1, 2$, for which $\frac{\lambda_1(x)}{r_1(x)} = \frac{\lambda_2(x)}{r_2(x)}, \forall x > 0$, thus extending results from Chapter 2 to queues with restricted accessibility. These relationships will allow us to obtain results for a whole class of models from the analysis of one particular model. This is particularly useful in considering performance measures such as steady-state workload densities and loss probabilities.

Our second goal is to obtain an explicit (formal) expression for the cycle maximum in an M/G/1 queue with workload-dependent service and arrival rate. This may be useful to determine the maximum buffer size. Exact results for such systems are hardly known; we refer to Asmussen [7] for an overview on cycle maxima.

Third, we derive a formal solution for the steady-state workload density in finite-buffer M/G/1 systems. The density may be expressed as the solution of a Volterra integral equation. In some special cases, this reduces to an analytically tractable expression. Otherwise, numerical methods are widely available, see e.g. [88, 116]. Another tool to solve the workload density is the proportionality of the workload distribution of systems with finite and infinite buffer capacities. This relation is well-known for some traditional queueing models (where work is depleted at unit rate), see [33, 52, 56, 85] and specifically Theorem 1.6.3. Using a similar sample-path approach as in Subsection 1.6.3, the proportionality relation is extended to similar systems with workload-dependent arrival and service rate.

In ordinary queueing systems, the workload just before a customer arrival represents a waiting time, and the workload right after an arrival instant may be identified with a sojourn time. For such models, the rejection rules have a direct interpretation. Our first discipline, the finite dam (scenario f), represents a system where every customer has a bounded sojourn time; a rejected customer may enter the system but its sojourn time is restricted such that it exactly equals the buffer size (i.e., partial rejection). This model is also commonly used in the context of inventory and storage processes. Due to the above-mentioned proportionality, the finite dam is closely related to the infinite-buffer version of the model and a detailed analysis can be found in, e.g., [52, 56].

Under the second rejection discipline, scenario i, customers are only willing to wait a limited amount of time. Results are also well-known for ordinary queueing models, see e.g. [33, 49, 61, 62]. In queues with general service speeds, the workload found upon arrival does in general not equal the waiting time. However, these two quantities are closely related and the admission may depend on the workload upon arrival.

Finally, the third discipline, scenario c, also considers the case in which customers have a limited sojourn time. In contrast to scenario f, rejected customers are completely discarded and do not join the queue. Results are only known for the M/M/1 and M/D/1 case (see e.g. [49, 81]), and the Ph/Ph/1 case [117]. Asymptotics for more general models are obtained in [178].

This chapter is organized as follows. In Section 3.2 we introduce the general model. The results for the Markovian case of Chapter 2 are summarized and extended to finite-buffer queues in Section 3.3. In Section 3.4, the finite dam (scenario f) is studied and the proportionality relation between finite and infinite-buffer systems is presented. We also briefly consider scenarios i and c. First-exit probabilities and cycle maxima are considered in Section 3.5, and we conclude with some examples in Section 3.6.

3.2 Model description and preliminaries

In this section we introduce the general model and obtain some preliminary results. Some examples of canonical finite-buffer models are given at the end of the section.

We first describe the general system. Customers arrive at a queueing system according to a Poisson process with arrival rate $\lambda(x)$ when the workload equals

$x, x \geq 0$; in other words, the probability of an arrival in some interval $(t, t + h)$ equals $\lambda(x)h + o(h)$ for $h \downarrow 0$ when the work present at time t equals x. We assume that $\lambda(\cdot)$ is nonnegative, left-continuous and has a right limit on $[0, \infty)$. The service requirement of customer n is denoted by $B_n, n = 1, 2, \ldots$, where $B_1, B_2, \ldots$ are assumed to be independent, identically distributed with distribution $B(\cdot)$, independent of the sequence of interarrival times.

Depending on the service requirement and the amount of work found upon arrival, customers may or may not be (fully) accepted. In particular, if the workload just before an arrival equals w, and the service requirement is b, then the amount of work right after the arrival instant is $g(w, b, K)$. We assume that $w \leq g(w, b, K) \leq w + b$, where K represents the size of a possibly finite buffer (see the end of this section for some examples).

We allow the server to operate according to a general service-rate (speed) function, a function of the amount of work present. We denote the service-rate function by $r : [0, \infty) \to [0, \infty)$, assume that $r(0) = 0$ and that $r(\cdot)$ is strictly positive, left-continuous, and has a right limit on $(0, \infty)$.

In the general model, we define V_t^g as the workload at time t (with distribution function $V_t^g(\cdot)$) and let W_n^g be the workload immediately before the n-th arrival epoch. We denote the steady-state random variables corresponding to V_t^g and W_n^g by V^g and W^g, respectively, and let $V^g(\cdot)$ and $W^g(\cdot)$ denote their distributions, and $v^g(\cdot)$ and $w^g(\cdot)$ their densities. In the present chapter, it is assumed that $\lambda(\cdot), r(\cdot), B(\cdot)$ are chosen such that the steady-state distribution of the infinite-buffer version, that is, for $g(w, b, K) = w + b$, exists (and then for all $g(\cdot, \cdot, \cdot)$). For details on stability and existence of steady-state distributions, we refer to [45, 46].

Define

$$R(x) := \int_0^x \frac{1}{r(y)} \mathrm{d}y, \qquad 0 < x < \infty,$$

representing the time required for the system to become empty in the absence of any arrivals, starting with workload x. Note that $R(x) < \infty$, for all $x > 0$, means that state zero can be reached in a finite amount of time from any state $x > 0$. A related quantity is

$$\Lambda(x) := \int_0^x \frac{\lambda(y)}{r(y)} \mathrm{d}y, \qquad 0 < x < \infty,$$

which determines whether the workload process of the infinite-buffer version of the queue has an atom at state zero. In case of finite buffers, some modification is required to regulate the workload behavior for states that can not be attained. Specifically, set $r(x) \equiv 1$ and $\lambda(x) \equiv \lambda$ for all $x > 0$ for which $\mathbb{P}(g(y, B, K) > x) = 0$, for all $0 \leq y < x$. Then the workload process has indeed an atom at state zero if and only if $\Lambda(x) < \infty$ for all $0 < x < \infty$, as in the infinite-buffer queue, see Section 2.2. If we take $\lambda(\cdot)$ fixed ($\lambda(x) \equiv \lambda$), then $\Lambda(x) = \lambda R(x)$ and we refer to Asmussen [10, Ch. XIV] and Brockwell *et al.* [45] for more details.

Furthermore, consider the interarrival time and its corresponding workload decrement, i.e., the amount of work served during the interarrival time. Denote by A_y the conditional workload decrement during the interarrival interval starting with workload y, i.e., the event $\{A_y > v\}$ means that the workload is smaller than $y - v$ upon the arrival of the next customer. For convenience, we remove the superscript w of Chapter 2 that was intended to stress the fact that we consider a workload decrement instead of an interarrival interval. Note that the time required to move from y down to v in the absence of any arrivals equals $R(y) - R(v)$. Since $r(x) > 0$ for all $x > 0$, it follows that $R(\cdot)$ is strictly increasing, which implies a one-to-one correspondence between the interarrival time and its corresponding workload decrement.

The conditional distribution of the workload decrement during an interarrival interval was obtained in Chapter 2. For completeness, we recall Proposition 2.3.1:

Proposition 3.2.1 *Let the workload just after an arrival be y ($g(w, b, K) = y$); then, for $y > v$,*

$$\mathbb{P}(A_y > v) = e^{-\int_{u=y-v}^{y} \frac{\lambda(u)}{r(u)} \mathrm{d}u}.$$

Returning to the workload process $\{V_t^g, t \geq 0\}$, we may define the process right before jump epochs recursively, by

$$W_{n+1}^g = \max(g(W_n^g, B_n, K) - A_{n,g(W_n^g, B_n, K)}, 0), \tag{3.1}$$

where $A_{n,g(\cdot,\cdot,\cdot)}$ is the workload decrement between the arrival of the n-th and $(n+1)$-th customer, depending on the workload right after the n-th jump epoch. In between jumps, the workload process is governed by the service rate function, and satisfies

$$\frac{\mathrm{d}V_t^g}{\mathrm{d}t} = -r(V_t^g).$$

We refer to Harrison and Resnick [83] for a further discussion of the system dynamics.

Kolmogorov equations

The Kolmogorov forward equations of the workload process can now be constructed using a similar approach as sketched in Subsection 1.6.1, see also [159], [56], p. 263. First, note that in the interval $(t, t + \Delta t)$ a new customer may arrive with probability $\lambda(V_t^g)\Delta t + o(\Delta t)$ as $\Delta t \downarrow 0$. Then, conditioning on the amount of work at time t, we deduce for $x, t > 0$ and some finite constant $\theta \geq 0$,

$$\begin{aligned} V_{t+\Delta t}^g(x) &= \int_{0^-}^{x} (1 - \lambda(y + r(y)\Delta t)\Delta t)\mathrm{d}_y V_t^g(y + r(y)\Delta t) \\ &\quad + \int_{0^-}^{x} \lambda(y + \theta\Delta t)\Delta t \mathbb{P}(g(y, B, K) \leq x)\mathrm{d}_y V_t^g(y + \theta\Delta t) + o(\Delta t). \end{aligned}$$

Now, write for $\Delta t \to 0$ (see also [56, 159]),

$$\begin{aligned} V_{t+\Delta t}(x) &= V_t(x) + \Delta t \frac{\partial}{\partial t} V_t(x) + o(\Delta t), \\ V_t(x + r(x)\Delta t) &= V_t(x) + r(x)\Delta t \frac{\partial}{\partial x} V_t(x) + o(\Delta t). \end{aligned}$$

Using similar arguments as in Subsection 1.6.1, we find the Kolmogorov forward equation of the process:

$$\frac{\partial}{\partial t}V_t^g(x) = r(x)\frac{\partial}{\partial x}V_t^g(x) - \int_{0^-}^{x} \lambda(y)\mathbb{P}(g(y,B,K) > x)\mathrm{d}_y V_t^g(y).$$

Letting $t \to \infty$ and using the fact that $v^g(\cdot)$ denotes a density (see [45, 46] for details on existence and uniqueness of a steady-state solution in case of an infinite buffer), we have for $x > 0$,

$$r(x)v^g(x) = \lambda(0)V^g(0)\mathbb{P}(g(0,B,K) > x) + \int_{0^+}^{x} \lambda(y)v^g(y)\mathbb{P}(g(y,B,K) > x)\mathrm{d}y. \tag{3.2}$$

This equation is also well-known as the level crossing equation, see [68] and Section 2.5 for alternative proofs. It reflects the fact that the rate of crossing level x from above should equal, in steady-state, the rate of crossing level x from below.

Special cases

As mentioned in Section 3.1, three important special cases of the general model are finite-buffer dams (scenario f), systems with customer impatience (scenario i), and queues regulated by the complete-rejection discipline (scenario c). The finite-buffer dam, regulated by the partial-rejection discipline, originates from the study of water dams. The content of a dam is finite and additional water just overflows. In the context of queueing, this implies that the n-th arriving customer is admitted to the system if and only if $W_n + B_n \leq K$. However, a partially rejected (not fully accepted) customer may enter the system, but with restricted service requirement $K - W_n$.

Models with customer impatience stem from ordinary queueing systems, with a server working at unit speed. In that case, the workload upon arrival identifies a waiting time and the impatience is represented by the fact that customers are only willing to wait a limited amount of time K. In case of general service speeds, the n-th arriving customer is accepted if $W_n \leq K$ and fully rejected otherwise (see [33] for some potential applications). Finally, in the system with complete rejections, the n-th customer is admitted if $W_n + B_n \leq K$, and totally rejected otherwise.

A further special case is the queue with infinite buffer. This model is discussed in Chapter 2 and is simply the model where every customer is completely accepted.

Summarizing, these four scenarios may be represented as follows:

$$g(w,b,K) = \begin{cases} w + b, & \text{infinite-buffer queue,} \\ \min(w+b, K), & \text{scenario } f\text{; finite dam,} \\ w + bI(w \leq K), & \text{scenario } i\text{; customer impatience,} \\ w + bI(w + b \leq K), & \text{scenario } c\text{; complete rejection discipline.} \end{cases}$$

Here $I(\cdot)$ denotes the indicator function. Finally, we indicate the notational conventions arising from the models. If we consider an arbitrary $g(\cdot,\cdot,\cdot)$, we add an index g. The infinite-buffer system is denoted by just omitting the g from the definitions of the general model (as in Chapter 2). The finite-buffer dam may be obtained by substituting K for g. The models with customer impatience and complete rejection are given by writing K, i and K, c for g, respectively.

3.3 Relations between two finite-buffer queues

In this section, we analyze the workload relations between two (general) finite-buffer queues that have the same ratio between arrival and service rate. The present section also provides some results for infinite-buffer queues that we earlier found in Chapter 2. Specifically, we extend the relations of Theorems 2.3.1–2.3.3 to queues with finite buffer. As described above, the infinite-buffer queue of Chapter 2 is just a special case of the general setting studied here: Choose $g(w, b, K) = w + b$, for all $w, b \geq 0$. In addition, the formal solution of the steady-state workload density is considered. However, we start by studying the relation between workloads at arrival instants and arbitrary epochs.

In view of loss probabilities, the relation between the workload at jump epochs and arbitrary epochs is significant. The following theorem extends Theorem 2.3.2 to the general setting.

Theorem 3.3.1 *Define the average arrival rate as* $\bar{\lambda}^g := \int_{0+}^{\infty} \lambda(x) v^g(x) \mathrm{d}x + \lambda(0) V^g(0)$. *Then,* $W^g(0) = \lambda(0) V^g(0) / \bar{\lambda}^g$ *and, for all* $x > 0$,

$$w^g(x) = \frac{1}{\bar{\lambda}^g} \lambda(x) v^g(x).$$

Proof Observe that $g(w, b, K) \leq w + b$ ensures that the expected cycle length is finite and the workload process is thus ergodic. By level crossing theory, it then follows that the workload density is well-defined. Moreover, $g(w, b, K) \geq w$ rules out scenarios of work removal. Now, substitute $g(W^g, B, K)$ for every $W + B$ in Theorem 2.5.1 and the results easily follow. □

Note that $\lambda(x) \equiv \lambda$ would yield the PASTA property, which states that the workload at an arbitrary time and the workload at an arrival epoch have the same distribution. With respect to workload as the key performance measure, Theorem 3.3.1 may be viewed as a generalization of the PASTA property.

Theorems 2.3.1 and 2.3.3 indicate that two infinite-buffer models, with identical ratios between arrival and release rate, can be related. This relationship between two different M/G/1 queues can be extended to the general model, as presented in the next theorem.

Theorem 3.3.2 *Consider two queueing models as defined in Section 3.2, to be denoted as Models* 1 *and* 2*, such that* $\lambda_1(x)/r_1(x) = \lambda_2(x)/r_2(x)$*, for all* $x > 0$. *Then,* $W_1^g(0) = W_2^g(0)$*, and for all* $x > 0$,

$$w_1^g(x) = w_2^g(x). \tag{3.3}$$

Also,

$$\frac{v_1^g(x)}{v_2^g(x)} = C\frac{r_2(x)}{r_1(x)}, \tag{3.4}$$

with $C = \frac{\lambda_1(0)V_1^g(0)}{\lambda_2(0)V_2^g(0)}$ *if* $\Lambda_i(x) < \infty$ *for all* $0 < x < \infty$*, and* $C = 1$ *if* $\Lambda_i(x) = \infty$ *for some* $0 < x < \infty$*.*

Before we prove the above theorem, we first derive the steady-state workload density. Besides the formal solution of this density being a slight extension of infinite-buffer results, it turns out to be a useful tool to express the workload density in a more elegant form in some special cases. Moreover, Equation (3.4) follows then directly by division.

Now, consider either of the two models and assume for the moment that the workload process has an atom at state 0. The Kolmogorov forward equations (in the thesis also referred to as level crossing equations) of the workload process are given by (3.2). Note that in many finite-buffer systems, the workload is bounded by the capacity K, as in scenarios f and c. In that case, $\mathbb{P}(g(y,B,K) > K) = 0$, and we only have to consider $0 < x \leq K$. In scenario i, for example, cases with workloads above K may exist. However, jumps occur only from workloads smaller than K, and the range of integration can be modified to $(0, \min(x, K)]$.

Define $z(y) := \lambda(y)v^g(y)$ and multiply both sides of (3.2) by $\lambda(x)/r(x)$. We then obtain

$$z(x) = \frac{\lambda(x)}{r(x)}\lambda(0)V^g(0)\mathbb{P}(g(0,B,K) > x) + \int_{0^+}^{x} \frac{\lambda(x)}{r(x)} z(y)\mathbb{P}(g(y,B,K) > x)\mathrm{d}y. \tag{3.5}$$

We now proceed as in Subsection 1.6.1 (see also [83]): Define the kernel $K^g(x,y) := K_1^g(x,y) := \mathbb{P}(g(y,B,K) > x)\lambda(x)/r(x), 0 \leq y < x < \infty$, and let its iterates be defined by

$$K_{n+1}^g(x,y) := \int_y^x K^g(x,z)K_n^g(z,y)\mathrm{d}z.$$

Note that in, for instance, the infinite-buffer system, $\mathbb{P}(g(y,B,K) > x) = 1 - B(x-y)$, and thus $K(x,y) = (1-B(x-y))\lambda(x)/r(x)$ similar to Section 2.3. Moreover, observe that (3.5) is a Volterra integral equation of the second kind, and rewrite it as

$$z(x) = \lambda(0)V^g(0)K^g(x,0) + \int_{0^+}^{x} z(y)K^g(x,y)\mathrm{d}y. \tag{3.6}$$

Iterate this relation $N-1$ times, for some $N \in \mathbb{N}$, (see [83]):

$$z(x) = \lambda(0)V^g(0)\sum_{n=1}^{N} K_n^g(x,0) + \int_{0^+}^{x} z(y)K_N^g(x,y)\mathrm{d}y.$$

Finally, define

$$K^{g,*}(x,y) := \sum_{n=1}^{\infty} K_n^g(x,y). \tag{3.7}$$

If this sum is well-defined, we have $z(x) = \lambda(0)V^g(0)K^{g,*}(x,0)$. However, we may use the obvious bound $K^g(x,y) \leq \lambda(x)/r(x)$ to show inductively that $K^g_{n+1}(x,y) \leq (\Lambda(x) - \Lambda(y))^n\lambda(x)/(r(x)n!)$. Hence, the infinite sum is indeed well-defined and $\int_{0+}^{x} z(y)K^g_N(x,y)\mathrm{d}y \to 0$ as $N \to \infty$. Now use the definition of $z(\cdot)$ to obtain

$$v^g(x) = \frac{\lambda(0)V^g(0)K^{g,*}(x,0)}{\lambda(x)}, \tag{3.8}$$

where $V^g(0)$ follows from normalization. The steady-state workload density in our general setting is presented in the following lemma.

Lemma 3.3.1 *If* $\Lambda(x) < \infty$ *for all* $0 < x < \infty$*, then*

$$v^g(x) = \frac{\lambda(0)V^g(0)K^{g,*}(x,0)}{\lambda(x)}, \tag{3.9}$$

where $V^g(0) = \left[1 + \lambda(0)\int_0^\infty \frac{K^{g,*}(x,0)}{\lambda(x)}\mathrm{d}x\right]^{-1}$.

As indicated earlier, the workload density in the infinite-buffer case may be obtained by defining $K(x,y) = (1 - B(x-y))\lambda(x)/r(x)$, see also Section 2.3. In the remainder, we refer to this kernel as the basic kernel as it appears in many queueing systems. The iterates $K_n(x,y)$ and the infinite sum $K^*(x,y)$ are defined accordingly.

In Section 3.4, we indicate how this general approach may be applied to some finite-buffer queues. In Section 3.6, the infinite sum of Volterra kernels is explicitly calculated for some special cases. But, first we use this lemma to derive Equation (3.4).

Proof of Theorem 3.3.2 Observe that, by Proposition 3.2.1 and (3.1), the dynamics of both systems are equivalent when $\frac{\lambda_1(x)}{r_1(x)} = \frac{\lambda_2(x)}{r_2(x)}$. Hence, using a stochastic coupling argument, the first part of the theorem, that is (3.3), easily follows.

We now turn to (3.4). Note that $\Lambda_1(x) = \Lambda_2(x)$, implying that either the workload processes in both systems have an atom at state zero, or in neither system. If $\Lambda_i(x) = \infty$ $(i = 1,2)$ for some $x > 0$, then $V^g_i(0) = 0$ and (3.4) follows directly from (3.5) and the definition of $z(\cdot)$. So, assume that $V^g_i(0) > 0$. We use the derivation of the steady-state workload density as described above. First, observe that the kernels $K^g(x,y) = \mathbb{P}(g(y,B,K) > x)\lambda(x)/r(x)$ are the same in both models, and hence, the iterated kernels and their infinite sums are equal. Now, use (3.9) and divide $v^g_1(x)$ by $v^g_2(x)$ to obtain

$$\frac{v^g_1(x)}{v^g_2(x)} = \frac{\lambda_1(0)V^g_1(0)}{\lambda_2(0)V^g_2(0)}\frac{\lambda_2(x)}{\lambda_1(x)}, \qquad x > 0.$$

Substituting $\lambda_2(x)/\lambda_1(x) = r_2(x)/r_1(x)$ completes the proof. □

Remark 3.3.1 There are alternative ways to solve (3.2). We may also divide by $r(x)$ on both sides of Equation (3.2), or define $z(x) := r(x)v^g(x)$. The technique to solve the integral equation remains the same, however, with slightly different kernels. ◇

3.4 Finite-buffer queues

In this section, we study the steady-state workload distribution in some finite-buffer M/G/1 systems with general arrival rate and service speed. In the first part, we consider the M/G/1 dam (scenario f) and show that the workload distribution is proportional to the workload distribution in its infinite-buffer counterpart. In the second part, we note that the same proportionality also holds for scenario i, and we conclude with some remarks about scenario c.

3.4.1 Finite-buffer dam

Consider the M/G/1 queue under the partial rejection discipline (scenario f), that is, take $g(w, b, K) = \min(w + b, K)$. For convenience, we also refer to scenario f as the finite-buffer queue or dam.

First, we show that the steady-state workload distributions in the finite and infinite-buffer dam are proportional. For instance, Hooghiemstra [85] based his proof for the ordinary M/G/1 queue on the insight that the finite and infinite-buffer queue follow similar sample paths below workload level K, see also Subsection 1.6.3 for an explanation. He observed that at a downcrossing of level K in the infinite-buffer queue, the time until the next arrival epoch is independent of the previous arrival, and hence, the residual interarrival time behaves like an ordinary one. As required in the sample path comparison, we show that this lack of memory also holds for general M/G/1-type queues with state-dependent arrival and service rates. After making some comments about regenerative properties, we extend the result of Hooghiemstra to our system, using similar arguments. Moreover, the steady-state workload distribution at arrival epochs is considered. This is no longer necessarily equal to the workload distribution at arbitrary epochs, since the classical PASTA property no longer holds.

Second, we give the steady-state workload density for the finite dam. Third, we consider the long-run fraction of not fully accepted customers, denoting this performance measure by P_K.

The next preparatory lemma presents the lack-of-memory property of the workload decrement during an interarrival interval.

Lemma 3.4.1 (**Memoryless property**). *The residual workload decrement at a downcrossing of level x in an M/G/1 queue with arrival rate $\lambda(\cdot)$ and service rate $r(\cdot)$ is independent of the finished amount of work during the elapsed interarrival time, i.e.,*

$$\mathbb{P}(A_{x+y} > y + v | A_{x+y} > y) = \mathbb{P}(A_x > v), \qquad x, y, v > 0, \quad x > v.$$

Proof Using a simple conditioning argument and Proposition 3.2.1, it follows that

$$\begin{aligned}\mathbb{P}(A_{x+y} > y+v | A_{x+y} > y) &= e^{-\int_{x-v}^{x+y} \frac{\lambda(u)}{r(u)}\mathrm{d}u} e^{\int_{x}^{x+y} \frac{\lambda(u)}{r(u)}\mathrm{d}u} \\ &= e^{-\int_{x-v}^{x} \frac{\lambda(u)}{r(u)}\mathrm{d}u} \\ &= \mathbb{P}(A_x > v).\end{aligned}$$

Notice that $\mathbb{P}(A_{x+y} > y+v | A_{x+y} > y)$ is independent of y, representing the lack of memory. □

Next, we state our main proportionality result.

Theorem 3.4.1 *For* $0 \leq x \leq K$,

$$\mathbb{P}(V^K \leq x) = \frac{\mathbb{P}(V \leq x)}{\mathbb{P}(V \leq K)}, \tag{3.10}$$

while at arrival epochs,

$$\mathbb{P}(W^K \leq x) = \frac{\mathbb{P}(W \leq x)}{\mathbb{P}(W \leq K)}.$$

We like to emphasize that $\mathbb{P}(V \leq x)$ and $\mathbb{P}(W \leq x)$ refer to the unbounded case, cf. Chapter 2.

Before proving the theorem, we first make some general remarks about regenerative processes. Instead of applying level crossing arguments and using Lemma 3.3.1, it is also possible to make a direct comparison between the finite and infinite-buffer queues as depicted in Subsection 1.6.3. We apply the latter approach. Following Asmussen [10], we exploit the regenerative nature of the workload process and let the downcrossings of workload level K be its regeneration points. Note that this is possible due to the memoryless property (Lemma 3.4.1). Furthermore, this choice allows queueing systems where empty queues cannot occur. Denote the length of a regeneration cycle in the finite and infinite-buffer queues and the number of arrivals during this cycle, by τ^K, τ, N^K, and N, respectively. Then, the distributions of V and V^K are given by, cf. [52],

$$\mathbb{P}(V \leq x) = \frac{1}{\mathbb{E}\tau}\mathbb{E}\left[\int_0^{\tau} I(V_t \leq x)\mathrm{d}t\right], \tag{3.11}$$

$$\mathbb{P}(V^K \leq x) = \frac{1}{\mathbb{E}\tau^K}\mathbb{E}\left[\int_0^{\tau^K} I(V_t^K \leq x)\mathrm{d}t\right]. \tag{3.12}$$

The distributions of W and W^K can be obtained in a similar fashion, cf. [52],

$$\begin{aligned}\mathbb{P}(W \leq x) &= \frac{1}{\mathbb{E}N}\mathbb{E}\left[\sum_{i=1}^{N} I(W_i \leq x)\right], \\ \mathbb{P}(W^K \leq x) &= \frac{1}{\mathbb{E}N^K}\mathbb{E}\left[\sum_{i=1}^{N^K} I(W_i \leq x)\right].\end{aligned}$$

We are now ready to prove our main theorem.

Proof of Theorem 3.4.1 Consider the stochastic process $\{V_t, t \geq 0\}$. We construct a stochastic process $\hat{V}_t^K$ directly from V_t and show that $\hat{V}_t^K$ and V_t^K are governed by the same probabilistic laws. First, take an arbitrary sample path of V_t. We leave the parts below level K unchanged and delete the parts of the sample path between each upcrossing and a subsequent downcrossing of level K. Connecting the remaining parts, we obtain the process $\hat{V}_t^K$. By the lack-of-memory property, the workload decrement of $\hat{V}_t^K$ after hitting K behaves like an ordinary workload decrement. Thus $\hat{V}_t^K$ and V_t^K have the same statistical properties and we may simplify notation by identifying the process $\{V_t^K, t \geq 0\}$ with $V_t^K := \hat{V}_t^K$, $t \geq 0$.

Clearly, $\mathbb{E}\left[\int_0^\tau I(V_t \leq x)\mathrm{d}t\right]$ and $\mathbb{E}\left[\int_0^{\tau^K} I(V_t^K \leq x)\mathrm{d}t\right]$ are equal. Observe that $\frac{\mathbb{E}\tau^K}{\mathbb{E}\tau}$ represents the long-run fraction of time that the workload process of the infinite-buffer queue is below level K and, by (3.11) and (3.12), we have shown the first part of the theorem. The second part follows directly from the same sample path construction and the observation that $\frac{\mathbb{E}N^K}{\mathbb{E}N}$ equals the long-run fraction of arrivals of the infinite-buffer queue finding the workload below level K. This completes the proof. □

Remark 3.4.1 Theorem 3.4.1 remains valid for any other model where the virtual waiting time process remains unchanged below level K (Hooghiemstra [85] noted this already for the ordinary M/G/1 queue). Specifically, the theorem applies for any function $g(w, b, K)$, if for all $w, b, K \geq 0$ the function satisfies

$$\begin{aligned} g(w,b,K) &= w + b, \quad \text{if } w + b \leq K, \\ g(w,b,K) &\geq K, \qquad\;\; \text{if } w + b > K. \end{aligned} \tag{3.13}$$

◇

In the remainder of the chapter, we assume that the workload process has an atom at state 0, i.e., $\Lambda(x) < \infty$, for all $0 < x < \infty$. Under this assumption, using Theorem 3.4.1, the workload density for the finite dam may be obtained from the workload density in the infinite-buffer version. Hence, applying Theorem 3.4.1 and the result for the infinite-buffer queue (see, e.g., Sections 3.3 or 2.3), we have

$$v^K(x) = \frac{\lambda(0)V^K(0)K^*(x,0)}{\lambda(x)}, \qquad 0 < x \leq K, \tag{3.14}$$

where $V^K(0)$ follows from normalization.

Note that (3.10) may also be derived from the formal solution of the density (Lemma 3.3.1). However, we believe that the derivation of Theorem 3.4.1 is especially insightful as it brings out the typical sample-path relation between the infinite-buffer queue and the finite dam (scenario f).

We conclude by analyzing the probability that a customer cannot be completely accepted, also referred to as loss probability. It follows directly from a regenerative argument, see also [176], that

$$P_K = \mathbb{P}(W^K + B > K).$$

Condition on W^K and apply Theorem 3.3.1 to obtain the following corollary:

Corollary 3.4.1 *The loss probability in scenario f is given by*

$$P_K = \frac{1}{\bar{\lambda}^K}\left[\lambda(0)V^K(0)(1-B(K)) + \int_{0^+}^{K} \lambda(y)v^K(y)(1-B(K-y))\mathrm{d}y\right],$$

with $v^K(\cdot)$ given in (3.14) and $V^K(0)$ equal to $1-\int_0^K v^K(x)\mathrm{d}x$.

Other performance measures may be directly obtained from the workload density and Theorem 3.3.1.

3.4.2 Other finite-buffer systems

Two other finite-buffer systems of importance are models with customer impatience (scenario i) and queues governed by the complete-rejection discipline (scenario c). In this subsection, we first examine scenario i and observe that for workloads less than K the proportionality relation (Theorem 3.4.1) holds. To determine the density for workloads larger than K, in addition to the normalizing constant, we apply level crossings and the successive-substitution method for Volterra integral equations. We conclude the study of scenario i by considering the loss probability. Turning to the second model, scenario c, we derive a formal solution for the steady-state workload density using similar techniques as for scenario i. Finally the loss probability in scenario c is considered.

For scenario i, let $g(w,b,K) = w + bI(w < K)$ in the general set-up. Hence, customers join the queue only when the workload just before arrival is smaller than K. This resembles the impatience of arriving customers: They are only willing to wait a maximum (stochastic) amount of time. As noted in [33, 85], the virtual waiting time process below level K remains unchanged for this model. This intuitive statement can be made rigorous by observing that for all $0 \leq w \leq K$, $g(w,b,K) = w + b$ and thus (3.13) is satisfied. We consequently have the following (see also Remark 3.4.1):

Corollary 3.4.2 *For $0 \leq x \leq K$, we have,*

$$\mathbb{P}(V^{K,i} \leq x) = c^{K,i}\mathbb{P}(V \leq x), \tag{3.15}$$

with $c^{K,i}$ some normalizing constant, $\mathbb{P}(V \leq K) \leq (c^{K,i})^{-1} \leq 1$, while at arrival epochs of accepted customers (thus given $W^{K,i} \leq K$),

$$\mathbb{P}(W^{K,i} \leq x | W^{K,i} \leq K) = \frac{\mathbb{P}(W \leq x)}{\mathbb{P}(W \leq K)}.$$

Remark 3.4.2 By a simple division and using (3.15) and (3.10) twice, we may alternatively write, for $0 \leq x, y \leq K$,

$$\frac{\mathbb{P}(V \leq x)}{\mathbb{P}(V \leq y)} = \frac{\mathbb{P}(V^{K,i} \leq x)}{\mathbb{P}(V^{K,i} \leq y)} = \frac{\mathbb{P}(V^K \leq x)}{\mathbb{P}(V^K \leq y)}. \tag{3.16}$$

◇

However, the workload distribution in the infinite-buffer case does not completely determine the workload distribution in scenario i. The normalizing constant can only be obtained by knowledge of the workload behavior on all possible levels of the process. For $x > K$, we apply level crossings (3.2) and the result for $x \in [0, K)$ to express the workload density in terms of the basic kernel.

Next, we derive the steady-state workload distribution for all $x \geq 0$ in scenario i, using the general approach described in Section 3.3. If the workload upon arrival is below level K, thus $0 \leq w \leq K$, then we just have $\mathbb{P}(g(w, B, K) > x) = 1 - B(x - w)$, while $\mathbb{P}(g(w, B, K) > w) = 0$ otherwise. The general level crossing equations can now be rewritten into a more appealing expression: For $0 < x \leq K$, we have

$$r(x)v^{K,i}(x) = \lambda(0)V^{K,i}(0)(1 - B(x)) + \int_{0^+}^{x} \lambda(y)v^{K,i}(y)(1 - B(x - y))\mathrm{d}y,$$

and for $x > K$,

$$r(x)v^{K,i}(x) = \lambda(0)V^{K,i}(0)(1-B(x))+\int_{0^+}^{K} \lambda(y)v^{K,i}(y)(1-B(x-y))\mathrm{d}y. \quad (3.17)$$

These equations can be solved using Lemma 3.3.1, by defining the kernel $K^i(x, y) := I(y < K)(1 - B(x - y))\lambda(x)/r(x)$, for $0 \leq y < x < \infty$. In case $0 < y \leq K$, we just obtain the basic kernel $K(x, y)$ of Section 3.4. By Lemma 3.3.1, it is thus evident that, for $0 \leq x \leq K$,

$$z^{K,i}(x) = \lambda(0)V^{K,i}(0)K^*(x, 0), \quad (3.18)$$

where $z^{K,i}(x) := \lambda(x)v^{K,i}(x)$. The same result can be deduced from Theorem 3.4.1 and (3.14).

The case $x > K$ may be derived in a slightly more elegant fashion; rewrite (3.17) into

$$z^{K,i}(x) = \lambda(0)V^{K,i}(0)K(x, 0) + \int_{0^+}^{K} z^{K,i}(y)K(x, y)\mathrm{d}y. \quad (3.19)$$

Substituting the result of $z^{K,i}(y)$ for $y \leq K$ in (3.19), we obtain

$$z^{K,i}(x) = \lambda(0)V^{K,i}(0)\left[K(x, 0) + \int_{0}^{K} K(x, y)K^*(y, 0)\mathrm{d}y\right],$$

after which $v^{K,i}(x) = z^{K,i}(x)/\lambda(x)$ and $V^{K,i}(0)$ can be determined by normalization.

For completeness, we give the resulting normalizing constant $c^{K,i}$ in general terms (take $y = 0$ in (3.16)):

$$\frac{1 + \int_0^\infty \frac{\lambda(0)}{\lambda(x)}K^*(x, 0)\mathrm{d}x}{1 + \int_0^K \frac{\lambda(0)}{\lambda(x)}K^*(x, 0)\mathrm{d}x + \int_K^\infty \left[\frac{\lambda(0)}{r(x)}\overline{B}(x) + \int_0^K \frac{\lambda(0)}{r(x)}\overline{B}(x-y)K^*(y, 0)\mathrm{d}y\right]\mathrm{d}x},$$

with $\overline{B}(x) := 1 - B(x)$.

Remark 3.4.3 The cases $0 < x \leq K$ and $x > K$ may be combined by writing

$$v^{K,i}(x) = \frac{\lambda(0)V^{K,i}(0)}{\lambda(x)}\left[K(x,0) + \int_0^{x\wedge K} K(x,y)K^*(y,0)\mathrm{d}y\right]. \tag{3.20}$$

Equation (3.18) can then be recovered by using $K^*(y,0) = \sum_{n=1}^{\infty} K_n(y,0)$ and interchanging integral and sum. ◇

Finally, it is an easy exercise to determine the long-run fraction of rejected customers P_K^i. After all, the customers that are rejected are just those that arrive while the workload is above level K, or more formally $P_K^i = \mathbb{P}(W^{K,i} > K)$. Apply Theorem 3.3.1 to see that

$$\begin{aligned} P_K^i &= \int_K^\infty w^{K,i}(x)\mathrm{d}x \qquad (3.21)\\ &= \frac{1}{\overline{\lambda}^{K,i}}\int_K^\infty \lambda(x)v^{K,i}(x)\mathrm{d}x. \end{aligned}$$

We now turn to scenario c. This system is also a special case of the general set-up and is obtained by taking $g(w,b,K) = w + bI(w+b \leq K)$. Note that there is no $w \in [0,K]$ and $b, K \geq 0$ such that $g(w,b,K) > K$. This implies that, starting from initial workload below K, the workload process is bounded by the buffer size and we only have to analyze workloads below K.

The proportionality result, as presented in Theorem 3.4.1, does not hold for this scenario. Combined with Remark 3.4.1, this is obvious from the fact that $g(w,b,K) = w < K$ for $w \in [0,K)$ with $w + b > K$. Intuitively, the workload process below level K is indeed affected if a customer arrives that would cause a workload above the buffer size (in which case that customer is completely rejected). However, we can still solve the level crossing equations to determine the steady-state workload density.

Denote the steady-state workload density by $v^{K,c}(\cdot)$. Observe that an up-crossing of level x occurs, if at levels $y < x$ a customer arrives that has a service requirement larger than $x - y$, but smaller than $K - y$. Specifically, $\mathbb{P}(g(y,B,K) > x) = 1 - B(x-y) - (1 - B(K-y)) = B(K-y) - B(x-y)$. The level crossing equation may then be rewritten as follows. For $0 < x < K$,

$$\begin{aligned} r(x)v^{K,c}(x) &= \lambda(0)V^{K,c}(0)(B(K) - B(x))\\ &\quad + \int_{0+}^{x} \lambda(y)v^{K,c}(y)(B(K-y) - B(x-y))\mathrm{d}y. \qquad (3.22) \end{aligned}$$

In view of (3.6), we define the Volterra kernel as $K^c(x,y) := (B(K-y) - B(x-y))\frac{\lambda(x)}{r(x)}$, $0 \leq y < x < K$. Using Lemma 3.3.1 (with respective iterates and infinite sum), we can directly write

$$v^{K,c}(x) = \frac{\lambda(0)V^{K,c}(0)K^{c,*}(x,0)}{\lambda(x)}, \qquad 0 < x < K.$$

Determining $V^{K,c}(0)$ by normalization completes the derivation of the steady-state workload distribution.

Finally, we focus on the long-run fraction of rejected customers, P_K^c. By definition, a customer is rejected if, upon arrival, the workload present in addition to the service requirement exceeds the buffer capacity K. Then, conditioning on the workload just before a customer arrival and using Theorem 3.3.1 in the second equation, we have

$$\begin{aligned} P_K^c &= W^{K,c}(0)(1-B(K)) + \int_{0^+}^{K} w^{K,c}(x)(1-B(K-x))\mathrm{d}x \qquad (3.23) \\ &= \frac{1}{\bar{\lambda}^{K,c}}\left[\lambda(0)V^{K,c}(0)(1-B(K)) + \int_{0^+}^{K} \lambda(x)v^{K,c}(x)(1-B(K-x))\mathrm{d}x\right]. \end{aligned}$$

Remark 3.4.4 Note that the loss probabilities P_K, P_K^i, and P_K^c only depend on the ratio between $\lambda(\cdot)$ and $r(\cdot)$. This is a direct consequence of Theorem 3.3.2. At an intuitive level, this is evident from the fact that changing between Models 1 and 2 (in which $\lambda_1(x)/r_1(x) = \lambda_2(x)/r_2(x)$, for all $x > 0$) is just a rescaling of time. So, without loss of generality we may assume that the arrival rate is fixed. ◇

3.5 First-exit probabilities and cycle maxima

In this section, we focus on queues with infinite buffer capacity and determine first exit probabilities and the distribution of the cycle maximum. To do so, we use to a large extent the finite-buffer dam, analyzed in Subsection 3.4.1. Moreover, we show that first-exit probabilities are related to the dual of a finite dam. Also observe that, for well-chosen K, first-exit probabilities are the same for a range of finite-buffer models, such as the scenario f (use Remark 3.4.1).

Consider the model with arrival rate 1, and release rate $\hat{r}(x) := \frac{r(x)}{\lambda(x)}$ when the workload equals x. Theorem 3.3.1 shows that both models have the same workload density at arrival epochs, $w(\cdot)$. As a consequence, the amounts of work just after an arrival instant follow the same distribution as well. Also, observe that the workload process $\{V_t, t \geq 0\}$ attains local minima just before a jump and local maxima right after a jump. Considering first-exit probabilities, it then easily follows that we may consider (without loss of generality) a model with arrival rate 1, and release rate $\hat{r}(x)$. In fact, the same argument holds for cycle maxima, as it may be considered to be a special case of a first-exit probability. So, in this section we often assume, without loss of generality, that the arrival rate equals 1.

Starting with first-exit probabilities, we assume that $0 \leq a < b < \infty$, and let $\tau(a) := \inf\{t > 0 | V_t \leq a\}$ and $\tau(b) := \inf\{t > 0 | V_t \geq b\}$ correspond to the first-exit times. Note that we use b here in a different fashion than in Sections 3.2-3.4. An alternative notation for b could be K, but we decided to follow the literature on first-exit probabilities and use b in the context of hitting times. Starting from x, we denote the probability that the workload process hits state b before state

a by $U(x)$, i.e., $U(x) := \mathbb{P}_x(\tau(b) < \tau(a))$. Now, the first-exit probabilities can be obtained from those in models with constant arrival rate (in particular [83]) and the observation above. Define

$$\alpha(a, b) := \left[1 + \frac{r(b)}{\lambda(b)} \int_a^b \frac{\lambda(x)}{r(x)} K^*(b, x)\mathrm{d}x\right]^{-1}. \tag{3.24}$$

We obtain the following lemma:

Lemma 3.5.1 *We have,*

$$U(x) = \begin{cases} 0, & \text{if } 0 \le x \le a, \\ \int_a^x u(y)\mathrm{d}y, & \text{if } a < x \le b, \\ 1, & \text{if } x > b, \end{cases}$$

where $u(x) = \alpha(a, b)r(b)\lambda(x)K^*(b, x)/(\lambda(b)r(x))$ *for* $x \in (a, b)$.

Proof Apply [83, Theorem 3] to the dam with release rate $\hat{r}(\cdot)$. □

Remark 3.5.1 In fact, first-exit probabilities with $a > 0$ may be reduced to a similar first-exit probability with $a = 0$. Modify the system to a finite-buffer dam of capacity $b - a$ and with release rate $\breve{r}(x) := r(x + a)$ when the workload equals x. Denote the modified first hitting times by $\breve{\tau}(0)$ and $\breve{\tau}(b - a)$, and let, for $x \in (0, b - a]$, $\breve{U}(x) := \mathbb{P}_x(\breve{\tau}(b - a) < \breve{\tau}(0))$ be the probability that the modified system hits state $b - a$ before state 0, starting from x. Then, apply Lemma 3.5.1 to the modified system (thus with release rate $\breve{r}(\cdot)$). Note that $\breve{K}(x, y) = (1 - B(x - y))/\breve{r}(x) = K(x + a, y + a)$, and it can be easily shown (by induction) that $\breve{K}_n(x, y) = K_n(x + a, y + a)$. Now, it is just a straightforward calculation to show that $\breve{U}(x - a) = U(x)$. ◇

Concerning cycle maxima, we assume that at time 0 a customer enters an empty system and define $C_{\max} := \sup\{V_t, 0 \le t \le \tau(0)\}$. Denote $\tilde{r}(x) := \hat{r}(b - x) = r(b - x)$ and let $P_b^{\tilde{r}(\cdot)}$ be the loss probability in a finite dam (scenario f) with release rate $\tilde{r}(\cdot)$. The following relationship between cycle maxima and loss probabilities is obtained in Chapter 4, Theorem 4.3.1:

Lemma 3.5.2 *We have,*

$$\mathbb{P}(C_{\max} \ge b) = P_b^{\tilde{r}(\cdot)}.$$

Motivated by this relation, we first analyze scenario f with arrival rate 1 and release rate $\tilde{r}(\cdot)$ in more detail. This turns out to be a useful tool to determine the distribution of the cycle maximum in general terms.

Let $\tilde{v}(\cdot)$ denote the steady-state workload density of the model with arrival rate 1 and release rate $\tilde{r}(\cdot)$. Using level crossing arguments, we have, for $0 < x < b$,

$$\tilde{r}(x)\tilde{v}(x) = \tilde{V}(0)(1 - B(x)) + \int_{0^+}^x \tilde{v}(y)(1 - B(x - y))\mathrm{d}y.$$

Define $z(x) := \tilde{r}(x)\tilde{v}(x)$, and Volterra kernel $\tilde{K}(x,y) := (1-B(x-y))/\tilde{r}(y)$, for $0 < y < x < b$, and $\tilde{K}(x,0) := 1 - B(x)$ for $0 < x < b$. Observe that we can relate $\tilde{K}(x,y)$ to the basic kernel in Section 3.4. Specifically, for $0 < y < x < b$,

$$\tilde{K}(x,y) = (1 - B(b-y-(b-x)))/r(b-y) = K(b-y, b-x),$$

and for $0 < x < b$,

$$\tilde{K}(x,0) = 1 - B(b-(b-x)) = K(b, b-x)r(b).$$

Now, using the successive-substitution method for Volterra kernels as in Section 3.3, yields (for $0 < x < b$),

$$\begin{aligned} z(x) &= \tilde{V}(0)\tilde{K}(x,0) + \int_{0^+}^{x} z(y)\tilde{K}(x,y)\mathrm{d}y \\ &= \tilde{V}(0)K(b,b-x)r(b) + \int_{0^+}^{x} z(y)K(b-y,b-x)\mathrm{d}y \\ &= \tilde{V}(0)K(b,b-x)r(b) + \int_{0^+}^{x} [K(b,b-y)K(b-y,b-x)r(b)\tilde{V}(0) \\ &\qquad + \int_0^y K(b-u,b-y)K(b-y,b-x)z(u)\mathrm{d}u]\mathrm{d}y \\ &= K_1(b,b-x)r(b)\tilde{V}(0) + K_2(b,b-x)r(b)\tilde{V}(0) \\ &\qquad + \int_0^x z(u)K_2(b-u,b-x)\mathrm{d}u, \end{aligned}$$

where the last equality follows from Fubini's theorem and

$$\begin{aligned} \int_u^x K(b-u,b-z)K_n(b-z,b-x)\mathrm{d}z &= \int_{b-x}^{b-u} K(b-u,z)K_n(z,b-x)\mathrm{d}z \\ &= K_{n+1}(b-u,b-x). \end{aligned}$$

Iterating this argument gives

$$z(x) = r(b)\tilde{V}(0)K^*(b,b-x).$$

Finally, use the definition of $z(\cdot)$ to express the steady-state density of the model with release rate $\tilde{r}(\cdot)$ into the original model (with release rate $r(\cdot)$):

$$\tilde{v}(x) = \frac{\tilde{V}(0)r(b)K^*(b,b-x)}{r(b-x)}, \tag{3.25}$$

where $\tilde{V}(0)$ follows from normalization.

Returning to the cycle maximum of our original model, we now have the following theorem:

Theorem 3.5.1 *For the cycle maximum in an M/G/1-type dam, with arrival rate $\lambda(\cdot)$ and release rate $r(\cdot)$, we have*

$$\mathbb{P}(C_{\max} \geq b) = \tilde{V}(0)(1 - B(b)) + \int_0^b \tilde{v}(x)(1 - B(b - x))\mathrm{d}x, \tag{3.26}$$

where

$$\tilde{v}(x) = \frac{\tilde{V}(0)r(b)K^*(b, b - x)\lambda(b - x)}{\lambda(b)r(b - x)}, \tag{3.27}$$

and $\tilde{V}(0) = [1 + \int_0^b r(b)K^*(b, b - x)\lambda(b - x)(\lambda(b)r(b - x))^{-1}\mathrm{d}x]^{-1}$.

We give two different proofs of the above theorem; the first one uses the equivalence between cycle maxima and loss probabilities, and the second exploits knowledge of first-exit probabilities.

Proof I *(via $P_b^{\tilde{r}(\cdot)}$).* To prove Theorem 3.5.1, we use the relation between loss probabilities and cycle maxima, Lemma 3.5.2 (we refer to Chapter 4 for a discussion of this relationship). We already analyzed $P_b^{r(\cdot)}$ in Section 3.4 (Corollary 3.4.1). Use the fact that the cycle maximum only depends on $\lambda(\cdot)$ and $r(\cdot)$ via their ratio and note that the steady-state density for the model with release rate $\tilde{r}(x) = \frac{r(b-x)}{\lambda(b-x)}$ is then given by (3.25). Applying Corollary 3.4.1 to the model with $\lambda = 1$ and release rate $\tilde{r}(\cdot)$ gives the result. □

Proof II *(via $U(x)$).* First note that $\alpha(0, b) = \tilde{V}(0)$, by substituting $u = b - x$ in (3.24). Then, given the service requirement of a customer entering an empty system, the cycle maximum may be rewritten as a first-exit probability. Specifically, condition on the service requirement of the "first customer", and use Lemma 3.5.1 in the third equality:

$$\begin{aligned}
\mathbb{P}(C_{\max} \geq b) &= \int_{x=0}^{\infty} U(x)\mathrm{d}B(x) \\
&= \int_{x=0}^{b} \int_{y=0}^{x} u(y)\mathrm{d}y\mathrm{d}B(x) + (1 - B(b)) \\
&= \int_{y=0}^{b} \alpha(0, b)\frac{r(b)\lambda(y)K^*(b, y)}{r(y)\lambda(b)} \int_{x=y}^{b} \mathrm{d}B(x)\mathrm{d}y + (1 - B(b)) \\
&= \int_{y=0}^{b} \alpha(0, b)\frac{r(b)\lambda(y)K^*(b, y)}{r(y)\lambda(b)}(1 - B(y))\mathrm{d}y + \alpha(0, b)(1 - B(b)),
\end{aligned}$$

where the final step follows from (cf. (3.24))

$$1 = \alpha(0, b) + \int_0^b \alpha(0, b)\frac{r(b)\lambda(y)K^*(b, y)}{r(y)\lambda(b)}\mathrm{d}y.$$

The theorem now follows by straightforward substitution. □

Alternatively, the first-exit probabilities, given by Harrison and Resnick [83], may also be related to a finite dam with release rate $\tilde{r}(x) = \hat{r}(b-x)$ when the workload equals x. For $a = 0$ and $0 \leq x \leq b$, we use the steady-state workload density (3.27) directly:

$$\begin{aligned} 1 - \tilde{V}(x) &= \int_{y=x}^{b} \tilde{v}(y)\mathrm{d}y \\ &= \int_{y=x}^{b} \alpha(0,b)\frac{r(b)K^*(b,b-y)\lambda(b-y)}{\lambda(b)r(b-y)}\mathrm{d}y \\ &= \int_{u=0}^{b-x} \alpha(0,b)\frac{r(b)K^*(b,u)\lambda(u)}{\lambda(b)r(u)}\mathrm{d}u = U(b-x), \qquad (3.28) \end{aligned}$$

where $V(0) = \alpha(0,b)$ follows from Lemma 3.5.1 and Theorem 3.5.1. Using Remark 3.5.1, we may generalize this equivalence relation to cases with $a > 0$.

Lemma 3.5.3 *Let $\tilde{V}(\cdot)$ be the workload distribution of the finite-buffer system (scenario f) of capacity $b-a$ and release rate $\tilde{r}(\cdot)$. Then, for $x \in [0, b-a]$,*

$$1 - \tilde{V}(x) = U(b-x).$$

Proof Consider the system with finite buffer $b-a$ and release rate $\tilde{r}(x) = r(b-x)$ when the workload equals x. Note that a modification of $\tilde{r}(\cdot)$ to the case $a = 0$ (as in Remark 3.5.1) is not required, since $\tilde{r}(x) = \hat{r}((b-a-x)+a)$.

Although the workload is bounded from above by $b-a$, we can use exactly the same analysis as if it was bounded by b, and express the steady-state workload density as follows:

$$\tilde{v}(x) = \frac{\tilde{V}(0)\hat{r}(b)K^*(b,b-x)}{\hat{r}(b-x)},$$

where

$$\tilde{V}(0) = \left[1 + \int_{x=0}^{b-a} \frac{\hat{r}(b)K^*(b,b-x)}{\hat{r}(b-x)}\mathrm{d}x\right]^{-1}, \qquad (3.29)$$

follows from normalization. Substitute $y = b-x$ and $\hat{r}(x) = \frac{r(x)}{\lambda(x)}$ in (3.29), to see that

$$\tilde{V}(0) = \left[1 + \int_{y=a}^{b} \frac{r(b)K^*(b,y)\lambda(y)}{\lambda(b)r(y)}\mathrm{d}y\right]^{-1} = \alpha(a,b).$$

Now, using the same argument as in (3.28), with substitution $\hat{r}(y) = r(y)/\lambda(y)$ and $u = b-y$, completes the proof. □

Remark 3.5.2 We conjecture that (a modified version of) Lemma 3.5.3 also holds in case of general i.i.d. interarrival times and a general service rate function $r(\cdot)$ if we start with a regular interarrival interval. Denote by $\tilde{W}(\cdot)$ the workload distribution right before an arrival instant in the finite-buffer system of capacity $b-a$ and service rate $\tilde{r}(\cdot)$. It is then possible to show that

$$1 - \tilde{W}(x) = U(b-x),$$

using the machinery of monotone stochastic recursions [15] and a similar construction as in Chapter 4. ◇

3.6 Some examples and extensions

In Sections 3.3-3.5 we expressed the steady-state workload densities, first-exit probabilities, and cycle maxima in terms of an infinite sum of Volterra kernels. Numerical methods to compute these sums are widely available, see for example [88, 116]. Since we obtained closed-form expressions for the performance measures of interest, this concludes our analysis from a practical point of view. However, for some special cases, the Volterra integral equations reduce to analytically tractable expressions.

In this section, we discuss some special cases and show that several known results can be recovered from the Volterra kernels. In addition, we derive some results that appear to be new. We first discuss the case of constant arrival and service rates and then continue with the case of exponential service requirements. We conclude with a remark on the extension to rejection rules based on a stochastic barricade.

3.6.1 Constant arrival and service rates

Suppose that $r(x) \equiv r > 0$ and $\lambda(x) \equiv \lambda r > 0$. Observe that, using Theorems 3.3.1 and 3.3.2, we may assume that $r(x) \equiv 1$ and the model thus reduces to an ordinary M/G/1 queue. Denote the arrival rate by λ, the mean service requirement by β, and let $\rho := \lambda\beta$ be the load of the system. Further, let

$$H(x) := \beta^{-1} \int_0^x (1 - B(y))\mathrm{d}y,$$

denote the stationary residual service requirement distribution with density $h(\cdot)$.

In the M/G/1 case, the basic kernel $K(x, y)$ reduces to $\lambda(1 - B(x - y))$ and it is well-known, see for instance [83], that

$$K^*(x, y) = \sum_{n=1}^{\infty} \rho^n h_n(x - y). \tag{3.30}$$

Here, $h_n(\cdot)$ is the density of the n-fold convolution $H_n(\cdot)$. Now, combine Lemma 3.3.1 with (3.30) to obtain the famous Pollaczek-Khinchine formula Theorem 1.6.1. The finite-dam is just the truncated version, see Theorems 3.4.1 and 1.6.3.

Turning to the model with customer impatience (scenario i), the normalizing constant in (3.15) may be determined using (3.16) and an application of Little's law. First apply (3.16) with $y = 0$:

$$\mathbb{P}(V^{K,i} \leq x) = \mathbb{P}(V \leq x)\frac{V^{K,i}(0)}{V(0)}.$$

Then, use 'Little' in the first and $P_K^i = \mathbb{P}(W^{K,i} > K)$ and the PASTA property in the second equation (see also [62]), to obtain

$$V^{K,i}(0) = 1 - \rho(1 - P_K^i) = 1 - \rho V^{K,i}(K).$$

Apply (3.16) again to $V^{K,i}(K)$ (and use $V(0) = 1 - \rho$), then, after some rewriting, we may express the steady-state workload density for scenario i in terms of the ordinary M/G/1 queue (see also [33, 62]):

$$V^{K,i}(x) = \frac{V(x)}{1 - \rho + \rho V(K)}.$$

Finally, the first-exit probabilities follow from a direct computation, see [83]. Also, Takács' formula for cycle maxima [160], that is Theorem 1.6.4, may be easily recovered from Theorem 3.5.1 and the truncation property for finite dams (Theorem 3.4.1).

3.6.2 Exponential service requirements

Suppose that $1 - B(x) = e^{-\mu x}$, meaning that the service requirements are exponentially distributed with mean $1/\mu$. For the basic kernel, we then may write $K(x,y) = e^{-\mu(x-y)}\lambda(x)/r(x)$, and we can explicitly compute (similar to [83])

$$K^*(x,y) = \frac{\lambda(x)}{r(x)} \exp\{-\mu(x-y) + \Lambda(x) - \Lambda(y)\}. \tag{3.31}$$

Using Lemma 3.3.1, the familiar steady-state workload density in the infinite-buffer queue directly appears (see e.g. [45, 83], [10], p. 388, or Section 2.4):

$$v(x) = \frac{\lambda(0)V(0)}{r(x)} \exp\{-\mu x + \Lambda(x)\}. \tag{3.32}$$

The explicit form in (3.31) also allows us to evaluate $v^{K,i}(\cdot)$. After lengthy calculations, we deduce the following:

Corollary 3.6.1 *For the M/M/1 queue with customer impatience (scenario i), arrival rate $\lambda(\cdot)$, and service rate $r(\cdot)$, we have*

$$v^{K,i}(x) = \frac{\lambda(0)V^{K,i}(0)}{r(x)} \exp\{-\mu x + \Lambda(x \wedge K)\},$$

where $V^{K,i}(0)$ follows by normalization.

Turning to scenario c (complete rejection), we obtain the following corollary:

Corollary 3.6.2 *For the M/M/1 queue with complete rejection (scenario c), arrival rate $\lambda(\cdot)$, and service rate $r(\cdot)$, we have*

$$v^{K,c}(x) = \frac{V^{K,c}(0)\lambda(0)(1 - e^{-\mu(K-x)})}{r(x)} \exp\{-\mu x + \Lambda^c(x)\},$$

where $\Lambda^c(x) = \int_0^x \frac{\lambda(y)(1-e^{-\mu(K-y)})}{r(y)} \mathrm{d}y$ and $V^{K,c}(0)$ follows by normalization.

Proof Note that, by conditioning on $B > x - y$, $\mathbb{P}(g(y, B, K) > x)$ may be rewritten as $e^{-\mu(x-y)}(1 - e^{-\mu(K-x)})$. Thus, by substitution in (3.2), we have

$$\begin{aligned} r(x)v^g(x) &= \lambda(0)V^g(0)e^{-\mu x}(1 - e^{-\mu(K-x)}) \\ &\qquad + \int_{0^+}^{x} \lambda(y)v^g(y)e^{-\mu(x-y)}(1 - e^{-\mu(K-x)})\mathrm{d}y. \end{aligned}$$

Divide both sides by $1-e^{-\mu(K-x)}$. Then, comparing with (3.2) in the finite-dam case (scenario f), it follows that scenario c is equivalent to scenario f, but with $r(x)$ replaced by $r^c(x) := r(x)(1 - e^{-\mu(K-x)})^{-1}$. Appropriately adjusting $\Lambda(\cdot)$, resulting in $\Lambda^c(\cdot)$, and applying (3.32) gives the result. □

The result for the standard M/M/1-queue with complete rejection [81] can now easily be recovered from Corollary 3.6.2.

The first-exit probabilities may be deduced from Lemma 3.5.1. Alternatively, the first-exit probabilities may also be obtained from the steady-state workload density in the finite dam with arrival and release rate $\lambda(b-x)$ and $r(b-x)$ when the workload equals x. The cycle maximum can be derived in the same way.

Corollary 3.6.3 *For the cycle maximum in an M/M/1 queue, with arrival rate $\lambda(\cdot)$ and service rate $r(\cdot)$, we have*

$$\mathbb{P}(C_{\max} > b) = \tilde{V}(0)\exp\{\Lambda(b) - \mu b\},$$

where $\tilde{V}(0) = \left[1 + \int_0^b \frac{\lambda(x)}{r(x)} \exp\{-\mu(b - x) + \Lambda(b) - \Lambda(x)\}\mathrm{d}x\right]^{-1}$.

Proof Combining (3.27) with (3.31) and some rewriting yields

$$\tilde{v}(x) = \tilde{V}(0)\frac{\lambda(x)}{r(x)} \exp\{-\mu x + \Lambda(b) - \Lambda(b - x)\}.$$

$\tilde{V}(0)$ now follows directly by normalization. Moreover, using (3.26),

$$\begin{aligned} \mathbb{P}(C_{\max} > b) &= \tilde{V}(0)e^{-\mu b} + \int_0^b \tilde{V}(0)e^{-\mu b}\frac{\lambda(b - x)}{r(b - x)}e^{\Lambda(b)-\Lambda(b-x)}\mathrm{d}x \\ &= \tilde{V}(0)e^{-\mu b}\left[1 + e^{\Lambda(b)}\int_0^b \frac{\lambda(x)}{r(x)}e^{-\Lambda(x)}\mathrm{d}x\right] \\ &= \tilde{V}(0)e^{-\mu b}\left[1 + e^{\Lambda(b)}\left(1 - e^{-\Lambda(b)}\right)\right] \\ &= \tilde{V}(0)e^{-\mu b+\Lambda(b)}, \end{aligned}$$

which completes the proof. □

3.6.3 Stochastic barricade

In this chapter we considered an M/G/1-type model with finite buffer, in which the rejection rule is based on a deterministic barricade. This may be extended by replacing K by a random variable, see for instance [61, 137]. One now speaks of a stochastic barricade. This extension can easily be included into our framework. Replace at the n-th arrival epoch K by the random variable U_n, with distribution $F_U(\cdot)$ (independent of the service and arrival processes). The acceptance of the n-th customer in the scenarios of Section 3.2 is now determined as follows, see also [137]:

$$g(W_n, B_n, U_n) = \begin{cases} W_n + \min(W_n + B_n, (U_n - W_n)^+), & \text{scenario } f, \\ W_n + B_n I(W_n < U_n), & \text{scenario } i, \\ W_n + B_n I(W_n + B_n \leq U_n), & \text{scenario } c. \end{cases}$$

Note that in case $\lambda(\cdot)$ and $r(\cdot)$ are fixed, U_n represents the maximum waiting time (scenario i), or sojourn time (scenarios f and c). This model with stochastic impatience is well-known and studied in, e.g., [61, 137].

In case of a stochastic barricade, we again obtain a Volterra integral equation of the second kind. For the given examples, we have the following Volterra kernels, where $0 \leq y < x < \infty$ (see [137]):

$$K^g(x,y) = \begin{cases} (1 - B(x-y))(1 - F_U(x))\lambda(x)/r(x), & \text{scenario } f, \\ (1 - B(x-y))(1 - F_U(y))\lambda(x)/r(x), & \text{scenario } i, \\ \frac{\lambda(x)}{r(x)} \int_x^\infty (B(z-y) - B(x-y))\mathrm{d}F_U(z), & \text{scenario } c. \end{cases}$$

Even though these kernels might be difficult to determine in general, we may apply Lemma 3.3.1 to express the steady-state workload density in terms of these kernels. Some examples can be found in [137] in case of exponential service requirements and either exponential or deterministic barricades.

Finally, consider two (general) finite-buffer queues governed by the same distributions $B(\cdot)$ and $F_U(\cdot)$, but with arrival and service rates $\lambda_i(\cdot)$ and $r_i(\cdot)$, $i = 1, 2$, such that $\frac{\lambda_1(x)}{r_1(x)} = \frac{\lambda_2(x)}{r_2(x)}$, for all $x > 0$. The queues are then related in the same way as the queues in Section 3.3. From the discussion of Volterra kernels above, it is evident that (3.4) still holds in this broader context. Note that the probabilistic laws in both systems are still identical, resulting in the generalization of (3.3). More generally, Theorems 3.3.1 and 3.3.2 also hold in this framework.

Chapter 4

On an equivalence between loss rates and cycle maxima

4.1 Introduction

In Chapter 2 we analyzed an infinite-buffer G/G/1 queue with workload-dependent service rate and interarrival times. We specifically addressed the relation between workloads at arbitrary instants and at embedded epochs. In the present chapter we consider the G/G/1 queue under the partial-rejection discipline with both fixed and workload-dependent service rates. The focus in this chapter is, however, on the relation between the loss probability and the excess distribution of the cycle maximum in the infinite-buffer scenario. Finite-buffer queues and cycle maxima were also subject of study in the previous chapter under the assumption of a compound Poisson arrival process with workload-dependent arrival rate.

Queueing models with finite buffers are useful to model systems where losses are of crucial importance, as in inventory theory and telecommunications. Unfortunately, finite-buffer queues are often more difficult to analyze than their infinite-buffer counterparts. An important exception is the G/G/1 queue where the total amount of work is bounded from above by K and customers are rejected under the partial-rejection discipline. This rejection discipline operates such that if a customer's sojourn time would exceed K, then the customer only receives a fraction of its service requirement to make its sojourn time equal to K. This model is also known as the finite dam; see Section 4.2 for a precise description of the dynamics of this queue.

We consider the probability P_K that a customer gets partially rejected when entering the system in steady state. It is readily seen that

$$P_K = \mathbb{P}(W^K + S \geq K), \tag{4.1}$$

with W^K being the steady-state waiting time, and S a generic service requirement. Thus, information about P_K can be recovered from the distribution of W^K.

Cohen [56], Chapter III.6, analyzed the distribution of W^K in the case that both the interarrival times and service requirements have a rational LT. For the M/G/1 queue with traffic intensity $\rho < 1$ the distribution of W^K can be written in an elegant form, i.e., in terms of the steady-state waiting-time distribution of the M/G/1 queue with infinite buffer size. This result is already known since Takács [161]. Using this result, Zwart [176] showed that P_K can be identified with Takács' formula [161] for the tail distribution of the cycle maximum in the M/G/1 queue, i.e.,

$$P_K = \mathbb{P}(C_{\max} \geq K). \tag{4.2}$$

For the G/G/1 queue with light-tailed service times, Van Ommeren and De Kok [168] derived exact asymptotics for P_K as $K \to \infty$. From their main result, it immediately follows that

$$P_K \sim \mathbb{P}(C_{\max} \geq K),$$

as $K \to \infty$. This naturally leads to the conjecture that (4.2) can be extended to the G/G/1 queue. Unfortunately, the proof in [176] can not be extended to renewal arrivals, as it relies on exact computations for both P_K and the distribution of $C_{\max}$.

This brings us to the main goal of the chapter: Our aim is to show that (an appropriate modification of) (4.2) is valid for a large class of queueing models. In particular, we establish this equivalence for any positive ρ without the need to compute both sides of (4.2) separately. Instead, the proof method in the present chapter relates the distribution of $W^K + S$ to a first-passage probability, which is in turn related to the distribution of $C_{\max}$. We will also give another proof based on a regenerative argument.

Both proof techniques strongly rely on a powerful duality theory for stochastic recursions, which has been developed by Asmussen and Sigman [15], and dates back to Lindley [115], Loynes [118], and Siegmund [155]. For a recent textbook treatment, see Asmussen [10], Section IX.4. This type of duality, also known as Siegmund duality, relates the stationary distribution of a given model to the first-passage time of a dual model. Thus, Siegmund duality provides the right framework for proving (4.2). In its most simple form, Siegmund duality yields the well-known relationship between waiting-time probabilities for infinite-buffer queues and ruin probabilities.

This chapter is organized as follows. We treat the G/G/1 queue in Section 4.2. Section 4.3 extends the results of Section 4.2 to queues with state-dependent service rates. The final result for this class of models is somewhat more complicated than (4.2). In both sections, we give two proofs. These two proofs lead to different identities in Section 4.3. In Section 4.4 we show that (4.2) is not only useful to derive new results for the loss probability P_K, but also for the distribution of $C_{\max}$. Our main results in this section are: (i) a substantially shorter proof of the light-tailed asymptotics for P_K derived in [168], (ii) asymptotics of P_K for heavy-tailed service requirements, and (iii) an extension of Takács' formula for $\mathbb{P}(C_{\max} \geq K)$ to M/G/1 queues with state-dependent release rates. Concluding remarks can be found in Section 4.5.

Notation. Contrary to other chapters of this thesis, in the present chapter a generic service requirement is denoted by S and a generic interarrival time is denoted by T. (Consequently, the n-th service requirement and interarrival time are given by S_n and T_n, respectively). Thus, we stay close to notation that is common in much of the literature on random walks, in particular [15]. Moreover, in this chapter we adapted the notation in the paper [28] for the workload process and its dual to be consistent with the notation for the main performance measure in this thesis.

4.2 The G/G/1 queue

In this section we consider the G/G/1 queue with partial rejection, which is also known as the finite G/G/1 dam. Before we present our main result, we first introduce some notation and give a detailed model description.

Let $T_1, T_2, \ldots$ be the interarrival times of customers and denote the n-th arrival epoch after time 0 by $\bar{T}_n$, i.e., $\bar{T}_n = \sum_{k=1}^{n} T_k$. We assume that the interarrival times form an i.i.d. sequence and that $\mathbb{E}[T_1] < \infty$. The service requirement of the n-th customer is denoted by S_n, $n = 1, 2, \ldots$, where $S_1, S_2, \ldots$ are also assumed to be independent, identically distributed. We assume that the sequences of interarrival intervals and service requirements are independent. Define $\rho := \mathbb{E}[S_1]/\mathbb{E}[T_1]$ as the load of the system. We like to emphasize that ρ may take any (positive) value. To obtain a non-trivial model though, we additionally assume that $\mathbb{P}(T_1 > S_1) > 0$.

The workload process $\{V_t, t \in \mathbb{R}\}$ is now defined recursively by, cf. [56],

$$V_t = \max(\min(V_{\bar{T}_k^-} + S_k, K) - (t - \bar{T}_k), 0), \qquad t \in [\bar{T}_k, \bar{T}_{k+1}).$$

Since the workload in the system is uniformly bounded, the process $\{V_t, t \in \mathbb{R}\}$ is regenerative with customer arrivals into an empty system being regeneration points, independent of the load of the system. Let a regeneration cycle start at time 0, and define the first return time to state 0 by

$$\tau_0 := \inf\{t > 0 : V_t = 0\}.$$

Furthermore, let $C_{\max}$ be the cycle maximum of a busy cycle, or, more formally,

$$C_{\max} := \sup\{V_t, 0 \leq t \leq \tau_0\}.$$

Observe that, for $x \leq K$, $\mathbb{P}(C_{\max} \geq x)$ is the same for the finite dam and its infinite-buffer counterpart. So, without affecting the results, we will henceforth adopt the above definition of $C_{\max}$ when we consider the cycle maximum in the G/G/1 queue with infinite buffer capacity. Note that $\mathbb{P}(C_{\max} = \infty) > 0$ if $K = \infty$ and $\rho > 1$.

From the workload process in the finite G/G/1 dam we construct a “dual” process $\{D_t, t \in \mathbb{R}\}$, as in [138], by defining

$$D_t := K - V_t.$$

Away from the boundaries, this process increases linearly at rate 1 and negative jumps occur at times $\bar{T}_n$ of size S_n, $n = 1, 2, \ldots$. By definition, jumps below 0 are truncated and if the process hits state K, it remains in K until the next (downward) jump (see Figure 4.1 for an illustration). In fact, we are only interested in the behavior of D_t until it hits one of the boundaries and in this region the process $\{D_t, t \in \mathbb{R}\}$ shows strong resemblence with a risk process (where 0 is supposed to be an absorbing state).

Due to the finite capacity, the process $\{D_t, t \in \mathbb{R}\}$ is also regenerative and regeneration points in the process correspond to downward jump epochs from level K. Hence, τ_0 can be alternatively defined by $\tau_0 := \inf\{t > 0 : D_t \geq K\}$.

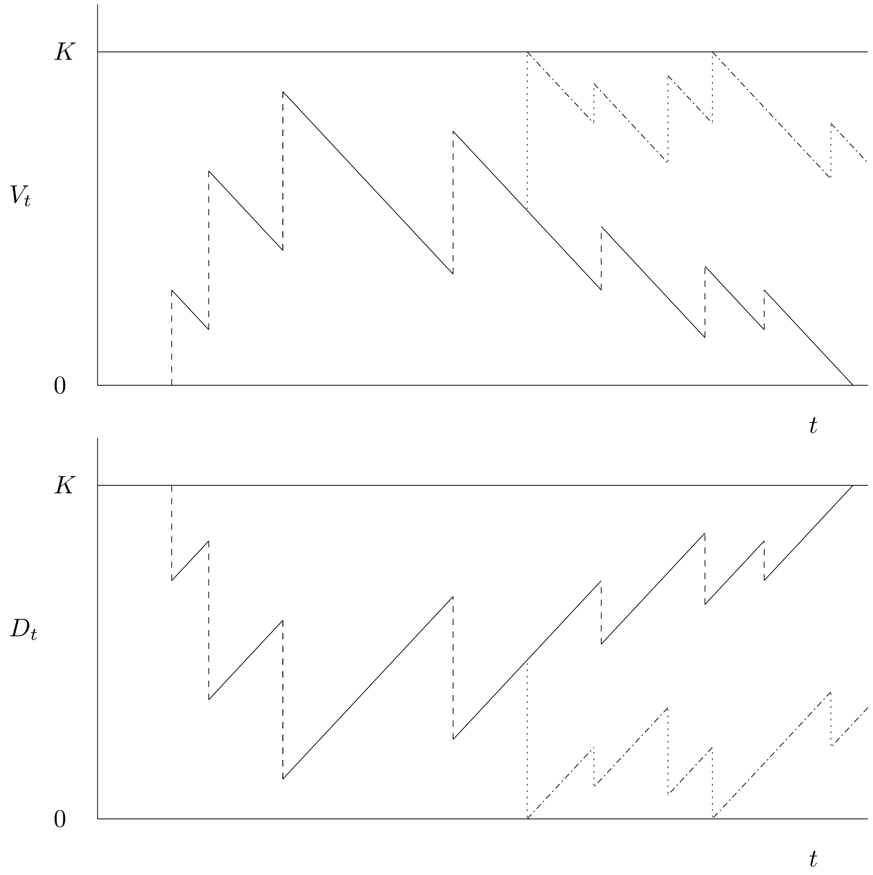

Figure 4.1: Two sample paths of V_t until it hits one of the boundaries, with corresponding D_t.

Recall that P_K is the steady-state probability that an arriving customer is (partially) rejected. The main result in this section is the following theorem.

Theorem 4.2.1 *For the G/G/1 queue we have*

$$P_K = \mathbb{P}(C_{\max} \geq K).$$

In the remaining part of this section we present two proofs of Theorem 4.2.1. In the first proof, to be presented in Subsection 4.2.1, we take a direct approach, using the representation $P_K = \mathbb{P}(W^K + S \geq K)$ and the abovementioned definition of the cycle maximum. Equivalence is then shown using the machinery developed in [15].

The second proof, given in Subsection 4.2.2, establishes a link between the loss rate and the cycle maximum using an insightful regenerative argument. In particular, we utilize the fact that the number of losses in a cycle, given that at least one loss occurs, is geometrically distributed. The main step in this approach is the computation of the success parameter of that distribution. This is again established by results in [15].

4.2.1 Direct approach

To determine the tail distribution of the cycle maximum in an infinite-capacity model, we may also assume that the workload is uniformly bounded as described above. So, consider one regeneration cycle of the process $\{V_t, t \in \mathbb{R}\}$ (or equivalently $\{D_t, t \in \mathbb{R}\}$) and let a customer enter the system at time 0. Let N_K be the number of arrivals in $[0, \tau_0]$. Since the workload process has peaks at time epochs just after an arrival instant, we may write

$$\begin{aligned}\mathbb{P}(C_{\max} \geq K) &= \mathbb{P}(\exists n \leq N_K : W_n + S_n \geq K) \\ &= \mathbb{P}(\exists n \leq N_K : D_{\bar{T}_n^-} - S_n \leq 0). \qquad (4.3)\end{aligned}$$

Observe that the right-hand side of (4.3) corresponds to a hitting probability; starting in state K, (4.3) may be interpreted as the probability that state 0 is reached before D_t hits state K again. Note that the process $\{D_t, t \in \mathbb{R}\}$ embedded at points $\bar{T}_n$ is also recursively defined by the interarrival times and the service requirements. Since we are both studying the process in continuous time, i.e., D_t, and at embedded epochs, we add a hat for the latter quantity, denoting it by $\hat{D}_n$. The above two observations allow us to rewrite this embedded process as a monotone stochastic recursion with two absorbing states (0 and K): We define $\hat{D}_0 = K, \hat{D}_{n+1} = g(\hat{D}_n, U_n)$, where $U_n := (S_{n+1}, T_n)$ and

$$g(x,s,t) = \begin{cases} 0, & \text{if } x = 0 \text{ or if } x \in (0,K] \text{ and } s \geq x, \\ x - (s-t), & \text{if } x \in (0,K] \text{ and } s < x, \\ \infty, & \text{if } x > K. \end{cases}$$

Thus, we start our recursion with initial reserve K, after which it evolves as an unrestricted random walk, until it leaves $(0, K]$. Moreover, it is always checked ahead whether a downward jump will cause a negative value of the process, leading to absorption in state 0.

Now, Example 4 of Asmussen and Sigman [15] gives the corresponding dual stochastic recursion $\{\hat{V}_n\}$ which is defined as $\hat{V}_{n+1} = f(\hat{V}_n, S_{n+1}, T_n)$, where

$$f(y, s, t) = \min(((y - t)^+ + s), K).$$

This recursion corresponds to the workload right *after* a jump, or the sojourn time, in a finite G/G/1 dam. Under i.i.d. assumptions, $\hat{V}_n$ weakly converges to a random variable $\hat{V}$ as $n \to \infty$, see for example Chapter III.6 in Cohen [56]. Let

$$\gamma(x, K) := \min\{n \geq 1 : \hat{D}_0 = x, \hat{D}_n \notin (0, K]\},$$

denote the first-exit time of $(0, K]$ starting in x. Then, Corollary 3.1 of [15] yields the following fundamental result:

$$\mathbb{P}(\hat{V} \geq x) = \lim_{n\to\infty} \mathbb{P}(\hat{D}_n \leq 0 \mid \hat{D}_0 = x) = \mathbb{P}(\hat{D}_{\gamma(x,K)} \leq 0). \tag{4.4}$$

Thus, the distribution of $\hat{V}$ can be written as a first-passage probability. Using (4.3) and taking $x = K$ in (4.4), we have

$$\begin{aligned} \mathbb{P}(C_{\max} \geq K) &= \mathbb{P}(\hat{D}_{\gamma(K,K)} \leq 0) \\ &= \mathbb{P}(\hat{V} \geq K). \end{aligned}$$

Hence,

$$P_K = \mathbb{P}(W^K + S \geq K) \equiv \mathbb{P}(\hat{V} \geq K) = \mathbb{P}(C_{\max} \geq K),$$

which completes the proof.

4.2.2 Regenerative approach

Let L_K be the number of not fully accepted customers, and recall that N_K is the total number of customer arrivals during a regeneration cycle. A basic regenerative argument yields

$$P_K = \frac{\mathbb{E}L_K}{\mathbb{E}N_K}. \tag{4.5}$$

The denominator follows easily by

$$\mathbb{P}(W^K = 0) = \frac{1}{\mathbb{E}N_K}\mathbb{E}\left[\sum_{i=1}^{N_K} I(W_i^K = 0)\right] = \frac{1}{\mathbb{E}N_K}, \tag{4.6}$$

where $I(\cdot)$ is the indicator function.

The numerator may be rewritten as follows

$$\begin{aligned} \mathbb{E}L_K &= \mathbb{E}[L_K I(L_K \geq 1)] \\ &= \mathbb{E}[L_K \mid L_K \geq 1]\mathbb{P}(L_K \geq 1) \\ &= \mathbb{E}[L_K \mid L_K \geq 1]\mathbb{P}(C_{\max} \geq K). \end{aligned} \tag{4.7}$$

Moreover, observe that whenever the workload reaches level K and a customer is (partially) rejected, the process continues from level K starting with a new interarrival time, which clearly is independent of the past. Then, the probability of an additional customer loss in the regeneration cycle is equal to the probability that the workload process reaches level K again before the end of the busy cycle. Denoting $\tau_K := \inf\{t > 0 : V_t \geq K \mid V_0 = K\}$, this leads to

$$\begin{aligned} \mathbb{P}(L_K \geq n+1 \mid L_K \geq n) &= \mathbb{P}(\tau_K < \tau_0 \mid V_0 = K) \\ &=: 1 - q_K. \end{aligned} \tag{4.8}$$

Iterating this argument, we conclude that $L_K \mid L_K \geq 1$ is geometrically distributed with success parameter $1 - q_K$. Since the expectation of such a geometric distribution equals $1/(1 - q_K)$, we have to show that $q_K = \mathbb{P}(W^K > 0)$ to complete the proof.

To do so, we use a similar construction of the "dual risk-type" process $\{D_t, t \in \mathbb{R}\}$ as in the first proof. Note that (4.8) corresponds to the probability that from initial level 0, D_t reaches level 0 again before it hits level K. Again, this can be transformed into a monotone stochastic recursion with two absorbing barriers, 0 and K: Define $\hat{D}_{n+1} = g(\hat{D}_n, S_{n+1}, T_n)$, with

$$g(x, s, t) = \begin{cases} 0, & \text{if } x = 0 \text{ or if } 0 < x < s - t, \\ x - (s - t), & \text{if } 0 < s - t \leq x \leq K - t, \\ \infty, & \text{if } x + t > K. \end{cases}$$

Thus, starting from level 0, $\hat{D}_n$ evolves as an unrestricted random walk until it leaves $(0, K]$. Note that it is indeed checked ahead whether the workload increases above level K before the next downward jump.

Now, another example of Asmussen and Sigman [15] provides the dual stochastic recursion $\{\hat{V}_n\}$. In particular, Example 3 of [15] gives the dual function

$$f(y, s, t) = (\min(y + s, K) - t)^+,$$

defining the dual recursion $\hat{V}_{n+1} = f(\hat{V}_n, S_{n+1}, T_n)$. This recursion corresponds to the workload right *before* a jump (or the waiting time) in a finite G/G/1 dam. Use Corollary 3.1 of [15] and take $x = \epsilon > 0$ in (4.4) to show that

$$\begin{aligned} q_K &= \lim_{\epsilon \downarrow 0} \mathbb{P}(\hat{D}_{\gamma(\epsilon,K)} \leq 0) \\ &= \lim_{\epsilon \downarrow 0} \mathbb{P}(\hat{V} \geq \epsilon) = \mathbb{P}(\hat{V} > 0). \end{aligned} \tag{4.9}$$

Recall that the $\hat{V}_n$ corresponds to the waiting time of the n-th customer, and $\hat{V}$ thus represents the *waiting* time in steady-state. Combining (4.5)-(4.9) completes the proof.

Remark 4.2.1 Both proofs rely on computing the dual of a recursion driven by a specific function $f(x, z)$, which is monotone in x for every z. In general, the driving function f and its dual g are related by

$$\begin{aligned} g(x,z) &= \inf\{y : f(y,z) \geq x\}, \\ f(y,z) &= \inf\{x : g(x,z) \geq y\}. \end{aligned}$$

We refer to [15] (in particular Equation (2.4) of [15]) for details. ◇

4.3 Dams with state-dependent release rates

In this section we consider the G/G/1 dam with general release rate. We start with introducing some definitions and a description of the driving sequence of the queueing process. Next, we state the main result and give two proofs, analogous to the proofs in Section 4.2.

Consider the model of Section 4.2, but let the release rate be $r(x)$ when the workload equals x. We assume that $r(0) = 0$ and that $r(\cdot)$ is strictly positive, left-continuous, and has a strictly positive right limit on $(0,\infty)$. Also, define

$$R(x) := \int_0^x \frac{1}{r(y)} \mathrm{d}y, \qquad 0 < x < \infty,$$

representing the time required for a workload x to drain in the absence of any arrivals. We assume that $R(x) < \infty$, $0 < x < \infty$, indicating that state 0 can be reached in a finite amount of time. This ensures that $C_{\max}$ is well-defined. Note that $R(\cdot)$ is strictly increasing and we can thus unambiguously speak of $R^{-1}(\cdot)$. Similar to [83, 138], we define

$$q(u,t) := R^{-1}(R(u) - t).$$

Then $q(u,t)$ represents the workload level at time t if we start from level u at time 0 and no arrivals have taken place in between, see also Section 2.5.

Denote the workload process of the G/G/1 queue with finite buffer K and general release-rate function $r(\cdot)$ by $\{V_t^{r(\cdot)}, t \in \mathbb{R}\}$. Let $T_0 = 0$ and $V_0^{r(\cdot)} = x$. Between jump epochs, the workload process is defined recursively by, cf. [138],

$$V_t^{r(\cdot)} = q(V_{\bar{T}_k^-}^{r(\cdot)}, t), \qquad \bar{T}_k \leq t < \bar{T}_{k+1},$$

and at the $(k+1)$-th jump epoch after time 0,

$$V_{\bar{T}_{k+1}}^{r(\cdot)} = \min\left(q(V_{\bar{T}_k}^{r(\cdot)}, T_{k+1}) + S_{k+1}, K\right).$$

To exclude trivial cases where the workload is bounded from below, we assume that $\mathbb{P}(q(x + S_1, T_1) < x) > 0$, for all $x > 0$. Combined with the finite capacity and $R(x) < \infty$ for all finite x, this ensures that the workload process $\{V_t^{r(\cdot)}, t \in \mathbb{R}\}$ is still regenerative with customer arrivals into an empty system as regeneration points.

Define $\tilde{r}(x) := r(K - x)$, for $0 \leq x \leq K < \infty$, and let all random variables $X^{r(\cdot)}, X^{\tilde{r}(\cdot)}$ correspond to the model with release rate $r(x), \tilde{r}(x)$, respectively, if the process is at level x. Similar to Section 4.2, we construct a "dual risk-type" process $\{D_t^{\tilde{r}(\cdot)}, t \in \mathbb{R}\}$, by taking $D_t^{\tilde{r}(\cdot)} = K - V_t^{r(\cdot)}$. In between the (downward) jumps, the newly defined process is governed by the input rate function $\tilde{r}(x) = r(K-x)$, and satisfies

$$\frac{\mathrm{d}D_t^{\tilde{r}(\cdot)}}{\mathrm{d}t} = \tilde{r}(D_t^{\tilde{r}(\cdot)}).$$

Also, the process starts at $D_0^{\tilde{r}(\cdot)} = K - V_0^{r(\cdot)}$. In addition, if $\{D_t^{\tilde{r}(\cdot)}, t \in \mathbb{R}\}$ starts at y and no jumps occur for t time units, its value increases, similar to the decrease in the workload process, to

$$\tilde{q}(y,t) := \tilde{R}^{-1}(\tilde{R}(y) + t).$$

Here, $\tilde{R}(x) := \int_0^x (\tilde{r}(y))^{-1}\mathrm{d}y$ represents the time required to move from 0 to x in the absence of any negative jumps, with inverse $\tilde{R}^{-1}(\cdot)$. Note that, for finite K, $\int_0^x (\tilde{r}(y))^{-1}\mathrm{d}y < \infty$, meaning that any state x can be reached from state zero in a finite amount of time and the cycle maximum is also well-defined in this case.

Theorem 4.3.1 *For the G/G/1 queue with general release rate we have*

$$P_K^{r(\cdot)} = \mathbb{P}(C_{\max}^{\tilde{r}(\cdot)} \geq K), \tag{4.10}$$

or alternatively,

$$P_K^{r(\cdot)} = \frac{\mathbb{P}(W^{K,r(\cdot)} = 0)}{\mathbb{P}(W^{K,\tilde{r}(\cdot)} = 0)} \mathbb{P}(C_{\max}^{r(\cdot)} \geq K). \tag{4.11}$$

We use a direct approach to show (4.10), thereby extending the proof in Subsection 4.2.1. To show (4.11), we follow the lines of Subsection 4.2.2, using an insightful regenerative argument and noting that the number of losses in a cycle, given that at least one loss occurs, has a geometric distribution. Let us start with (4.10).

Proof of (4.10) As noted earlier, the workload process $\{V_t^{r(\cdot)}, t \in \mathbb{R}\}$ is still regenerative with customer arrivals into an empty system as regeneration points. The observation that the workload process has local peaks at epochs right after an arrival instant, together with (4.3) and the construction of the "dual" process $\{D_t^{\tilde{r}(\cdot)}, t \in \mathbb{R}\}$, leads to

$$\mathbb{P}(C_{\max}^{r(\cdot)} \geq K) = \mathbb{P}(\exists n \leq N_K : D_{T_n^-}^{\tilde{r}(\cdot)} - S_n \leq 0). \tag{4.12}$$

The probability in (4.12) can be interpreted as the probability that state 0 is reached before $D_t^{\tilde{r}(\cdot)}$ hits state K again, starting from level K. Define $\hat{D}_0^{\tilde{r}(\cdot)} = K$ and $\hat{D}_{n+1}^{\tilde{r}(\cdot)} = g(\hat{D}_n^{\tilde{r}(\cdot)}, S_{n+1}, T_n)$, with

$$g(x,s,t) = \begin{cases} 0, & \text{if } x = 0 \text{ or if } x \in (0,K] \text{ and } s \geq x, \\ \tilde{R}^{-1}(\tilde{R}(x-s)+t), & \text{if } x \in (0,K] \text{ and } s < x, \\ \infty, & \text{if } x > K. \end{cases}$$

Following [15], we construct the dual function corresponding to the described process $\{D_t^{\tilde{r}(\cdot)}, t \in \mathbb{R}\}$, yielding

$$f(y, s, t) = \min\left(\tilde{R}^{-1}(\tilde{R}(y) - t) + s, K\right),$$

and define $\{\hat{V}_n^{\tilde{r}(\cdot)}\}$ recursively by $\hat{V}_{n+1}^{\tilde{r}(\cdot)} = f(\hat{V}_n^{\tilde{r}(\cdot)}, S_{n+1}, T_n)$. This process corresponds to a G/G/1 queue with release rate $\tilde{r}(x) = r(K - x)$ if the workload equals x, embedded at epochs right *after* a jump. We now complete the proof of (4.10) by combining the duality (4.4) between storage and risk processes with the expression (4.1) for P_K:

$$\begin{aligned}
\mathbb{P}(C_{\max}^{r(\cdot)} \geq K) &= \mathbb{P}(\hat{D}_{\gamma(K,K)}^{\tilde{r}(\cdot)} \leq 0) \\
&= \mathbb{P}(\hat{V}^{\tilde{r}(\cdot)} \geq K) \\
&= \mathbb{P}(W^{K,\tilde{r}(\cdot)} + S \geq K) \\
&= P_K^{\tilde{r}(\cdot)},
\end{aligned}$$

which completes the proof. □

Next we turn to (4.11), which we show following the lines of Subsection 4.2.2.

Proof of (4.11) As mentioned above, the workload process is still regenerative, and we consider the total number of (partially) rejected customers during a regeneration cycle. We apply the same regenerative argument as in Subsection 4.2.2 and note that customers are rejected if and only if the process reaches level K before the end of the cycle (which happens with probability $\mathbb{P}(C_{\max}^{r(\cdot)} \geq K)$). Moreover, after a customer rejection, the process continues from level K, starting with a new interarrival time. This implies that the probability of an additional customer loss is independent of the past, or equivalently, that K is also a regeneration point. Therefore, we may conclude that, given that at least one loss occurs and the process starts from level K, the additional number of customer rejections is geometrically distributed with success parameter $1 - q_K = \mathbb{P}(\tau_K < \tau_0 \mid V_0^{r(\cdot)} = K)$. Thus, we have to show that $q_K = \mathbb{P}(W^{K,\tilde{r}(\cdot)} > 0)$ and combine (4.5)–(4.8) to complete the proof.

We start with the construction of the "dual risk-type" process $\{D_t^{\tilde{r}(\cdot)}, t \in \mathbb{R}\}$ defined at the beginning of the section. We rewrite $1 - q_K$ as the probability that, starting from level 0, $D_t^{\tilde{r}(\cdot)}$ hits level 0 again before it reaches level K. Interpreting our process as a monotone stochastic recursion with two absorbing barriers, we define $\hat{D}_{n+1}^{\tilde{r}(\cdot)} = g(\hat{D}_n^{\tilde{r}(\cdot)}, S_{n+1}, T_n)$, where

$$g(x, s, t) = \begin{cases} 0, & \text{if } x = 0 \text{ or if } \tilde{R}(x) < \tilde{R}(s) - t, \\ \tilde{R}^{-1}(\tilde{R}(x) + t) - s, & \text{if } \tilde{R}(s) - t < \tilde{R}(x) < \tilde{R}(K) - t, \\ \infty, & \text{if } \tilde{R}(x) + t > \tilde{R}(K). \end{cases}$$

Again, using [15] it can be seen that the dual recursion is defined as $\hat{V}_{n+1}^{\tilde{r}(\cdot)} = f(\hat{V}_n^{\tilde{r}(\cdot)}, S_{n+1}, T_n)$, with

$$f(y, s, t) = \tilde{R}^{-1}(\tilde{R}(\min(y + s, K)) - t).$$

The latter recursion corresponds to the workload at time epochs right *before* a jump. As the speed of the server is determined by the general release function, this does not equal the waiting time.

Finally, using Corollary 3.1 of [15] once more, we obtain

$$\begin{aligned} q_K &= \lim_{\epsilon \downarrow 0} \mathbb{P}(\hat{D}_{\gamma(\epsilon,K)}^{\tilde{r}(\cdot)} \leq 0) \\ &= \lim_{\epsilon \downarrow 0} \mathbb{P}(\hat{V}^{\tilde{r}(\cdot)} \geq \epsilon) = \mathbb{P}(W_k^{\tilde{r}(\cdot)} > 0). \end{aligned} \tag{4.13}$$

Hence, by combining (4.5)-(4.8), and (4.13) we also have shown the second part of the result. □

Remark 4.3.1 The constant $\mathbb{P}(W^{K,r(\cdot)} = 0)/\mathbb{P}(W^{K,\tilde{r}(\cdot)} = 0)$ in (4.10) can easily be interpreted. As the interarrival times in both systems follow the same distribution, using (4.6), the constant equals the ratio of the respective mean cycle lengths. ◇

Remark 4.3.2 A sample-path argument can also provide some intuition into the equivalence between (4.10) and (4.11). First, the process $\{D_t^{\tilde{r}(\cdot)} \mid t \geq 0\}$ can easily be interpreted as the available buffer capacity of a dam with release rate $r(x)$ when the content equals x. Second, to convert the risk-type process into a queueing process again, we use a reversibility argument, as in [11, 13]. The sample path of this queueing process can essentially be obtained by time-reversing the sample path of $\{D_t^{\tilde{r}(\cdot)} \mid t \geq 0\}$, resulting in a queueing process with service speed $\tilde{r}(x)$ when the workload equals x. ◇

4.4 Applications

In this section we state some exact and asymptotic results for P_K, by applying results for $C_{\max}$ which are available in the literature. Given the results derived before, this leads to more transparent proofs of existing results, and also yields some results that are new.

4.4.1 Exact expressions for P_K

In the literature, there are several studies devoted to the distribution of $C_{\max}$ for a variety of queueing models. We refer to Asmussen [7] for a survey of these results. The M/G/1 case has already been treated in Zwart [176]. Here, we give an analogous result for the G/M/1 queue.

Corollary 4.4.1 *Consider the finite G/M/1 dam with $\rho < 1$ and service rate μ. Then*

$$P_K = \frac{1}{G(K)},$$

where $G(x), x \geq 0$, is a function with LST

$$\frac{1}{s - \mu(1 - \alpha(s))},$$

with $\alpha(s)$ the LST of the interarrival time distribution.

Proof The result follows immediately from Theorem 4.2.1 and formula (7.76) of Cohen [56] stating that, for the G/M/1 queue,

$$\mathbb{P}(C_{\max} \geq K) = \frac{1}{G(K)}$$

with $G(x), x \geq 0$, defined as above. □

4.4.2 Asymptotics

Van Ommeren and De Kok [168] derive asymptotics for P_K in the G/G/1 queue under light-tailed assumptions. After a lengthy argument they find the asymptotics of P_K, from which it immediately follows that (under their assumptions) $P_K \sim \mathbb{P}(C_{\max} > K)$.

Asymptotics for the latter are due to Iglehart [86]: Under certain regularity conditions (see [86]), it holds that

$$\mathbb{P}(C_{\max} \geq K) \sim De^{-\gamma K}, \tag{4.14}$$

for certain positive constants γ and D. Using Theorem 4.2.1, the proof of the main result of [168] is now trivial: Just combine Theorem 4.2.1 with (4.14) to (re-)obtain

$$P_K \sim De^{-\gamma K}.$$

For more details concerning specific assumptions and expressions for γ and D we refer to [86] and [168].

We conclude by giving results for the heavy-tailed case: Consider again the G/G/1 queue, but assume now that service requirements belong to the subclass $\mathcal{S}^*$ of the class of subexponential distributions (see, e.g., Embrechts *et al.* [74] for a definition). This class contains all heavy-tailed distributions of main interest, such as the Pareto, lognormal, and certain Weibull distributions.

Asymptotics for the cycle maximum are presented in Theorem 1.6.6 (see also [8]). If we combine these asymptotics with Theorem 4.2.1 we obtain (with N being the number of customers served in one busy cycle in the infinite-buffer version of the G/G/1 queue):

Corollary 4.4.2 *If $\rho < 1$ and the service requirement $S \in \mathcal{S}^*$, then*

$$P_K \sim \mathbb{E}N\mathbb{P}(S \geq K).$$

Also, in case of Poisson arrivals, Theorem 1.6.8 extends the result to queues with general service speeds, see [8] for details. Note that (4.10) and (4.11), combined with Remark 4.3.1, indeed result in the same asymptotics.

4.4.3 Poisson arrivals and Takács' formula

The equivalence in Theorem 4.2.1 can also be used the other way around: Given information on P_K, we derive a new identity for the distribution of $C_{\max}$ for queues with general release rate. For the special M/G/1 case, the distribution of $C_{\max}$ is known through Takács' formula. We combine the results of Section 4.3 with an identity for P_K which is valid under the additional assumptions of Poisson arrivals and a stationary (embedded) workload distribution in case of infinite buffer capacity (see e.g. [10, 46] for details). Under these assumptions, Theorem 3.4.1 shows that the steady-state distribution of the amount of work in the system found by a customer $W^{K,r(\cdot)}$ satisfies the following *proportionality* property:

$$\mathbb{P}(W^{K,r(\cdot)} \leq x) = \frac{\mathbb{P}(W^{r(\cdot)} \leq x)}{\mathbb{P}(W^{r(\cdot)} \leq K)}. \tag{4.15}$$

Here, $W^{r(\cdot)}$ is the steady-state amount of work in the system with $K = \infty$ (assuming it exists). For similar proportionality relations in the ordinary M/G/1 queue, see for example Takács [161], Cohen [56] and Hooghiemstra [85]; see Asmussen [10], Chapter XIV, Proposition 3.1, in case of a general release rate.

Writing $1 - P_K = \mathbb{P}(W^{K,r(\cdot)} + S < K)$, conditioning on S, applying (4.15), and deconditioning on S then results in

$$\begin{aligned} P_K^{r(\cdot)} &= 1 - \mathbb{P}(W^{K,r(\cdot)} + S < K) \\ &= \frac{\mathbb{P}(W^{r(\cdot)} + S \geq K) - \mathbb{P}(W^{r(\cdot)} > K)}{\mathbb{P}(W^{r(\cdot)} \leq K)}. \end{aligned}$$

Combining this result with (4.11) then results in the following corollary.

Corollary 4.4.3 *Assume that the M/G/1 queue with infinite buffer size and general release rate has a stationary (embedded) workload distribution. Then,*

$$\mathbb{P}(C_{\max}^{r(\cdot)} \geq x) = \frac{\mathbb{P}(W^{x,\tilde{r}(\cdot)} = 0)}{\mathbb{P}(W^{x,r(\cdot)} = 0)} \frac{\mathbb{P}(W^{r(\cdot)} + S \geq x) - \mathbb{P}(W^{r(\cdot)} > x)}{\mathbb{P}(W^{r(\cdot)} \leq x)}.$$

This is an extension of Theorem 1.6.4, the classical formula for the distribution of $C_{\max}$ in the M/G/1 queue, which is due to Takács [161] (see also Cohen [52], and Asmussen and Perry [12] for alternative proofs). That result can be easily recovered from Corollary 4.4.3, since, for the M/G/1 queue, we have $r(x) \equiv \tilde{r}(x) \equiv 1$. This yields the well-known formula (see also Theorem 1.6.4)

$$\mathbb{P}(C_{\max} < x) = \frac{\mathbb{P}(W + S < x)}{\mathbb{P}(W \leq x)}.$$

Explicit results for $C_{\max}^{r(\cdot)}$ in terms of Volterra functions were presented in Chapter 3. Chapter 3 also contained related results for first-exit probabilities, as well as expressions for the distribution of $W^{r(\cdot)}$ in terms of Volterra functions, which can also be found in Harrison and Resnick [83]. Although Corollary 4.4.3 does not give a very explicit formula for the distribution of $C_{\max}$, in contrast to Theorem 3.5.1, we expect that this representation may be useful to obtain asymptotics and/or bounds. Asymptotic results in the light-tailed case are hardly known; see Asmussen [7, 8].

4.5 Conclusion

We have considered several queueing models which operate under the partial-rejection mechanism. For these models, we have shown that the loss probability of a customer can be identified with the tail distribution of the cycle maximum.

This chapter raises several questions that could be interesting for further research. First of all, we believe that an appropriate modification of Theorem 4.2.1 still holds for other queueing models, such as queueing models with Markov-modulated input. This is potentially useful, since the distribution of the cycle maximum is known for a large class of such models; see Asmussen and Perry [12].

Furthermore, we expect that Siegmund duality and related results can also be fruitful in other queueing problems. In the context of the present chapter, we believe that an analogue of (4.2) can be shown for queues which can be modeled as birth-and-death processes: Siegmund-type duality results for birth-and-death processes have been derived by Dette *et al.* [65].

Chapter 5

Optimal admission control in queues with workload-dependent service rates

5.1 Introduction

In the previous chapters we studied various queueing systems with workload-dependent service rates. In the present chapter we specifically consider the case in which the service rate is first increasing and then decreasing as a function of the amount of work present. In addition, the amount of work is controlled by an admission policy for accepting or rejecting arriving jobs, depending on the state of the system. We seek an admission policy that maximizes the long-run throughput, and show that, under certain conditions, a threshold policy is optimal. Because the workload process under the threshold policy is identical to the workload process of a queueing system with customer impatience, we can apply results from Chapter 3 to determine the optimal threshold value.

The study of queues with state-dependent rates is motivated in Chapter 1. A typical application for the model in this chapter concerns production systems where the productivity of the shop floor personnel depends on the level of work-in-process (workload). In particular, the productivity, i.e., the speed of the server, first increases when the workload is low until a certain optimum is attained and then decreases when the system reaches overload (caused by, e.g., stress factors), see Section 1.3. The latter qualitative behavior is quite characteristic of efficiency patterns observed in many practical scenarios.

The above-described M/G/1 queue with admission control may be modeled as a semi-Markov Decision Process (MDP). Most of the theory on MDP's concerns models with finite or countable state spaces. Because in the present queueing model both the admission policy and the service speed depend on the workload, we are dealing with an MDP with uncountable state space $[0, \infty)$. See for instance [145, 147, 150] for some general MDP's with infinite state spaces. To derive structural properties of the optimal policy, a commonly used approach in MDP's is the construction of value functions that possess certain concavity properties. Because the value functions in our model typically do not exhibit such behavior, we apply sample-path techniques to compare different policies.

An interesting related study is [67], where the author considers an M/G/1 queueing system with continuous-time arrival control and a fixed reward rate R when the server is busy and holding cost rate cx when the workload is x. Hence, such an M/G/1 queue can also be modeled as an MDP with the admission control depending on the system state, while the state space is infinite $[0, \infty)$. Using sample-path arguments and general theory on continuous-time MDP's developed in [66], the author proves the average-cost optimality of threshold policies.

Another branch of single-server queues with uncountable state spaces concerns M/G/1 queues with service control. Specifically, the service speed may be continuously adapted over time based on the residual amount of work. In [53], the service speed equals r_1 when the workload is less than some fixed level K and r_2 when the workload exceeds K. For some fairly general cost functions, the author determines the optimal switching level K. In [75, 162], the server works at constant speed, but can be switched on and off. The cost function includes holding cost and switching cost for turning the server on. The average-cost optimality of D-policies is shown in [75, 162]. In D-policies, the server is turned off only when the system becomes empty (while the server was on) and the server is turned on only when the workload exceeds level D (and the server was off).

This chapter is organized as follows. We give a detailed model description and several representations of the throughput in Section 5.2. In Section 5.3, the optimality of threshold policies under Assumption 5.2.1 (see Section 5.2) is shown. A criterion for the optimal threshold value is derived in Section 5.4. In Section 5.5 we present several examples of (combinations of) service speed functions and service requirement distributions satisfying Assumption 5.2.1. We explicitly determine the optimal threshold value and corresponding throughput in Section 5.6 for some special cases. Some concluding remarks and suggestions for further research are given in Section 5.7.

5.2 Model description

We consider an M/G/1 queue with a workload-dependent service rate. The customers (or jobs) arrive according to a Poisson process of rate λ. The service requirement of the n-th customer is B_n, $n = 1, 2, \ldots$, where the B_n are assumed to be independent, identically distributed copies of a random variable B with distribution $B(\cdot)$ and mean β. We also assume that the sequences of interarrival intervals and service requirements are independent.

The server works at a rate that depends on the amount of work in the system as described by some function $r(\cdot)$, i.e., the service rate is $r(x)$ when the amount of work is x. As in previous chapters and [83], we assume that $r(0) = 0$ and that $r(\cdot)$ is strictly positive, left-continuous, and has a right limit on $(0, \infty)$. In addition, we specifically focus on the case that $r(\cdot)$ is increasing on $(0, r_{\max}]$ and decreasing on $(r_{\max}, \infty)$ for some $r_{\max} \geq 0$.

The admission of work into the system is governed by a control policy which prescribes whether arriving customers are accepted or rejected, depending on the state of the system. We assume that the service requirement of a customer only becomes known after the acceptance decision, see Section 5.7 for a further discussion. Thus, the admission control policy may equivalently be interpreted as a rule for closing or opening access to the system.

We seek an admission control policy that maximizes the long-run throughput. The long-run throughput under policy π is defined as

$$TH^{\pi} := \lim_{t\to\infty} \frac{B^{\pi}(0,t)}{t},$$

assuming the limit to exist. Here $B^{\pi}(0,t)$ denotes the amount of work completed during $[0,t]$ under policy π. A policy π^* is said to be (strictly) optimal if $TH^{\pi^*} \geq TH^{\pi}$ ($TH^{\pi^*} > TH^{\pi}$) for all policies $\pi \neq \pi^*$.

For now, we restrict the attention to the class of stationary and deterministic policies that base their actions on the current amount of work in the system only. For a given policy π, we use $\pi(x) = 1$ to denote that it accepts a customer that arrives when the workload equals x and write $\pi(x) = 0$ otherwise. Later we will show that the found optimal policy is in fact optimal within a broader class that includes non-stationary and randomized policies as well.

Let V_t^{π} be the workload at time t and let W_n^{π} be the workload just before the n-th arrival epoch. Denote by V^{π} and W^{π} the random variables with the corresponding steady-state distributions, if they exist, and let $v^{\pi}(\cdot)$ be the density of V^{π}.

We first consider the case $\lambda\beta < r_{\infty}$, with $r_{\infty} := \lim_{x\to\infty} r(x)$. In that case, the system remains stable under the greedy policy that always accepts customers. Thus, the throughput achieved under the latter policy equals $\lambda\beta$, which is optimal, since the maximum achievable long-run throughput is bounded by the offered traffic load.

In the remainder of the chapter we focus on the case $\lambda\beta > r_{\infty}$. (The boundary case $\lambda\beta = r_{\infty}$ is rather delicate, and a full analysis is beyond the scope of the present chapter.) In that case, the system is unstable under the greedy policy that always accepts customers. Henceforth, we restrict the attention to policies π such that $\pi(x) = 0$ for all $x > M$ for some large M, which ensures the existence of the steady-state workload distribution. Even though the policy that always accepts customers may continue to be optimal, the maximum achievable throughput can be approached arbitrarily close for sufficiently large M.

Since the steady-state workload distribution exists, the throughput TH^{π} under policy π as defined above may in fact be expressed in several alternative ways. Observing that $B^{\pi}(0,t) = \int_0^t r(V_u^{\pi})\mathrm{d}u$, the throughput may be equivalently written as

$$TH^{\pi} = \lim_{t\to\infty} \frac{1}{t}\int_0^t r(V_u^{\pi})\mathrm{d}u = \mathbb{E}[r(V^{\pi})] = \int_0^{\infty} r(x)v^{\pi}(x)\mathrm{d}x.$$

Invoking the further identity relation (with $A^{\pi}(0,t)$ denoting the amount of work accepted during $[0,t]$ under policy π)

$$B^{\pi}(0,t) = V_0^{\pi} + A^{\pi}(0,t) - V_t^{\pi},$$

and noting that $V_t^\pi/t \to 0$ as $t \to \infty$, we observe that the throughput may also be expressed as

$$TH^\pi = \beta \lim_{t\to\infty} \frac{N^\pi(0,t)}{t},$$

where $N^\pi(0,t)$ denotes the number of accepted customers during $[0,t]$ under policy π. Using the PASTA property, the above expression may be further rewritten as

$$TH^\pi = \lambda\beta\mathbb{P}(\pi(V^\pi)=1) = \lambda\beta\left(\pi(0)\mathbb{P}(V^\pi=0) + \int_0^\infty \pi(x)v^\pi(x)\mathrm{d}x\right). \qquad (5.1)$$

Finally, we introduce some additional notation. Define

$$R(x) := \int_0^x \frac{1}{r(y)}\mathrm{d}y, \qquad 0 < x < \infty,$$

representing the time required for the system to empty in the absence of any arrivals, starting from workload x. In order to avoid technicalities, we assume that $R(x) < \infty$ for all $x > 0$, as in [83] and parts of Chapters 2–4. Moreover, we assume that $\mathbb{E}[R(x+B+\delta)-R(x+B)] \to 0$, as $\delta \to 0$. The latter condition only rules out cases where the workload process is being absorbed in some positive workload level and is satisfied if, for instance, $r_\infty > 0$ or if B has finite support. Further define

$$Z(x) := \int_0^\infty \int_x^{x+y} \frac{1}{r(z)}\mathrm{d}z\mathrm{d}B(y) = \int_x^\infty \frac{1}{r(z)}(1-B(z))\mathrm{d}z, \qquad (5.2)$$

representing the expected time required for the system to return to workload x after a customer has been accepted, in the absence of any further arrivals. In the remainder of the chapter, we make the following assumption with regard to $Z(x)$.

Assumption 5.2.1 *There exists some* $z_{\min} \geq 0$ *such that* $Z(x)$ *is decreasing on* $[0, z_{\min}]$ *and increasing on* $[z_{\min}, \infty)$.

The above assumption is satisfied for a wide class of M/G/1-type models with workload-dependent service rates. We give several illustrative examples in Section 5.5.

To provide some intuition, suppose that the system operates according to the Last-Come First-Served Preemptive-Resume (LCFS-PR) discipline, which does not affect the workload process in any way. With that view in mind, $Z(x)$ may be thought of as the expected service time of a customer that arrives when the workload equals x, and $z_{\min}$ represents the workload level at which arriving customers have the minimum expected service time. Thus, from the LCFS-PR perspective, the direct reward of accepting customers is first increasing (on $(0, z_{\min}]$) and then decreasing (on $(z_{\min}, \infty)$). However, the decision to either accept or reject also affects future rewards (service times). In Section 5.3, insights from the LCFS-PR discipline are applied to show that the optimal policy has a threshold structure when Assumption 5.2.1 is satisfied.

5.3 Optimality of threshold policies

In the first part of this section, we only consider stationary deterministic policies. Since the actions of the admission control policy then only depend on the workload level x, we will also for brevity refer to the value of x as the state of the system. An excursion from state x is then a period that starts with the acceptance of a customer in state x and ends with the first subsequent return to state x. For conciseness, we will frequently write that a policy accepts/rejects in an interval $[v, w]$ when it accepts/rejects customers that arrive when the workload is in the interval $[v, w]$. In the second part of this section, we show that the found optimal policy is in fact optimal within a broader class that also includes non-stationary and randomized policies.

Define $N^\pi(x)$ and $T^\pi(x)$ as follows:
$N^\pi(x) \equiv$ expected number of accepted customers during an excursion from state x under policy π.
$T^\pi(x) \equiv$ expected duration of an excursion from state x under policy π.
It may be verified that $N^\pi(x)$ and $T^\pi(x)$ are continuous, see also the proof of Lemma 5.3.2.

Consider an arbitrary policy π that rejects in $[x, x+\delta]$. Let π' be a modified policy, which does the same as π except that it accepts in $[x, x+\delta]$. Let $G^\pi(y)$ be the expected number of excursions during a busy cycle that start from a workload level below y under policy π, which are not part of an excursion starting from a level $z \in [x, y]$, $y \geq x$.

Lemma 5.3.1 *For some $\gamma \in (0, 1)$, we have*

$$TH^{\pi'} = (1-\gamma) TH^\pi + \gamma \frac{\int_x^{x+\delta} \beta N^{\pi'}(y) \mathrm{d}G^{\pi'}(y)}{\int_x^{x+\delta} T^{\pi'}(y) \mathrm{d}G^{\pi'}(y)}.$$

Proof By [148, Theorem 1], the throughput under policy π may be equivalently expressed as $TH^\pi = \mathbb{E}R^\pi / \mathbb{E}T^\pi$, where R^π is the reward (i.e., amount of work served) during a busy cycle and T^π is the cycle length under policy π. Consider a busy cycle and take an arbitrary sample path of the workload process $\{V_t^{\pi'}, t \geq 0\}$ under policy π'. We construct a stochastic process $\hat{V}_t$ by deleting the excursions from level $y \in [x, x+\delta]$ and pasting together the remaining parts. First note that the residual interarrival time at a downcrossing of y is still exponential (see, e.g., Lemma 3.4.1). Now, it may be readily checked that $\hat{V}_t$ and V_t^π have the same statistical properties. Thus, for the expected number of accepted customers during a busy cycle under policy π', $\mathbb{E}N^{\pi'}$, we have

$$\mathbb{E}N^{\pi'} = \mathbb{E}N^\pi + \int_x^{x+\delta} N^{\pi'}(y) \mathrm{d}G^{\pi'}(y),$$

and, equivalently, for the expected duration of a busy cycle

$$\mathbb{E}T^{\pi'} = \mathbb{E}T^\pi + \int_x^{x+\delta} T^{\pi'}(y) \mathrm{d}G^{\pi'}(y).$$

Using Wald's theorem, we derive

$$\begin{aligned}\frac{\mathbb{E}R^{\pi'}}{\mathbb{E}T^{\pi'}} &= \frac{\beta\mathbb{E}N^{\pi'}}{\mathbb{E}T^{\pi'}} = \frac{\beta(\mathbb{E}N^{\pi} + \int_x^{x+\delta} N^{\pi'}(y)\mathrm{d}G^{\pi'}(y))}{\mathbb{E}T^{\pi} + \int_x^{x+\delta} T^{\pi'}(y)\mathrm{d}G^{\pi'}(y)} \\ &= (1-\gamma)\frac{\mathbb{E}R^{\pi}}{\mathbb{E}T^{\pi}} + \gamma\frac{\int_x^{x+\delta} \beta N^{\pi'}(y)\mathrm{d}G^{\pi'}(y)}{\int_x^{x+\delta} T^{\pi'}(y)\mathrm{d}G^{\pi'}(y)},\end{aligned}$$

where $\gamma = \frac{\int_x^{x+\delta} T^{\pi'}(y)\mathrm{d}G^{\pi'}(y)}{\mathbb{E}T^{\pi}+\int_x^{x+\delta} T^{\pi'}(y)\mathrm{d}G^{\pi'}(y)}$ represents the fraction of time spent on excursions starting between x and $x+\delta$. This completes the proof. □

Let π^* denote an optimal policy, with corresponding throughput $TH^* = \mathbb{E}[r(V^{\pi^*})]$.

Lemma 5.3.2 *(Optimality properties)*

(i) it is strictly optimal to reject in $[v, w] \Longrightarrow \frac{\beta N^{\pi^*}(x)}{T^{\pi^*}(x)} < TH^*$, *for almost every* $x \in [v, w]$.

(ii) it is optimal to accept in $[v, w] \Longrightarrow \frac{\beta N^{\pi^*}(x)}{T^{\pi^*}(x)} \geq TH^*$, $\forall x \in [v, w]$.

Note that the inequality in (i) may hold with equality for some $x \in [v, w]$.

Proof We first prove that

$$\frac{\int_x^{x+\delta} N^{\pi'}(y)\mathrm{d}G^{\pi'}(y)}{\int_x^{x+\delta} T^{\pi'}(y)\mathrm{d}G^{\pi'}(y)} \to \frac{N^{\pi}(x)}{T^{\pi}(x)}, \qquad \text{as } \delta \downarrow 0. \tag{5.3}$$

For some small $\delta > 0$ and $y \in [x, x+\delta]$, we have

$$T^{\pi'}(y) - T^{\pi}(x+\delta) \leq (R(x+\delta) - R(y))(1 + \lambda \max_{y \leq u \leq x+\delta} T^{\pi'}(u)) \to 0, \qquad \text{as } \delta \downarrow 0,$$

where $R(x+\delta) - R(y)$ is the time required to go from $x+\delta$ to y in the absence of any arrivals. Similarly, as $\delta \downarrow 0$, $T^{\pi'}(y) - T^{\pi}(x+\delta)$ can be bounded from below by

$$-\mathbb{E}[R(x+B+\delta) - R(y+B)](1 + \lambda \max_{y+B \leq u \leq x+B+\delta} \pi'(u)T^{\pi'}(u)) \to 0.$$

Applying similar arguments to $N^{\pi'}(y)$ then yields (5.3). (Another way to see that (5.3) holds, is to observe that the density $\mathrm{d}G^{\pi}(\cdot)$ is well-defined.)

The remainder of the proof is by contradiction. For part (i), assume that the strictly optimal policy π^* rejects in $[v, w]$, but there is some interval $(u, u+\delta) \subseteq [v, w]$, with $\delta > 0$, such that $\beta N^{\pi^*}(x)/T^{\pi^*}(x) \geq TH^*$ for $x \in (u, u+\delta)$. Consider a modified policy π which accepts in $[u, u+\delta]$ and follows π^* otherwise. First using Lemma 5.3.1 and then letting $\delta \downarrow 0$ (and using (5.3)), it follows that $\mathbb{E}[r(V^{\pi})] \geq TH^*$, contradicting the strict optimality of π^*.

For part (ii), assume that the optimal policy π^* accepts in $[v, w]$ but, for some $x \in [v, w]$, $\beta N^{\pi^*}(x)/T^{\pi^*}(x) < TH^*$. Using (5.3), it follows that there is some interval $U := (u - \delta, u + \delta) \subseteq [v, w]$ such that $\beta N^{\pi^*}(x)/T^{\pi^*}(x) < TH^*$ for every $x \in U$. Consider the modified policy π that rejects in U and follows π^* otherwise. Using Lemma 5.3.1 (with $\pi' \equiv \pi^*$), it is easily seen that $\mathbb{E}[r(V^\pi)] > TH^*$, contradicting the optimality of π^*. □

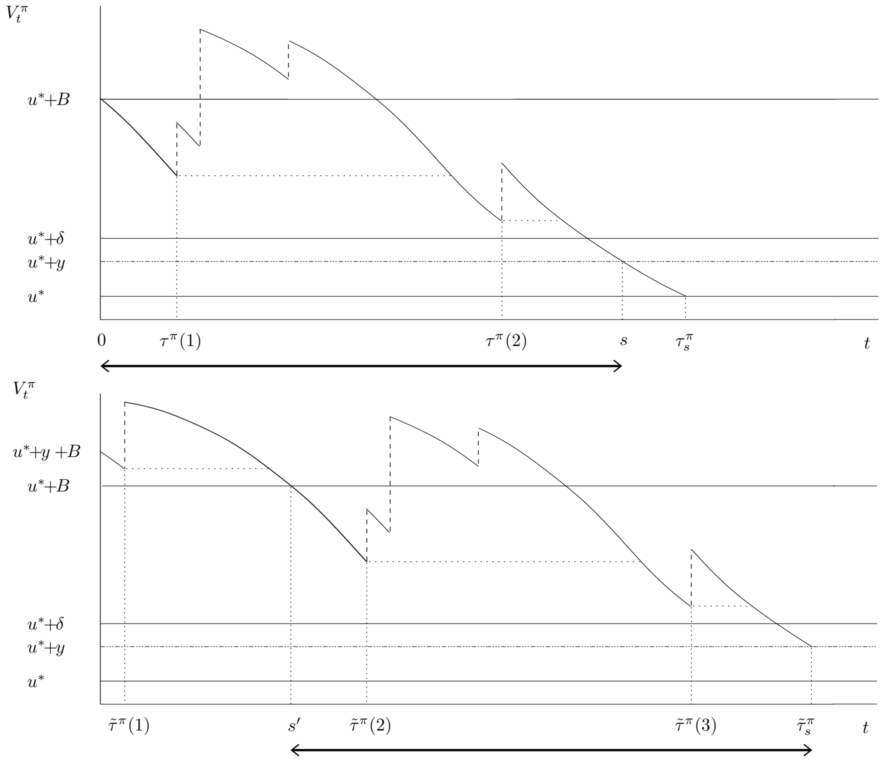

Figure 5.1: The sample paths of two excursions of V_t^π; one excursion from state u^* and one excursion from state $u^* + y$. In this example $N = 2$ and $M = 3$.

Lemma 5.3.3 *It is optimal to accept in* $[0, z_{\min}]$.

Proof It is obvious that it is optimal to accept in an empty system. Now, assume that it is not optimal to accept in $[0, z_{\min}]$. Then there is some policy π, such that $\pi(x) = 1$ for $x \in [0, u^*]$, but $\pi(x) = 0$ for $x \in (u^*, u^* + \delta]$, with $u^* + \delta \leq z_{\min}$ and $\delta > 0$, that is strictly optimal.

Take some arbitrary $0 < y < \delta$. In the proof, we compare $N^\pi(u^*)$ and $T^\pi(u^*)$ with $N^\pi(u^*+y)$ and $T^\pi(u^*+y)$. Using stochastic coupling, we show that $\beta N^\pi(u^* + y)/T^\pi(u^* + y)$ may be written as a combination of $\beta N^\pi(u^*)/T^\pi(u^*)$

and possibly contributions from some additional excursions. Since π is assumed to be optimal, both terms provide an average reward of at least TH^* by Lemma 5.3.2(ii). By Lemma 5.3.2(i), this contradicts the strict optimality of rejecting in $(u^*, u^*+\delta]$, because the coupling holds for any $y \in (0, \delta)$.

For the first part in the stochastic coupling, i.e., the part of the excursion from u^*+y related to $\beta N^\pi(u^*)/T^\pi(u^*)$, observe that it follows from Assumption 5.2.1 that $Z(u^*) \geq Z(u^*+y)$, implying that the direct reward of accepting customers at level u^*+y is at least as high as the direct reward of accepting at level u^*. For the second part, we use the fact that we only make additional excursions if they are advantageous.

First consider the expected duration of an excursion from level u^* under policy π, and the expected number of accepted customers during such an excursion (i.e., $N^\pi(u^*)$ and $T^\pi(u^*)$). Let the first jump, initiating an excursion, occur at time 0 and observe that the workload level right after the first jump equals u^*+B, i.e., $V^\pi_{0+} = u^*+B$. Note that the workload process attains local minima just before arrival instants at which customers are going to be accepted. Using terminology of random walks, define a stopping time $\tau^\pi_s := \inf\{t \geq 0 : V^\pi_t \leq u^*\}$, an equivalent notion measured in the number of arrivals $\tau^\pi := \inf\{k \geq 0 : W^\pi_k \leq u^*\}$, and a sequence of descending ladder epochs $\tau^\pi(1) < \cdots < \tau^\pi(N) < \tau^\pi$ with corresponding descending ladder heights $u^*+B > W^\pi_{\tau^\pi(1)} > \cdots > W^\pi_{\tau^\pi(N)} > u^*$, as follows: $\tau^\pi(1) := \inf\{0 \leq k \leq \tau^\pi : \pi(W^\pi_k) = 1\}$, and for $n = 2, \ldots, N$ (if $\tau^\pi(1) < \tau^\pi$)

$$\tau^\pi(n+1) := \inf\{\tau^\pi(n) < k < \tau^\pi : W^\pi_k < W^\pi_{\tau^\pi(n)}, \pi(W^\pi_k) = 1\}.$$

Note that $W^\pi_{\tau^\pi(N)} > u^*+\delta$, since $\pi(x) = 0$ for $x \in [u^*, u^*+\delta]$. A typical sample path in case $N = 2$ is depicted in the first part of Figure 5.1. Using the above, we may write

$$N^\pi(u^*) \quad = \quad 1 + \sum_{n=1}^{N} N^\pi(W^\pi_{\tau^\pi(n)}), \tag{5.4}$$

$$T^\pi(u^*) \quad = \quad Z(u^*) + \sum_{n=1}^{N} T^\pi(W^\pi_{\tau^\pi(n)}). \tag{5.5}$$

Now consider $N^\pi(u^*+y)$ and $T^\pi(u^*+y)$. In this case, at time 0 the workload jumps to u^*+y+B, i.e., $V^\pi_{0+} = u^*+y+B$. As defined above, we have a stopping time $\tilde{\tau}^\pi_s$, a discrete-time equivalent $\tilde{\tau}^\pi$, and a sequence of descending ladder epochs $0 < \tilde{\tau}^\pi(1) < \cdots < \tilde{\tau}^\pi(M) < \tilde{\tau}^\pi$ with corresponding descending ladder heights $u^*+y+B > W^\pi_{\tilde{\tau}^\pi(1)} > \cdots > W^\pi_{\tilde{\tau}^\pi(M)} > u^*+y$ (see the second part of Figure 5.1 for a typical realization). Observe that the residual interarrival time at a downcrossing of u^*+B is still exponential. Hence, using stochastic coupling and the fact that $W^\pi_{\tau^\pi(N)} > u^*+\delta$, the descending ladder epochs may be divided into two sets: (i) $\tilde{\tau}^\pi(1), \ldots, \tilde{\tau}^\pi(M-N)$ with $u^*+y+B > W^\pi_{\tilde{\tau}^\pi(1)} > \cdots > W^\pi_{\tilde{\tau}^\pi(M-N)} > u^*+B$; and (ii) $\tilde{\tau}^\pi(M-N+1), \ldots, \tilde{\tau}^\pi(M)$ such that $W^\pi_{\tilde{\tau}^\pi(n+M-N)} =^d W^\pi_{\tau^\pi(n)}$ for $n = 1, \ldots, N$. This coupling is illustrated in Figure 5.1 (with $N = 2$ and $M = 3$). In this figure, the sample paths in the range of the solid arrow (that is between $[0, s]$ and $[s', \tilde{\tau}^\pi_s]$ respectively) are identical. Using the arguments above, we have

$$N^\pi(u^* + y) = 1 + \sum_{n=1}^{M-N} N^\pi(W^\pi_{\tilde{\tau}^\pi(n)}) + \sum_{n=1}^{N} N^\pi(W^\pi_{\tau^\pi(n)}), \quad (5.6)$$

$$T^\pi(u^* + y) = Z(u^* + y) + \sum_{n=1}^{M-N} T^\pi(W^\pi_{\tilde{\tau}^\pi(n)}) + \sum_{n=1}^{N} T^\pi(W^\pi_{\tau^\pi(n)}). \quad (5.7)$$

Since $W^\pi_{\tilde{\tau}^\pi(n)}$, $n = 1, \dots, M - N$, are the workloads just before an arriving customer is accepted and π is the supposed optimal policy, Lemma 5.3.2 yields

$$\frac{\beta \sum_{n=1}^{M-N} N^\pi(W^\pi_{\tilde{\tau}^\pi(n)})}{\sum_{n=1}^{M-N} T^\pi(W^\pi_{\tilde{\tau}^\pi(n)})} \geq \beta \min_{n=1,\dots,M-N} \frac{N^\pi(W^\pi_{\tilde{\tau}^\pi(n)})}{T^\pi(W^\pi_{\tilde{\tau}^\pi(n)})} \geq TH^*. \quad (5.8)$$

Moreover, using (5.4) and (5.5) in addition to Assumption 5.2.1, we obtain

$$\frac{\beta(1 + \sum_{n=1}^{N} N^\pi(W^\pi_{\tau^\pi(n)}))}{Z(u^* + y) + \sum_{n=1}^{N} T^\pi(W^\pi_{\tau^\pi(n)})} \geq \frac{\beta N^\pi(u^*)}{T^\pi(u^*)} \geq TH^*, \quad (5.9)$$

where the second inequality relies on the fact that it is optimal to accept at level u^*. Combining (5.6)-(5.9) yields

$$\begin{aligned} &\frac{\beta N^{\pi'}(u^* + y)}{T^{\pi'}(u^* + y)} \\ \geq\ & \min\left\{ \frac{\beta \sum_{n=1}^{M-N} N^\pi(W^\pi_{\tilde{\tau}^\pi(n)})}{\sum_{n=1}^{M-N} T^\pi(W^\pi_{\tilde{\tau}^\pi(n)})}, \frac{\beta(1 + \sum_{n=1}^{N} N^\pi(W^\pi_{\tau^\pi(n)}))}{Z(u^* + y) + \sum_{n=1}^{N} T^\pi(W^\pi_{\tau^\pi(n)})} \right\} \\ \geq\ & TH^*. \end{aligned}$$

By Lemma 5.3.2 it can thus not be strictly optimal to reject at level $u^* + y$, $0 < y < \delta$. □

Theorem 5.3.1 *There exists a threshold policy that is optimal among the class of stationary deterministic policies.*

Proof It follows from Lemma 5.3.3 that it is optimal to accept when the workload is in $[0, z_{\min}]$. Suppose that a threshold policy is not optimal, i.e., there exists some policy π that is strictly better than any threshold policy. Let $n^\pi := \int_0^\infty \max(\pi(x^+) - \pi(x), 0)\mathrm{d}x$ be the number of "gaps" of policy π, i.e., the number of times $\pi(\cdot)$ switches from 0 to 1. Let π be an optimal policy, which is strictly better than any threshold policy, with the least number of gaps, that is, $\pi = \arg\min_{\pi \in \Pi^*} n^\pi$, with Π^* the class of optimal policies. This implies that there is some $u^* > z_{\min}$ and $\delta_2 > \delta_1 > 0$ such that $\pi(x) = 0$ on $(u^*, u^* + \delta_1)$ and

$\pi(x) = 1$ on $(u^* + \delta_1, u^* + \delta_2)$. We note that gaps consisting of singular points can be removed.

Take some arbitrary $0 < y < \delta_1$. In the proof, we consider $N^\pi(u^* + y)$ and $T^\pi(u^* + y)$. Using the fact that it is optimal to accept in $(u^* + \delta_1, u^* + \delta_2)$, we show that $\beta N^\pi(u^* + y)/T^\pi(u^* + y) \geq TH^*$ (contradicting the fact that π contains the least number of gaps among policies in Π^*). This follows from the fact that the direct reward of accepting at level $u^* + y$ exceeds the reward of accepting at any level $x > u^* + y$. Moreover, additional excursions are only made when they are advantageous.

Suppose that at time 0 an arriving customer with service requirement B is accepted when the workload equals $u^* + y$, i.e., $V^\pi_{0+} = u^* + y + B$. As in the proof of Lemma 5.3.3 (see also the first part of Figure 5.1, with $\delta_1 \equiv \delta$), we may define "stopping times" τ^π_s and τ^π and a sequence of descending ladder epochs $\tau^\pi(1) < \cdots < \tau^\pi(N) < \tau^\pi$ with corresponding descending ladder heights $u^* + y + B > W^\pi_{\tau^\pi(1)} > \cdots > W^\pi_{\tau^\pi(N)} > u^* + y$. Note that $W^\pi_{\tau^\pi(N)} > u^* + \delta_1$ (if $N > 0$), since $\pi(x) = 0$ for $x \in [u^*, u^* + \delta_1]$. Applying this construction yields

$$\begin{aligned} N^\pi(u^* + y) &= 1 + \sum_{n=1}^{N} N^\pi(W^\pi_{\tau^\pi(n)}), \\ T^\pi(u^* + y) &= Z(u^* + y) + \sum_{n=1}^{N} T^\pi(W^\pi_{\tau^\pi(n)}). \end{aligned}$$

By Lemma 5.3.2, $\beta \sum_{n=1}^{N} N^\pi(W^\pi_{\tau^\pi(n)}) / \sum_{n=1}^{N} T^\pi(W^\pi_{\tau^\pi(n)}) \geq TH^*$ since π is assumed to be an optimal policy. Moreover, using a similar ladder height construction, it may be easily checked (in general) that

$$\frac{N^\pi(x)}{T^\pi(x)} \leq \max_{v \geq x} \frac{1}{Z(v)}. \tag{5.10}$$

Hence, invoking Assumption 5.2.1 yields

$$\frac{\beta}{Z(u^* + y)} \geq \frac{\beta}{Z(u^* + \delta_1)} \geq TH^*.$$

Combining the above, we obtain $\beta N^\pi(u^* + y)/T^\pi(u^* + y) \geq TH^*$ for any $y \in (0, \delta_1)$. By Lemma 5.3.2, this contradicts the fact that policy π has the minimum number of gaps among the class of optimal policies Π^*. □

The ladder height construction in the proof of Theorem 5.3.1 allows us to generalize Relation (5.10):

Proposition 5.3.1 *For the throughput during an excursion from level x, we have the following bounds,*

$$\min_{v \geq x : \pi(v) = 1} \frac{1}{Z(v)} \leq \frac{N^\pi(x)}{T^\pi(x)} \leq \max_{v \geq x : \pi(v) = 1} \frac{1}{Z(v)}.$$

These bounds are especially natural from the perspective of the LCFS-PR discipline. In that view, the proposition simply states that the throughput during an excursion from level x is at least the minimum (and at most the maximum) of one over the mean service time of accepting at any level above x if policy π is applied.

Remark 5.3.1 The proof of Theorem 5.3.1 crucially depends on the fact that $Z(\cdot)$ has only one local minimum, i.e., Assumption 5.2.1. Suppose for the moment that $Z(\cdot)$ has L local minima. Thus, $Z(\cdot)$ is decreasing on $[z^k_{\max}, z^k_{\min})$ and increasing on $[z^k_{\min}, z^{k+1}_{\max})$, $k = 1, \ldots, L$, where $z^1_{\max} = 0$ and $z^{L+1}_{\max} = \infty$. Similar to the proof of Lemma 5.3.3, we deduce that if $\pi(x) = 1$ for some $x \in [z^k_{\max}, z^k_{\min})$, then $\pi(y) = 1$ for all $y \in [x, z^k_{\min})$ (note that $\pi(0) = 1$ and accepting is thus optimal in $[0, z^1_{\min})$). Also, it follows from the proof of Theorem 5.3.1 that if $\pi(x) = 1$ for some $x \in [z^L_{\max}, \infty)$, then $\pi(y) = 1$ for all $y \in [z^L_{\max}, x)$. However, the intervals $[z^k_{\min}, z^{k+1}_{\max})$, $k = 1, \ldots, L-1$, are not covered by the proof. In particular, the trade-off between direct and future rewards remains undecided there. ◇

Theorem 5.3.1 shows that the threshold policy is optimal among the class of stationary and deterministic policies. To prove that a (stationary and deterministic) threshold policy is also optimal within the broader class of policies considered in [148], we use insights from this section to construct an appropriate (value) function satisfying [148, Theorem 2]. The class of policies in [148] consists of all measurable decision rules, and includes non-stationary and non-deterministic policies.

Theorem 5.3.2 *There exists a threshold policy that is optimal within the class of policies considered in [148].*

Proof Let π be a threshold policy with threshold value x^* that is optimal within the class of stationary and deterministic policies. Now, define $n^\pi(x)$ and $t^\pi(x)$ as follows:
$n^\pi(x) \equiv$ expected amount of work served in a period starting with workload level x until the end of the busy cycle under policy π.
$t^\pi(x) \equiv$ expected length of a period starting with workload level x until the end of the busy cycle under policy π.
Similar to [67], let

$$\tilde{f}(x) := n^\pi(x) - TH^\pi t^\pi(x).$$

Consider $\mathbb{E}[\tilde{f}(x+B)]$ and divide the busy cycle in two parts; first we have an excursion from state x followed by the remaining part of the cycle starting with a downcrossing of level x. Hence (see also [67, Lemma 6.3]),

$$\mathbb{E}[\tilde{f}(x+B)] = \beta\,(N^\pi(x) - 1) - TH^\pi T^\pi(x) + \tilde{f}(x), \tag{5.11}$$

where the $N^\pi(x) - 1$ stems from the fact that the arrival in state x is not counted in $\mathbb{E}[\tilde{f}(x+B)]$.

Define, for $x \geq 0$,

$$f(x) := \begin{cases} \beta + \mathbb{E}[\tilde{f}(x+B)], & \text{for } 0 \leq x \leq x^*, \\ \tilde{f}(x), & \text{for } x > x^*, \end{cases} \tag{5.12}$$

where x is the state of the system just before a decision epoch. By conditioning on the first arrival, we also obtain the following relationship between $\tilde{f}(\cdot)$ and $f(\cdot)$:

$$\tilde{f}(x) = \int_0^\infty f(R^{-1}((R(x)-y)^+))\lambda e^{-\lambda y}\mathrm{d}y - \frac{TH^\pi}{\lambda}, \tag{5.13}$$

with $R^{-1}(\cdot)$ the inverse function of $R(\cdot)$, see e.g. Chapters 2 and 3 for details. Because π is assumed to be an optimal stationary deterministic policy, Lemma 5.3.2 yields that $\beta N^\pi(x) - TH^\pi T^\pi(x)$ is positive for $x \in [0, x^*)$, and non-positive for $x \in [x^*, \infty)$. Using the above in addition to (5.11) and (5.12), we obtain

$$f(x) = \max\{\beta N^\pi(x) - TH^\pi T^\pi(x), 0\} + \tilde{f}(x). \tag{5.14}$$

Combining (5.11) with (5.13), we may rewrite (5.14) into

$$\begin{aligned} f(x) \quad = \quad & \max\Bigg\{\beta + \int_0^\infty \int_0^\infty f(R^{-1}((R(x+z)-y)^+))\lambda e^{-\lambda y}\mathrm{d}y\mathrm{d}B(z), \\ & \int_0^\infty f(R^{-1}((R(x)-y)^+))\lambda e^{-\lambda y}\mathrm{d}y\Bigg\} - \frac{TH^\pi}{\lambda}. \end{aligned}$$

Thus the function $f(\cdot)$ satisfies the optimality equation for the average-cost criterion, i.e. Equation (3) in [148]. The theorem now follows directly from [148, Theorem 2]. □

5.4 Criterion for the optimal threshold

In Section 5.3 we showed that, if Assumption 5.2.1 is satisfied, a threshold policy is optimal. The derivation of that result also suggested the following criterion for the optimal threshold:

$$TH^{\pi_{\bar{x}}} = \frac{\beta}{Z(\bar{x})}, \tag{5.15}$$

where $\pi_{\bar{x}}$ denotes a threshold policy with parameter $\bar{x}$. The above criterion is intuitively appealing when we consider marginal arguments. Informally speaking, the optimal threshold will be chosen such that the throughput just equals the expected reward of customers accepted in state $\bar{x}$ (which has reward $\beta/Z(\bar{x})$).

Moreover, the above criterion allows us to deduce some properties of the optimal threshold value. Using a similar construction as in (some of) the proofs of Section 5.3, it may be shown that $TH^{\pi_{\bar{x}}}$ is increasing as a function of λ. (To see this, we note that a higher λ yields additional arrivals which are only accepted if the resulting excursions are advantageous.) Because $\beta/Z(\bar{x})$ is independent of λ we can directly conclude from (5.15) that the optimal threshold value is decreasing in λ. It may also easily be checked that the optimal threshold approaches $z_{\min}$ as $\lambda \to \infty$. This behavior of the optimal threshold reveals the typical trade-off between direct and future rewards; the upper bound for the throughput is attained by accepting customers in state $z_{\min}$, but the optimal policy anticipates decreasing arrival rates by starting to accept customers at increasing workload levels to compensate for the increased probability of reaching an empty system (where the server is idle).

In the remainder of this section, we use another method to derive a criterion for the optimal threshold value and give some properties of $TH^{\pi_{\bar{x}}}$ as a function of $\bar{x}$. Moreover, when $Z(\cdot)$ does not satisfy Assumption 5.2.1, we show that a similar criterion as (5.15) holds for the optimal threshold value, which provides the optimal policy within the class of threshold policies. (Note that a threshold strategy may then not be optimal among the class of stationary and deterministic policies). However, we start with the general form of the throughput under a threshold strategy with some fixed threshold $\bar{x}$.

Observe that, for fixed $\bar{x}$, the workload under policy $\pi_{\bar{x}}$ has the same dynamics as an M/G/1 queue with a general service rate and impatience of customers depending on the amount of work found upon arrival. Under policy $\pi_{\bar{x}}$ the model is in fact a special case of the finite-buffer queue in Chapter 3, with

$$v^{\pi_{\bar{x}}}(x) = \begin{cases} \mathbb{P}(V^{\pi_{\bar{x}}} = 0)K^*(x,0), & 0 < x \leq \bar{x}, \\ \mathbb{P}(V^{\pi_{\bar{x}}} = 0)\left[K(x,0) + \int_0^{\bar{x}} K(x,y)K^*(y,0)\mathrm{d}y\right], & x > \bar{x}, \end{cases}$$

where $\mathbb{P}(V^{\pi_{\bar{x}}} = 0)$ follows from normalization:

$$\begin{aligned} \mathbb{P}(V^{\pi_{\bar{x}}} = 0) \quad = \quad & \left[1 + \int_0^{\bar{x}} K^*(x,0)\mathrm{d}x + \int_{\bar{x}}^{\infty} K(x,0)\mathrm{d}x \right. \\ & \left. + \int_{\bar{x}}^{\infty}\int_0^{\bar{x}} K(x,y)K^*(y,0)\mathrm{d}y\mathrm{d}x\right]^{-1}. \end{aligned} \tag{5.16}$$

Here, the (iterated) kernels are defined as in Section 3.3 and [83]. That is, for $0 \leq y < x < \infty$, $K(x,y) := \lambda(1 - B(x-y))/r(x)$,

$$K_{n+1}(x,y) := \int_y^x K(x,z)K_n(z,y)\mathrm{d}z, \tag{5.17}$$

and $K^*(x,0) := \sum_{n=1}^{\infty} K_n(x,0)$. Using the representation in (5.1) for the throughput, we obtain

$$TH^{\pi_{\bar{x}}} = \lambda\beta\mathbb{P}(V^{\pi_{\bar{x}}} = 0)\left(1 + \int_0^{\bar{x}} K^*(x,0)\mathrm{d}x\right). \tag{5.18}$$

Note that $Z(\bar{x})$ and $TH^{\pi_{\bar{x}}}$ are continuous and differentiable functions of $\bar{x}$. In order to determine the optimal threshold, it is useful to consider the derivative of $TH^{\pi_{\bar{x}}}$ with respect to $\bar{x}$.

Lemma 5.4.1 *For the derivative of $TH^{\pi_{\bar{x}}}$, we have*

$$\frac{\mathrm{d}}{\mathrm{d}\bar{x}} TH^{\pi_{\bar{x}}} = \lambda\beta\mathbb{P}(V^{\pi_{\bar{x}}} = 0)K^*(\bar{x}, 0)\left[1 - TH^{\pi_{\bar{x}}} Z(\bar{x})/\beta\right].$$

Proof The proof is deferred to Appendix 5.A. □

Before we further discuss the optimal threshold criterion, we first derive some properties of $TH^{\pi_{\bar{x}}}$ as a function of $\bar{x}$. As in Lemma 5.3.1, consider a policy π that does not accept in $[a, b]$ and a modified policy π', which does the same as π except that $\pi'(x) = 1$ for $x \in [a, b]$. Then, the throughput under policy π' may be written as a convex combination of the throughput under policy π and the throughput due to excursions starting from levels in $[a, b]$ (see Lemma 5.3.1). This relation is particularly useful in studying the relationship between $TH^{\pi_{\bar{x}}}$ and $Z(\cdot)$.

Lemma 5.4.2 *Suppose that (i) $\mathrm{d}Z(x)/\mathrm{d}x \leq 0$, for $x \in [a, b]$, and (ii) $TH^{\pi_a} \leq \beta/Z(a)$. Then,*

$$TH^{\pi_x} \leq \frac{\beta}{Z(x)}, \qquad \textit{for all } x \in [a, b]. \tag{5.19}$$

If either (i) (for some $x \in [a, b]$) or (ii) holds with strict inequality, then (5.19) holds with strict inequality. Moreover, if the (strict) inequalities in (i) and (ii) are reversed, then the (strict) inequality in (5.19) is reversed.

Proof Fix an arbitrary $x \in (a, b]$. Lemma 5.3.1 yields that, for $\gamma \in (0, 1)$,

$$TH^{\pi_x} = (1 - \gamma)TH^{\pi_a} + \gamma \frac{\int_a^x \beta N^{\pi_x}(y)\mathrm{d}G^{\pi_x}(y)}{\int_a^x T^{\pi_x}(y)\mathrm{d}G^{\pi_x}(y)}.$$

From (i) and Proposition 5.3.1, we obtain $\beta N^{\pi_x}(y)/T^{\pi_x}(y) \leq \beta/Z(x)$ for every $y \in [a, x]$. Invoking (ii), it trivially follows that

$$TH^{\pi_x} \leq (1 - \gamma)\frac{\beta}{Z(a)} + \gamma\frac{\beta}{Z(x)} \leq \frac{\beta}{Z(x)}, \tag{5.20}$$

where the last step is due to (i) again. Now, if (i) holds with strict inequality for some $x \in [a, b]$ then the second inequality of (5.20) is strict, while the first one is strict if (ii) holds with strict inequality. The proof for the reversed signs is similar (use the lower bound in Proposition 5.3.1). □

We now derive a criterion for the optimal threshold. Let π^*_{th} denote the optimal threshold strategy. Define the set $A := \{x \geq 0 : TH^{\pi_x} = \beta/Z(x)\}$. Note that, in general, A is a collection of N disjoint closed intervals A_i, $i = 1, \ldots, N$, where each interval may be a singleton. However, if A_i is not a singleton, then it follows directly from Lemma 5.4.2 that $Z(\cdot)$ is constant on A_i.

Proposition 5.4.1 *If A is the empty set, then the greedy policy is optimal and $TH^{\pi^*_{th}} = r_\infty$. If A is non-empty,*

$$TH^{\pi^*_{th}} = \max\left\{ r_\infty, \max_{x\in A} \beta/Z(x) \right\},$$

where the greedy policy is optimal when $TH^{\pi^*_{th}} = r_\infty$ and the optimal (finite) threshold is given by any $\bar{x} \in \arg\max_{x\in A} \beta/Z(x)$ otherwise.

Proof For the threshold at 0 we have

$$TH^{\pi_0} = \frac{\beta}{Z(0)+\lambda} < \frac{\beta}{Z(0)}. \tag{5.21}$$

If A is the empty set, then we have from the continuity of $Z(\cdot)$ and $TH^{\pi_{\bar{x}}}$ that $TH^{\pi_x} < \beta/Z(x)$ for all x. Applying Lemma 5.4.1, we obtain that $\mathrm{d}TH^{\pi_x}/\mathrm{d}x > 0$ for all x and the greedy policy is thus optimal.

If $N > 0$, then it follows from Lemma 5.4.1 that A contains all points that satisfy $\mathrm{d}TH^{\pi_x}/\mathrm{d}x = 0$. Hence, A contains at least all extreme points. From (5.21) and Lemma 5.4.1 it follows that 0 is a local minimum. Moreover, $TH^{\pi_x} \to r_\infty$ as $x \to \infty$. Because $TH^{\pi_{\bar{x}}}$ is continuous, finding the global maximum of $TH^{\pi_{\bar{x}}}$ reduces to finding the maximum of $\beta/Z(x)$, $x \in A$, and comparing it with r_∞. □

Using Lemma 5.4.2 , some additional properties of $TH^{\pi_{\bar{x}}}$ as a function of $\bar{x}$ may be derived. For instance, it may be shown that if $Z(\cdot)$ has m local maxima, then $N \le 2m-1$. In particular, if Assumption 5.2.1 is satisfied, then $N \le 1$. This case is of special interest because a threshold policy is then optimal. Moreover, if in that case $N = 1$, then $A \cap [0, z_{\min})$ is empty and each value in A (possibly a singleton) is a global maximum of $TH^{\pi_{\bar{x}}}$. These arguments are summarized in the following corollary.

Corollary 5.4.1 *Suppose that Assumption 5.2.1 is satisfied. If A is the empty set, then the greedy policy is optimal. Otherwise, $N = 1$ and $x \in A$ is an optimal threshold with corresponding throughput*

$$TH^* = \beta/Z(x).$$

Finally, if $r(x)$ is constant for $x \ge L$ (and Assumption 5.2.1 holds), there is an easy way to determine directly whether the greedy policy is optimal or not. From Lemma 5.4.2 we then deduce the following:

Corollary 5.4.2 *Suppose Assumption 5.2.1 is satisfied and $r(x) = r_\infty$ for all $x \ge L$ for some $L > 0$. Then, the greedy policy is optimal if and only if*

$$TH^{\pi_L} \le \frac{\beta}{Z(L)} = r_\infty.$$

5.5 Assumption on $Z(x)$

Although Assumption 5.2.1 is quite natural, it involves the service-rate function as well as the distribution of the service requirement. Below, we give some examples satisfying this assumption, assuming that $r(\cdot)$ is increasing on $(0, r_{\max}]$ and decreasing on $(r_{\max}, \infty)$ for some $r_{\max} \geq 0$ (as described in Section 5.2). We consider both cases with general service requirement distributions and cases with a wide class of service-rate functions. In addition, we provide a natural example that does not have the desired properties. This case reveals the strong dependence on both the service-rate function and the service requirement distribution.

To show that Assumption 5.2.1 is satisfied, we frequently use the derivative of $Z(\cdot)$. Interchanging derivative and sum in addition to some rewriting, yields

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}Z(x) &= \int_0^\infty \frac{\mathrm{d}}{\mathrm{d}x}(R(x+b)-R(x))\mathrm{d}B(b)\\ &= \int_0^\infty \frac{1}{r(x+b)}\mathrm{d}B(b) - \frac{1}{r(x)} = \int_0^\infty \frac{r(x)-r(x+b)}{r(x+b)r(x)}\mathrm{d}B(b).\end{aligned}$$

For Assumption 5.2.1 to be satisfied, it remains to be shown that

$$\begin{aligned} r(x) - \mathbb{E}_B[r(x+B)] &\leq 0, \qquad x \in [0, z_{\min}],\\ r(x) - \mathbb{E}_B[r(x+B)] &\geq 0, \qquad x \in [z_{\min}, \infty).\end{aligned}$$

Example 5.5.1 Suppose that $r_{\max} = 0$, that is, $r(\cdot)$ is decreasing on the positive halfline. By definition, $r(y) \geq r(x)$, for $y > x$, and it is readily seen that Assumption 5.2.1 is satisfied. Also, $z_{\min} = 0$ in this case.

Example 5.5.2 Suppose that $B(x) = I(x \geq \beta)$, i.e., the service requirement is deterministic β. Observe that $\mathbb{E}_B[r(x+B)]$ is just the shifted $r(\cdot)$ function. Thus, $r(x) \leq \mathbb{E}_B[r(x+B)]$ if $x \in [0, (r_{\max} - \beta)^+]$ and $r(x) \geq \mathbb{E}_B[r(x+B)]$ if $x \in [r_{\max}, \infty)$. Moreover, $r(x)$ is increasing on $[(r_{\max} - \beta)^+, r_{\max}]$, while $\mathbb{E}_B[r(x+B)]$ is decreasing on the same interval. This directly yields the required property.

Example 5.5.3 Suppose that $B(x) = 1 - e^{-\mu x}$, meaning that the service requirement is exponentially distributed. Observe that $r(x) - \mathbb{E}_B[r(x+B)] \geq 0$ for $x \geq r_{\max}$. Now take some arbitrary x and y, with $0 < x < y \leq r_{\max}$. Conditioning on the service requirement in case a customer arrives at level x and using the memoryless property of the exponential distribution, we have

$$\begin{aligned}\mathbb{E}_B[r(x+B)] &= \int_x^y r(z)\mu e^{-\mu(z-x)}\mathrm{d}z + e^{-\mu(y-x)}\mathbb{E}_B[r(y+B)]\\ &\geq \int_x^y r(x)\mu e^{-\mu(z-x)}\mathrm{d}z + e^{-\mu(y-x)}\mathbb{E}_B[r(y+B)],\end{aligned}$$

and thus,

$$r(x) - \mathbb{E}_B[r(x+B)] \leq e^{-\mu(y-x)}r(x) - e^{-\mu(y-x)}\mathbb{E}_B[r(y+B)].$$

Note that if $r(y) - \mathbb{E}_B[r(y+B)] \leq 0$, then $r(x) - \mathbb{E}_B[r(x+B)] \leq 0$ (since $r(y) \geq r(x)$). Similarly, if $r(x) - \mathbb{E}_B[r(x+B)] \geq 0$, then $r(y) - \mathbb{E}_B[r(y+B)] \geq 0$. This directly gives the desired property, where $z_{\min} = \arg\inf\{h : r(h) \geq \mathbb{E}_B[r(h+B)]\}$.

Example 5.5.4 Suppose that $r(\cdot)$ is defined as follows:

$$r(x) = \begin{cases} r_1, & 0 < x \leq a, \\ \text{increasing and concave}, & a < x \leq r_{\max}, \\ \text{decreasing}, & x > r_{\max}. \end{cases}$$

In addition, assume that $r_\infty \geq r_1$. From the properties of $r(\cdot)$, it is obvious that $r(x) - \mathbb{E}_B[r(x+B)] \leq 0$ as $x \in (0, a]$ and $r(x) - \mathbb{E}_B[r(x+B)] \geq 0$ as $x \geq r_{\max}$. Hence, $a \leq z_{\min} \leq r_{\max}$. Now take arbitrary x, y, with $a \leq x < y \leq r_{\max}$. First, consider the following:

$$\begin{aligned} &\mathbb{E}_B[r(y+B) - r(x+B)] \\ &\quad = \int_0^{r_{\max}-x} (r(y+b) - r(x+b))\mathrm{d}B(b) + \int_{r_{\max}-x}^{\infty} (r(y+b) - r(x+b))\mathrm{d}B(b) \\ &\quad \leq \int_0^{r_{\max}-x} (r(y) - r(x))\mathrm{d}B(b) + 0 = (r(y) - r(x))B(r_{\max} - x), \end{aligned}$$

where we used that $r(\cdot)$ is concave on $[a, r_{\max}]$ and decreasing on $[r_{\max}, \infty)$ in the second step. Using the above, we obtain

$$\begin{aligned} &r(x) - \mathbb{E}_B[r(x+B)] \\ &\quad = r(y) - \mathbb{E}_B[r(y+B)] + r(x) - r(y) + \mathbb{E}_B[r(y+B)] - \mathbb{E}_B[r(x+B)] \\ &\quad \leq r(y) - \mathbb{E}_B[r(y+B)] + (r(x) - r(y))(1 - B(r_{\max} - x)). \end{aligned}$$

As in Example 5.5.3, note that if $r(y) - \mathbb{E}_B[r(y+B)] \leq 0$, then $r(x) - \mathbb{E}_B[r(x+B)] \leq 0$ (since $r(y) \geq r(x)$). Similarly, if $r(x) - \mathbb{E}_B[r(x+B)] \geq 0$, then $r(y) - \mathbb{E}_B[r(y+B)] \geq 0$. Hence, Assumption 5.2.1 is satisfied, with $z_{\min} = \arg\inf\{h : r(h) \geq \mathbb{E}_B[r(h+B)]\}$.

Finally, note that Example 5.5.1 is just a special case (take $a = r_{\max} = 0$). However, we believe that Example 5.5.1 is a natural special case, which admits an easy verification of Assumption 5.2.1.

Example 5.5.5 Here we provide an example for which Assumption 5.2.1 is not satisfied. For simplicity, we choose specific values for some model parameters. A slightly more general model could be constructed by leaving some parameters unspecified, while leaving the structure unaltered.

Consider the following service rate function:

$$r(x) = \begin{cases} r_1, & 0 < x \leq a, \\ (x-a)c + r_1, & a < x \leq r_{\max}, \\ (\hat{h} - x)c + r_1, & r_{\max} < x \leq b, \\ r_2 < r_1, & x > b, \end{cases}$$

with $\hat{h} = 2r_{\max} - a$, implying that $r(\hat{h}) = r_1$. Also, suppose that $B = a/3$ with probability 1/2 and $B = \hat{h} - a/3$ with probability 1/2, and take $c > 3(r_1 - r_2)/a$. After some calculations, we derive that $\mathrm{d}Z(x)/\mathrm{d}x$ is strictly positive on $(0, \frac{a}{3})$ and $(\frac{2a}{3} + \frac{r_1-r_2}{c}, \frac{4a}{3} - \frac{r_1-r_2}{c})$ and strictly negative on $(\frac{a}{3}, \frac{2a}{3} + \frac{r_1-r_2}{c})$ and $(\frac{4a}{3} - \frac{r_1-r_2}{c}, \infty)$. Clearly, $Z(\cdot)$ has two local minima and Assumption 5.2.1 is not satisfied in this case.

5.6 Some examples

In general, Expression (5.18) is suitable for a numerical calculation of the optimal threshold. Also, the characteristics of $TH^{\pi_{\bar{x}}}$ described in Section 5.4 suggest another numerical calculation of this optimal value, for instance, using a bisection method. In this section, we give some examples in which we obtain an analytically more tractable expression for the optimal throughput with corresponding optimal threshold value. In Subsection 5.6.1, we consider a two-level service rate: The service rate at time t is r_1 when $V_t^\pi \leq a$ and $r_2 < r_1$ when $V_t^\pi \geq a$ (see for instance [53]). In Subsection 5.6.2, we generalize the service rate to an arbitrary step function, but we restrict ourselves to exponential service requirements there.

In any case, if the greedy policy is not optimal, the optimal threshold value must satisfy Relation (5.15), see Proposition 5.4.1. Define

$$z(\bar{x}) := \left[\mathbb{P}(V^{\pi_{\bar{x}}} = 0)^{-1} - W(\bar{x})\lambda Z(\bar{x})\right], \tag{5.22}$$

where $W(x) := 1 + \int_0^x K^*(y, 0)\mathrm{d}y$ represents a non-normalized workload distribution. Using (5.18) and some straightforward manipulations, we may rewrite (5.15) into

$$\frac{\mathbb{P}(V^{\pi_{\bar{x}}} = 0)}{Z(\bar{x})} z(\bar{x}) = 0.$$

Note that both $\mathbb{P}(V^{\pi_{\bar{x}}} = 0) > 0$ and $Z(\bar{x}) > 0$ and finite. Finding the extremes of $TH^{\pi_{\bar{x}}}$ thus reduces to solving $z(\bar{x}) = 0$.

5.6.1 Two-level service rate

Suppose that the service rate is specified as

$$r(x) = \begin{cases} r_1, & \text{for } 0 < x \leq a, \\ r_2, & \text{for } x > a, \end{cases}$$

where $0 < r_2 < r_1$. Define $\rho_i := \lambda\beta/r_i$, $i = 1, 2$. Because the service-rate function is decreasing, we obtain from Example 5.5.1 that Assumption 5.2.1 is satisfied and a threshold policy is thus optimal. To determine the optimal threshold $\bar{x}$, we derive from Corollary 5.4.2 that we only need to consider $\bar{x} \leq a$.

Fix some $\bar{x} \in [0, a]$. Using results of Chapter 3 or [83], the stationary workload distribution may be easily reduced to a more tractable expression. Let

$$H(x) := \beta^{-1} \int_0^x (1 - B(y))\mathrm{d}y \tag{5.23}$$

be the stationary residual service requirement distribution with density $h(\cdot)$. For $x \leq a$, $K(x,y) = \rho_1 h(x-y)$ and it is well-known that $K^*(x,y) = \sum_{n=1}^{\infty} \rho_1^n h_n(x-y)$, where $h_n(\cdot)$ is the density of the n-fold convolution $H_n(\cdot)$ (see, e.g., [83] or Subsection 3.6.1).

Now we determine the three elements on the right-hand side of (5.22) separately, after which we combine them to determine $z(\bar{x})$. First consider $\lambda Z(\bar{x})$. Using the definitions of $Z(\cdot)$ and $H(\cdot)$, respectively (5.2) and (5.23), yields

$$\begin{aligned} \lambda Z(\bar{x}) &= \lambda \int_{\bar{x}}^{a} \frac{1}{r_1}(1 - B(x-\bar{x}))\mathrm{d}x + \lambda \int_{a}^{\infty} \frac{1}{r_2}(1 - B(x-\bar{x}))\mathrm{d}x \\ &= \rho_2 + (\rho_1 - \rho_2)H(a-\bar{x}). \end{aligned} \tag{5.24}$$

Second, consider the non-normalized workload distribution $W(\cdot)$. Interchanging integral and sum in addition to the results above, we immediately obtain for each $x \in [0, \bar{x}]$,

$$W(x) = 1 + \int_{0^+}^{x} \sum_{n=1}^{\infty} \rho_1^n h_n(y)\mathrm{d}y = \sum_{n=0}^{\infty} \rho_1^n H_n(x). \tag{5.25}$$

Remark 5.6.1 Note that $W(\cdot)/W(a)$ is the steady-state workload distribution in a finite dam with speed r_1 and buffer size a. In case $\rho_1 < 1$, it is an easy exercise to see that the Laplace-Stieltjes transform of $W(\cdot)$ provides the well-known Pollaczek-Khinchine formula. If $\rho_1 \geq 1$, $\int_0^a W(x)\mathrm{d}x$ is still finite and a steady-state workload distribution exists (see e.g. [83]). However, Cohen [52, 53] describes a more elegant way to determine $W(\cdot)$ in that case. $\diamond$

Finally, the first term of (5.22), that is the inverse of the normalizing constant $\mathbb{P}(V^{\pi_{\bar{x}}} = 0)$, is the most complicated one. Using the expression for the steady-state workload density in addition to the results above, we derive for $\bar{x} < x \leq a$,

$$\begin{aligned} V^{\pi_{\bar{x}}}(x) &= V^{\pi_{\bar{x}}}(\bar{x}) + \mathbb{P}(V^{\pi_{\bar{x}}} = 0)\Bigg[\int_{\bar{x}}^{x} \rho_1 h(y)\mathrm{d}y \\ &\qquad + \int_{\bar{x}}^{x} \int_{0^+}^{\bar{x}} \rho_1 h(y-u) \sum_{n=1}^{\infty} \rho_1^n h_n(u)\mathrm{d}u\mathrm{d}y\Bigg] \\ &= \mathbb{P}(V^{\pi_{\bar{x}}} = 0)\Bigg[W(\bar{x}) + \rho_1(H(x) - H(\bar{x})) \\ &\qquad + \rho_1 \int_{0^+}^{\bar{x}} (H(x-u) - H(\bar{x}-u))\mathrm{d}W(u)\Bigg] \\ &= \mathbb{P}(V^{\pi_{\bar{x}}} = 0)\left[W(\bar{x}) + \rho_1 \int_{0^-}^{\bar{x}} (H(x-u) - H(\bar{x}-u))\mathrm{d}W(u)\right], \end{aligned}$$

where we used $W(0) = 1$ in the final step. Note that, for $x > a$, $K(x,y) =$

$\rho_2 h(x-y)$. Using similar arguments, we obtain for $x > a$,

$$\begin{aligned} V^{\pi_{\bar{x}}}(x) &= V^{\pi_{\bar{x}}}(a) + \mathbb{P}(V^{\pi_{\bar{x}}} = 0)\Bigg[\int_a^x \rho_2 h(y)\mathrm{d}y \\ &\qquad + \int_a^x \int_{0+}^{\bar{x}} \rho_2 h(y-u) \sum_{n=1}^{\infty} \rho_1^n h_n(u)\mathrm{d}u\mathrm{d}y\Bigg] \\ &= \mathbb{P}(V^{\pi_{\bar{x}}} = 0)\Bigg[W(\bar{x}) + \rho_1 \int_{0^-}^{\bar{x}} (H(a-u) - H(\bar{x}-u))\mathrm{d}W(u) \\ &\qquad + \rho_2 \int_{0^-}^{\bar{x}} (H(x-u) - H(a-u))\mathrm{d}W(u)\Bigg]. \end{aligned}$$

By (5.25), we have $\rho_1 \int_0^{\bar{x}} H(\bar{x}-u)\mathrm{d}W(u) = W(\bar{x}) - 1$. Letting $x \to \infty$ and some rewriting then yields

$$\mathbb{P}(V^{\pi_{\bar{x}}} = 0)^{-1} = \rho_2 W(\bar{x}) + (\rho_1 - \rho_2)\int_{0^-}^{\bar{x}} H(a-u)\mathrm{d}W(u) + 1. \qquad (5.26)$$

It is now easy to get $z(\bar{x})$. Substituting (5.24)–(5.26) into (5.22) gives

$$z(\bar{x}) = 1 + (\rho_2 - \rho_1)\left[W(\bar{x})H(a-\bar{x}) - \int_{0^-}^{\bar{x}} H(a-u)\mathrm{d}W(u)\right]. \qquad (5.27)$$

Summarizing, Corollary 5.4.2 implies that the greedy policy is optimal if and only if

$$\rho_2 - (\rho_2 - \rho_1)W(a) > 0.$$

Otherwise, $TH^* = \rho/(\rho_2 + (\rho_1 - \rho_2)H(a - x^*))$, with $\rho := \lambda\beta$ and x^* is a solution to $z(x^*) = 0$.

In general, the convolution in (5.27) can only be determined numerically. However, if the service requirement follows a phase-type distribution, explicit expressions can be obtained. For instance, if $B(x) = 1 - e^{-\mu x}$ (see also Subsection 5.6.2), then after quite lengthy but standard calculations, it follows that, for $\rho_1 \neq 1$,

$$z(\bar{x}) = 1 + \frac{\rho_2 - \rho_1}{\rho_1 - 1} e^{-\mu a}\left(e^{\mu\bar{x}} - e^{\mu\rho_1\bar{x}}\right). \qquad (5.28)$$

In case $\rho_1 = 1$, we obtain

$$z(\bar{x}) = 1 - (\rho_2 - 1)\mu\bar{x}e^{-\mu(a-\bar{x})}.$$

5.6.2 Exponential service requirements

Suppose that the service requirements are exponentially distributed with mean $1/\mu$, i.e., $1 - B(x) = e^{-\mu x}$. Then, for fixed $\bar{x}$, the steady-state workload density is given in Corollary 3.6.1:

$$v^{\pi_{\bar{x}}}(x) = \frac{\lambda\mathbb{P}(V^{\pi_{\bar{x}}} = 0)}{r(x)} \exp\{-\mu x + \lambda R(x \wedge \bar{x})\}, \qquad (5.29)$$

where $\mathbb{P}(V^{\pi_{\bar{x}}} = 0)$ follows from normalization. In this subsection, we also assume that the service rate is a step function. More specifically, let $r(x) = r_i$ for $x \in [a_{i-1}, a_i)$, $i = 1, \ldots, N$ (where $a_0 = 0$), and let $r(x) = r_{N+1} < r_N$ for $x \geq a_N$. Denote $\rho_i = \lambda/(\mu r_i)$, $i = 1, \ldots, N+1$ and assume for simplicity that $\rho_i \neq 1$.

Example 5.5.3 shows that Assumption 5.2.1 is satisfied and a threshold policy is thus optimal. By Corollary 5.4.2, either the greedy policy is optimal, or the optimal threshold x^* is less than a_N. Let $\bar{x} \in [a_n, a_{n+1})$ for some $n \leq N-1$. Next, we consider each of the three elements of $z(\bar{x})$ separately, after which we combine them into an expression for x^* satisfying (5.15). However, for later use, we first define the following three constants. (In the sequel we follow the convention that empty sums are equal to 0.)

$$\begin{aligned}
\gamma_n &= \exp\left\{\sum_{k=1}^{n}\left(\frac{\lambda}{r_k} - \frac{\lambda}{r_{k+1}}\right) a_k\right\}, \\
C_n &= \frac{1}{1-\rho_1} + \sum_{k=1}^{n}\left(\frac{\rho_k}{\rho_k - 1} - \frac{\rho_{k+1}}{\rho_{k+1}-1}\right)\gamma_{k-1} e^{-\mu(1-\rho_k)a_k}, \\
D_n &= \sum_{k=n+1}^{N} (\rho_{k+1} - \rho_k) e^{-\mu a_k}.
\end{aligned}$$

First, consider $\lambda Z(\bar{x})$. Using (5.2) and rewriting the integral, we obtain

$$\begin{aligned}
\lambda Z(\bar{x}) &= \int_{\bar{x}}^{a_{n+1}} \frac{\lambda}{r_{n+1}} e^{-\mu(x-\bar{x})} \mathrm{d}x + \sum_{k=n+1}^{N-1} \int_{a_k}^{a_{k+1}} \frac{\lambda}{r_{k+1}} e^{-\mu(x-\bar{x})} \mathrm{d}x \\
&\quad + \int_{a_N}^{\infty} \frac{\lambda}{r_{N+1}} e^{-\mu(x-\bar{x})} \mathrm{d}x \\
&= \rho_{n+1}(1 - e^{-\mu(a_{n+1}-\bar{x})}) + \sum_{k=n+1}^{N-1} \rho_{k+1}(e^{-\mu(a_k-\bar{x})} - e^{-\mu(a_{k+1}-\bar{x})}) \\
&\quad + \rho_{N+1} e^{-\mu(a_N - \bar{x})} \\
&= \rho_{n+1} + D_n e^{\mu \bar{x}}. \qquad (5.30)
\end{aligned}$$

Second, consider $W(\cdot)$, i.e., the workload distribution ‘without normalization’. It is easily checked that the time to empty the system starting from a_i, $i = 1, \ldots, n$, in the absence of any arrivals (i.e., $R(a_i)$) equals $a_i/r_i + \sum_{k=1}^{i-1}(1/r_k - 1/r_{k+1})a_k$. Hence, for $x \in [a_i, a_{i+1})$, we may deduce that

$$\exp\{\lambda R(x)\} = \exp\left\{\frac{\lambda(x - a_i)}{r_{i+1}} + \lambda R(a_i)\right\} = \gamma_i \exp\left\{\frac{\lambda x}{r_{i+1}}\right\}. \qquad (5.31)$$

Now, for $i = 1, \ldots, n$, using (5.29) and (5.31), we obtain after some standard algebra, that

$$\begin{aligned}
&V^{\pi_{\bar{x}}}(a_i)/\mathbb{P}(V^{\pi_{\bar{x}}}=0)\\
&\quad= 1+\sum_{k=0}^{i-1}\int_{a_k}^{a_{k+1}}\frac{\lambda}{r_{k+1}}e^{-\mu x+\lambda R(x)}\mathrm{d}x\\
&\quad= 1+\sum_{k=0}^{i-1}\frac{\rho_{k+1}}{\rho_{k+1}-1}\gamma_k\left(e^{-\mu(1-\rho_{k+1})a_{k+1}}-e^{-\mu(1-\rho_{k+1})a_k}\right)\\
&\quad= \frac{1}{1-\rho_1}+\sum_{k=0}^{i-1}\frac{\rho_{k+1}}{\rho_{k+1}-1}\gamma_k e^{-\mu(1-\rho_{k+1})a_{k+1}}-\sum_{k=1}^{i-1}\frac{\rho_{k+1}}{\rho_{k+1}-1}\gamma_k e^{-\mu(1-\rho_{k+1})a_k}\\
&\quad= \frac{1}{1-\rho_1}+\sum_{k=1}^{i}\frac{\rho_k}{\rho_k-1}\gamma_{k-1}e^{-\mu(1-\rho_k)a_k}-\sum_{k=1}^{i-1}\frac{\rho_{k+1}}{\rho_{k+1}-1}\gamma_{k-1}e^{-\mu(1-\rho_k)a_k}\\
&\quad= C_{i-1}+\frac{\rho_i}{\rho_i-1}\gamma_{i-1}e^{-\mu(1-\rho_i)a_i},
\end{aligned}$$

where we used $\gamma_k e^{\mu\rho_{k+1}a_k}=\gamma_{k-1}e^{\mu\rho_k a_k}$ in the fourth equality. Thus, combining (5.29) and (5.31) with the above, we obtain, after similar manipulations,

$$\begin{aligned}
V^{\pi_{\bar{x}}}(\bar{x}) &= V^{\pi_{\bar{x}}}(a_n)+\mathbb{P}(V^{\pi_{\bar{x}}}=0)\int_{a_n}^{\bar{x}}\frac{\lambda}{r_{n+1}}e^{-\mu x+\lambda R(x)}\mathrm{d}x \qquad (5.32)\\
&= V^{\pi_{\bar{x}}}(a_n)+\mathbb{P}(V^{\pi_{\bar{x}}}=0)\frac{\rho_{n+1}}{\rho_{n+1}-1}\gamma_n\left(e^{-\mu(1-\rho_{n+1})\bar{x}}-e^{-\mu(1-\rho_{n+1})a_n}\right)\\
&= \mathbb{P}(V^{\pi_{\bar{x}}}=0)\left[C_n+\frac{\rho_{n+1}}{\rho_{n+1}-1}\gamma_n e^{-\mu(1-\rho_{n+1})\bar{x}}\right], \qquad (5.33)
\end{aligned}$$

which completes the calculation of $W(\cdot)$ (since $W(\bar{x})\mathbb{P}(V^{\pi_{\bar{x}}}=0)=V^{\pi_{\bar{x}}}(\bar{x})$).

For the first term on the right-hand side of (5.22), i.e., $\mathbb{P}(V^{\pi_{\bar{x}}}=0)^{-1}$, we use similar arguments as for the previous one. We first consider $V^{\pi_{\bar{x}}}(x)$ with $x>a_{n+1}$ and let $i=\arg\max\{a_i : a_i\le x\}$ be the largest a_i smaller than x. Using (5.29) and applying (5.31) to determine $\lambda R(\bar{x})$, we obtain after similar algebra as above, that

$$\begin{aligned}
V^{\pi_{\bar{x}}}(x) &= V^{\pi_{\bar{x}}}(\bar{x})+\gamma_n e^{\mu\rho_{n+1}\bar{x}}\mathbb{P}(V^{\pi_{\bar{x}}}=0)\bigg[\int_{\bar{x}}^{a_{n+1}}\frac{\lambda}{r_{n+1}}e^{-\mu y}\mathrm{d}y\\
&\qquad+\sum_{k=n+1}^{i-1}\int_{a_k}^{a_{k+1}}\frac{\lambda}{r_{k+1}}e^{-\mu y}\mathrm{d}y+\int_{a_i}^{x}\frac{\lambda}{r_{i+1}}e^{-\mu y}\mathrm{d}y\bigg]\\
&= \mathbb{P}(V^{\pi_{\bar{x}}}=0)\bigg[C_n+\frac{\rho_{n+1}^2}{\rho_{n+1}-1}\gamma_n e^{-\mu(1-\rho_{n+1})\bar{x}}\\
&\qquad+\gamma_n e^{\mu\rho_{n+1}\bar{x}}\sum_{k=n+1}^{i}(\rho_{k+1}-\rho_k)e^{-\mu a_k}-\gamma_n\rho_{i+1}e^{-\mu x+\mu\rho_{n+1}\bar{x}}\bigg].
\end{aligned}$$

Thus, letting $x \to \infty$, we obtain the normalizing constant:

$$\mathbb{P}(V^{\pi_{\bar{x}}} = 0)^{-1} = C_n + \frac{\rho_{n+1}^2}{\rho_{n+1} - 1}\gamma_n e^{-\mu(1-\rho_{n+1})\bar{x}} + \gamma_n D_n e^{\mu\rho_{n+1}\bar{x}}. \qquad (5.34)$$

The function $z(\bar{x})$ can now easily be rewritten into a more appealing expression. In particular, substituting (5.30), (5.34), and $W(\bar{x})$ resulting from (5.33) into (5.22) and some reordering of terms, yields

$$\begin{aligned} z(\bar{x}) &= (1-\rho_{n+1})C_n + D_n\gamma_n e^{\mu\rho_{n+1}\bar{x}} \\ &\quad -D_n e^{\mu\bar{x}}\left(C_n + \frac{\rho_{n+1}}{\rho_{n+1}-1}\gamma_n e^{-\mu(1-\rho_{n+1})\bar{x}}\right) \\ &= (1-\rho_{n+1})C_n - C_n D_n e^{\mu\bar{x}} - \frac{D_n\gamma_n}{\rho_{n+1}-1}e^{\mu\rho_{n+1}\bar{x}}. \end{aligned}$$

Solving $z(\bar{x}) = 0$ is thus remarkably simple in this case, since the variable $\bar{x}$ only appears in two of the exponents. Summarizing, we conclude that the optimal policy is of the threshold type where the optimal threshold value is given by the solution of $z(\bar{x}) = 0$. Moreover $TH^* = \beta/Z(\bar{x})$, where $\lambda Z(\bar{x})$ is given in (5.30).

Remark 5.6.2 It is easily checked that, in case $N = 1$, the formula for $z(\bar{x})$ indeed reduces to (5.28). ◇

5.7 Concluding remarks and further research

In the present chapter, we considered the problem of optimal admission control in a system with a workload-dependent service rate. We assumed that the service requirement only becomes known right after the decision of accepting or rejecting jobs. Our objective was to find a policy that maximizes the long-run throughput. Under some assumptions (in particular Assumption 5.2.1), we showed that a threshold policy for accepting jobs is optimal and derived a criterion for the optimal threshold value.

We note that our main assumption, i.e. Assumption 5.2.1, involves sufficient conditions for optimality of threshold policies. An interesting subject for further research is to examine the structure of the optimal policy when Assumption 5.2.1 is not satisfied.

Moreover, there are various interesting model variations. For instance, the analysis is significantly changed if information about the service requirement is available. In that case, the decision will not only depend on the workload level, but also on the size of the job, yielding a two-dimensional state space. A characterization of the optimal policy in that model might be a subject of further study. We note that a threshold policy will not be optimal in general. However, in some special cases, as for, e.g., deterministic service requirements or decreasing service rate functions, the optimal policy continues to be of the threshold type.

Other model variations are scenarios where jobs can be partly accepted (or rejected). The simplest version concerns a model where an infinite amount of work becomes available at Poisson instants and the policy prescribes the amount of work to accept. In some sense, this model is related to the case $\lambda \to \infty$ in the model of the present chapter, which may be interpreted as an infinite supply of jobs and the policy prescribes the time to accept a new job. More interesting are scenarios where the supply of work is bounded by the service requirements of arriving jobs and the decision is the amount of work to accept. In that case, the state space is two-dimensional and the action space is continuous. The structure of the optimal policy in the latter model is also left for future investigation.

Appendix

5.A Proof of Lemma 5.4.1

Lemma 5.4.1 *For the derivative of $TH^{\pi_{\bar{x}}}$, we have*

$$\frac{\mathrm{d}}{\mathrm{d}\bar{x}} TH^{\pi_{\bar{x}}} = \lambda\beta\mathbb{P}(V^{\pi_{\bar{x}}} = 0)K^*(\bar{x},0)\left[1 - TH^{\pi_{\bar{x}}} Z(\bar{x})/\beta\right].$$

Proof We first consider $\mathbb{P}(V^{\pi_{\bar{x}}} = 0)$. Observe that the double integration in (5.16) may be equivalently expressed as: $\int_{x=\bar{x}}^{\infty}\int_{y=0}^{\bar{x}} = \int_{y=0}^{\bar{x}}\int_{x=y}^{\infty} - \int_{x=0}^{\bar{x}}\int_{y=0}^{x}$. Using the definition of K^*, interchanging integral and sum and finally applying (5.17), we may write

$$\begin{aligned}\int_0^{\bar{x}} K(\bar{x},y)K^*(y,0)\mathrm{d}y &= \sum_{n=1}^{\infty}\int_0^{\bar{x}} K(\bar{x},y)K_n(y,0)\mathrm{d}y \\ &= \sum_{n=0}^{\infty} K_{n+1}(\bar{x},0) - K_1(\bar{x},0) \\ &= K^*(\bar{x},0) - K(\bar{x},0). \end{aligned} \tag{5.35}$$

Taking the derivative of $\mathbb{P}(V^{\pi_{\bar{x}}} = 0)$ with respect to $\bar{x}$, we obtain from (5.16) and the reordering of integration that

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}\bar{x}}\mathbb{P}(V^{\pi_{\bar{x}}} = 0) \\ &\quad= -\mathbb{P}(V^{\pi_{\bar{x}}} = 0)^2\Bigg(K^*(\bar{x},0) - K(\bar{x},0) \\ &\qquad\qquad + \int_{\bar{x}}^{\infty} K(x,\bar{x})K^*(\bar{x},0)\mathrm{d}x - \int_0^{\bar{x}} K(\bar{x},y)K^*(y,0)\mathrm{d}y\Bigg) \\ &\quad= -\mathbb{P}(V^{\pi_{\bar{x}}} = 0)^2 K^*(\bar{x},0)\int_{\bar{x}}^{\infty} K(x,\bar{x})\mathrm{d}x,\end{aligned}$$

where we used (5.35) in the second step. Now, invoking (5.18) and taking the derivative of $TH^{\pi_{\bar{x}}}$ with respect to $\bar{x}$ yields

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\bar{x}} TH^{\pi_{\bar{x}}} &= \lambda\beta\mathbb{P}(V^{\pi_{\bar{x}}}=0)K^*(\bar{x},0) + \lambda\beta\frac{\mathrm{d}}{\mathrm{d}\bar{x}}\mathbb{P}(V^{\pi_{\bar{x}}}=0)\left(1+\int_0^{\bar{x}} K^*(x,0)\mathrm{d}x\right) \\
&= \lambda\beta\mathbb{P}(V^{\pi_{\bar{x}}}=0)K^*(\bar{x},0) \\
&\quad \times\left[1-\mathbb{P}(V^{\pi_{\bar{x}}}=0)\int_{\bar{x}}^{\infty} K(x,\bar{x})\mathrm{d}x\left(1+\int_0^{\bar{x}} K^*(x,0)\mathrm{d}x\right)\right] \\
&= \lambda\beta\mathbb{P}(V^{\pi_{\bar{x}}}=0)K^*(\bar{x},0)\left[1-TH^{\pi_{\bar{x}}}Z(\bar{x})/\beta\right],
\end{aligned}$$

where the final step follows from (5.18) and the fact that $\lambda Z(\bar{x}) = \int_{\bar{x}}^{\infty} K(x,\bar{x})\mathrm{d}x$. This completes the proof. □

Chapter 6

Queues with adaptable service speed

6.1 Introduction

In the previous chapters we considered various queueing systems with workload-dependent service speeds. In those chapters, as in most of the literature on queues with state-dependent rates, it is assumed that the speed of the server is continuously adapted over time based on the buffer content. In many practical situations, though, service speed adaptations are only made at particular points in time, like arrival epochs. For example, feedback information about the buffer state may only be available at such epochs. Furthermore, continuously changing the service speed may come with certain costs.

In this chapter, we consider a single-server queue with adaptable service speed based on the amount of work right after customer arrivals. In between arrivals, the service speed is held fixed and may not be changed until the next customer arrival. The main aim of this chapter is to find the (LST of the) distribution of the steady-state workload embedded at epochs immediately after arrivals, and the steady-state workload distribution at arbitrary epochs.

Related literature

Models with continuously adaptable service speed originate from the study of dams and storage processes. For an overview of the literature on dams, storage systems, and queueing models with workload-dependent service speeds, we refer to Section 1.4. Furthermore, in [53, 80] and [56], p. 555-556, the authors consider a queueing system with a two-stage service rule: If the workload is less than K, then the service speed equals r_1, whereas the service speed equals r_2 when the workload exceeds K. Using an elegant technique for the convolution of two LSTs, they determine the steady-state workload distribution. In this chapter, we apply a similar method to obtain the LST of the workload at embedded epochs for the M/G/1 queue with service speeds only being changed at customer arrivals.

A related branch of literature considers queueing systems where the service speed depends not only on the buffer content, but also on the stage of the system. In particular, an (m, M) control rule prescribes to switch from stage 1 to stage 2 at an upcrossing of the workload of level M (and the stage is 1) and to switch back from stage 2 to stage 1 at a downcrossing of m (and the stage is 2), see also e.g. [17, 114, 164]. The control of the service speed may be realized by letting r_i be the service speed in stage i, $i = 1, 2$. In such control systems, usually costs are imposed including, e.g., holding costs and switchover costs. In [164], the long-run average costs per unit time for the (m, M)-policy are determined. Of special interest is the case when $m = 0$ which is commonly referred to as a D-policy (that is $(m, M) = (0, D)$). In [162], the author shows that the D-policy is average-cost optimal under the assumption that the workload can only be controlled at arrival epochs. In [75], the average-cost optimality of D-policies is rigorously proved in a more general setting.

Model description

We consider an M/G/1 queueing system where feedback information about the level of congestion is available right after arrival instants. The customers arrive to the system according to a Poisson process with rate λ. Let A_n, $n = 1, 2, \ldots$, denote the time between the arrival instants of customers n and $n + 1$. Also, denote by B_n, $n = 1, 2, \ldots$, the service requirement of customer n. We assume that $B_1, B_2, \ldots$ are i.i.d. copies of the generic random variable B with distribution $B(\cdot)$, mean β, and LST $\beta(\cdot)$. We also assume that the sequences of interarrival intervals and service requirements are independent.

When the amount of work right after an arrival instant equals x, the server works at constant speed $r(x)$ until the next customer arrival. Note that the service speed is thus only changed at discrete points in time. In this chapter, we specifically consider the case of a two-step service speed function: If the amount of work right after an arrival is smaller than (or equal to) a finite number K, then the server starts to work at speed r_1, whereas the service speed equals r_2 if the workload is larger than K. Later, we also consider the generalization to an N-step service-speed function (see Subsection 6.5.3).

Define $\rho_i := \lambda\beta/r_i$, $i = 1, 2$. Throughout, we assume that the system is stable, i.e., $\rho_2 < 1$. Let W_n and S_n be the workload just before, respectively right after, the arrival instant of customer n. We denote by W and S the steady-state random variables corresponding to W_n and S_n. We have the following recursion relation:

$$S_{n+1} = (S_n - r(S_n)A_n)^+ + B_{n+1}, \tag{6.1}$$

where $x^+ = \max(x, 0)$. Because of the trivial relation $S_n = W_n + B_n$, one also has $W_{n+1} = (S_n - r(S_n)A_n)^+$.

In queueing systems where the server always works at unit speed when there is any work in the system, W corresponds to a waiting time and S represents a customer's sojourn time. This equivalence no longer holds when the service speed varies with the amount of work present. For convenience, however, we often refer to W and S as the waiting and sojourn time, respectively.

Goal and organization

The main aim of this chapter is to find the distribution (and LST) of S, and then also of W. It should be observed that, due to PASTA, the distribution of W also equals the steady-state workload distribution.

The chapter is organized as follows. In Section 6.2 we derive two distinct equations for the LST of S and sketch a four-step procedure to determine its distribution. The first step of this procedure does not depend on the distribution of the service requirement and is analyzed in detail in Section 6.2. We give steps two to four in Section 6.3 in case the service requirements follow an exponential distribution. It turns out that the density of S is then a weighted combination of two exponentials for $x \leq K$, and is purely exponential for $x > K$. The M/M/1 case gives much insight into the structure of the solution for more general cases, like the M/G/1 case, which is addressed in Section 6.4. For expository reasons, we have chosen to treat these cases separately instead of all in one. Special cases and the extension to the N-step service rule are discussed in Section 6.5.

6.2 Sojourn times: Equations and general procedure

In this section, we first derive equations to determine the LST of S in case of the two-step service speed function. Secondly, we outline a four-step procedure to find the LST and distribution of S from the constructed equations, and describe the first step in detail.

For convenience, we recall the definition of the two-step service rule:

$$r(x) = \begin{cases} r_1, & \text{for } 0 < x \leq K, \\ r_2, & \text{for } x > K. \end{cases}$$

Denote the LST of S by

$$\phi(\omega) := \int_0^\infty e^{-\omega x} \mathrm{d}\mathbb{P}(S < x). \tag{6.2}$$

Also, define, for $i = 1, 2$ and $\rho_i \neq 1$,

$$F_i(\omega) := (1 - \rho_i) \frac{r_i \omega \beta(\omega)}{\omega r_i - \lambda + \lambda \beta(\omega)}. \tag{6.3}$$

Observe that $F_i(\omega)$ corresponds to the LST of the sojourn time in an M/G/1 queue with service speed r_i, $i = 1, 2$.

The equations for $\phi(\omega)$ are summarized in the following lemma:

Lemma 6.2.1 *$\phi(\omega)$ satisfies the following two equations, for* $\operatorname{Re} \omega \geq 0$*,*

$$\begin{aligned} \phi(\omega) &= F_2(\omega) \frac{W(0)}{1 - \rho_2} && (6.4) \\ &+ F_2(\omega) \frac{\lambda(\frac{r_1}{r_2} - 1)}{(\omega r_1 - \lambda)(1 - \rho_2)} \left[\int_0^K e^{-\omega x} \mathrm{d}\mathbb{P}(S < x) - \int_0^K e^{-\frac{\lambda}{r_1} x} \mathrm{d}\mathbb{P}(S < x) \right], \end{aligned}$$

with $W(0) := \mathbb{P}(W=0)$*. Also,*

$$\begin{aligned}\phi(\omega) &= F_1(\omega)\frac{W(0)}{1-\rho_1} \qquad (6.5)\\ &+ F_1(\omega)\frac{\lambda(1-\frac{r_2}{r_1})}{(\omega r_2-\lambda)(1-\rho_1)}\left[\int_K^\infty e^{-\frac{\lambda}{r_2}x}\mathrm{d}\mathbb{P}(S<x) - \int_K^\infty e^{-\omega x}\mathrm{d}\mathbb{P}(S<x)\right].\end{aligned}$$

Proof It follows after some straightforward calculations that, for $\omega \neq \lambda/r_1$ and $\omega \neq \lambda/r_2$,

$$\begin{aligned}\mathbb{E}\left[e^{-\omega(S_n-r(S_n)A_n)^+}|S_n=x\right] &= e^{-\omega x}\lambda\int_0^{x/r(x)} e^{(\omega r(x)-\lambda)y}\mathrm{d}y + e^{-\lambda x/r(x)}\\ &= \frac{\omega r(x)}{\omega r(x)-\lambda}e^{-\frac{\lambda}{r(x)}x} - \frac{\lambda}{\omega r(x)-\lambda}e^{-\omega x}. \qquad (6.6)\end{aligned}$$

Using the recursion (6.1), conditioning on S_n, and applying the above, yields

$$\begin{aligned}&\mathbb{E}\left[e^{-\omega S_{n+1}}\right]\\ &= \int_0^\infty \mathbb{E}\left[e^{-\omega S_{n+1}}|S_n=x\right]\mathrm{d}\mathbb{P}(S_n<x)\\ &= \beta(\omega)\Bigg[\frac{\omega r_1}{\omega r_1-\lambda}\int_0^K e^{-\frac{\lambda}{r_1}x}\mathrm{d}\mathbb{P}(S_n<x) - \frac{\lambda}{\omega r_1-\lambda}\int_0^K e^{-\omega x}\mathrm{d}\mathbb{P}(S_n<x)\\ &\quad + \frac{\omega r_2}{\omega r_2-\lambda}\int_K^\infty e^{-\frac{\lambda}{r_2}x}\mathrm{d}\mathbb{P}(S_n<x) - \frac{\lambda}{\omega r_2-\lambda}\int_K^\infty e^{-\omega x}\mathrm{d}\mathbb{P}(S_n<x)\Bigg]. \qquad (6.7)\end{aligned}$$

To analyze the steady-state behavior of S_n, we let $n \to \infty$. Furthermore, combining (6.2) and (6.7), in addition to some basic manipulations, we may obtain two alternative equations for $\phi(\omega)$: First,

$$\begin{aligned}\phi(\omega) &= \frac{F_2(\omega)}{(1-\rho_2)(\omega r_1-\lambda)}\Bigg[\frac{r_1}{r_2}(\omega r_2-\lambda)\int_0^K e^{-\frac{\lambda}{r_1}x}\mathrm{d}\mathbb{P}(S<x) \qquad (6.8)\\ &\quad +\lambda(\frac{r_1}{r_2}-1)\int_0^K e^{-\omega x}\mathrm{d}\mathbb{P}(S<x) + (\omega r_1-\lambda)\int_K^\infty e^{-\frac{\lambda}{r_2}x}\mathrm{d}\mathbb{P}(S<x)\Bigg],\end{aligned}$$

and second,

$$\begin{aligned}\phi(\omega) &= \frac{F_1(\omega)}{(1-\rho_1)(\omega r_2-\lambda)}\Bigg[(\omega r_2-\lambda)\int_0^K e^{-\frac{\lambda}{r_1}x}\mathrm{d}\mathbb{P}(S<x) \qquad (6.9)\\ &\quad -\lambda(1-\frac{r_2}{r_1})\int_K^\infty e^{-\omega x}\mathrm{d}\mathbb{P}(S<x) + \frac{r_2}{r_1}(\omega r_1-\lambda)\int_K^\infty e^{-\frac{\lambda}{r_2}x}\mathrm{d}\mathbb{P}(S<x)\Bigg].\end{aligned}$$

Now, Equations (6.4) and (6.5) follow from (6.8) and (6.9), respectively, and from the observation that

$$W(0) = \int_0^K e^{-\frac{\lambda}{r_1}x}\mathrm{d}\mathbb{P}(S<x) + \int_K^\infty e^{-\frac{\lambda}{r_2}x}\mathrm{d}\mathbb{P}(S<x). \qquad (6.10)$$

This completes the proof. □

Determining $\phi(\omega)$ from Equations (6.4) and (6.5) involves the more complicated part. We introduce a four-step procedure to determine the distribution of S. Below, we sketch each of the four steps. Because Step 1 is the only step that does not depend on the service requirement distribution, we analyze it in detail at the end of this section. Steps 2–4 are carried out in Section 6.3 in case the distribution of the service requirements is exponential. The general M/G/1 case is considered in Section 6.4. The procedure builds upon techniques applied in [53, 80] and [56], p. 556. It starts from the observation that a serious complication in determining $\phi(\omega)$ from (6.4) and (6.5) is that both equations involve the incomplete LST of S.

The basic algorithm to obtain $\mathbb{P}(S < x)$ is as follows:

Step 1 Rewrite Equation (6.5) such that the second term of (6.5) can be interpreted as the sum of (i) the LST of the convolution of $F_1(\cdot)$ with an exponential term, and (ii) a transform that only has points of increase on (K, ∞).

Step 2 Apply Laplace inversion to the reformulated Equation (6.5) resulting from Step 1, to determine $\mathbb{P}(S < x)$ for $x \in (0, K]$.

Step 3 By Step 2, we may now calculate $\int_0^K e^{-\omega x} \mathrm{d}\mathbb{P}(S < x)$. Substitution in (6.4) then directly provides $\phi(\omega)$. Applying Laplace inversion again, we determine $\mathbb{P}(S < x)$ for $x > K$.

Step 4 The remaining constants may be found by normalization.

The remainder of this section is devoted to the description of Step 1.

Step 1: *Rewriting (6.5)*
In this part, when considering the sojourn time of customer $n+1$, we distinguish between two cases: (i) $S_n \leq K$, and (ii) $S_n > K$. If $S_{n+1} \leq K$, this imposes for case (ii) that a downcrossing of level K occurs between the arrival instants of customers n and $n+1$. However, the residual interarrival time at a downcrossing of K is still exponential. Consequently, given a downcrossing of level K between the arrival epochs of customers n and $n+1$, the precise distribution of S_n on (K, ∞) does not affect the distribution of $S_{n+1} \leq K$, because W_{n+1} is simply distributed as $(K - r_2 A_n)^+$. The aim of this first step is to show that the second part of Equation (6.5) corresponds to case (ii) and to apply the intuitive arguments above in reformulating (6.5).

Denote by $I(\cdot)$ the indicator function. Using (6.6), we get

$$\mathbb{E}\left[e^{-\omega(S_n - r(S_n)A_n)^+} I(S_n > K)\right] - \int_K^\infty e^{-\frac{\lambda}{r_2}x} \mathrm{d}\mathbb{P}(S_n < x)$$
$$= \frac{\lambda}{\omega r_2 - \lambda}\left[\int_K^\infty e^{-\frac{\lambda}{r_2}x} \mathrm{d}\mathbb{P}(S_n < x) - \int_K^\infty e^{-\omega x} \mathrm{d}\mathbb{P}(S_n < x)\right].$$

Observe that the right-hand side (rhs) corresponds to the final part of the second term in (6.5). However, by conditioning on S_n, we may also rewrite this expression as

$$\begin{aligned}
&\mathbb{E}\left[e^{-\omega(S_n-r(S_n)A_n)^+}I(S_n>K)\right]-\int_K^\infty e^{-\frac{\lambda}{r_2}x}\mathrm{d}\mathbb{P}(S_n<x)\\
&= \int_K^\infty e^{-\omega(x-r_2A_n)^+}\mathrm{d}\mathbb{P}(S_n<x)-\int_K^\infty e^{-\frac{\lambda}{r_2}x}\mathrm{d}\mathbb{P}(S_n<x)\\
&= \int_K^\infty e^{-\omega(x-r_2A_n)}I(A_n\le(x-K)/r_2)\mathrm{d}\mathbb{P}(S_n<x)\\
&\qquad+\int_K^\infty\int_{(x-K)/r_2}^{x/r_2}\lambda e^{-\lambda y}e^{-\omega(x-r_2y)}\mathrm{d}y\mathrm{d}\mathbb{P}(S_n<x)\\
&= \mathbb{E}\left[e^{-\omega(S_n-r_2A_n)}I(S_n-r_2A_n>K)\right]\\
&\qquad+\frac{\lambda}{\omega r_2-\lambda}\left(1-e^{-\omega K+\frac{\lambda}{r_2}K}\right)\int_K^\infty e^{-\frac{\lambda}{r_2}x}\mathrm{d}\mathbb{P}(S_n<x).
\end{aligned}$$

Letting $n\to\infty$ and combining the above, Equation (6.5) reads,

$$\begin{aligned}
\phi(\omega) &= F_1(\omega)\frac{W(0)}{1-\rho_1}+F_1(\omega)\frac{(1-\frac{r_2}{r_1})}{1-\rho_1}\Big\{\mathbb{E}\left[e^{-\omega(S-r_2A)}I(S-r_2A>K)\right]\\
&\qquad+\frac{\lambda}{\omega r_2-\lambda}\left(1-e^{-\omega K+\frac{\lambda}{r_2}K}\right)\int_K^\infty e^{-\frac{\lambda}{r_2}x}\mathrm{d}\mathbb{P}(S<x)\Big\}. \qquad (6.11)
\end{aligned}$$

The second and third term on the rhs of (6.11) directly correspond to the intuitive observations made above. The second one provides the LST of W when $W>K$. The third one involves the LST of $K-r_2A$ (with A a generic interarrival time) multiplied by a constant (see Section 6.4 for an interpretation).

6.3 Exponential service requirements

In this section, we assume that $B(x)=1-e^{-\mu x}$, i.e., the service requirements are exponentially distributed with mean $1/\mu$. Applying the procedure described in Section 6.2, we explicitly determine the steady-state “sojourn time” distribution. We have chosen to treat the M/M/1 case first, because the structure of the density of S is here readily exposed, yielding insight into the solution for the M/G/1 case. Moreover, the solutions reduce to nice analytical expressions in the M/M/1 case.

Because the interpretation of Step 1 is valid independently of $B(\cdot)$, the starting point of the algorithm is Equation (6.11).

Step 2: *Sojourn time density on* $(0,K]$
Using the construction of Step 1, we apply Laplace inversion to determine the density $f_S(x)$ of S for $0<x\le K$. In the exponential case, we easily obtain for the first transform in (6.11),

$$F_1(\omega)=(1-\rho_1)\frac{r_1\mu}{r_1(\omega+\mu)-\lambda}.$$

Laplace inversion provides the familiar M/M/1 term for queues with constant service speed r_1,

$$s_1(x) = \mu(1-\rho_1)e^{(\frac{\lambda}{r_1}-\mu)x}, \qquad \text{for } \; x > 0,$$

where $s_1(\cdot)$ denotes the density of a random variable with LST $F_1(\cdot)$.

The inversion of the second transform in (6.11) is based on an observation made in [53, 56, 80]. First, consider

$$F_1(\omega)\mathbb{E}\left[e^{-\omega(S-r_2A)}I(S-r_2A>K)\right].$$

This term involves a product of two LST, corresponding to the sum of a random variable with mass on $[0,\infty)$, and one with mass on $[K,\infty)$. Hence, that sum has no mass on $[0,K]$.

Second, consider

$$F_1(\omega)\frac{\lambda}{\omega r_2-\lambda}\left(1-e^{-\omega K+\frac{\lambda}{r_2}K}\right). \tag{6.12}$$

It is readily checked that the latter part, $\frac{\lambda}{\omega r_2-\lambda}\left(1-e^{-\omega K+\frac{\lambda}{r_2}K}\right)$, is the Laplace Transform of the function

$$f(x) = \begin{cases} \frac{\lambda}{r_2}e^{\frac{\lambda}{r_2}x}, & \text{for } 0 < x \leq K, \\ 0, & \text{for } x > K. \end{cases} \tag{6.13}$$

Thus, (6.12) represents the convolution of $s_1(\cdot)$ with $f(\cdot)$. By applying (6.11) and combining the above, we obtain after lengthy calculations the following "sojourn time" density $f_S(x)$, for $0 < x \leq K$,

$$\begin{aligned} f_S(x) &= s_1(x)\frac{W(0)}{1-\rho_1} + \frac{1-\frac{r_2}{r_1}}{1-\rho_1}\int_K^\infty e^{-\frac{\lambda}{r_2}y}\mathrm{d}\mathbb{P}(S<y)\int_0^x s_1(y)\frac{\lambda}{r_2}e^{\frac{\lambda}{r_2}(x-y)}\mathrm{d}y \\ &= Q_1 e^{(\frac{\lambda}{r_1}-\mu)x} + Q_2 e^{\frac{\lambda}{r_2}x}, \end{aligned} \tag{6.14}$$

with

$$\begin{aligned} Q_1 &= \mu\int_0^K e^{-\frac{\lambda}{r_1}y}\mathrm{d}\mathbb{P}(S<y) \\ &\qquad + \frac{r_1 r_2\mu^2}{\lambda(r_1-r_2)+r_1r_2\mu}\int_K^\infty e^{-\frac{\lambda}{r_2}y}\mathrm{d}\mathbb{P}(S<y), \end{aligned} \tag{6.15}$$

$$Q_2 = \frac{\lambda\mu(r_1-r_2)}{\lambda(r_1-r_2)+r_1r_2\mu}\int_K^\infty e^{-\frac{\lambda}{r_2}y}\mathrm{d}\mathbb{P}(S<y). \tag{6.16}$$

Because we have determined the density of S on $(0,K]$ up to some constants, this concludes Step 2.

Step 3: *Sojourn time density on* (K,∞)
In this step, we first determine $\phi(\omega)$ using (6.4) and then apply Laplace inversion

once more to obtain the density of S on (K,∞). From the final result of Step 2, we deduce,

$$\int_0^K e^{-\omega x}\mathrm{d}\mathbb{P}(S<x)=\frac{Q_1}{\omega+\mu-\lambda/r_1}(1-e^{(\frac{\lambda}{r_1}-\mu-\omega)K})+\frac{Q_2}{\omega-\lambda/r_2}(1-e^{(\frac{\lambda}{r_2}-\omega)K}).\tag{6.17}$$

Substitution in (6.4) then immediately yields $\phi(\omega)$.

Next, to obtain $f_S(x)$ for $x>K$, we invert $\phi(\omega)$ on the corresponding interval. Similar to $F_1(\omega)$ in Step 2, we have

$$F_2(\omega)=(1-\rho_2)\frac{r_2\mu}{r_2(\omega+\mu)-\lambda}.$$

Laplace inversion provides the expression of an M/M/1 queue with service speed r_2:

$$s_2(x)=\mu(1-\rho_2)e^{(\frac{\lambda}{r_2}-\mu)x},\qquad \text{for } x>0,\tag{6.18}$$

where $s_2(\cdot)$ represents the density of a random variable with LST $F_2(\cdot)$.

By (6.17), it follows that the second term of Equation (6.4) constitutes a Laplace transform having four poles. We observe that the zero in the denominator of $\lambda/(\omega r_1-\lambda)$ is a removable zero. The expression in (6.17) is the LST of a density on $(0,K]$. Hence, the only pole contributing on (K,∞) is the zero in the denominator of $F_2(\omega)$, that is, $\eta=\lambda/r_2-\mu$. Since the first term of (6.4) provides the same pole, we immediately deduce that

$$f_S(x)=Q_3e^{(\frac{\lambda}{r_2}-\mu)x},\qquad \text{for } x>K.\tag{6.19}$$

We note that the terms with removable singularities in $\lambda/(\omega r_1-\lambda)$ and (6.17) do affect the constant Q_3. However, Q_3 is determined in Step 4 using the expressions for Q_1, Q_2, and the normalizing condition, and there is thus no need to specify Q_3 any further.

Step 4: *Determination of the constants*
In this final step, we use the normalizing condition $\int_0^\infty \mathrm{d}\mathbb{P}(S<x)=1$ to determine the constants Q_1, Q_2, and Q_3. In particular, combining normalization with (6.15) and (6.16) we obtain a set of three equations with the above three unknowns (hence, there is indeed no need to give Q_3 explicitly in Step 3).

Substituting (6.19) in (6.16) and calculating the integral yields

$$Q_2=Q_3\frac{\lambda(r_1-r_2)}{\lambda(r_1-r_2)+r_1r_2\mu}e^{-\mu K}.\tag{6.20}$$

Also, substitution of both (6.14) and (6.19) in (6.15) and performing the integrations, yield, for $r_1\neq r_2$,

$$Q_1=Q_1(1-e^{-\mu K})+Q_2\frac{r_1r_2\mu}{\lambda(r_1-r_2)}(e^{(\frac{\lambda}{r_2}-\frac{\lambda}{r_1})K}-1)+Q_3\frac{r_1r_2\mu}{\lambda(r_1-r_2)+r_1r_2\mu}e^{-\mu K}.$$

Consequently, using the expression for Q_2 in (6.20) in addition to some rewriting, we express Q_1 in terms of Q_3 as

$$Q_1 = Q_3 \frac{r_1 r_2 \mu}{\lambda(r_1 - r_2) + r_1 r_2 \mu} e^{(\frac{\lambda}{r_2} - \frac{\lambda}{r_1})K}. \tag{6.21}$$

From the normalizing condition $\int_0^\infty f_S(x)\mathrm{d}x = 1$, we obtain an additional equation. Using the densities of (6.14) and (6.19) and determining the integrals yields (for $\lambda \neq r_1\mu$, with an obvious modification when $\lambda = r_1\mu$):

$$\frac{Q_1 r_1}{\lambda - r_1\mu}(e^{(\frac{\lambda}{r_1}-\mu)K} - 1) + \frac{Q_2 r_2}{\lambda}(e^{\frac{\lambda}{r_2}K} - 1) + \frac{Q_3 r_2}{\lambda - r_2\mu} e^{(\frac{\lambda}{r_2}-\mu)K} = 1.$$

Now, substituting (6.21) and (6.20) in the above in addition to some manipulations, gives

$$\begin{aligned} Q_3 \quad = \quad & \left[\left(\frac{r_2}{\lambda - r_1\mu} - \frac{r_2}{\lambda - r_2\mu} \right) e^{(\frac{\lambda}{r_2}-\mu)K} - \frac{r_2(r_1 - r_2)}{\lambda(r_1 - r_2) + r_1 r_2 \mu} e^{-\mu K} \right. \\ & \left. - \frac{r_1^2 r_2 \mu}{(\lambda(r_1 - r_2) + r_1 r_2 \mu)(\lambda - r_1\mu)} e^{(\frac{\lambda}{r_2} - \frac{\lambda}{r_1})K} \right]^{-1}. \end{aligned} \tag{6.22}$$

The expressions for Q_1 and Q_2 follow directly from (6.21) and (6.20).

Summarizing, we have found that, in the M/M/1 queue with a two-step service speed function, the density of the "sojourn time" is given by (6.14) and (6.19), the constants Q_1, Q_2, Q_3 being specified by (6.20), (6.21) and (6.22). Observing that $S_n = W_n + B_n$, where W_n and B_n are independent, now yields the distribution of W, and hence, using PASTA, the steady-state workload distribution. We give $\mathbb{P}(W = 0)$ and the density $f_W(x)$, $x > 0$:

$$\mathbb{P}(W = 0) = \frac{Q_1 + Q_2}{\mu}, \tag{6.23}$$

$$f_W(x) = \begin{cases} Q_1 \rho_1 \mathrm{e}^{(\frac{\lambda}{r_1}-\mu)x} + Q_2(1 + \rho_2)\mathrm{e}^{\frac{\lambda}{r_2}x}, & \text{for } 0 < x \leq K, \\ Q_3 \rho_2 \mathrm{e}^{(\frac{\lambda}{r_2}-\mu)x}, & \text{for } x > K. \end{cases} \tag{6.24}$$

Remark 6.3.1 Note that the above equations reduce to familiar results for the M/M/1 queue with service speed r_2 in case either $K = 0$, or $r_1 = r_2$. In particular, we then have

$$f_S(x) = \mu(1 - \rho_2)e^{-\mu(1-\rho_2)x}, \qquad \text{for } x > 0,$$

corresponding to $s_2(x)$ in (6.18). ◇

6.4 General service requirements

In this section we apply the procedure described in Section 6.2 to the general M/G/1 queue. The basic ideas are similar as in the M/M/1 case of Section 6.3.

Again, we start the algorithm with Equation (6.11), which is the result of Step 1 in Section 6.2.

Step 2: *Sojourn time distribution on* $(0, K]$
The transforms in this step can be treated in a similar manner as the transforms in the exponential case of Section 6.3. First, to describe the inverse of $F_1(\omega)$, we recall that

$$H(x) = \beta^{-1} \int_0^x (1 - B(y))\mathrm{d}y$$

represents the stationary residual service requirement distribution. Similar to [52, 53], let $\delta_1 = 0$ for $\rho_1 \leq 1$ and for $\rho_1 > 1$ let δ_1 be the unique positive zero of the function

$$\int_0^\infty e^{-xy} \rho_1 \mathrm{d}H(y) - 1.$$

Then, for $x > 0$, define

$$L(x) := \int_0^x e^{-\delta_1 y} \rho_1 \mathrm{d}H(y),$$

and

$$W_1(x) := \int_{0-}^x e^{\delta_1 y} \mathrm{d} \left\{ \sum_{n=0}^\infty L^{n^*}(y) \right\},$$

where $L^{n^*}(\cdot)$ denotes the n-fold convolution of $L(\cdot)$ with itself (which notation, in this chapter, is more convenient than L_n). Finally, let

$$S_1(x) := (1 - \rho_1) \int_0^x B(x - y)\mathrm{d}W_1(y),$$

be the convolution of $(1 - \rho_1)W_1(\cdot)$ with $B(\cdot)$. It may be checked that, as in [52, 53], the LST of $S_1(\cdot)$ equals $F_1(\omega)$, that is, Equation (6.3) with $i = 1$.

For $\rho_1 < 1$, we note that $(1-\rho_1)W_1(\cdot)$ and $S_1(\cdot)$ are the steady-state waiting-time and sojourn-time distributions in an M/G/1 queue with service speed r_1. In case $\rho_1 \geq 1$, $W_1(\cdot)$ may be interpreted in terms of a dam with release rate r_1 and capacity K. Specifically, the stationary waiting-time distribution for such a dam equals $W_1(\cdot)/W_1(K)$, see for instance [52], or [56], p. 536.

To obtain the sojourn time distribution on $(0, K]$, we apply Laplace inversion to each of the transforms in (6.11) as in Section 6.3. The inverse of the first LST $F_1(\omega)$ is described above. For the second transform

$$F_1(\omega)\mathbb{E}\left[e^{-\omega(S - r_2 A)} I(S - r_2 A > K)\right],$$

we recall that this involves a product of two LSTs, corresponding to the sum of a random variable with mass on $[0, \infty)$, and one with mass on $[K, \infty)$. Thus the sum has no mass on $(0, K]$. Using (6.13) for the third transform in (6.11) as in Section 6.3, we obtain, for $\rho_1 \neq 1$,

$$\mathbb{P}(S < x) = \frac{W(0)}{1-\rho_1} S_1(x) + \frac{(1-\frac{r_2}{r_1})}{1-\rho_1} \int_K^\infty e^{-\frac{\lambda}{r_2}y} \mathrm{d}\mathbb{P}(S < y) \int_0^x S_1(x-y) f(y) \mathrm{d}y. \tag{6.25}$$

The above equation may be rewritten into an intuitively more appealing expression by using the interpretation of $f(\cdot)$. As discussed in Step 1, the event $S_n \leq K$ implies that either the previous sojourn time was also at or below K, or a downcrossing has occurred between the two consecutive arrivals. Denote the probability of a downcrossing of K between two successive arrivals by $P_{\downarrow K}$. Then, obviously,

$$P_{\downarrow K} = \int_K^\infty e^{-\frac{\lambda}{r_2}(y-K)} \mathrm{d}\mathbb{P}(S < y).$$

Let A_λ be a generic exponential random variable with mean $1/\lambda$. It is then easily seen that

$$\mathbb{E}\left[e^{-\omega(K-A_{\lambda/r_2})^+}\right] = \frac{\lambda}{r_2\omega - \lambda}\left(e^{-\frac{\lambda}{r_2}K} - e^{-\omega K}\right) + e^{-\frac{\lambda}{r_2}K}.$$

In case $\rho_1 < 1$, let $\hat{S}_1$ denote a generic sojourn time in an M/G/1 queue with service rate r_1. Combining the above directly gives, for $x \in (0, K]$ and $\rho_1 < 1$,

$$\mathbb{P}(S < x) = \frac{Q}{1-\rho_1}\mathbb{P}(\hat{S}_1 < x) + \frac{1-\frac{r_2}{r_1}}{1-\rho_1} P_{\downarrow K} \mathbb{P}(\hat{S}_1 + (K - A_{\lambda/r_2})^+ < x), \tag{6.26}$$

where

$$Q := \int_0^K e^{-\frac{\lambda}{r_1}y} \mathrm{d}\mathbb{P}(S < y) + \frac{r_2}{r_1}\int_K^\infty e^{-\frac{\lambda}{r_2}y} \mathrm{d}\mathbb{P}(S < y).$$

To provide some insight, let a cycle be the sample path in $(0, K]$ starting when the workload process enters $(0, K]$ and ending when it leaves $(0, K]$. Then, the two probabilities in (6.26) have a direct interpretation: The first probability stems from sojourn times of customers arriving in cycles starting from the empty system, while the second term is due to cycles starting with a downcrossing of K. The sum with $(K - A_{\lambda/r_2})^+$ in the second probability corresponds to the first "waiting time" after such a downcrossing.

Finally, in case $\rho_1 \geq 1$ the intuitive form may be expressed in a similar way as (6.26). In that case, let $\hat{W}_1$ be a generic waiting time in an M/G/1 dam with service speed r_1 and finite buffer K and let B be a generic service requirement. Expression (6.26) then holds upon replacing $\hat{S}_1$ by $\hat{W}_1 + B$ and $1/(1-\rho_1)$ by $W_1(K)$.

Step 3: *Sojourn time distribution on* (K, ∞)
Taking the LST of (6.25) on $(0, K]$ and substituting the result in (6.4) yields $\phi(\omega)$. Below, we apply Equation (6.4) directly though to derive the sojourn time distribution on (K, ∞).

First, define

$$W_2(x) := (1-\rho_2)\sum_{n=0}^{\infty}\rho_2^n H^{n^*}(x).$$

Because $\rho_2 < 1$, $W_2(\cdot)$ corresponds to the steady-state waiting-time distribution in an M/G/1 queue with service speed r_2, see for instance Theorem 1.6.1. Let $S_2(x) = W_2(x) * B(x)$ be the stationary sojourn time distribution in such a queue, with generic random variable $\hat{S}_2$. As is well-known, $F_2(\omega)$ in (6.3) is the LST of $S_2(\cdot)$.

For convenience, denote $\gamma(\omega) := \int_0^K e^{-\omega x}\mathrm{d}\mathbb{P}(S < x)$. Using standard algebra, we deduce

$$\lambda\frac{\gamma(\lambda)-\gamma(\omega)}{\omega-\lambda} = \mathbb{E}\left[e^{-\omega(S-A_\lambda)^+}I(S \le K)\right] - \gamma(\lambda). \qquad (6.27)$$

Define, for $0 \le x \le K$,

$$\begin{aligned}\tilde{S}(x) &:= \mathbb{P}((S-A_{\lambda/r_1})^+ I(S \le K) \le x)\\ &= \int_0^x \tilde{s}(y)\mathrm{d}y + \tilde{S}(0),\end{aligned}$$

where $\tilde{S}(0) = \int_0^K e^{-\frac{\lambda}{r_1}y}\mathrm{d}\mathbb{P}(S < y)$, which is also equal to $\gamma(\lambda/r_1)$, and

$$\tilde{s}(x) := \int_x^K \frac{\lambda}{r_1}e^{-\frac{\lambda}{r_1}(y-x)}\mathrm{d}\mathbb{P}(S < y).$$

Combining the above with (6.4) rewritten as

$$\phi(\omega) = F_2(\omega)\frac{W(0)}{1-\rho_2} + F_2(\omega)\frac{1-\frac{r_1}{r_2}}{1-\rho_2}\frac{\lambda/r_1}{\omega-\lambda/r_1}\left(\gamma\left(\lambda/r_1\right)-\gamma(\omega)\right),$$

we obtain, for $x > K$,

$$\mathbb{P}(S < x) = \frac{W(0)}{1-\rho_2}S_2(x) + \frac{1-\frac{r_1}{r_2}}{1-\rho_2}\int_0^K S_2(x-y)\tilde{s}(y)\mathrm{d}y. \qquad (6.28)$$

Alternatively, using that

$$W(0) = \frac{r_1}{r_2}Q + (1-\frac{r_1}{r_2})\gamma(\lambda/r_1),$$

the sojourn time distribution may be expressed as

$$\mathbb{P}(S < x) = \frac{\frac{r_1}{r_2}Q}{1-\rho_2}\mathbb{P}(\hat{S}_2 < x) + \frac{1-\frac{r_1}{r_2}}{1-\rho_2}\mathbb{P}(\hat{S}_2 + (S-A_{\lambda/r_1})^+ I(S \le K) < x). \qquad (6.29)$$

Here, the first probability relates to busy cycles in which all "sojourn times" are larger than K. In that case, the system is identical to an M/G/1 queue with service speed r_2. In case $S_n \le K$ before the end of the busy cycle, the sample path above level K in the subsequent part of the busy cycle is initiated by $S - A_{\lambda/r_1}$ with $S \leq K$, as is reflected in the second term. Note that Equation (2.15) in [53] has a similar structure.

Step 4: *Determination of the constants*
Using the fact that $\lim_{x\to\infty} \mathbb{P}(S < x) = 1$ and $\lim_{x\to\infty} S_2(x) = 1$, we deduce from (6.28) that

$$W(0) = 1 - \rho_2 - (1 - \frac{r_1}{r_2}) \left(\mathbb{P}(S < K) - \int_0^K e^{-\frac{\lambda}{r_1}y} \mathrm{d}\mathbb{P}(S < y) \right). \tag{6.30}$$

Moreover, substituting $x = K$ in (6.25) yields

$$\mathbb{P}(S < K) = \frac{W(0)}{1-\rho_1} S_1(K) + \frac{(1 - \frac{r_2}{r_1})}{1-\rho_1} \int_K^\infty e^{-\frac{\lambda}{r_2}y} \mathrm{d}\mathbb{P}(S < y) \int_0^K S_1(K-y) f(y) \mathrm{d}y. \tag{6.31}$$

The constants $\int_0^K e^{-\frac{\lambda}{r_1}y}\mathrm{d}\mathbb{P}(S < y)$ and $\int_K^\infty e^{-\frac{\lambda}{r_2}y}\mathrm{d}\mathbb{P}(S < y)$ can be determined in terms of $W(0)$ and $\mathbb{P}(S < K)$ using Equations (6.10) and (6.25). Hence, using (6.30) and (6.31), we find after lengthy calculations that

$$\begin{aligned} W(0) &= \frac{(1-\rho_1)(1-\rho_2)(D_1 + e^{-\frac{\lambda}{r_1}K} f_2)}{(1-\frac{r_1}{r_2})S_1(K)D_2 + D_3 + (1-\rho_1)\frac{r_1}{r_2}e^{-\frac{\lambda}{r_1}K} f_2}, & (6.32) \\ \int_K^\infty e^{-\frac{\lambda}{r_2}y}\mathrm{d}\mathbb{P}(S < y) &= W(0) \frac{D_1 - e^{-\frac{\lambda}{r_1}K} S_1(K)}{D_1 + e^{-\frac{\lambda}{r_1}K} f_2}, & (6.33) \end{aligned}$$

where

$$\begin{aligned} f_i &:= \int_0^K \frac{\lambda}{r_i} e^{\frac{\lambda}{r_i}(K-y)} S_1(y) \mathrm{d}y, \qquad i = 1, 2, \\ D_1 &:= 1 - \rho_1 - e^{-\frac{\lambda}{r_1}K} f_1, \\ D_2 &:= D_1 + e^{-\frac{\lambda}{r_1}K} \left(\frac{r_2}{r_1} f_2 - (1 - \rho_1) \right), \\ D_3 &:= D_1 \left(1 - \rho_1 + (1 - \frac{r_1}{r_2})(1 - \frac{r_2}{r_1}) f_2 \right). \end{aligned}$$

Summarizing, the density of the "sojourn time" is given by (6.25) and (6.28) (see (6.26) and (6.29) for another representation), where the main constants are given by (6.32) and (6.33). Because $S_n = W_n + B_n$, where W_n and B_n are independent, we also directly obtain the "waiting-time" distribution and, applying PASTA, the steady-state workload distribution. In particular, for $x \in (0, K]$, we have

$$\mathbb{P}(W < x) = W(0) W_1(x) + (1 - \frac{r_2}{r_1}) \int_K^\infty e^{-\frac{\lambda}{r_2}y} \mathrm{d}\mathbb{P}(S < y) \int_0^x W_1(x-y) f(y) \mathrm{d}y, \tag{6.34}$$

and for $x > K$,

$$\mathbb{P}(W < x) = \frac{W(0)}{1-\rho_2} W_2(x) + \frac{1-\frac{r_1}{r_2}}{1-\rho_2} \int_0^K W_2(x-y)\tilde{s}(y)\mathrm{d}y.$$

Note that we may determine the density $\tilde{s}(y)$, $0 < y \leq K$, up to some constants, once we have found the workload distribution on $(0, K]$.

6.5 Special cases and extensions

In this section we first consider some special cases of the model with a two-step service rule and conclude with the extension to the N-step service speed function. The case of exponentially distributed service requirements and the two-step service rule has already been treated in Section 6.3. In Subsection 6.5.1 we focus on service requirements with a rational LST to provide some structural properties. Furthermore, by allowing general service requirements, but letting $r_2 \to \infty$ we obtain an M/G/1 queue with disasters (clearings) at level crossings in Subsection 6.5.2. Finally, in Subsection 6.5.3 we analyze the M/G/1 queue with an N-step service speed function.

6.5.1 Service requirements with rational LST

In this subsection we assume that the LST $\beta(\omega)$ is a rational function of ω. This allows us to obtain some structural properties of the steady-state sojourn time distribution. In particular, let

$$\beta(\omega) = \frac{\beta_1(\omega)}{\beta_2(\omega)},$$

where $\beta_1(\omega)$ and $\beta_2(\omega)$ are polynomials in ω with $\beta_2(\omega)$ of degree n and $\beta_1(\omega)$ of degree strictly less than n (in other words, we assume $B(0^+) = 0$). This class includes, for instance, phase-type distributions. We use the notation M/K_n/1 to denote single-server queues where the service requirements have such rational LSTs.

The inverse of $F_i(\omega)$, $i = 1, 2$, can now be given more explicitly. Rewrite (6.3) as

$$F_i(\omega) = (1-\rho_i)\frac{r_i\beta_1(\omega)}{r_i\beta_2(\omega) - \lambda(\beta_2(\omega) - \beta_1(\omega))/\omega}.$$

Let $\delta_2 := 0$ and $\epsilon > 0$ be arbitrary small. It then follows from Rouché's theorem applied to the function $r_i\beta_2(\omega) - \lambda(\beta_2(\omega) - \beta_1(\omega))/\omega$ for Re $\omega \leq \delta_i + \epsilon$, $i = 1, 2$, that the function has exactly n zeros in the plane with Re $\omega < \delta_i + \epsilon$ (see for instance [56], p. 323, in case $\rho_i < 1$).

For ease of presentation, we assume that the function $r_i\beta_2(\omega) - \lambda(\beta_2(\omega) - \beta_1(\omega))/\omega$, $i = 1, 2$, has one zero of multiplicity m_i, $m_i = 2, 3, \ldots, n$, while the other $n - m_i$ zeros are simple, i.e., have multiplicity one. Let $\omega_i(1)$ be the non-simple zero and $\omega_i(m_i+1), \ldots, \omega_i(n)$ be the distinct simple zeros. By a partial-fraction expansion and Laplace inversion of $F_i(\omega)$, we have

$$s_i(x) = \sum_{k=1}^{m_i} \tilde{Q}_i(k) x^k e^{\omega_i(1)x} + \sum_{k=m_i+1}^{n} \tilde{Q}_i(k) e^{\omega_i(k)x},$$

for some constants $\tilde{Q}_i(k)$, $i = 1, 2$ and $k = 1, \ldots, n$. In other words, the density of the sojourn time in the M/K_n/1 queue with service speed r_i may be written as the mixture of m_i Erlang densities with scale parameter $\omega_i(1)$ and $n - m_i$ exponential terms.

It now follows from the general expressions in Section 6.4 that the "sojourn time" density has a similar structure. First consider $0 < x \leq K$. Note that the convolution of an Erlang(k, μ) distribution with an exponential term is a mixture of Erlang(i, μ), $i = 1, \ldots, k$, distributions and the same exponential. Using (6.25), we obtain, for $0 < x \leq K$,

$$f_S(x) = \sum_{k=1}^{m_1} Q_1(k) x^k e^{\omega_1(1)x} + \sum_{k=m_1+1}^{n} Q_1(k) e^{\omega_1(k)x} + Q_0 e^{\frac{\lambda}{r_2}x}.$$

Observe that $f_S(x)$ has the same Erlang and exponential terms as the sojourn time density in an ordinary M/K_n/1 queue with service speed r_1 (for $\rho_1 < 1$) plus one additional exponential $\exp(x\lambda/r_2)$ (but with different constants). Further observe that $\omega_i(k)$, $i = 1, 2$, $k = m_i + 1, \ldots, n$, might be complex, in which case its complex conjugate will also appear, leading to an exponential times a cosine, respectively, sine function.

Second, for $x > K$, we use the fact that the conditional sojourn time density of $\hat{S}_2$ has the same structure as the density of $\hat{S}_2$ itself, i.e.,

$$s_2(x + y | \hat{S}_2 > y) = \sum_{k=1}^{m_2} \hat{Q}_2(k) x^k e^{\omega_2(1)x} + \sum_{k=m_2+1}^{n} \hat{Q}_2(k) e^{\omega_2(k)x},$$

for some constants $\hat{Q}_2(k)$, $k = 1, \ldots, n$ (which depend on y). Combining the above with (6.28), we deduce that

$$f_S(x) = \sum_{k=1}^{m_2} Q_2(k) x^k e^{\omega_2(1)x} + \sum_{k=m_2+1}^{n} Q_2(k) e^{\omega_2(k)x}.$$

Finally, using the normalization condition $\int_0^\infty f_S(x)\mathrm{d}x = 1$ together with the definitions of $Q_i(k)$, $i = 1, 2$ and $k = 1, \ldots, n$, provides $2n + 1$ equations for determining the $2n + 1$ constants Q_0, $Q_i(k)$, for $i = 1, 2$ and $k = 1, \ldots, n$.

6.5.2 Disasters at level crossings

A special case of the model discussed in Section 6.4 is an M/G/1 queue with disasters at level crossings, see e.g. [42]. In such a model, the system is immediately cleared when the workload exceeds some level K, that is, the residual amount of work is removed from the system when the workload becomes larger than K. In case $r_2 \to \infty$ in our model, the available amount of work is not removed but served instantaneously when the workload upcrosses K. However, both interpretations of the work present after such an upcrossing result in identical mathematical models.

First, we note that the workload embedded at epochs right after arrival instants may be larger than K in our model (with $r_2 \to \infty$). In terms of clearing processes, this embedded workload may be considered as the overshoot (and thus the amount of work lost) rather than the actual amount of work present. Letting $r_2 \to \infty$ in (6.28) yields, for $x > K$,

$$\mathbb{P}(S < x) = W(0)B(x) + \int_0^K B(x-y)\tilde{s}(y)\mathrm{d}y,$$

where $\tilde{s}(\cdot)$ may, for instance, be determined by letting $r_2 \to \infty$ in (6.25).

For clearing models, the workload might be a more natural performance measure than the "sojourn time". In particular, we have $\mathbb{P}(W \leq K) = 1$ and, for $x \in (0, K)$, Equation (6.34) reduces to

$$\mathbb{P}(W < x) = W(0)W_1(x) - \mathbb{P}(S > K)\frac{\lambda}{r_1}\int_0^x W_1(y)\mathrm{d}y.$$

By letting $r_2 \to \infty$ in (6.32) and (6.33), we obtain the two main constants

$$\begin{aligned}
W(0) &= \frac{(1-\rho_1)D_1}{S_1(K)\left(D_1 - e^{-\frac{\lambda}{r_1}K}D_4\right) + D_1D_4},\\
\mathbb{P}(S > K) &= W(0)\frac{D_1 - e^{-\frac{\lambda}{r_1}K}S_1(K)}{D_1},
\end{aligned}$$

where

$$D_4 = 1 - \rho_1 - \frac{\lambda}{r_1}\int_0^K S_1(y)\mathrm{d}y.$$

Observe that Equations (6.10) and (6.30) are identical when $r_2 \to \infty$. Because $\mathbb{P}(S > K)$ equals $\int_K^\infty e^{-\frac{\lambda}{r_2}y}\mathrm{d}\mathbb{P}(S < y)$ in that case, the three constants can also be found from the three independent equations as discussed in Section 6.4.

Remark 6.5.1 In the M/M/1 case with $r_1 = 1$, it may be checked that (6.23) and (6.24) for $r_2 \to \infty$, or the expressions given above, indeed reduce to the workload density and the probability of an empty system of [42, Theorem 3]. $\diamond$

6.5.3 N-step service rule

In this subsection we extend the analysis to an N-step service rule. Specifically, let $r(x) = r_i$ for $x \in (K_{i-1}, K_i]$, $i = 1, \ldots, N$ (where $K_0 = 0$ and $K_N = \infty$). Also, define $\rho_i := \lambda\beta/r_i$. For stability, we require that $\rho_N < 1$. The basic ideas are now similar to the case $N = 2$ discussed in Section 6.4.

Below, we give the derivation of the "sojourn time" distribution for the N-step service rule along similar lines as the four-step procedure described in Section 6.2. That is, we first present N different equations for $\phi(\omega)$. Second, we use a similar interpretation as in Step 1 to rewrite the N equations. Third, similar to Step 2 in Section 6.4 we analyze $\mathbb{P}(S < x)$ for $x \in (0, K_1]$. Then, we recursively determine $\mathbb{P}(S < x)$ for $x \in (K_{i-1}, K_i]$, $i = 2, \ldots, N$ (comparable with Step 3). We conclude with some remarks about the determination of the constants.

Concerning the equations for $\phi(\omega)$, it follows from (6.1), (6.6), and conditioning on S_n that

$$\begin{aligned}\mathbb{E}\left[e^{-\omega S_{n+1}}\right] &= \int_0^\infty \mathbb{E}\left[e^{-\omega S_{n+1}} | S_n = x\right] \mathrm{d}\mathbb{P}(S_n < x) \\ &= \beta(\omega) \sum_{j=1}^N \Bigg[\frac{\omega r_j}{\omega r_j - \lambda} \int_{K_{j-1}}^{K_j} e^{-\frac{\lambda}{r_j} x} \mathrm{d}\mathbb{P}(S_n < x) \\ &\qquad\qquad - \frac{\lambda}{\omega r_j - \lambda} \int_{K_{j-1}}^{K_j} e^{-\omega x} \mathrm{d}\mathbb{P}(S_n < x) \Bigg],\end{aligned}$$

with obvious modification for $\omega = \lambda / r_j$, $j = 1, \ldots, N$. Using similar manipulations as in the proof of Lemma 6.2.1, we obtain N alternative equations for $\phi(\omega)$; for $i = 1, \ldots, N$, we have

$$\begin{aligned}\phi(\omega) &= F_i(\omega) \frac{W(0)}{1 - \rho_i} \qquad (6.35) \\ &+ \frac{F_i(\omega)}{1 - \rho_i} \sum_{j=i+1}^N \Bigg[\frac{\lambda(1 - \frac{r_j}{r_i})}{\omega r_j - \lambda} \Bigg(\int_{K_{j-1}}^{K_j} e^{-\frac{\lambda}{r_j} x} \mathrm{d}\mathbb{P}(S < x) \\ &\qquad\qquad - \int_{K_{j-1}}^{K_j} e^{-\omega x} \mathrm{d}\mathbb{P}(S < x) \Bigg) \Bigg] \\ &+ \frac{F_i(\omega)}{1 - \rho_i} \sum_{j=1}^{i-1} \Bigg[\frac{\lambda(1 - \frac{r_j}{r_i})}{\omega r_j - \lambda} \Bigg(\int_{K_{j-1}}^{K_j} e^{-\frac{\lambda}{r_j} x} \mathrm{d}\mathbb{P}(S < x) \\ &\qquad\qquad - \int_{K_{j-1}}^{K_j} e^{-\omega x} \mathrm{d}\mathbb{P}(S < x) \Bigg) \Bigg],\end{aligned}$$

with obvious notation for $F_i(\omega)$ and

$$W(0) = \sum_{j=1}^N \int_{K_{j-1}}^{K_j} e^{-\frac{\lambda}{r_j} x} \mathrm{d}\mathbb{P}(S < x). \qquad (6.36)$$

In the remainder, we follow the convention that empty sums are equal to zero.

Step 1: *Rewriting (6.35)*
Fix some $i = 1, \ldots, N$ and consider the second term on the rhs of (6.35). As in Step 1 of Section 6.2, $S_n > K_i$ and $S_{n+1} \leq K_i$ means that a downcrossing of level K_i occurs between the arrival epochs of customers n and $n+1$. Again, the residual interarrival time at a downcrossing of K_i is still exponential, but the service speed now depends on the value of S_n. In particular, the precise distribution of S_n on (K_i, ∞) does not directly affect the distribution of $S_{n+1} \leq K_i$ but determines the service speed until the next arrival epoch. Using similar calculations as in Step 1 of Section 6.2, we obtain

$$\begin{aligned}
&\sum_{j=i+1}^{N} \left[\frac{\lambda}{\omega r_j - \lambda} \left(\int_{K_{j-1}}^{K_j} e^{-\frac{\lambda}{r_j} x} \mathrm{d}\mathbb{P}(S_n < x) - \int_{K_{j-1}}^{K_j} e^{-\omega x} \mathrm{d}\mathbb{P}(S_n < x) \right) \right] \\
&\quad = \quad \mathbb{E}\left[e^{-\omega(S_n - r(S_n)A_n)^+} I(S_n > K_i) \right] - \sum_{j=i+1}^{N} \int_{K_{j-1}}^{K_j} e^{-\frac{\lambda}{r_j} x} \mathrm{d}\mathbb{P}(S_n < x) \\
&\quad = \quad \mathbb{E}\left[e^{-\omega(S_n - r(S_n)A_n)} I(S_n - r(S_n)A_n > K_i) \right] \\
&\qquad\qquad + \sum_{j=i+1}^{N} \frac{\lambda}{\omega r_j - \lambda} \left(1 - e^{-\omega K_i + \frac{\lambda}{r_j} K_i} \right) \int_{K_{j-1}}^{K_j} e^{-\frac{\lambda}{r_j} x} \mathrm{d}\mathbb{P}(S_n < x).
\end{aligned}$$

For convenience, we define the quantity

$$C_j := \int_{K_{j-1}}^{K_j} e^{-\frac{\lambda}{r_j} x} \mathrm{d}\mathbb{P}(S < x),$$

which is clearly independent of ω. Then, by letting $n \to \infty$, we may rewrite (6.35) as

$$\begin{aligned}
\phi(\omega) \quad &= \quad F_i(\omega) \frac{W(0)}{1 - \rho_i} \qquad\qquad (6.37) \\
&+ \quad \frac{F_i(\omega)}{1 - \rho_i} \mathbb{E}\left[e^{-\omega(S - r(S)A)} I(S - r(S)A > K_i) \right] \\
&+ \quad \frac{F_i(\omega)}{1 - \rho_i} \sum_{j=i+1}^{N} (1 - \frac{r_j}{r_i}) C_j \frac{\lambda}{\omega r_j - \lambda} \left(1 - e^{-\omega K_i + \frac{\lambda}{r_j} K_i} \right), \\
&+ \quad \frac{F_i(\omega)}{1 - \rho_i} \sum_{j=1}^{i-1} \left[\frac{\lambda(1 - \frac{r_j}{r_i})}{\omega r_j - \lambda} \left(\int_{K_{j-1}}^{K_j} e^{-\frac{\lambda}{r_j} x} \mathrm{d}\mathbb{P}(S < x) \right. \right. \\
&\qquad\qquad\qquad \left. \left. - \int_{K_{j-1}}^{K_j} e^{-\omega x} \mathrm{d}\mathbb{P}(S < x) \right) \right] \\
&=: \quad I + II + III + IV.
\end{aligned}$$

Note that the intuitive observations made above are reflected in Terms II and III.

Step 2: *Sojourn time distribution on* $(0, K_1]$
First we consider $i = 1$, i.e., the interval $(0, K_1]$. Note that this implies that $IV = 0$.

As in Step 2 of Section 6.4 we now apply Laplace inversion to each of the Terms I, II, and III separately. Again, $S_1(\cdot)W(0)/(1-\rho_1)$ is the inverse of Term I, see also Section 6.4. Term II involves the convolution of two random variables, one with mass on $[0,\infty)$ and one with mass on (K_1,∞). Hence, the sum clearly has no mass on $(0, K_1]$.

For Term III, we note that $\frac{\lambda}{\omega r_j - \lambda}\left(1 - e^{-\omega K_i + \frac{\lambda}{r_j}K_i}\right)$ is the Laplace Transform of the function

$$f_{i,j}(x) = \begin{cases} \frac{\lambda}{r_j} e^{\frac{\lambda}{r_j}x}, & \text{for } 0 < x \leq K_i, \\ 0, & \text{for } x > K_i. \end{cases}$$

To provide some intuition, suppose that $S_n \in (K_{j-1}, K_j]$ and a downcrossing of level $K_i \leq K_{j-1}$ occurs in the subsequent interarrival time, which has stationary probability

$$P^j_{\downarrow K_i} = \int_{K_{j-1}}^{K_j} e^{-\frac{\lambda}{r_j}(y-K_i)} \mathrm{d}\mathbb{P}(S < y).$$

Then $P^j_{\downarrow K_i} f_{i,j}$ may be interpreted as C_j times the "density" of $(K_i - A_{\lambda/r_j})^+$ (in fact, $(K_i - A_{\lambda/r_j})^+$ has a defective distribution with an atom in 0).

Combining the above and applying Laplace inversion provides an extension of Equation (6.25) to the case of an N-step service rule, with $0 < x \leq K_1$,

$$\mathbb{P}(S < x) = \frac{W(0)}{1-\rho_1} S_1(x) + \frac{1}{1-\rho_1} \sum_{j=2}^{N} (1 - \frac{r_j}{r_1}) C_j \int_{0^+}^{x} S_1(x-y) f_{1,j}(y) \mathrm{d}y. \quad (6.38)$$

Note that the difference with $N = 2$ is the fact that the service speed now depends on the previous "sojourn time" in case of a downcrossing of K_1. This naturally leads to a mixture of convolutions of $S_1(\cdot)$ with various exponential functions depending on the service speed in the second part of (6.38).

Step 3: *Sojourn time distribution on* $(K_{i-1}, K_i]$
In Step 2 we obtained the "sojourn time" distribution on the first interval $(0, K_1]$. We may now recursively determine the "sojourn time" distribution on the remaining intervals. That is, suppose that $\mathbb{P}(S < x)$ is known for $x \in (K_{j-1}, K_j]$, $j = 1, \ldots, i-1$, with $i = 2, \ldots, N$ (the case $i = 1$ corresponds to Step 2). Using (6.37), we then find $\mathbb{P}(S < x)$ for $x \in (K_{i-1}, K_i]$.

To do so, we apply Laplace inversion again to each of the four terms in (6.37). Terms I, II, and III can be treated as in Step 2, with obvious notation for $W_i(\cdot)$, $i = 2, \ldots, N$. For the fourth term, we apply similar arguments as in Step 3 of Section 6.4, in particular Equation (6.27). Thus,

$$IV = \frac{F_i(\omega)}{1-\rho_i} \sum_{j=1}^{i-1} (1 - \frac{r_j}{r_i}) \left(\mathbb{E}\left[e^{-\omega(S - A_{\lambda/r_j})^+} I(K_{j-1} < S \leq K_j) \right] - C_j \right).$$

Note again that $(S - A_{\lambda/r_j})^+ I(K_{j-1} < S \leq K_j)$ has a defective distribution function with an atom at zero, $\tilde{S}_j(0) := C_j$. Moreover, the density reads, for $0 < x < K_j$,

$$\tilde{s}_j(x) := \int_{\max(x,K_{j-1})}^{K_j} \frac{\lambda}{r_j} e^{-\frac{\lambda}{r_j}(y-x)} \mathrm{d}\mathbb{P}(S < y).$$

Because we assumed that $\mathbb{P}(S < x)$ is known on $(0, K_{i-1}]$, $\tilde{s}_j(x)$ is computable for every $j = 1, \ldots, i-1$.

Now, combining the above and applying Laplace inversion to (6.37) yields, for $K_{i-1} < x \leq K_i$, $i = 1, \ldots, N$,

$$\begin{aligned}\mathbb{P}(S < x) \quad &= \quad \frac{W(0)}{1-\rho_i} S_i(x) + \frac{1}{1-\rho_i} \sum_{j=i+1}^{N} (1 - \frac{r_j}{r_i}) C_j \int_{0^+}^{x} S_i(x-y) f_{i,j}(y) \mathrm{d}y \\ &\quad + \frac{1}{1-\rho_i} \sum_{j=1}^{i-1} (1 - \frac{r_j}{r_i}) \int_{0^+}^{K_j} S_i(x-y) \tilde{s}_j(y) \mathrm{d}y. \end{aligned} \qquad (6.39)$$

The $S_i(\cdot)$ term and the convolution of $S_i(\cdot)$ with $\tilde{s}_j(\cdot)$ are similar to the case $N = 2$, see (6.28). For $i = 1, \ldots, N-1$, we just have an additional convolution of $S_i(\cdot)$ with $f_{i,j}(\cdot)$, which is the consequence of "sojourn times" after a downcrossing of K_i, as discussed in Step 2.

Step 4: *Determination of the constants*
Taking $i = N$ and letting $x \to \infty$ in (6.39), yields

$$W(0) = 1 - \rho_N - \sum_{j=1}^{N-1} (1 - \frac{r_j}{r_N}) \left(\mathbb{P}(K_{j-1} \leq S < K_j) - C_j\right). \qquad (6.40)$$

Moreover, (6.39) can be used to give expressions for $\mathbb{P}(S < K_i)$ and C_i, $i = 1, \ldots, N-1$. To obtain the latter $N-1$ constants, differentiate (6.39) with respect to x, multiply by $\exp(-\lambda x/r_i)$, and integrate over the interval $(K_{i-1}, K_i]$. Together with (6.36) and (6.40), this provides $2N$ independent equations to determine the $2N$ unknowns: $W(0)$, $\mathbb{P}(S < K_i)$ for $i = 1, \ldots, N-1$, and C_i, $i = 1, \ldots, N$.

Chapter 7

Adaptive protocols for the integration of streaming and heavy-tailed elastic traffic

7.1 Introduction

In Chapters 2–5 we analyzed various queueing systems with workload-dependent service (and arrival) rates. Moreover, in Chapter 6 the service speed is adjusted only at embedded epochs based on the amount of work present. In those studies, workload typically was one of the main subjects of consideration. In this chapter, we analyze the workload distribution in a queueing model with a general (non-decreasing) input process and a varying service speed that is determined by the state of a random environment. More specifically, the random environment consists of a second class of (*elastic*) customers with heavy-tailed characteristics sharing the service capacity with the first (*streaming*) class according to the PS discipline. In contrast to previous chapters, in this chapter we mainly focus on asymptotic results as the workload gets large.

The analysis of the present chapter has been motivated by applications in communication networks. Where the queueing systems of previous chapters may, for instance, be applied to model the packet-level dynamics, we now focus on the flow level (of the elastic flows), see Section 1.5 for details. In particular, we consider a fixed number of streaming users and a dynamic population of elastic flows sharing the bandwidth in a bottleneck link. Under some assumptions we derive various performance measures for the elastic flows and the workload asymptotics for the streaming class (the main result of the chapter). The latter is especially relevant because the workload can be interpreted as the shortfall in received amount of service compared to a nominal service target. This interpretation, in addition to an elaborate model description, is discussed in Section 7.3.

The remainder of the chapter is organized as follows. We give some background information on the integration of streaming and elastic traffic on a common infrastructure in Section 7.2. In Section 7.3 we present a detailed model description. In Section 7.4 we analyze the delay and workload performance of the elastic flows by exploiting a useful relationship with an M/G/1 PS model with permanent customers. The main result is presented in Section 7.5, where we consider the workload asymptotics of the streaming users for the case of constant-rate traffic. Besides a heuristic interpretation of the result, we also give some preliminaries and an outline of the proof, which involves lower and upper bounds that asymptotically coincide. The proofs of the lower and upper bounds may be found in Sections 7.6 and 7.7, respectively. We extend the results to the case of variable-rate streaming traffic in Section 7.8. In addition, we consider the tail asymptotics of the joint workload distribution of the K individual streaming users. In Section 7.9 we make some concluding remarks.

7.2 Integration of elastic and streaming traffic

Over the past decade, TCP has become the most prominent congestion control mechanism in the Internet. While TCP is adequate for best-effort elastic traffic, such as file transfers and Web browsing sessions, it is less suitable for supporting delay-sensitive streaming applications. In particular, real-time streaming applications are extremely vulnerable to the fluctuations in the window size that are characteristic for TCP. As a potential alternative, UDP could be used to avoid the wild oscillations in the transmission rate. Since UDP does not respond to congestion, it may cause severe packet losses however, and give rise to unfairness in the competition for bandwidth with TCP-controlled flows.

Discriminatory packet scheduling mechanisms provide a further alternative to achieve some form of prioritization of streaming applications. However, the implementation of scheduling mechanisms involves major complexity and scalability issues. In addition, prioritization of streaming applications may cause performance degradation and even starvation of TCP-controlled flows that back off in response to congestion. Evidently, the latter issue gains importance as the amount of streaming traffic in the Internet grows.

The above considerations have motivated an interest in *TCP-friendly* or *equation-based* rate control protocols for streaming applications [77, 134, 142]. The key goal is to eliminate severe fluctuations in the window size and adjust the transmission rate in a smoother manner. In order to ensure fairness with competing TCP-controlled flows, the specific aim is to set the transmission rate to the 'fair' bandwidth share, i.e., the throughput that a long-lived TCP flow would receive under similar conditions.

Various methods have been proposed for determining the fair bandwidth share in an accurate and robust manner. Typical methods involve measuring the packet loss rate and round-trip delay (e.g. by running a low-rate connection to identify the network conditions). The corresponding throughput may then be estimated from equations that express the throughput of a TCP-controlled flow in terms of the packet loss rate and round-trip delay, see for instance [122, 133]. Obviously, the adaptation mechanism faces the usual trade-off between responsiveness and smoothness, which is worsened by the fact that the estimation procedure relies on measurements of noisy traffic.

In the present chapter we explore the performance of streaming applications under such TCP-friendly rate control protocols. As mentioned, we consider a fixed number of streaming sessions which share a bottleneck link with a dynamic population of elastic flows. The assumption of persistent streaming users is motivated by the separation of time scales between the typical duration of streaming sessions (minutes to hours) and that of the majority of elastic flows (seconds to minutes). We assume that the sizes of the elastic flows exhibit heavy-tailed characteristics. The latter assumption is based on extensive measurement studies which show that file sizes in the Internet, and hence the volumes of elastic transfers, commonly have heavy-tailed features, see for instance [60].

As mentioned above, the design and implementation of TCP-friendly mechanisms is a significant challenge. In the present chapter we leave implementation issues aside though, and investigate the performance under idealizing assumptions. Specifically, we assume that the rate control mechanism reacts instantly and perfectly accurately to changes in the population of elastic flows, and maintains a constant rate otherwise. This results – at the flow level – in a fair sharing of the link rate in a PS manner. The PS discipline has emerged as a useful paradigm for modeling the bandwidth sharing among dynamically competing TCP flows, see for instance [29, 121, 130]. Although the PS paradigm may not be entirely justified for short flows, inspection of the proofs suggests that this assumption is actually not that crucial for the asymptotic results to hold. The effect of oscillations, inaccuracies and delays in the estimation procedure on the performance remains as a subject for future research.

We consider the probability that a possible deficit in service received by the streaming sessions compared to a nominal service target exceeds a certain threshold. The latter probability provides a measure for the quality of the connection experienced by the streaming users. We determine the asymptotic behavior of the service deficit (or *workload*) probability for a large value of the threshold. The results yield useful qualitative insight into the occurrence of persistent quality disruption for the streaming users. We furthermore examine the delay performance of the elastic flows.

In [106], the authors consider a mixture of elastic transfers and streaming users sharing the network bandwidth according to weighted α-fair rate algorithms. Weighted α-fair allocations include various common fairness notions, such as max-min fairness and proportional fairness, as special cases. They also provide a tractable theoretical abstraction of the throughput allocations under decentralized feedback-based congestion control mechanisms such as TCP, and in particular cover TCP-friendly rate control protocols. In a recent paper [32], the authors derive various performance bounds for a related model with a combination of elastic flows and streaming traffic sharing the link bandwidth in a fair manner. The latter papers however focus on other performance metrics than in the present chapter.

7.3 Model description

We consider two traffic classes sharing a link of unit rate. Class 1 consists of a static population of $K \geq 1$ statistically identical streaming sessions. These sessions stay in the system indefinitely. Class 2 consists of a dynamic configuration of elastic flows. These users arrive according to a renewal process with mean interarrival time $1/\lambda$, and have service requirements with distribution $B(\cdot)$ and mean $\beta < \infty$.

The elastic flows are TCP-controlled, while the transmission rates of the streaming sessions are adapted in a TCP-friendly fashion. Abstracting from packet-level details, we assume that this results in a fair sharing of the link rate according to the PS discipline. Thus, when there are $N(u)$ elastic flows in the system at time u, the available service rate for each of the users – either elastic or streaming – is $1/(K + N(u))$. Denote by $C_1(u) := K/(K + N(u))$ the total available service rate for the streaming traffic at time u. Define $C_1(s,t) := \int_{u=s}^{t} C_1(u)\mathrm{d}u$ as the total amount of service available for the streaming sessions during the time interval $[s,t]$.

In the present chapter, we will mainly be interested in the quantity $V_1(t) := \sup\limits_{s \leq t}\{A_1(s,t) - C_1(s,t)\}$, where $A_1(s,t)$ denotes the amount of service which ideally should be available for the streaming traffic during the interval $[s,t]$. For example, $A_1(s,t)$ may be taken as the amount of streaming traffic that would nominally be generated during the interval $[s,t]$ if there were ample bandwidth. Thus, $V_1(t)$ may be interpreted as the shortfall in service for the streaming traffic at time t compared to what should have been available in ideal circumstances. For conciseness, we will henceforth refer to $V_1(t)$ as the *workload* of the streaming traffic at time t (see (1.8) for a representation of the steady-state workload). Throughout the chapter, we also often refer to $A_1(s,t)$ as the amount of streaming traffic generated. It is worth emphasizing though that $A_1(s,t)$ represents just the amount of traffic which ideally should have been served, and not the amount of traffic that is actually generated, which is primarily governed by the fair service rates as described above. Thus, $V_1(t)$ provides just a virtual measure of a service deficit compared to an ideal environment, and by no means corresponds to the backlog or buffer content in an actual system.

In Sections 7.4–7.7 we will focus on the 'constant-rate' case $A_1(s,t) \equiv Kr(t - s)$, which amounts to a fixed target service rate r per streaming session. We will extend the analysis in Section 7.8 to the 'variable-rate' case where $A_1(s,t)$ is a general stochastic process with stationary increments.

We will also consider the quantity $V_2(t) := \sup\limits_{s \leq t}\{A_2(s,t) - C_2(s,t)\}$, where $A_2(s,t)$ denotes the amount of elastic traffic generated during the time interval $[s,t]$, and $C_2(s,t)$ represents the amount of service available for the elastic flows during $[s,t]$. By definition, $C_2(s,t) := \int_{u=s}^{t} C_2(u)\mathrm{d}u$, with $C_2(u)$ denoting the total available service rate for the elastic traffic at time u. Evidently, $C_2(u) \geq 1 - C_1(u)$, with equality in case the streaming sessions always claim the full service rate available. For the elastic traffic, the latter case is equivalent to a

G/G/1 PS queue with K permanent customers, accounting for the presence of the competing streaming sessions.

However, we allow for possible strict inequality in case the streaming sessions do not always consume the full service rate available, and the unused surplus is granted to the elastic class, i.e., $C_2(s,t) = t - s - B_1(s,t)$, with $B_i(s,t) \leq C_i(s,t)$ denoting the actual amount of service received by class i, $i = 1,2$, during the interval $[s,t]$. For example, when the 'workload' of the streaming sessions is zero, the actual service rate may be set to the minimum of the aggregate input rate and the total service rate available. In particular, in the 'constant-rate' case the actual service rate per streaming session at time u is then only $\min\{r, 1/(K+N(u))\}$ when $V_1(u) = 0$. Note that the total service rate is thus used at time u as long as $V_1(u) + V_2(u) > 0$, which implies that $V_1(t) + V_2(t) = \sup_{s \leq t}\{A_1(s,t) + A_2(s,t) - (t-s)\}$. Hence, the case $C_2(s,t) = t - s - B_1(s,t)$ will be termed the *work-conserving* scenario, whereas the case $C_2(u) = 1 - C_1(u) = N(u)/(K+N(u))$ will be referred to as the *permanent-customer* scenario. It may be checked that the work-conserving and permanent-customer scenarios provide lower and upper bounds for the general case with $t - s - C_1(s,t) \leq C_2(s,t) \leq t - s - B_1(s,t)$.

Define $\rho := \lambda\beta$ as the traffic intensity of class 2. Without proof, we claim that $\rho < 1$ is a necessary and sufficient condition for class 2 to be stable. While the former is obvious, the latter may be concluded from the comparison with the G/G/1 PS queue with K permanent customers mentioned above (see [166] for the case of Poisson arrivals). For class 1 to be stable as well, we need to assume that $\rho + Kr < 1$, with $\mathbb{E}\{A(0,1)\} = Kr$. Here class 1 is said to be stable if the 'workload' $V_1(t)$ converges to a finite random variable as $t \to \infty$. Denote by V_i a random variable with the steady-state distribution of $V_i(t)$, $i = 1,2$. In Sections 7.5–7.8, we additionally assume that $(K+1)r > 1 - \rho$, which implies that the system is critically loaded in the sense that one extra streaming session – or a 'persistent' elastic flow – would cause instability. Combined, the above two assumptions give $Kr < 1 - \rho < (K+1)r$.

We finally introduce some additional notation. Let B be a random variable distributed as the generic service requirement of an elastic user, and let B^r be a random variable distributed as the residual lifetime of B, i.e., $B^r(x) = \mathbb{P}\{B^r < x\} = \frac{1}{\beta}\int_0^x (1 - B(y))\mathrm{d}y$. We assume that the service requirement distribution is regularly varying of index $-\nu$ (denoted as $B(\cdot) \in \mathcal{R}_{-\nu}$), i.e., $1 - B(x) \sim L(x)x^{-\nu}, \nu > 1$, with $L(x)$ some slowly varying function, see also Definition 1.6.1. As in Subsection 1.6.4, we use the notation $f(x) \sim g(x)$ to indicate that $f(x)/g(x) \to 1$ as $x \to \infty$. (A function $L(\cdot)$ is called slowly varying if $L(\eta x) \sim L(x)$ for all $\eta > 1$.) It follows from Karamata's Theorem [31, Theorem. 5.1.11] that $x\mathbb{P}\{B > x\} \sim (\nu - 1)\beta\,\mathbb{P}\{B^r > x\}$, so that $B^r(\cdot) \in \mathcal{R}_{1-\nu}$.

Remark 7.3.1 The analysis may be generalized to the case of *Discriminatory* Processor Sharing (DPS), that is, when the rate share per streaming session is $w/(wK + N(u))$ rather than $1/(K + N(u))$, for some positive weight factor w. In the 'constant-rate' case, it is in fact easily verified that the workload for

K streaming sessions each with weight w and target rate r is equivalent to that in a model with $K' = wK$ streaming sessions each with unit weight and target rate $r' = r/w$ (with some abuse of terminology when wK is not integer). For notational transparency, we henceforth focus on the case $w = 1$. ◇

7.4 Delay performance of the elastic flows

As mentioned earlier, our model shows strong resemblance with a G/G/1 PS queue with K permanent customers [166]. The permanent customers play the role of the persistent streaming users in our model, while the regular (non-permanent) customers correspond to the elastic flows, inheriting the same arrival process and service requirement distribution $B(\cdot)$. It may be checked that the service rate available for the elastic class in our model is always at least that in the model with K permanent customers. Hence, the number of elastic flows, their individual residual service requirements, their respective delays (sojourn times), and the workload of the elastic class are stochastically dominated by the corresponding quantities in the model with permanent customers. This may be formally shown using similar arguments as in the proof of Lemma 4 in [35]. The stochastic ordering between the two models is particularly useful, since it provides upper bounds for several performance measures of interest in our model in terms of the model with permanent customers. In order for the bounds to be analytically tractable, we assume in the remainder of the section that the elastic flows arrive according to a Poisson process of rate λ.

Remark 7.4.1 As noted earlier, in the special case where the service rate of the elastic class is always $C_2(t) \equiv \frac{N(t)}{K+N(t)}$ (which we named the *permanent-customer* scenario), the two models are actually equivalent in terms of the number of elastic users and their respective residual service requirements. In that case, the inequalities in Equations (7.2)-(7.6) below hold with equality. ◇

The M/G/1 PS queue with permanent customers is a special case of the model studied in [55], where each customer receives service at rate $f(n)$, $0 \leq f(n) < \infty$, when there are n customers. To obtain the model with K permanent customers, we take $f(n) = \frac{1}{K+n}$. Let $N_{(K)}$ be the number of regular customers in the model with K permanent customers and, given $N_{(K)} = n$, let $\hat{B}_1, \ldots, \hat{B}_n$ be their residual service requirements. Then, according to [55],

$$\begin{aligned}\mathbb{P}\left\{N_{(K)} = n; \hat{B}_1 > x_1; \ldots; \hat{B}_n > x_n\right\}& \\ = \quad (1-\rho)^{K+1}\rho^n \binom{n+K}{n} \prod_{m=1}^{n} \mathbb{P}\left\{B^r > x_m\right\}. & \qquad (7.1)\end{aligned}$$

(When $w \neq 1$ and wK is not integer, the above formula remains valid upon substituting wK for K and replacing the factorial function in the binomial coefficients by the Gamma function.) We thus obtain an upper bound for the probability that the service rate of the streaming users is below a given desired rate s:

$$\begin{aligned}\mathbb{P}\left\{\frac{1}{K+N} < s\right\} &\leq \mathbb{P}\left\{N_{(K)} > \lfloor 1/s - K \rfloor\right\} \\ &= \sum_{j=0}^{K} \binom{\lfloor 1/s \rfloor + 1}{j} (1-\rho)^j \rho^{\lfloor 1/s \rfloor + 1 - j}. \end{aligned} \tag{7.2}$$

As mentioned above, the delay (sojourn time) of elastic users in our model (denoted by S_2) is stochastically dominated by the corresponding quantity in the model with permanent customers. In the M/G/1 PS queue with m permanent customers, let $S_{(m)}$ be the delay and $S_{(m)}(x)$ be the conditional sojourn time *given* that the service requirement of the customer is x. It is known that this random variable is the $(m+1)$-fold convolution of the distribution of $S_{\mathrm{PS}}(x)$, the conditional sojourn time in the standard M/G/1 PS queue [166]:

$$\mathbb{P}\left\{S_{(m)}(x) \leq t\right\} = \mathbb{P}\left\{\sum_{j=1}^{m+1} S_{\mathrm{PS},j}(x) \leq t\right\},$$

where $S_{\mathrm{PS},j}$, $j = 1, \dots, m+1$, represent i.i.d. copies of S_{PS}. (It is worth emphasizing that the unconditional sojourn time does not allow for a similar decomposition.) In particular, using that $\mathbb{E}S_{\mathrm{PS}}(x) = \frac{x}{1-\rho}$, we obtain an upper bound for the conditional mean delay of elastic users in our model (denoted as $S_2(x)$):

$$\mathbb{E}S_2(x) \leq \mathbb{E}S_{(K)}(x) = (K+1)\frac{x}{1-\rho}, \tag{7.3}$$

and, hence, the (unconditional) mean delay satisfies

$$\mathbb{E}S_2 \leq (K+1)\frac{\beta}{1-\rho}. \tag{7.4}$$

We now turn to the tail asymptotics for the unconditional sojourn time. The next proposition shows that the exact asymptotics of S_2 depend on the assumptions on $C_2(s,t)$ in case $B_1(s,t) < C_1(s,t)$. As observed in Remark 7.4.1, in case $C_2(t) \equiv \frac{N(t)}{K+N(t)}$, the model is equivalent to the M/G/1 PS queue with K permanent customers. Asymptotically, the equivalence also continues to hold when the system is critically loaded, i.e., $(K+1)r > 1-\rho$, which implies that class-1 users will be rarely non-backlogged over the course of a long sojourn time. However, the sojourn time asymptotics change when the system is below critical load, i.e., $(K+1)r < 1-\rho$, and the elastic flows receive (part of) the capacity left over by the streaming users, i.e., $C_2(t) > \frac{N(t)}{K+N(t)}$.

Proposition 7.4.1 *If $B(\cdot) \in \mathcal{R}_{-\nu}$ and $(K+1)r > 1-\rho$ or $C_2(t) \equiv \frac{N(t)}{K+N(t)}$, or both, then*

$$\mathbb{P}\left\{S_2 > x\right\} \sim \mathbb{P}\left\{S_{(K)} > x\right\} \sim \mathbb{P}\left\{B > \frac{(1-\rho)x}{K+1}\right\}.$$

In contrast, if $(K+1)r < 1-\rho$ *and* $C_2(s,t) \equiv t - s - B_1(s,t)$*, then*

$$\mathbb{P}\{S_2 > x\} \sim \mathbb{P}\{B > (1-\rho-Kr)x\}.$$

Proof The asymptotics for $S_{(K)}$ (and, thus, for S_2 in the permanent-customer scenario) follow from [82]. As noted above, the service rate of a customer is $f(n) = \frac{1}{K+n}$ when there are n non-permanent customers in the system. We can therefore apply [82, Theorem 3] to obtain $\gamma^f = \frac{1-\rho}{K+1}$ and the desired result follows.

For the remainder of the proof we only provide an intuitive sketch; we refer to Appendix 7.C for a detailed proof. In both cases, a large delay of an elastic flow is due to a large service requirement of the flow itself, and the ratio between the two quantities is simply the average service rate received by the large flow. Over the duration of the large flow, the other elastic flows continue to produce traffic at an average rate ρ, and also receive service at an average rate ρ. The remaining service capacity is shared among the large elastic flow and the streaming users, each entitled to a fair share $(1-\rho)/(K+1)$. In case $(K+1)r > 1-\rho$, the fair share of the streaming users is below their average input rate r. Thus, the streaming users will be almost constantly backlogged, and the average service rate for the large elastic flow is just $(1-\rho)/(K+1)$. In case $(K+1)r < 1-\rho$, the fair share of the streaming users exceeds their average 'input rate' r. Hence, the streaming users will only claim an average service rate Kr, and the average service rate left for the large elastic flow is $1-\rho-Kr$ now. □

In case the system is not critically loaded and $t - s - C_1(s,t) < C_2(s,t) < t - s - B_1(s,t)$ for at least some s and t, we obtain the immediate bound

$$\mathbb{P}\{S_2 > x\} \leq (1+o(1))\mathbb{P}\left\{B > \frac{(1-\rho)x}{K+1}\right\}, \qquad \text{as } x \to \infty. \tag{7.5}$$

Remark 7.4.2 The result for the *permanent customer* scenario is formulated for regularly varying service requirements, but it may readily be extended (following the proof of [132, Theorem 4.1]) to the slightly larger class of *intermediately* regularly varying distributions. ◇

Finally, we turn to the workload of the elastic class which is also stochastically dominated by the corresponding quantity in the model with permanent customers. Again, we first state a result for the M/G/1 PS queue with permanent customers.

Proposition 7.4.2 *If* $B(\cdot) \in \mathcal{R}_{-\nu}$*, then* $V_{(m)}$*, the workload in the M/G/1 PS queue with* m *permanent customers, satisfies*

$$\mathbb{P}\left\{V_{(m)} > x\right\} \sim \mathbb{E}N_{(m)}\,\mathbb{P}\{B^r > x\} = \frac{(m+1)\rho}{1-\rho}\,\mathbb{P}\{B^r > x\}.$$

Proof From (7.1) we observe that, given $N_{(m)} = n$, $\hat{B}_1, \ldots, \hat{B}_n$ are i.i.d. copies of B^r. Using [158] together with $V_{(m)} = \sum_{i=1}^{N_{(m)}} \hat{B}_i$, and the fact that $\mathbb{P}\{N > n\}$ decays geometrically fast when $n \to \infty$, we obtain the desired equivalence. □

As an immediate corollary, we derive

$$\mathbb{P}\left\{V_2 > x\right\} \leq (1+o(1))\frac{(K+1)\rho}{1-\rho}\,\mathbb{P}\left\{B^r > x\right\}, \qquad \text{as } x \to \infty. \tag{7.6}$$

7.5 Workload asymptotics of the streaming traffic

In this section we turn the attention to the workload distribution of class 1. For convenience, we assume that each class-1 source generates traffic at a constant rate r. The latter assumption is however not essential for the asymptotic results to hold, and in Section 7.8 we extend the results to the case of variable-rate class-1 traffic. In the remainder of the chapter, we assume that $\rho + Kr < 1$ to ensure stability. In addition, we impose the condition that $(K+1)r > 1-\rho$, i.e., the system is critically loaded. Thus, $Kr < 1-\rho < (K+1)r$.

The next theorem provides the main result of the present chapter.

Theorem 7.5.1 *If $B(\cdot) \in \mathcal{R}_{-\nu}$ and $Kr < 1-\rho < (K+1)r$, then*

$$\mathbb{P}\left\{V_1 > x\right\} \sim \frac{\rho}{1-\rho-Kr}\mathbb{P}\left\{B^r > \frac{x\frac{1-\rho}{K+1}}{K(r-\frac{1-\rho}{K+1})}\right\}. \tag{7.7}$$

The proof of the above theorem involves asymptotic lower and upper bounds which will be provided in Sections 7.6 and 7.7, respectively. In this section, we sketch a heuristic derivation of the result, which will also serve as an outline for the construction of the lower bound in Subsection 7.6.1. The heuristic arguments are in essence similar to the arguments given in Subsection 1.6.4 for the classical M/G/1 queue. In addition, we give an intuitive interpretation, which provides the basis for the lower bound in Subsection 7.6.2 and the upper bound in Section 7.7. First, however, we give some basic relations between traffic processes, amounts of service and workloads, and state a few preliminary results.

Preliminary results

The amounts of service satisfy the following simple inequality

$$B_1(s,t) + B_2(s,t) \leq t-s. \tag{7.8}$$

For the workloads, the following obvious identity relation holds for $i = 1, 2$ and $s < t$,

$$V_i(t) = V_i(s) + A_i(s,t) - B_i(s,t). \tag{7.9}$$

As mentioned in Section 7.3, in the work-conserving scenario, i.e., $C_2(s,t) \equiv t-s-B_1(s,t)$, the system is equivalent in terms of the total workload to a single queue of unit rate fed by the aggregate class-1 and class-2 traffic processes,

$$V_1(t) + V_2(t) = \sup_{s \leq t}\{A_1(s,t) + A_2(s,t) - (t-s)\}. \tag{7.10}$$

In particular, in the constant-rate case,

$$\begin{aligned} V_1(t) + V_2(t) &= \sup_{s \leq t} \{Kr(t-s) + A_2(s,t) - (t-s)\} \\ &= \sup_{s \leq t} \{A_2(s,t) - (1-Kr)(t-s)\} \\ &= V_2^{1-Kr}(t), \end{aligned} \tag{7.11}$$

with $V_2^c(t)$ the workload at time t in an isolated queue with service rate c fed by class 2 only. For any $\rho < c$, let V_2^c be its steady-state version. The asymptotic tail distribution of the latter quantity is given by the next theorem (see Theorem 1.6.5 in case $c = 1$), which is originally due to Cohen [50], and has been extended to subexponential distributions by Pakes [135].

Theorem 7.5.2 *Assume that $\rho < c$. Then, $B(\cdot) \in \mathcal{R}_{-\nu}$ iff $\mathbb{P}\{V_2^c < \cdot\} \in \mathcal{R}_{1-\nu}$, and then*

$$\mathbb{P}\{V_2^c > x\} \sim \frac{\rho}{c-\rho}\mathbb{P}\{B^r > x\}.$$

The same relation holds when V_2^c represents the workload distribution at arrival epochs of class 2.

Relation (7.11) plays a central role in the proof of Theorem 7.5.1. Throughout we will consider several extensions of the basic model, allowing the system to be non-work-conserving (e.g., the *permanent-customer* scenario) and having variable-rate streaming traffic (with mean Kr). In those cases, (7.11) does not hold as a sample path identity, but (under some assumptions) V_1+V_2 and V_2^{1-Kr} are *asymptotically* equivalent in the following sense (similar reduced-load type of equivalences may be found in, e.g., [3, 91, 179]):

$$\mathbb{P}\{V_1 + V_2 > x\} \sim \mathbb{P}\left\{V_2^{1-Kr} > x\right\}. \tag{7.12}$$

The main intuitive idea is that a large total workload is most likely due to the arrival of a large class-2 user. Since the system is critically loaded, the class-1 workload builds up in the presence of the large class-2 user, so that the full service capacity is used and the system behaves as if it were work-conserving. The detailed proof of (7.12) is deferred to Appendix 7.A (Proposition 7.A.1).

Heuristic arguments

In queueing systems with heavy-tailed characteristics, rare events tend to occur as a consequence of a single most-probable cause. We will specifically show that in the present context the most likely way for a large class-1 workload V_1 to occur arises from the arrival of a class-2 user with a large service requirement B_{tag}, while the system shows average behavior otherwise. We will refer to the class-2 user as the “tagged” user.

Define $B_{\text{tag}}(s,t)$ as the amount of service received by the tagged user in $(s,t]$. In addition, denote by $B_2^-(s,t)$ the amount of service received by class-2 users in the time interval $(s,t]$, except for the tagged user. Then (7.8) may be rewritten as follows

$$B_1(s,t) + B_{\mathrm{tag}}(s,t) + B_2^-(s,t) \leq t - s. \tag{7.13}$$

Suppose that the tagged user arrives at time $-y - z_0$, with $z_0 = \frac{x}{K(r - \frac{1-\rho}{K+1})}$, $B_{\mathrm{tag}} \geq x + (1-\rho-Kr)(y+z_0)$, and $y \geq 0$. The amount of class-2 traffic generated during the time interval $[-y-z_0, 0]$ is close to average, i.e., $A_2(-y-z_0,0) \approx \rho(y+z_0)$. Since class 2 is stable, regardless of the presence of the tagged user, the amount of service received roughly equals the amount of class-2 traffic generated during the time interval $[-y-z_0,0]$, i.e., $B_2^-(-y-z_0,0) \approx \rho(y+z_0)$. The cumulative amount of service received by the tagged user up to time 0 is either $B_1(-y-z_0,0)/K$ or B_{tag}, depending on whether the user is still present at time 0 or not.

Using the inequality (7.13), the amount of service received by class 1 is approximately

$$\begin{aligned} B_1(-y-z_0,0) &\leq y + z_0 - B_{\mathrm{tag}}(-y-z_0,0) - B_2^-(-y-z_0,0) \\ &\approx (1-\rho)(y+z_0) - \min\{B_{\mathrm{tag}}, B_1(-y-z_0,0)/K\}. \end{aligned}$$

Thus,

$$\begin{aligned} B_1(-y-z_0,0) &\leq \max\{(1-\rho)(y+z_0) - B_{\mathrm{tag}}, \frac{K}{K+1}(1-\rho)(y+z_0)\} \\ &\leq \max\{Kr(y+z_0) - x, \frac{K}{K+1}(1-\rho)(y+z_0)\}. \end{aligned}$$

Using the above inequality and the identity relation (7.9), the class-1 workload at time 0 is

$$\begin{aligned} V_1(0) &\geq A_1(-y-z_0,0) - B_1(-y-z_0,0) \\ &\geq Kr(y+z_0) - \max\{Kr(y+z_0) - x, \frac{K}{K+1}(1-\rho)(y+z_0)\} \\ &= \min\{x, K(r - \frac{1-\rho}{K+1})(y+z_0)\} \geq \min\{x, K(r - \frac{1-\rho}{K+1})z_0\} = x. \end{aligned}$$

In the case of Poisson arrivals of class 2 we obtain (by integrating with respect to y and neglecting the probability of two or more "large" users)

$$\mathbb{P}\{V_1 > x\} \geq \int_{y=0}^{\infty} \lambda \mathbb{P}\left\{B_{\mathrm{tag}} > \frac{1-\rho}{K+1} z_0 + (1-\rho-Kr)y\right\} \mathrm{d}y,$$

which agrees with the right-hand side of (7.7).

Of course, there are alternative scenarios that could potentially lead to a large class-1 workload. Theorem 7.5.1 thus indirectly indicates that these are extremely unlikely compared to the one described above, as will be rigorously shown in Section 7.7.

A formal proof based on the above heuristics (in case of renewal arrivals of class 2) may be found in Subsection 7.6.1. The arrival of a class-2 user with a large service requirement in fact also results in a large total amount of work in the system after its arrival. We will use this alternative interpretation of the dominant scenario in Subsection 7.6.2 to derive a lower bound in case of renewal class-2 arrivals and in Section 7.7 to obtain an upper bound. In particular, we will show that the event $V_1(-t_1) + V_2(-t_1) \geq x + (1 - \rho - Kr)t_1$, with $t_1 := \frac{x}{K(r - \frac{1-\rho}{K+1})}$, corresponds to the dominant scenario described above. Using Proposition 7.A.1 and Theorem 7.5.2, we then obtain that the probability of the latter event coincides with the right-hand side of (7.7).

Finally, note that the dominant scenario crucially depends on the critical load, i.e., $1 - \rho < (K + 1)r$. Section 7.9 briefly discusses the case of a non-critically loaded system.

7.6 Lower bound

In this section we derive asymptotic lower bounds for $\mathbb{P}\{V_1 > x\}$ using two different approaches. In Subsection 7.6.1, we explicitly use the arrival of a class-2 user with a large service requirement (as described in the heuristics in Section 7.5) as the most likely way for a large class-1 workload to occur. We believe that this approach is especially insightful, as it brings out the typical cause of a large class-1 workload. In Subsection 7.6.2, we provide a proof based on the alternative characterization of the dominant scenario in Section 7.5. This approach is consistent with the derivation of the upper bound in Section 7.7. Moreover, it allows for modifications to include variable-rate class-2 traffic.

7.6.1 Approach 1

To obtain a lower bound for $\mathbb{P}\{V_1 > x\}$, we start by deriving a sufficient sample-path condition for the event $V_1(0) > x$ to occur (Lemma 7.6.1). Next, we convert the sample-path statement into a probabilistic lower bound which can be used to determine the asymptotic tail behavior of $\mathbb{P}\{V_1 > x\}$ (Proposition 7.6.1).

Consider the following three events.

1. $\exists y \geq 0$ such that at time $-t_0$, with $t_0 := \frac{x(1+K\epsilon+K\gamma)}{K(r - \frac{1-\rho+\delta}{K+1})} + y$, a tagged class-2 user arrives with service requirement

$$B_{\text{tag}} \geq \frac{x(1 + K\epsilon + K\gamma)}{K(r - \frac{1-\rho+\delta}{K+1})} \frac{1 - \rho + \delta}{K + 1} + y(1 - \rho + \delta - Kr) + (\epsilon + \gamma)x \quad (7.14)$$

2. For the amount of class-2 traffic arriving in the interval $(-t_0, 0]$ it holds that

$$A_2(-t_0, 0) \geq (\rho - \delta)t_0 - (K + 1)\gamma x \quad (7.15)$$

3. The amount of class-2 work at time 0, except from the tagged user, satisfies

$$V_2^-(0) \leq (K+1)\epsilon x \tag{7.16}$$

We first prove the next sample-path relation.

Lemma 7.6.1 *If the events (7.14)-(7.16) occur simultaneously with $\delta \leq (K+1)r - (1-\rho)$, then $V_1(0) > x$.*

Proof We distinguish between two cases: (i) the large tagged user is still present in the system at time 0; and (ii) the tagged user already left before time 0.

First consider case (i) and denote by $B_1^+(s,t)$ the amount of service received by the class-1 users and the large tagged class-2 user together in the interval $(s,t]$. Then, using (7.8) and (7.9),

$$\begin{aligned} B_1^+(-t_0,0) &\leq t_0 - V_2(-t_0) - A_2(-t_0,0) + V_2^-(0) \\ &\leq t_0 - A_2(-t_0,0) + V_2^-(0) \\ &\leq (1-\rho+\delta)t_0 + (K+1)(\epsilon+\gamma)x, \end{aligned} \tag{7.17}$$

where we used (7.15) and (7.16) in the third inequality. Because of the PS discipline, we have $B_1(-t_0,0) \leq \frac{K}{K+1}B_1^+(-t_0,0)$. Combining this with (7.17) and using (7.9) yields

$$\begin{aligned} V_1(0) &\geq A_1(-t_0,0) - B_1(-t_0,0) \\ &\geq Krt_0 - \frac{K}{K+1}\left[(1-\rho+\delta)t_0 + (K+1)(\epsilon+\gamma)x\right] \\ &= K\left(r - \frac{1-\rho+\delta}{K+1}\right)t_0 - K(\epsilon+\gamma)x \\ &\geq K\left(r - \frac{1-\rho+\delta}{K+1}\right)\frac{x(1+K\epsilon+K\gamma)}{K(r-\frac{1-\rho+\delta}{K+1})} - K(\epsilon+\gamma)x \\ &= x, \end{aligned}$$

where we used $\delta \leq (K+1)r - (1-\rho)$ in the fourth step.

Next, consider case (ii). From (7.8) and (7.9), we obtain

$$\begin{aligned} B_1(-t_0,0) &\leq t_0 - V_2(-t_0) - A_2(-t_0,0) + V_2^-(0) - B_{\text{tag}} \\ &\leq t_0 - A_2(-t_0,0) + V_2^-(0) - B_{\text{tag}} \\ &\leq (1-\rho+\delta)t_0 + (K+1)(\epsilon+\gamma)x - B_{\text{tag}}, \end{aligned}$$

where we used (7.15) and (7.16) in the final inequality. Applying similar arguments as in case (i) yields

$$\begin{aligned} V_1(0) &\geq A_1(-t_0,0) - B_1(-t_0,0) \\ &\geq (Kr-1+\rho-\delta)t_0 - (K+1)(\epsilon+\gamma)x + B_{\text{tag}} \\ &\geq (Kr-1+\rho-\delta)\left[\frac{x(1+K\epsilon+K\gamma)}{K(r-\frac{1-\rho+\delta}{K+1})} + y\right] - (K+1)(\epsilon+\gamma)x \\ &\quad + \frac{x(1+K\epsilon+K\gamma)}{K(r-\frac{1-\rho+\delta}{K+1})}\frac{1-\rho+\delta}{K+1} + y(1-\rho+\delta-Kr) + (\epsilon+\gamma)x \\ &= x, \end{aligned}$$

where we used (7.14) in the third inequality. This completes the proof. □

We now use the sample-path relation of Lemma 7.6.1 to prove the next asymptotic lower bound for the class-1 workload distribution.

Proposition 7.6.1 *(lower bound) Assume the class-2 arrivals follow a renewal process with mean interarrival time $\alpha = 1/\lambda = \beta/\rho$. If $B(\cdot) \in \mathcal{R}_{-\nu}$ and $Kr < 1-\rho < (K+1)r$, then*

$$\liminf_{x\to\infty} \frac{\mathbb{P}\{V_1 > x\}}{\frac{\rho}{1-\rho-Kr}\mathbb{P}\left\{B^r > \frac{x\frac{1-\rho}{K+1}}{K(r-\frac{1-\rho}{K+1})}\right\}} \geq 1.$$

Proof Let $-t_0 - \tau_{-m}$ be the arrival epoch of the $(m+1)$-th class-2 user before time $-t_0$ (counting backwards). In particular, τ_0 is the backward recurrence time of the class-2 arrival process at time $-t_0$. The corresponding service requirements are denoted by B_{-m}, $m \geq 0$.

In the following γ, δ, ϵ, κ and ζ are all small, but positive real numbers. Denote

$$g(\gamma,\delta,\epsilon,\kappa) := \frac{(1+K\epsilon+K\gamma)}{K(r-\frac{1-\rho+\delta}{K+1})}\frac{1-\rho+\delta}{K+1} + (\epsilon+\gamma) + (1-\rho+\delta-Kr)\kappa,$$

and rewrite (7.14) into $B_{-m} > g(\gamma,\delta,\epsilon,0)x + (1-\rho+\delta-Kr)\tau_{-m}$ for some $m \geq 0$. To bound the probability of (7.16), we apply the model with $K+1$ permanent customers, giving $V_2^-(0) \leq V_{(K+1)}(0)$. Now, using Lemma 7.6.1 yields

$$\begin{aligned} &\mathbb{P}\{V_1(0) > x\} \\ &\geq \mathbb{P}\{A_2(-t_0,0) \geq (\rho-\delta)t_0 - (K+1)\gamma x; V_2^-(0) \leq (K+1)\epsilon x; \\ &\qquad \exists m \geq 0 : B_{-m} > g(\gamma,\delta,\epsilon,0)x + (1-\rho+\delta-Kr)\tau_{-m}; \\ &\qquad \forall k \geq 0 : \tau_{-k} \leq k(\alpha+\zeta) + \kappa x\} \\ &\geq \mathbb{P}\{\exists m \geq 0 : B_{-m} > g(\gamma,\delta,\epsilon,\kappa)x + (1-\rho+\delta-Kr)m(\alpha+\zeta)\} \\ &\quad \times\mathbb{P}\{A_2(-t_0,0) \geq (\rho-\delta)t_0 - (K+1)\gamma x; V_{(K+1)}(0) \leq (K+1)\epsilon x; \\ &\qquad \forall k \geq 0 : \tau_{-k} \leq k(\alpha+\zeta) + \kappa x\}. \end{aligned} \tag{7.18}$$

We study each of the two probabilities separately. First note that

$$\begin{aligned}
&\mathbb{P}\{\exists m \geq 0 : B_{-m} > g(\gamma,\delta,\epsilon,\kappa)x + (1-\rho+\delta-Kr)m(\alpha+\zeta)\} \\
\geq\ & \sum_{m=0}^{\infty} \mathbb{P}\{B_{-m} > g(\gamma,\delta,\epsilon,\kappa)x + (1-\rho+\delta-Kr)m(\alpha+\zeta)\} \\
& - \sum_{m=0}^{\infty}\sum_{n=m+1}^{\infty} \mathbb{P}\{B_{-m} > g(\gamma,\delta,\epsilon,\kappa)x + (1-\rho+\delta-Kr)m(\alpha+\zeta), \\
& \qquad B_{-n} > g(\gamma,\delta,\epsilon,\kappa)x + (1-\rho+\delta-Kr)n(\alpha+\zeta)\} \\
\sim\ & (1+o(1))\frac{\beta/(\alpha+\zeta)}{1-\rho-Kr+\delta}\mathbb{P}\{B^r > g(\gamma,\delta,\epsilon,\kappa)x\}, \qquad (7.19)
\end{aligned}$$

where we used similar arguments as in [39] in the final step. As for the second probability in (7.18), observe that the τ_{-k}, $A_2(-t_0,0)$, and $V_{(K+1)}(0)$ are not independent. However, we may write

$$\begin{aligned}
&\mathbb{P}\left\{A_2(-t_0,0) \geq (\rho-\delta)t_0 - (K+1)\gamma x; V_{(K+1)}(0) \leq (K+1)\epsilon x;\right. \\
&\qquad \forall k \geq 0 : \tau_{-k} \leq k(\alpha+\zeta) + \kappa x\} \\
\geq\ & \mathbb{P}\left\{A_2(-t_0,0) \geq (\rho-\delta)t_0 - (K+1)\gamma x\right\} - \mathbb{P}\left\{V_{(K+1)}(0) > (K+1)\epsilon x\right\} \\
& -\mathbb{P}\left\{\exists k \geq 0 : \tau_{-k} > k(\alpha+\zeta) + \kappa x\right\}.
\end{aligned}$$

Now, $\mathbb{P}\left\{A_2(-t_0,0) \geq (\rho-\delta)t_0 - (K+1)\gamma x\right\} \to 1$ as $x \to \infty$ (and thus $t_0 \to \infty$). Moreover, since $V_{(K+1)}(0)$ has a proper distribution, we have

$$\lim_{x\to\infty} \mathbb{P}\left\{V_{(K+1)}(0) > (K+1)\epsilon x\right\} = 0,$$

and by the Strong Law of Large Numbers (the backward recurrence time at time $-t_0$ has a proper distribution because the renewal process has finite mean),

$$\lim_{x\to\infty} \mathbb{P}\left\{\exists k \geq 0 : \tau_{-k} > k(\alpha+\zeta) + \kappa x\right\} = 0.$$

Observing that the system is in steady state and using (7.19), we have

$$\liminf_{x\to\infty} \frac{\mathbb{P}\left\{V_1 > x\right\}}{\frac{\beta/(\alpha+\zeta)}{1-\rho-Kr+\delta}\mathbb{P}\left\{B^r > g(\gamma,\delta,\epsilon,\kappa)x\right\}} \geq 1.$$

Finally, use the fact that $B^r(\cdot) \in \mathcal{R}_{1-\nu}$ to obtain

$$\begin{aligned}
&\liminf_{x\to\infty} \frac{\mathbb{P}\left\{V_1 > x\right\}}{\frac{\rho}{1-\rho-Kr}\mathbb{P}\left\{B^r > \frac{x}{K(r-\frac{1-\rho}{K+1})}\frac{1-\rho}{K+1}\right\}} \qquad (7.20) \\
\geq\ & \liminf_{x\to\infty} \frac{\mathbb{P}\left\{V_1 > x\right\}}{\frac{\beta/(\alpha+\zeta)}{1-\rho-Kr+\delta}\mathbb{P}\left\{B^r > g(\gamma,\delta,\epsilon,\kappa)x\right\}} \frac{\frac{\beta/(\alpha+\zeta)}{1-\rho-Kr+\delta}\mathbb{P}\left\{B^r > g(\gamma,\delta,\epsilon,\kappa)x\right\}}{\frac{\rho}{1-\rho-Kr}\mathbb{P}\left\{B^r > g(0,0,0,0)x\right\}} \\
\geq\ & \liminf_{x\to\infty} \frac{\frac{\beta/(\alpha+\zeta)}{1-\rho-Kr+\delta}\mathbb{P}\left\{B^r > g(\gamma,\delta,\epsilon,\kappa)x\right\}}{\frac{\rho}{1-\rho-Kr}\mathbb{P}\left\{B^r > g(0,0,0,0)x\right\}} \uparrow 1, \qquad \gamma,\delta,\epsilon,\kappa,\zeta \downarrow 0.
\end{aligned}$$

□

7.6.2 Approach 2

As in Subsection 7.6.1, we start by deriving a sufficient sample-path condition for the event $V_1(0) > x$ to occur, but now based on the alternative characterization of the dominant scenario in Section 7.5 (Lemma 7.6.2). Then, we translate the sample-path statement into a probabilistic lower bound which can be used to determine the asymptotic tail behavior of $\mathbb{P}\{V_1 > x\}$ (Proposition 7.6.2).

We first introduce some additional notation and terminology. In the proof we frequently use the notion of "small" users. A user is called "small" if its (initial) service requirement does not exceed κx, for some $\kappa > 0$ independent of x. Denote by $N^{(u,v]}(t)$ the number of class-2 users in the system at time t that arrived during $(u, v]$, and add the subscript $\leq \kappa x$ when only "small" class-2 users are considered. Define $t_0 := \frac{x(1+\gamma+M_0\kappa)}{K(r-\frac{1-\rho+\delta}{K+1})}$, and fix $L_0 \geq \frac{1+K\rho}{1-\rho}$ and $M_0 \geq \max\{L_0, \frac{\rho(K+L_0)}{1-\rho}\}$. In the proof, users arriving before time $-t_0$ are referred to as "old" users, while users arriving after time $-t_0$ are called "new". Let $-u_0$, $u_0 := \sup\{0 \leq t \leq t_0 : N^{(-\infty,-t_0]}(-t) \leq L_0\}$, be the first epoch after time $-t_0$ that there are less than L_0 "old" class-2 users. Similarly, let $-s_0$, $s_0 := \inf\{0 \leq t \leq t_0 : N^{(-t_0,-t]}_{\leq\kappa x}(-t) < M_0\}$, be the last epoch before time 0 that there are less than M_0 "new small" class-2 users in the system.

Now, for fixed $\delta, \epsilon, \kappa, L_0, M_0 > 0$, consider the following two events.

1. At time $-t_0$, the total amount of work in the system satisfies
$$V_1(-t_0) + V_2(-t_0) \geq x(1 + \gamma + M_0\kappa) - (Kr + \rho - 1 - \delta)t_0 \qquad (7.21)$$

2. For the amount of "small" class-2 traffic arriving in $(-t_0, -s_0]$ it holds that
$$A_{2,\leq\kappa x}(-t_0, -s_0) \geq (\rho - \delta)(t_0 - s_0) - \gamma x \qquad (7.22)$$

We first prove the next sample-path relation.

Lemma 7.6.2 *If the above events (7.21) and (7.22) occur simultaneously, then* $V_1(0) > x$.

Proof We distinguish between two cases, depending on whether $u_0 \leq s_0$ or $u_0 > s_0$. First, we consider the 'easy' case $u_0 \leq s_0$ (or alternatively $-u_0 \geq -s_0$). Observe that during the entire interval $(-t_0, 0]$ there are at least L_0 class-2 users in the system (either "old" or "new"). Thus, $B_2(-t_0, 0) \geq \frac{L_0}{K} B_1(-t_0, 0)$, so that $B_1(-t_0, 0) \leq \frac{K}{K+L_0} t_0$. Using the above in addition to (7.9), we obtain

$$\begin{aligned} V_1(0) &\geq A_1(-t_0, 0) - B_1(-t_0, 0) \geq Krt_0 - \frac{K}{K + L_0} t_0 \\ &\geq K(r - \frac{1}{K + \frac{1+K\rho}{1-\rho}}) \frac{x(1 + \gamma + M_0\kappa)}{K(r - \frac{1-\rho+\delta}{K+1})} > x, \end{aligned}$$

where we used the definition of t_0 and the fact that $L_0 \geq \frac{1+K\rho}{1-\rho}$ in the third step.

Now consider the 'hard' case $u_0 > s_0$ (or $-u_0 < -s_0$). Denote by $B_2^{(u,v]}(s,t)$ the amount of service received during $(s,t]$ by class-2 users arriving in the interval $(u,v]$ (again, add the subscript $\leq \kappa x$ when only "small" class-2 users are considered). Using (7.9), the amount of service received during $(-t_0,-s_0]$ by the "new" class-2 users is bounded from below by

$$\begin{aligned} B_2^{(-t_0,0]}(-t_0,-s_0) &\geq B_{2,\leq\kappa x}^{(-t_0,-s_0]}(-t_0,-s_0) \\ &\geq A_{2,\leq\kappa x}(-t_0,-s_0) - V_{2,\leq\kappa x}^{(-t_0,-s_0]}(-s_0) \\ &\geq (\rho-\delta)(t_0-s_0) - \gamma x - M_0\kappa x, \end{aligned}$$

where $V_{2,\leq\kappa x}^{(u,v]}(t)$ denotes the workload at time t associated with "small" class-2 users arriving in $(u,v]$. Note that the final step follows from (7.22) and the definition of s_0. Since $M_0 \geq \frac{\rho(K+L_0)}{1-\rho}$, we also have

$$B_2^{(-t_0,0]}(-s_0,0) \geq \frac{M_0}{M_0+K+L_0}s_0 \geq (\rho-\delta)s_0.$$

Hence,

$$B_2^{(-t_0,0]}(-t_0,0) \geq (\rho-\delta)t_0 - \gamma x - M_0\kappa x. \tag{7.23}$$

Next, denote by $n \geq 0$ the number of "old" class-2 users present at time 0. We distinguish between two cases: (i) $n = 0$; and (ii) $n \geq 1$.

First, consider case (i). Note that $B_2^{(-\infty,-t_0]}(-t_0,0) = V_2(-t_0)$ and rewrite (7.8) into

$$B_1(-t_0,0) \leq t_0 - B_2^{(-\infty,-t_0]}(-t_0,0) - B_2^{(-t_0,0]}(-t_0,0). \tag{7.24}$$

Using (7.9), (7.21), (7.23), and (7.24), we deduce

$$\begin{aligned} V_1(0) &= V_1(-t_0) + A_1(-t_0,0) - B_1(-t_0,0) \\ &\geq V_1(-t_0) + V_2(-t_0) + Krt_0 - t_0 + (\rho-\delta)t_0 - (\gamma+M_0\kappa)x \\ &\geq x(1+\gamma+M_0\kappa) - (Kr+\rho-1-\delta)t_0 \\ &\qquad +Krt_0 - (1-\rho+\delta)t_0 - (\gamma+M_0\kappa)x \\ &= x. \end{aligned}$$

Second, consider case (ii). Because of the PS discipline, it follows from (7.8)

$$B_1(-t_0,0) \leq \frac{K}{K+1}[t_0 - B_2^{(-t_0,0]}(-t_0,0)]. \tag{7.25}$$

Now, combining (7.9), (7.23), and (7.25) yields

$$\begin{aligned} V_1(0) &\geq A_1(-t_0,0) - B_1(-t_0,0) \\ &\geq Krt_0 - \frac{K}{K+1}[(1-\rho+\delta)t_0 + (\gamma+M_0\kappa)x] \\ &= [Kr - \frac{K}{K+1}(1-\rho+\delta)]\frac{x(1+\gamma+M_0\kappa)}{K(r-\frac{1-\rho+\delta}{K+1})} - \frac{K}{K+1}(\gamma+M_0\kappa)x \\ &> x, \end{aligned}$$

where we used that $\gamma, \kappa, M_0 > 0$. This completes the proof. □

We now exploit the sample-path relation in Lemma 7.6.2 to establish the next asymptotic lower bound for the class-1 workload distribution.

Proposition 7.6.2 *(lower bound) If $B(\cdot) \in \mathcal{R}_{-\nu}$ and $Kr < 1-\rho < (K+1)r$, then*

$$\liminf_{x\to\infty} \frac{\mathbb{P}\{V_1 > x\}}{\frac{\rho}{1-\rho-Kr}\mathbb{P}\left\{B^r > \frac{x\frac{1-\rho}{K+1}}{K(r-\frac{1-\rho}{K+1})}\right\}} \geq 1.$$

Proof First observe that the events (7.21) and (7.22) are not independent. However, $V_1(-T_0) + V_2(-T_0)$ and $A_{2,\leq\kappa x}(-t_0, -s_0)$ are independent, with $-T_0$ representing the last arrival epoch of class 2 before time $-t_0$. Note that

$$V_1(-t_0) + V_2(-t_0) \geq V_1(-T_0) + V_2(-T_0) - \tau_0,$$

where τ_0 represents the backward recurrence time of the class-2 arrival process at time $-t_0$ (see also Subsection 7.6.1), which is independent of $V_1(-T_0)+V_2(-T_0)$ as well. Using Lemma 7.6.2 and the above, we obtain

$$\begin{aligned}
&\mathbb{P}\{V_1(0) > x\} \\
\geq\; &\mathbb{P}\{V_1(-T_0) + V_2(-T_0) > x(1+\gamma+M_0\kappa) - (Kr+\rho-1-\delta)t_0 + \tau_0; \\
&\qquad A_{2,\leq\kappa x}(-t_0,-s_0) \geq (\rho-\delta)(t_0-s_0) - \gamma x\} \\
\geq\; &\mathbb{P}\{V_1(-T_0) + V_2(-T_0) > x(1+\gamma+M_0\kappa+\epsilon) - (Kr+\rho-1-\delta)t_0\} \\
&\times\left[\mathbb{P}\left\{\sup_{0\leq t\leq t_0}\{(\rho-\delta)(t_0-t) - A_{2,\leq\kappa x}(-t_0,-t)\} \leq \gamma x\right\} - \mathbb{P}\{\tau_0 > \epsilon x\}\right].
\end{aligned}$$

Now, first invoking Proposition 7.A.1 in Appendix 7.A and then Theorem 7.5.2 yields

$$\begin{aligned}
&\mathbb{P}\{V_1(-T_0) + V_2(-T_0) > x(1+\gamma+M_0\kappa+\epsilon) - (Kr+\rho-1-\delta)t_0\} \\
&\quad\sim\; \frac{\rho}{1-\rho-Kr}\mathbb{P}\left\{B^r > \frac{x(1+\gamma+M_0\kappa)\frac{1-\rho+\delta}{K+1}}{K(r-\frac{1-\rho+\delta}{K+1})} + \epsilon x\right\}. \qquad (7.26)
\end{aligned}$$

Because τ_0 has a proper distribution, we have $\lim_{x\to\infty}\mathbb{P}\{\tau_0 > \epsilon x\} = 0$. Moreover, for $u > 0$ sufficiently large so that $\sup_{0\leq t}\{(\rho-\delta)t - A_{2,\leq u}(0,t)\}$ has a proper distribution, we have

$$\begin{aligned}
&\lim_{x\to\infty}\mathbb{P}\left\{\sup_{0\leq t\leq t_0}\{(\rho-\delta)(t_0-t) - A_{2,\leq\kappa x}(-t_0,-t)\} \leq \gamma x\right\} \\
\geq\; &\lim_{x\to\infty}\mathbb{P}\left\{\sup_{0\leq t\leq t_0}\{(\rho-\delta)(t_0-t) - A_{2,\leq u}(-t_0,-t)\} \leq \gamma x\right\} \\
\geq\; &\lim_{x\to\infty}\mathbb{P}\left\{\sup_{t\geq 0}\{(\rho-\delta)t - A_{2,\leq u}(0,t)\} \leq \gamma x\right\} = 1.
\end{aligned}$$

Combining the above arguments and applying (7.26), we obtain

$$\liminf_{x\to\infty} \frac{\mathbb{P}\{V_1 > x\}}{\frac{\rho}{1-Kr-\rho}\mathbb{P}\left\{B^r > \frac{x(1+\gamma+M_0\kappa)\frac{1-\rho+\delta}{K+1}}{K(r-\frac{1-\rho+\delta}{K+1})} + \epsilon x\right\}} \geq 1.$$

The proof may then be readily completed along the lines of (7.20). □

7.7 Upper bound

In this section we derive an asymptotic upper bound for $\mathbb{P}\{V_1 > x\}$. In the proof we frequently use the notion of a "large" user. A user is called "large" if its (initial) service requirement exceeds the value κx, for some fixed $\kappa > 0$ independent of x. Also, let $N_{>b}(s,t)$ be the number of class-2 users arriving during the time interval $(s,t]$ whose service requirement exceeds the value b. In particular, let $N(s,t) := N_{>0}(s,t)$ be the total number of class-2 users arriving in the interval $(s,t]$.

To handle scenarios in which the system is not work-conserving, we introduce the epoch $s^* := \inf\{t \geq 0 : V_1(-t) = 0\}$, which represents the last epoch before time 0 that the class-1 workload was zero. Note that $V_1(t) > 0$ for $t \in (-s^*, 0]$, and the system thus uses the full service capacity during the given interval. For epochs at which $V_1(t) = 0$, we make the following observation.

Observation 7.7.1 If $V_1(t) = 0$, then the available service rate for class 1 at time t is at least Kr, hence $\frac{K}{K+N(t)} \geq Kr$. Rewriting the inequality gives that $N(t) \leq M$, with $M := \lfloor \frac{1}{r} \rfloor - K$. ◇

We are now ready to prove the upper bound for $\mathbb{P}\{V_1 > x\}$.

Proposition 7.7.1 *(upper bound) If $B(\cdot) \in \mathcal{R}_{-\nu}$ and $Kr < 1-\rho < (K+1)r$, then*

$$\limsup_{x\to\infty} \frac{\mathbb{P}\{V_1 > x\}}{\frac{\rho}{1-\rho-Kr}\mathbb{P}\left\{B^r > \frac{x\frac{1-\rho}{K+1}}{K(r-\frac{1-\rho}{K+1})}\right\}} \leq 1.$$

Proof Let $t_1 := \frac{x(1-\epsilon)}{K(r-\frac{1-\rho-\delta}{K+1})}$. Then, for $\delta > 0, 0 < \epsilon < 1$,

$$\begin{aligned}
&\mathbb{P}\{V_1(0) > x\} \\
&\leq \mathbb{P}\{V_1(-t_1) + V_2(-t_1) > x(1-\epsilon) - (Kr+\rho+\delta-1)t_1\} \qquad (7.27)\\
&\quad + \mathbb{P}\{V_1(-t_1) + V_2(-t_1) \leq x(1-\epsilon) - (Kr+\rho+\delta-1)t_1; V_1(0) > x\}.
\end{aligned}$$

First, we determine the asymptotic behavior of the first probability on the rhs of (7.27). Then we show that the second probability on the rhs of (7.27) is negligible compared to the first one as $x \to \infty$. This way, we prove that the scenario described in Section 7.5 is indeed the dominant one.

Let us start with the former and note that the system at time $-t_1$ is in steady state. First, use Proposition 7.A.1 and then Theorem 7.5.2 to obtain that the first probability on the rhs of (7.27) behaves as

$$\begin{aligned}
&\mathbb{P}\{V_1(-t_1) + V_2(-t_1) > x(1-\epsilon) - (Kr+\rho+\delta-1)t_1\} \\
&\sim \quad \frac{\rho}{1-\rho-Kr}\mathbb{P}\left\{B^r > \frac{x(1-\epsilon)\frac{1-\rho-\delta}{K+1}}{K(r-\frac{1-\rho-\delta}{K+1})}\right\}.
\end{aligned}$$

Using the fact that $B^r(\cdot) \in \mathcal{R}_{1-\nu}$ (and letting $\delta, \epsilon \downarrow 0$), it easily follows that

$$\limsup_{x\to\infty} \frac{\mathbb{P}\left\{V_1(-t_1)+V_2(-t_1) > x(1-\epsilon)-(Kr+\rho+\delta-1)t_1\right\}}{\mathbb{P}\left\{B^r > \frac{x\frac{1-\rho}{K+1}}{K(r-\frac{1-\rho}{K+1})}\right\}} \leq 1.$$

To prove that any alternative scenario is highly unlikely compared to the dominant one, we show that, for $0<\delta<1-\rho-Kr$ and $0<\epsilon<1$,

$$\limsup_{x\to\infty} \frac{\mathbb{P}\left\{V_1(-t_1)+V_2(-t_1) \leq x(1-\epsilon)-(Kr+\rho+\delta-1)t_1; V_1(0)>x\right\}}{\mathbb{P}\left\{B^r > \frac{x\frac{1-\rho}{K+1}}{K(r-\frac{1-\rho}{K+1})}\right\}} = 0.$$

To do so, we split the second probability on the rhs of (7.27) by distinguishing between 0, 1, and 2 or more large user arrivals during $(-t_1,0]$, respectively. More specifically, write

$$\begin{aligned}
&\mathbb{P}\left\{V_1(-t_1)+V_2(-t_1) \leq x(1-\epsilon)-(Kr+\rho+\delta-1)t_1; V_1(0)>x\right\}\\
=\;& \mathbb{P}\{V_1(-t_1)+V_2(-t_1) \leq x(1-\epsilon)-(Kr+\rho+\delta-1)t_1;\\
&\qquad N_{>\kappa x}(-t_1,0)=0; V_1(0)>x\}\\
&+\mathbb{P}\{V_1(-t_1)+V_2(-t_1) \leq x(1-\epsilon)-(Kr+\rho+\delta-1)t_1;\\
&\qquad N_{>\kappa x}(-t_1,0)=1; V_1(0)>x\}\\
&+\mathbb{P}\{V_1(-t_1)+V_2(-t_1) \leq x(1-\epsilon)-(Kr+\rho+\delta-1)t_1;\\
&\qquad N_{>\kappa x}(-t_1,0)\geq 2; V_1(0)>x\}\\
=:\;& I+II+III. \qquad (7.28)
\end{aligned}$$

In the remainder of the proof we show that each of the three terms is negligible compared to the dominant scenario.

Term I

To bound term I, we consider the total workload at time 0. Recall that s^* represents the last epoch before time 0 that the class-1 workload was zero, and define $s' := \min\{s^*, t_1\}$, so that $V_1(t)>0$ for $t\in(-s',0]$. Then, using (7.9) and the fact that the system is work-conserving during $(-s',0]$, we have

$$\begin{aligned}
&V_1(0)+V_2(0)\\
=\;& V_1(-s')+V_2(-s')+Krs'+A_2(-s',0)-s'\\
=\;& V_1(-s')+V_2(-s')-(1-Kr-\rho-\delta)s'+A_2(-s',0)-(\rho+\delta)s'\\
\leq\;& \max\{V_1(-t_1)+V_2(-t_1)-(1-Kr-\rho-\delta)t_1, V_2(-s^*)\}\\
&+\sup_{0\leq s\leq t_1}\{A_2(-s,0)-(\rho+\delta)s\},
\end{aligned}$$

where we choose $0<\delta<1-Kr-\rho$. Moreover, take $\kappa>0$ such that $M\kappa<1$.

Then, combining the above and using Observation 7.7.1 yields

$$\begin{aligned} I &\leq \mathbb{P}\{\max\{V_1(-t_1)+V_2(-t_1)-(1-Kr-\rho-\delta)t_1, V_2(-s^*)\} \\ &\quad + \sup_{0\leq s\leq t_1}\{A_2(-s,0)-(\rho+\delta)s\} > x; \\ &\quad\ V_1(-t_1)+V_2(-t_1) < x(1-\epsilon)-(Kr+\rho+\delta-1)t_1; N_{>\kappa x}(-t_1,0)=0\} \\ &\leq \mathbb{P}\Bigg\{\max\{(1-\epsilon)x, M\kappa x\} + \sup_{0\leq s\leq t_1}\{A_2(-s,0)-(\rho+\delta)s\} > x \\ &\qquad\qquad \Bigg| \ N_{>\kappa x}(-t_1,0)=0\Bigg\} \\ &\leq \mathbb{P}\left\{\sup_{0\leq s\leq t_1}\{A_2(-s,0)-(\rho+\delta)s\} > \xi x \ \middle|\ N_{>\kappa x}(-t_1,0)=0\right\}, \end{aligned}$$

where $\xi := \min\{\epsilon, 1-M\kappa\}$. Lemma 7.B.4 in Appendix 7.B implies that $I = o(\mathbb{P}\{B^r > x\})$, as $x \to \infty$.

Term II
By conditioning on $V_1(-t_1)+V_2(-t_1)$, we obtain

$$\begin{aligned} II &= \mathbb{P}\{V_1(-t_1)+V_2(-t_1) < \eta x; N_{>\kappa x}(-t_1,0)=1; V_1(0) > x\} \qquad (7.29) \\ &\quad +\mathbb{P}\{\eta x < V_1(-t_1)+V_2(-t_1) < x(1-\epsilon)-(Kr+\rho+\delta-1)t_1; \\ &\qquad\qquad N_{>\kappa x}(-t_1,0)=1; V_1(0)>x\}. \end{aligned}$$

Again by Theorem 7.5.2 and Proposition 7.A.1, in addition to Lemma 7.B.3 with $t_1 = \gamma x$, we can control the second probability on the rhs of (7.29) as a "combination of two unlikely events". Specifically, the probability is bounded by

$$\mathbb{P}\{V_1(-t_1)+V_2(-t_1) > \eta x\}\,\mathbb{P}\left\{I(B>\kappa x) + \tilde{N}_{>\kappa x}(-t_1,0) \geq 1\right\},$$

which is bounded by $o(\mathbb{P}\{B^r > x\})$, as $x \to \infty$. Here $I(\cdot)$ is the indicator function, and $\tilde{N}_{>\kappa x}(-t_1,0)$ has the same distribution as $N_{>\kappa x}(-t_1,0)$, but is independent of $V_1(-t_1)+V_2(-t_1)$.

For the first probability on the rhs of (7.29), we use $s' = \min\{s^*, t_1\}$ (as in term I), so that $V_1(t) > 0$ for $t \in (-s', 0]$. Also, we tag the user with service requirement larger than κx, and let $V_2^-(t)$ be the class-2 workload at time t, excluding the tagged class-2 user. As in Section 7.5, denote by $B_2^-(s,t)$ the amount of service received by class 2 in the interval $(s,t]$, except for the tagged user. Then, using (7.9) in the first step and Observation 7.7.1 in the second, we find

$$B_2^-(-s',0) = V_2^-(-s') + A_2^-(-s',0) - V_2(0) \leq \zeta x + A_2^-(-s',0),$$

where $A_2^-(-s',0)$ denotes the amount of class-2 traffic generated during $(-s',0]$ excluding the tagged user, and $\zeta := \max\{\eta, M\kappa\}$. The large user together with the class-1 users receive the remaining amount of service: $B_1^+(-s',0) \geq$

$s' - A_2^-(-s',0) - \zeta x$. Because of the PS discipline, $B_1(-s',0) \geq \frac{K}{K+1} B_1^+(-s',0)$. Thus, using the above and applying (7.9),

$$\begin{aligned} V_1(0) &= V_1(-s') + A_1(-s',0) - B_1(-s',0) \\ &\leq \max\{V_1(-t_1), V_1(-s^*)\} + Krs' - \frac{K(s' - A_2^-(-s',0) - \zeta x)}{K+1} \\ &\leq \zeta x + \sup_{0 \leq s \leq t_1} \left\{ Krs - \frac{K(s - A_2^-(-s,0) - \zeta x)}{K+1} \right\}. \end{aligned}$$

Thus,

$$\begin{aligned} II &\leq \mathbb{P}\left\{ \zeta x + \sup_{0 \leq s \leq t_1} \left\{ Krs - \frac{K(s - A_2^-(-s,0) - \zeta x)}{K+1} \right\} > x \right. \\ &\qquad \left. \middle| \, N_{>\kappa x}(-t_1,0) = 1 \right\} + o(\mathbb{P}\{B^r > x\}), \end{aligned}$$

as $x \to \infty$. Choose η, κ such that $\max\{\eta, M\kappa\} \leq \frac{K+1}{K+3}\epsilon$. Then, using $r > \frac{1-\rho}{K+1}$ in the second inequality and substituting $x = \frac{t_1 K(r - \frac{1-\rho-\delta}{K+1})}{1-\epsilon}$ yields

$$\begin{aligned} &\mathbb{P}\left\{ \zeta x + \sup_{0 \leq s \leq t_1} \left\{ Krs - \frac{K(s - A_2^-(-s,0) - \zeta x)}{K+1} \right\} > x \,\middle|\, N_{>\kappa x}(-t_1,0) = 1 \right\} \\ &= \mathbb{P}\left\{ \sup_{0 \leq s \leq t_1} \left\{ Krs - \frac{K(s - A_2^-(-s,0))}{K+1} \right\} > x(1 - \frac{3K+1}{K+1}\zeta) + \frac{K}{K+1}\zeta x \right. \\ &\qquad \left. \middle| \, N_{>\kappa x}(-t_1,0) = 1 \right\} \\ &\leq \mathbb{P}\left\{ \sup_{0 \leq s \leq t_1} \left\{ Krs - \frac{K(s - A_2^-(-s,0))}{K+1} \right\} - t_1 K(r - \frac{1-\rho-\delta}{K+1}) > \frac{K}{K+1}\zeta x \right. \\ &\qquad \left. \middle| \, N_{>\kappa x}(-t_1,0) = 1 \right\} \\ &\leq \mathbb{P}\left\{ \sup_{0 \leq s \leq t_1} \left\{ A_2^-(-s,0) - (\rho+\delta)s \right\} > \zeta x \,\middle|\, N_{>\kappa x}(-t_1,0) = 1 \right\} \\ &\leq \mathbb{P}\left\{ \sup_{0 \leq s \leq t_1} \left\{ A_2(-s,0) - (\rho+\delta)s \right\} > \zeta x \,\middle|\, N_{>\kappa x}(-t_1,0) = 0 \right\}, \end{aligned}$$

which can be controlled using Lemma 7.B.4. This completes the estimation of term II.

Term III
It follows directly from Lemma 7.B.3 that $III = o(\mathbb{P}\{B^r > x\})$, as $x \to \infty$.

The proof is now completed by first letting $x \to \infty$, then $\eta, \kappa \downarrow 0$, and finally $\delta, \epsilon \downarrow 0$. □

7.8 Generalization to variable-rate streaming traffic

As mentioned earlier, the assumption that class 1 generates traffic at a constant rate Kr is actually not crucial. In this section, we show that our results remain valid in case class 1 generates traffic according to a general stationary process, provided that deviations from the mean are sufficiently unlikely. In such a scenario, the variations in class-1 traffic do not matter asymptotically, because they average out.

First, in Subsection 7.8.1 we consider the total workload of class 1 and extend Theorem 7.5.1 to the case of variable-rate streaming sources. Second, in Subsection 7.8.2 we consider the tail asymptotics of the joint workload distribution of individual class-1 users. Note that the individual class-1 workloads are not necessarily equal, since the traffic rates of the individual streaming sources also vary.

7.8.1 Total workload

In this subsection, we show that our results remain valid in case class 1 generates traffic according to a general stationary process with mean rate $\mathbb{E}[A_1(t,t+1)] = Kr$, provided that significant deviations from the mean are sufficiently unlikely. More specifically, we assume that the class-1 traffic satisfies the following assumption:

Assumption 7.8.1 *For all $\phi > 0$ and $\psi > 0$,*

$$\mathbb{P}\left\{\sup_{t\geq 0}\{A_1(-t,0) - K(r+\psi)t\} > \phi x\right\} = o(\mathbb{P}\{B^r > x\}), \qquad \text{as } x \to \infty.$$

Note that Assumption 7.8.1 holds for all $\phi > 0$ whenever it holds for one such value. Assumption 7.8.1 serves to ensure that the likelihood that rate variations in class-1 traffic cause a large workload is asymptotically negligible compared to scenarios with a large class-2 user described earlier. Also, observe that it may be equivalently expressed as

$$\mathbb{P}\left\{V_1^{K(r+\psi)} > \phi x\right\} = o(\mathbb{P}\{B^r > x\}), \qquad \text{as } x \to \infty, \tag{7.30}$$

where V_1^c denotes the steady-state workload in a system with service capacity c fed by class 1 only. Assumption 7.8.1 is satisfied by a wide range of traffic processes, as illustrated by the next two examples.

Example 7.8.1 (Instantaneous bursts) Let each class-1 user generate instantaneous bursts according to a renewal process, and let the burst sizes have distribution $F_1(\cdot)$, with mean σ_1. Let the interarrival times between bursts also be generally distributed with mean σ_1/r. Assume that $1-F_1(x) = o(\mathbb{P}\{B > x\})$ as $x \to \infty$. Then, it follows from [14, Theorem 4.1] that Assumption 7.8.1 is satisfied.

Example 7.8.2 (On-Off source) Let each class-1 user generate traffic according to an On-Off process, alternating between On- and Off-periods. The On-periods have general distribution $F_1(\cdot)$ with mean σ_1, and the Off-periods also follow a general distribution with mean $1/\lambda_1$. A class-1 user produces traffic at a constant rate r_{on} while On, and generates traffic at rate r_{off} while Off, $r_{\text{off}} < r < r_{\text{on}}$ (including the important special case in which $r_{\text{off}} = 0$), with $r(1+\lambda_1\sigma_1) = r_{\text{off}} + r_{\text{on}}\lambda_1\sigma_1$.

Moreover, assume that $1 - F_1(x) = o(\mathbb{P}\{B > x\})$ as $x \to \infty$. Now, asymptotic results for a fluid queue fed by multiple homogeneous On-Off sources (in particular [69], [179, Corollary 3.1] with $N^* = 1$), imply that Assumption 7.8.1 is satisfied.

In the remainder of the section, we show that our results remain valid under Assumption 7.8.1. In particular, we prove that Theorem 7.5.1 still holds. We add the superscript 'var' to indicate quantities corresponding to the scenario with variable-rate streaming sources.

Theorem 7.8.1 *Suppose that the process $\{A_1(-t,0), t \geq 0\}$ satisfies Assumption 7.8.1. If $B(\cdot) \in \mathcal{R}_{-\nu}$ and $Kr < 1-\rho < (K+1)r$, then*

$$\mathbb{P}\{V_1^{\text{var}} > x\} \sim \frac{\rho}{1-\rho-Kr}\mathbb{P}\left\{B^r > \frac{x\frac{1-\rho}{K+1}}{K(r-\frac{1-\rho}{K+1})}\right\}.$$

As before, the proof of Theorem 7.8.1 involves lower and upper bounds. In fact, the lower bound largely follows the lines of the proof of Proposition 7.6.2 (in Section 7.6), and is hardly affected by the variable rate of class 1. Informally speaking, the idea is to replace $A_1(s,t)$ by $K(r-\psi)(t-s) - \phi x$, and then use $\mathbb{E}[A_1(t,t+1)] = Kr$ to show that the correction terms $K\psi(t-s)$ and ϕx can be asymptotically neglected. More specifically, because the process $\{K(r-\psi)t - A_1(-t,0), t \geq 0\}$ has negative drift, for all $\phi, \psi > 0$,

$$\mathbb{P}\left\{\sup_{t\geq 0}\{K(r-\psi)t - A_1(-t,0)\} > \phi x\right\} \to 0, \qquad \text{as } x \to \infty. \tag{7.31}$$

Note that the above expression relates to long periods with less than average class-1 input, as opposed to Assumption 7.8.1 where periods with more than average class-1 traffic are considered.

Before describing the modifications of Subsection 7.6.2 required to handle variable-rate class-1 traffic, we note that a slightly more substantial modification is needed, to obtain an equivalence between $V_1^{\text{var}} + V_2^{\text{var}}$ and V_2^{1-Kr}. Moreover, in the lower bound we encounter the difficulty that $V_1^{\text{var}}(-t_0) + V_2^{\text{var}}(-t_0)$ and $A_1(-t_0,0)$ may no longer be independent. These issues are addressed in the proof of Proposition 7.D.1 in Appendix 7.D, where we extend relation (7.12) to the case of variable-rate class-1 traffic. Proposition 7.D.1 may also be of independent interest. In addition, we show in the proposition that relation (7.12) remains valid for a non-critically loaded *work-conserving* system.

For the upper bound, we provide a proof based on a comparison with a leaky-bucket system and use results of Section 7.5 (in particular Theorem 7.5.1).

We now give the proofs of the lower and upper bounds, together yielding Theorem 7.8.1.

Proposition 7.8.1 *(lower bound) If $B(\cdot) \in \mathcal{R}_{-\nu}$ and $Kr < 1-\rho < (K+1)r$, then*

$$\liminf_{x\to\infty} \frac{\mathbb{P}\left\{V_1^{\mathrm{var}} > x\right\}}{\frac{\rho}{1-\rho-Kr}\mathbb{P}\left\{B^r > \frac{x\frac{1-\rho}{K+1}}{K(r-\frac{1-\rho}{K+1})}\right\}} \geq 1.$$

Proof In view of the similarities with Subsection 7.6.2, we only give an outline of the proof. As in Subsection 7.6.2, consider the following three events:

- At time $-t_0$, with $t_0 := \frac{x(1+\gamma+M_0\kappa+\phi)}{K(r-\psi-\frac{1-\rho+\delta}{K+1})}$, the total amount of work in the system satisfies
$$V_1^{\mathrm{var}}(-t_0)+V_2^{\mathrm{var}}(-t_0) \geq x(1+\gamma+M_0\kappa+\phi)-(K(r-\psi)+\rho-1-\delta)t_0 \quad (7.32)$$
- The event (7.22), which we repeat for convenience,
$$A_{2,\leq\kappa x}(-t_0,-s_0) \geq (\rho-\delta)(t_0-s_0)-\gamma x$$
- For the amount of class-1 traffic arriving in the interval $(-t_0,0]$ it holds that
$$A_1(-t_0,0) \geq K(r-\psi)t_0-\phi x \quad (7.33)$$

Some calculations similar to the proof of Lemma 7.6.2 show that, if the events (7.32), (7.22), and (7.33) occur simultaneously, then $V_1^{\mathrm{var}}(0) > x$. As in Subsection 7.6.2, let $-T_0$ be the last class-2 arrival epoch before time $-t_0$. Denoting $\check{r} = r-\psi$ and $\check{\gamma} = \gamma+\phi$, we may write

$$\begin{aligned}
&\mathbb{P}\left\{V_1^{\mathrm{var}}(0) > x\right\}\\
\geq\;& \mathbb{P}\{V_1^{\mathrm{var}}(-T_0)+V_2^{\mathrm{var}}(-T_0) > x(1+\check{\gamma}+M_0\kappa+\epsilon)-(K\check{r}+\rho-1-\delta)T_0;\\
&\qquad A_1(-T_0,0) \geq K(r-\psi)T_0-\phi x;\\
&\qquad A_{2,\leq\kappa x}(-T_0,-s_0) \geq (\rho-\delta)(T_0-s_0)-\gamma x; \tau_0 \leq \epsilon x\}\\
\geq\;& \mathbb{P}\{V_1^{\mathrm{var}}(-T_0)+V_2^{\mathrm{var}}(-T_0) > x(1+\check{\gamma}+M_0\kappa+\epsilon)-(K\check{r}+\rho-1-\delta)T_0;\\
&\qquad \overrightarrow{U}_1^{K(r-\psi)}(-T_0) \leq \phi x\}\\
&\times\left[\mathbb{P}\left\{\sup_{0\leq t\leq T_0}\{(\rho-\delta)(T_0-t)-A_{2,\leq\kappa x}(-T_0,-t)\} \leq \gamma x\right\}-\mathbb{P}\left\{\tau_0 > \epsilon x\right\}\right],
\end{aligned}$$

where $\overrightarrow{U}_1^c(-T_0) := \sup_{0\leq t\leq T_0}\{c(T_0-t)-A_1(-T_0,-t)\}$. The second and third probabilities can be treated as in Subsection 7.6.2. For the first probability, apply (7.44) (see the proof of Proposition 7.D.1 in the Appendix) and then Theorem 7.5.2, to obtain

$$\liminf_{x\to\infty} \frac{\mathbb{P}\left\{V_1^{\mathrm{var}} > x\right\}}{\frac{\rho}{1-\rho-Kr}\mathbb{P}\left\{B^r > \frac{x(1+\gamma+M_0\kappa+\phi)\frac{1-\rho+\delta}{K+1}}{K(r-\psi-\frac{1-\rho+\delta}{K+1})} + \epsilon x\right\}} \geq 1.$$

Proposition 7.8.1 follows from the fact that $B^r(\cdot) \in \mathcal{R}_{1-\nu}$ (let $\gamma, \delta, \epsilon, \kappa, \phi, \psi \downarrow 0$). □

For the proof of the upper bound we compare the class-1 workload in the scenario with variable-rate streaming traffic to that in a scenario with constant-rate streaming traffic. Suppose we feed the variable-rate streaming traffic into a system (the leaky bucket) that drains at constant rate $K(r+\psi)$ into a second resource that is shared with the elastic class according to $C_2(t) = N_{(K)}(t)/(N_{(K)}(t)+K)$ (see Section 7.4). Because the drain rate of the first resource never exceeds $K(r+\psi)$, the second resource is closely related to the class-1 workload in the case of constant-rate traffic (in fact, the *permanent-customer* scenario provides an upper bound). The total class-1 workload at the first and second resources at time t is an upper bound for $V_1^{\mathrm{var}}(t)$ (see Equation (7.34) below). The proof is then established by using Theorem 7.5.1.

Proposition 7.8.2 *(upper bound) Suppose that the process* $\{A_1(-t,0), t \geq 0\}$ *satisfies Assumption 7.8.1. If* $B(\cdot) \in \mathcal{R}_{-\nu}$ *and* $Kr < 1-\rho < (K+1)r$*, then*

$$\limsup_{x\to\infty} \frac{\mathbb{P}\left\{V_1^{\mathrm{var}} > x\right\}}{\frac{\rho}{1-\rho-Kr}\mathbb{P}\left\{B^r > \frac{x\frac{1-\rho}{K+1}}{K(r-\frac{1-\rho}{K+1})}\right\}} \leq 1.$$

Proof Let $\psi > 0$. Using the definition of $V_1(t)$ in Section 7.3, we obtain the following representation

$$V_1^{\mathrm{var}}(t) = \sup_{s\leq t}\{A_1(s,t) - C_1(s,t)\} = \sup_{s\leq t}\{A_1(s,t) - \int_s^t \frac{K}{K+N^{\mathrm{var}}(u)}\mathrm{d}u\},$$

where the integral represents the amount of service available for class 1. Then,

$$\begin{aligned} V_1^{\mathrm{var}}(t) &= \sup_{s\leq t}\Bigg\{A_1(s,t) - K(r+\psi)(t-s) \\ &\qquad +K(r+\psi)(t-s) - \int_s^t \frac{K}{K+N^{\mathrm{var}}(u)}\mathrm{d}u\Bigg\} \\ &\leq \sup_{s\leq t}\{A_1(s,t) - K(r+\psi)(t-s)\} \\ &\qquad +\sup_{s\leq t}\Bigg\{K(r+\psi)(t-s) - \int_s^t \frac{K}{K+N^{\mathrm{var}}(u)}\mathrm{d}u\Bigg\}. \end{aligned}$$

Let $V_1^{\mathrm{cst},\psi}(t) = \sup_{s\leq t}\{K(r+\psi)(t-s) - \int_s^t \frac{K}{K+N_{(K)}(u)}\mathrm{d}u\}$ be the class-1 workload in a scenario with constant rate $r+\psi$ per streaming user and $C_2(t) \equiv$

$N_{(K)}(t)/(N_{(K)}(t)+K)$ (independent of the class-1 workload; this corresponds to the *permanent-customer* scenario discussed in Section 7.4). Similar to the constant-rate model, $N^{\mathrm{var}}(t) \leq N_{(K)}(t)$. Thus,

$$\int_s^t \frac{K}{K+N^{\mathrm{var}}(u)}\mathrm{d}u \geq \int_s^t \frac{K}{K+N_{(K)}(u)}\mathrm{d}u,$$

so that

$$V_1^{\mathrm{var}}(t) \leq V_1^{K(r+\psi)}(t) + V_1^{\mathrm{cst},\psi}(t). \tag{7.34}$$

For any $\xi > 0$, this sample-path relation implies

$$\mathbb{P}\left\{V_1^{\mathrm{var}} > x\right\} \leq \mathbb{P}\left\{V_1^{K(r+\psi)} > \xi x\right\} + \mathbb{P}\left\{V_1^{\mathrm{cst},\psi} > (1-\xi)x\right\},$$

where $V_1^{K(r+\psi)}$ and $V_1^{\mathrm{cst},\psi}$ have the limiting distributions of $V_1^{K(r+\psi)}(t)$ and $V_1^{\mathrm{cst},\psi}(t)$ for $t \to \infty$. The first term can be controlled by (7.30). For the second term, apply Theorem 7.5.1, use the fact that $B^r(\cdot) \in \mathcal{R}_{1-\nu}$, and let $\xi, \psi \downarrow 0$. This gives the desired result. □

7.8.2 Individual workloads

In this subsection we consider the asymptotics of the simultaneous workload distribution of the individual streaming users. In Subsection 7.8.1, we showed that a large service deficit for the K class-1 users together is most likely due to the arrival of a large class-2 user. Using similar arguments, we now also argue that the service deficits of the individual class-1 users are approximately equal after the arrival of a large class-2 user.

Denote by $A_{1,k}(s,t)$, $k = 1,\ldots,K$, the total traffic of streaming user k during the interval $[s,t]$ with mean rate $\mathbb{E}[A_{1,k}(t,t+1)] = r$. We make a similar assumption for the individual class-1 traffic processes as for the total traffic process in Subsection 7.8.1 (Assumption 7.8.1):

Assumption 7.8.2 *For all $\phi > 0$ and $\psi > 0$, $k = 1, \ldots K$,*

$$\mathbb{P}\left\{\sup_{t\geq 0}\{A_{1,k}(-t,0) - (r+\psi)t\} > \phi x\right\} = o(\mathbb{P}\left\{B^r > x\right\}), \qquad \text{as } x \to \infty.$$

Assumption 7.8.2 serves to ensure that the likelihood that rate variations in traffic of individual class-1 users cause a large workload is asymptotically negligible compared to scenarios with a large class-2 user as described earlier.

Similar to $V_1(t)$, define $V_{1,k}^{\mathrm{var}}(t) := \sup_{s\leq t}\{A_{1,k}(s,t) - C_{1,k}(s,t)\}$, where $C_{1,k}(s,t)$ denotes the total available service rate for streaming user k during the time interval $[s,t]$. Again, we added the superscript 'var' to indicate that the quantity corresponds to the scenario with variable-rate streaming sources. Note that $C_{1,k}(s,t) \geq \int_s^t 1/(K+N(u))\mathrm{d}u$ and also $\sum_{k=1}^K C_{1,k}(s,t) = C_1(s,t)$. The first relation holds with equality in case the streaming users always claim the full service rate available. However, we may allow for strict inequality in case several class-1 users do not always consume the service rate available and the unused surplus is redistributed among the other class-1 and class-2 users. Observe that the exact definition of $C_{1,k}(s,t)$ is not crucial in case $Kr < 1-\rho < (K+1)r$, because the workload of each class-1 user builds up in the presence of the large class-2 user, and each class-1 user will thus use its full service capacity.

Finally, denote the vectors $\boldsymbol{V_1}^{\text{var}} = (V_{1,1}^{\text{var}}, \cdots, V_{1,K}^{\text{var}})$, with $V_{1,k}^{\text{var}}$ the steady-state version of $V_{1,k}^{\text{var}}(t)$, and $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K)$. Moreover, let $\alpha^* := \max \alpha_k$. Then, we may derive a similar upper bound as in Propositions 7.7.1 and 7.8.2.

Proposition 7.8.3 *(upper bound) Suppose that the processes $\{A_{1,k}(-t,0), t \geq 0\}$, $k = 1, \ldots, K$, satisfy Assumption 7.8.2. If $B(\cdot) \in \mathcal{R}_{-\nu}$ and $Kr < 1-\rho < (K+1)r$, then*

$$\limsup_{x\to\infty} \frac{\mathbb{P}\{\boldsymbol{V_1}^{\text{var}} > \boldsymbol{\alpha} x\}}{\mathbb{P}\{V_1^{\text{var}} > K\alpha^* x\}} = \limsup_{x\to\infty} \frac{\mathbb{P}\{\boldsymbol{V_1}^{\text{var}} > \boldsymbol{\alpha} x\}}{\frac{\rho}{1-\rho-Kr}\mathbb{P}\left\{B^r > \frac{K\alpha^* x \frac{1-\rho}{K+1}}{K(r-\frac{1-\rho}{K+1})}\right\}} \leq 1.$$

Proof Let $k^* := \arg\max \alpha_k$, and note that

$$\mathbb{P}\{\boldsymbol{V_1}^{\text{var}} > \boldsymbol{\alpha} x\} \leq \mathbb{P}\left\{V_{1,k^*}^{\text{var}} > \alpha^* x\right\}.$$

Using a similar construction for streaming user k^* as in the proof of Proposition 7.8.2 (i.e., the leaky bucket), we obtain the following sample path relation

$$\begin{aligned} V_{1,k^*}^{\text{var}}(t) &\leq \sup_{s\leq t}\left\{A_{1,k^*}(s,t) - \int_s^t \frac{1}{K+N^{\text{var}}(u)}\text{d}u\right\} \\ &\leq \sup_{s\leq t}\{A_{1,k^*}(s,t) - (r+\psi)(t-s)\} \\ &\qquad + \sup_{s\leq t}\left\{(r+\psi)(t-s) - \int_s^t \frac{1}{K+N^{\text{var}}(u)}\text{d}u\right\} \\ &\leq \sup_{s\leq t}\{A_{1,k^*}(s,t) - (r+\psi)(t-s)\} + V_1^{\text{cst},\psi}/K, \end{aligned}$$

where we used the *permanent-customer* scenario as an upper bound in the final step. Combining the arguments above, the proof may be finished along similar lines as the proof of Proposition 7.8.2. □

For the lower bound, modifications to one of the proofs in Section 7.6 would imply that we have to keep track of all individual workloads and received amounts of services. In view of the exceedingly large amount of details and notational complexity, we present the next result as a conjecture:

Conjecture 7.8.1 *Suppose that the processes $\{A_{1,k}(-t,0), t \geq 0\}$, with $k = 1, \ldots, K$, satisfy Assumption 7.8.2. If $B(\cdot) \in \mathcal{R}_{-\nu}$ and $Kr < 1-\rho < (K+1)r$, then*

$$\begin{aligned} \mathbb{P}\{\boldsymbol{V_1}^{\text{var}} > \boldsymbol{\alpha} x\} &\sim \mathbb{P}\{V_1^{\text{var}} > K\alpha^* x\} \\ &\sim \frac{\rho}{1-\rho-Kr}\mathbb{P}\left\{B^r > \frac{x\frac{1-\rho}{K+1}}{K(r-\frac{1-\rho}{K+1})}\right\}. \end{aligned}$$

Conjecture 7.8.1 implies that the asymptotic tail probability of the K-dimensional random vector $\boldsymbol{V}_{\boldsymbol{1}}^{\text{var}}$ can be reduced to the tail probability of a 1-dimensional random variable B^r. In other words, we conclude that the workloads of the K individual class-1 users can only simultaneously grow large, requiring the presence of a large class-2 user.

7.9 Concluding remarks

We considered a bottleneck link shared by heavy-tailed TCP-controlled elastic flows and streaming sessions regulated by a TCP-friendly rate control protocol. We determined the asymptotic tail distribution of a possible shortfall in service received by the streaming users compared to a nominal service target. We showed that the distribution inherits the heavy-tailed behavior of the residual service requirement of an elastic flow.

In the case that the elastic flows arrive according to a Poisson process, we further derived bounds for performance measures for both classes of traffic by exploiting a relationship with the M/G/1 PS queue with permanent customers. In particular, we obtained bounds for the probability that the rate of the streaming applications falls below a given target rate, as well as for the delay and workload distributions of the elastic flows.

Besides the bounds provided by the M/G/1 PS queue with permanent customers, we also determined the exact delay asymptotics of the elastic flows, suggesting a certain dichotomy in the tail asymptotics, depending on whether the system is critically loaded or not.

The service deficit distribution of the streaming users was derived for critical load, i.e., an additional 'persistent' elastic flow would cause instability of the streaming class. In general, the most likely scenario for the class-1 workload to grow large involves the simultaneous presence of $l \geq 1$ large class-2 users, where $l := \min\left\{a \in \mathbb{N} : \frac{1-\rho}{K+a} < r\right\}$ is the number of 'persistent' elastic flows required to cause instability of the streaming class (class 1). This gives rise to the following conjecture:

Conjecture 7.9.1 *If $B(\cdot) \in \mathcal{R}_{-\nu}$ and $\rho + Kr < 1$, then*

$$\mathbb{P}\{V_1 > x\} = O(\mathbb{P}\{B^r > x\}^l), \qquad \text{as } x \to \infty.$$

Guillemin *et al.* [82] obtained similar asymptotics for the distribution of the available amount of service during an interval of length x in PS queues. However, obtaining exact asymptotics is a difficult task in this case as witnessed by [179].

Several other interesting issues remain for further research, e.g., transient performance measures, scenarios with finite buffers and/or dynamic populations of streaming sessions, and the performance impact of oscillations, inaccuracies, and delays in the estimation of the fair bandwidth share.

Appendix

7.A Proof of (7.12) for constant-rate streaming traffic

As mentioned previously, the asymptotic relation (7.12) plays a key role in our proofs, and is valid for several model extensions. To keep the presentation transparent, we first prove this relation in the next proposition for the case of constant-rate streaming traffic (assuming critical load). Appendix 7.D extends this result to variable-rate streaming traffic (as well as work-conserving, but possibly non-critically loaded, scenarios).

Proposition 7.A.1 *If $B(\cdot) \in \mathcal{R}_{-\nu}$ and $Kr < 1 - \rho < (K+1)r$, then,*

$$\mathbb{P}\left\{V_1 + V_2 > x\right\} \sim \mathbb{P}\left\{V_2^{1-Kr} > x\right\}.$$

This asymptotic relation also holds when $V_1 + V_2$ and V_2^{1-Kr} represent the workloads embedded at class-2 arrival epochs rather than at arbitrary instants.

Proof First observe that

$$\begin{aligned} \mathbb{P}\left\{V_1(0) + V_2(0) > x\right\} &\geq \mathbb{P}\left\{\sup_{t \geq 0}\{A_1(-t,0) + A_2(-t,0) - t\} > x\right\} \\ &= \mathbb{P}\left\{V_2^{1-Kr} > x\right\}. \end{aligned}$$

It remains to be shown that

$$\limsup_{x\to\infty} \frac{\mathbb{P}\left\{V_1 + V_2 > x\right\}}{\mathbb{P}\left\{V_2^{1-Kr} > x\right\}} \leq 1. \tag{7.35}$$

As defined in Section 7.7, $s^* := \inf\{t > 0 : V_1(-t) = 0\}$ is the last epoch before time 0 that the class-1 workload was zero. Hence, $V_1(t) > 0$ for $t \in (-s^*, 0]$, implying that the system operates at the full service rate during that interval. Now, as described in Section 7.5, the idea of the proof is that a large total workload is most likely caused by the arrival of a large class-2 user. In particular, the class-1 workload starts to build in the presence of a persistent class-2 user, and it may be shown that time s^* is close to the arrival epoch of the large user.

More formally, we split the class-2 workload at time t into workloads contributed by users with initial service requirements smaller than (or equal to) ϵx ($V_{2,\leq \epsilon x}(t)$), and those with initial service requirements larger than ϵx ($V_{2,>\epsilon x}(t)$). Moreover, let $V^c_{2,\leq \epsilon x}(t)$, $V^c_{2,>\epsilon x}(t)$ be the workloads in an isolated queue fed by class-2 traffic of users with service requirements smaller than and larger than ϵx, respectively. Then, use (7.9), apply Observation 7.7.1 to bound $V_{2,\leq \epsilon x}(-s^*)$

and Lemma 7.B.1 (stated below) to bound $V_{2,>\epsilon x}(-s^*)$:

$$\begin{aligned} V_1(0)+V_2(0) &= V_1(-s^*)+V_{2,\leq\epsilon x}(-s^*)+V_{2,>\epsilon x}(-s^*) \\ &\quad +A_1(-s^*,0)+A_{2,\leq\epsilon x}(-s^*,0)+A_{2,>\epsilon x}(-s^*,0)-s^* \\ &\leq 0+M\epsilon x+A_{2,\leq\epsilon x}(-s^*,0)-(\rho+\delta)s^* \\ &\quad +V_{2,>\epsilon x}^{1-Kr-\rho-\delta}(-s^*)+A_{2,>\epsilon x}(-s^*,0)-(1-Kr-\rho-\delta)s^* \\ &\leq M\epsilon x+V_{2,\leq\epsilon x}^{\rho+\delta}(0)+V_{2,>\epsilon x}^{1-Kr-\rho-\delta}(0). \end{aligned}$$

Converting this sample-path relation into a probabilistic upper bound gives (take $\epsilon < 1/M$)

$$\begin{aligned} \mathbb{P}\{V_1+V_2>x\} &\leq \mathbb{P}\left\{V_{2,\leq\epsilon x}^{\rho+\delta}(0)+V_{2,>\epsilon x}^{1-Kr-\rho-\delta}(0)>(1-M\epsilon)x\right\} \\ &\leq \mathbb{P}\left\{V_{2,\leq\epsilon x}^{\rho+\delta}(0)>\xi(1-M\epsilon)x\right\} \\ &\quad +\mathbb{P}\left\{V_{2,>\epsilon x}^{1-Kr-\rho-\delta}(0)>(1-\xi)(1-M\epsilon)x\right\}. \end{aligned}$$

The first term can be made sufficiently small for any fixed δ, ϵ, ξ, using similar arguments as in [36]. For the second term, we first apply Lemma 7.B.2 (given below) and Theorem 7.5.2, and then use the fact that $B^r(\cdot)\in\mathcal{R}_{1-\nu}$, and let δ, ξ, $\epsilon\downarrow 0$.

Note that the above proof applies regardless of whether 0 is an arbitrary instant or a class-2 arrival epoch. □

7.B Technical lemmas

Lemma 7.B.1 *For $1-\rho<(K+1)r$, $\epsilon>0$, and $\delta>0$,*

$$V_{2,>\epsilon x}(-s^*)\leq V_{2,>\epsilon x}^{r}(-s^*)\leq V_{2,>\epsilon x}^{1-Kr-\rho-\delta}(-s^*).$$

Proof Denote by $u^*:=\inf\{u\geq s^*:V_{2,>\epsilon x}(-u)=0\}$ the last epoch before time $-s^*$ that no large class-2 user was present. Hence, $N_{>\epsilon x}(t)\geq 1$ for $t\in(-u^*,-s^*]$. Observe that the amount of service received by the large users during $(-u^*,-s^*]$ then satisfies

$$B_{2,>\epsilon x}(-u^*,-s^*)\geq\int_{-u^*}^{-s^*}N_{>\epsilon x}(t)c_1(t)\mathrm{d}t\geq\int_{-u^*}^{-s^*}c_1(t)\mathrm{d}t\geq r(u^*-s^*),$$

where $c_1(t)$ is the service rate of an individual streaming user at time t. Here, the final step follows from the fact that $V_1(-s^*)=0$ and the service received during $(-u^*,-s^*]$ exceeds the amount of traffic generated. Using the above in the second step and (7.9) in the first and final one, gives

$$\begin{aligned} V_{2,>\epsilon x}(-s^*) &= V_{2,>\epsilon x}(-u^*)+A_{2,>\epsilon x}(-u^*,-s^*)-B_{2,>\epsilon x}(-u^*,-s^*) \\ &\leq A_{2,>\epsilon x}(-u^*,-s^*)-r(u^*-s^*) \\ &\leq V_{2,>\epsilon x}^{r}(-u^*)+A_{2,>\epsilon x}(-u^*,-s^*)-r(u^*-s^*) \\ &\leq V_{2,>\epsilon x}^{r}(-s^*). \end{aligned}$$

Finally, $V^r_{2,>\epsilon x}(-s^*) \leq V^{1-Kr-\rho-\delta}_{2,>\epsilon x}(-s^*)$ follows from $\delta \geq 0$ and $1-\rho < (K+1)r$. □

Lemma 7.B.2 *For all $c, \epsilon > 0$,*

$$\mathbb{P}\left\{V^c_{2,>\epsilon x} > x\right\} \leq (1+o(1))\frac{\rho}{c}\mathbb{P}\left\{B^r > x\right\} \sim \mathbb{P}\left\{V^{c+\rho}_2 > x\right\}, \qquad \text{as } x \to \infty.$$

Proof Fix L, $0 < L < \infty$, and consider an isolated system of capacity c, where only class-2 users with service requirements larger than L are admitted. The system load then equals $\rho_L := \lambda\mathbb{P}\{B > L\}\mathbb{E}[B|B > L]$. Moreover, using Theorem 7.5.2 (take L large enough, such that $\rho_L < c$), yields

$$\mathbb{P}\left\{V^c_{2,>L} > x\right\} \sim \frac{\rho_L}{c-\rho_L}\mathbb{P}\left\{B^r_{>L} > x\right\}.$$

For $x > L$, the probability on the right-hand side may be rewritten as follows

$$\begin{aligned}\mathbb{P}\left\{B^r_{>L} > x\right\} &= \frac{1}{\mathbb{E}[B|B > L]}\int_x^\infty \mathbb{P}\left\{B > y|B > L\right\}\mathrm{d}y \\ &= \frac{1}{\mathbb{E}[B|B > L]}\int_x^\infty \frac{\mathbb{P}\left\{B > y\right\}}{\mathbb{P}\left\{B > L\right\}}\mathrm{d}y \\ &= \frac{\mathbb{P}\left\{B^r > x\right\}\mathbb{E}B}{\mathbb{P}\left\{B > L\right\}\mathbb{E}[B|B > L]} = \frac{\rho}{\rho_L}\mathbb{P}\left\{B^r > x\right\}.\end{aligned}$$

Combining the above gives

$$\mathbb{P}\left\{V^c_{2,>L} > x\right\} \sim \frac{\rho}{c-\rho_L}\mathbb{P}\left\{B^r > x\right\}. \tag{7.36}$$

Now, observe that for $x \geq L/\epsilon$, we have $V^c_{2,>\epsilon x}(t) \leq V^c_{2,>L}(t)$, so that the first part of the result may be obtained from (7.36), letting $L \to \infty$, and observing that $\rho_L \to 0$ as $L \to \infty$. The second part follows from Theorem 7.5.2. □

Lemma 7.B.3 *For all $k \in \mathbb{N}$, $\kappa > 0$ (fixed), and $\gamma > 0$,*

$$\mathbb{P}\left\{N_{>\kappa x}(-\gamma x, 0) \geq k\right\} = O(\mathbb{P}\left\{B^r > x\right\}^k), \qquad \text{as } x \to \infty.$$

Proof Consider the time interval $(-t, 0)$ and denote by $T_{>\kappa x}(n)$ the interarrival time between the $(n-1)$-th and n-th user arrival after time $-t$ with service requirement larger than κx (with the natural amendment that the 0-th arrival represents the last arrival before time $-t$ with service requirement larger than κx). Also, let $T^r_{>\kappa x}(n)$ denote its residual interarrival time and let τ be an arbitrary class-2 arrival epoch. We first prove the lemma for $k = 1$. Note that

$$\begin{aligned}\mathbb{P}\left\{N_{>\kappa x}(-t, 0) \geq 1\right\} &\leq \mathbb{E}[N_{>\kappa x}(-t, 0)] \\ &= \mathbb{E}[N(-t, 0)]\mathbb{P}\left\{B > \kappa x\right\} = \lambda t\mathbb{P}\left\{B > \kappa x\right\}.\end{aligned}$$

In the final step we use that $-t$ is an arbitrary time instant, so that $N(-t, 0)$ is a stationary renewal process [10]. The statement of the lemma now follows for $k = 1$ by taking $t = \gamma x$ and using the fact that $B(\cdot)$ is regularly varying.

To extend this result to $k \geq 2$, note that, for all n,

$$\begin{aligned}\mathbb{P}\{T_{>\kappa x}(n) \leq t\} &= \mathbb{P}\{N_{>\kappa x}(\tau, \tau + t) \geq 1\} \\ &\leq \mathbb{E}[N_{>\kappa x}(\tau, \tau + t)] = \mathbb{E}[N(\tau, \tau + t)]\mathbb{P}\{B > \kappa x\}.\end{aligned}$$

By the Elementary Renewal Theorem [10], $\frac{1}{t}\mathbb{E}[N(\tau, \tau + t)] \to \lambda$ as $t \to \infty$, so that for any $\delta > 0$ there exists a $\bar{t}$ such that $\mathbb{E}[N(\tau, \tau + t)] \leq (\lambda + \delta)t$ for all $t \geq \bar{t}$.

Note that the following two events are equivalent for $k \geq 1$ (where we define the empty sum equal to 0 in case $k = 1$).

$$\{N_{>\kappa x}(-t, 0) \geq k\} = \{T^r_{>\kappa x}(1) + \sum_{n=2}^{k} T_{>\kappa x}(n) \leq t\}.$$

Thus, for $k \geq 2$ and $t \geq \bar{t}$,

$$\begin{aligned}\mathbb{P}\{N_{>\kappa x}(-t, 0) \geq k\} &\leq \mathbb{P}\left\{T^r_{>\kappa x}(1) + \sum_{n=2}^{k-1} T_{>\kappa x}(n) \leq t\right\}\mathbb{P}\{T_{>\kappa x}(k) \leq t\} \\ &\leq \mathbb{P}\{N_{>\kappa x}(-t, 0) \geq k - 1\}(\lambda + \delta)t\mathbb{P}\{B > \kappa x\}.\end{aligned}$$

By induction on k we obtain, for $k \geq 2$ and $t \geq \bar{t}$,

$$\mathbb{P}\{N_{>\kappa x}(-t, 0) \geq k\} \leq ((\lambda + \delta)t\mathbb{P}\{B > \kappa x\})^k.$$

Again, by taking $t = \gamma x$ (for large enough x) and using the fact that $B(\cdot)$ is regularly varying, the lemma follows. □

In case the class-2 users arrive according to a Poisson process, Lemma 7.B.3 can also be shown more directly. The crucial observation is that the number of class-2 arrivals with a service requirement larger than κx also follows a Poisson process, however with parameter $\lambda\mathbb{P}\{B > \kappa x\}$. Using the Poisson distribution function and taking the sum of a geometric series then completes the proof.

Lemma 7.B.4 *There exists a $\kappa^* > 0$ such that for all $\kappa \in (0, \kappa^*]$,*

$$\mathbb{P}\left\{\sup_{0 \leq s \leq \gamma x}\{A_2(-s, 0) - (\rho + \delta)s\} > \epsilon x \mid N_{>\kappa x}(-\gamma x, 0) = 0\right\} = o(\mathbb{P}\{B^r > x\}).$$

Proof Denote the interarrival time between the $(n-1)$-th and n-th user by T_n, and the service requirement of the n-th user by B_n. Let $S_n := X_1 + \ldots + X_n$ be a random walk with step sizes $X_m := B_m - (\rho + \delta)T_m$, with $\delta > 0$. Since $\rho = \mathbb{E}B_m/\mathbb{E}T_m$, we have $\mathbb{E}X_m < 0$, i.e., the random walk has negative drift. Observe that by the saw-tooth nature of the process $\{A_2(-s, 0) - (\rho + \delta)s\}$, the process attains a local maximum at epochs right after a jump, thus,

$$\sup_{0 \leq s \leq \gamma x}\{A_2(-s, 0) - (\rho + \delta)s\} \leq B_1 + \sup_{1 \leq n \leq N(-\gamma x, 0)} S_n.$$

Then, conditioning on the total number of class-2 arrivals in $(-\gamma x, 0)$ yields

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{0\le s\le\gamma x}\{A_2(-s,0)-(\rho+\delta)s\}>\epsilon x \,\middle|\, N_{>\kappa x}(-\gamma x,0)=0\right\}\\
&= \sum_{n=0}^{\infty}\mathbb{P}\Bigg\{\sup_{0\le s\le\gamma x}\{A_2(-s,0)-(\rho+\delta)s\}>\epsilon x\\
&\qquad\qquad \Bigg| N_{>\kappa x}(-\gamma x,0)=0; N(-\gamma x,0)=n\Bigg\}\times\mathbb{P}\left\{N(-\gamma x,0)=n\right\}\\
&\le \sum_{n=0}^{\bar{M}x}\mathbb{P}\left\{B_1+\sup_{0\le m\le n}\left\{\sum_{i=1}^{m}X_i\right\}>\epsilon x \,\middle|\, X_i<\kappa x, i=1,\ldots,n\right\}\\
&\quad \times\mathbb{P}\left\{N(-\gamma x,0)=n\right\}+\sum_{n=\bar{M}x+1}^{\infty}\mathbb{P}\left\{N(-\gamma x,0)=n\right\}\\
&\le \max_{0\le n\le\bar{M}x}\mathbb{P}\left\{\sup_{0\le m\le n}S_m>(\epsilon-\kappa)x \,\middle|\, X_i<\kappa x, i=1,\ldots,n\right\}\\
&\qquad\qquad +\mathbb{P}\left\{N(-\gamma x,0)>\bar{M}x\right\}\\
&\le \mathbb{P}\left\{\sup_{0\le m\le\bar{M}x}S_m>(\epsilon-\kappa)x \,\middle|\, X_i<\kappa x, i=1,\ldots,n\right\}\\
&\qquad\qquad +\mathbb{P}\left\{N(-\gamma x,0)>\bar{M}x\right\}, \qquad (7.37)
\end{aligned}$$

where the third inequality follows from the fact that $B_1 \le \epsilon x$. The second term of (7.37) decays exponentially fast in x when $\bar{M} > \lambda\gamma$. The first term may be rewritten as follows:

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{0\le m\le\bar{M}x}S_m>(\epsilon-\kappa)x \mid X_i<\kappa x, i=1,\ldots,n\right\}\\
&\qquad\le \sum_{m=0}^{\bar{M}x}\mathbb{P}\left\{S_m>(\epsilon-\kappa)x \mid X_i<\kappa x, i=1,\ldots,n\right\}.
\end{aligned}$$

This can be made sufficiently small by employing a powerful lemma of Resnick & Samorodnitsky [143]. According to this lemma, there exists a $\kappa^* > 0$ and a function $\phi(\cdot) \in \mathcal{R}_{-\alpha}$, with $\alpha > \nu$, such that for all $\kappa \in (0, \kappa^*]$ the first term of (7.37) can be bounded by $\bar{M}x\phi(x)$. Take $\phi(x) = x^{-1-\zeta}\mathbb{P}\{B^r > x\}$, with $\zeta = \alpha - \nu$, and note that $\bar{M}x\phi(x) = o(\mathbb{P}\{B^r > x\})$ to complete the proof. □

7.C Proof of Proposition 7.4.1

Proposition 7.4.1 *If* $B(\cdot) \in \mathcal{R}_{-\nu}$ *and either* $(K+1)r > 1-\rho$ *or* $C_2(t) \equiv \frac{N(t)}{K+N(t)}$ *or both, then*

$$\mathbb{P}\{S_2 > x\} \sim \mathbb{P}\left\{B > \frac{(1-\rho)x}{K+1}\right\}. \qquad (7.38)$$

In contrast, if $(K+1)r < 1-\rho$ and $C_2(s,t) \equiv t - s - B_1(s,t)$, then

$$\mathbb{P}\{S_2 > x\} \sim \mathbb{P}\{B > (1-\rho-Kr)x\}. \tag{7.39}$$

Proof First, the case $C_2(t) \equiv \frac{N(t)}{K+N(t)}$ follows directly from [82]. This result also directly provides the desired upper bound in case the system is critically loaded, i.e., $(K+1)r > 1-\rho$. The lower bound for (7.38) and the proof of (7.39) are somewhat similar to proofs of delay asymptotics in [35, 44, 82].

Let B_0 be the service requirement of a class-2 user arriving at time 0 and denote by S_0 its sojourn time. Also, let $B_0(0,t)$ be the amount of service received during $(0,t]$ if it had an infinite service requirement. Now, observe that an actual user arriving at time 0 would receive the same amount of service $B_0(0,t)$ if it is still present at time t. Thus, assume that at time 0 a persistent class-2 user arrives. Then,

$$\mathbb{P}\{S_0 > t\} = \mathbb{P}\{B_0 > B_0(0,t)\}. \tag{7.40}$$

For conciseness, we now first give the proof of (7.39) and then provide the lower bound for (7.38).

Proof of (7.39). We apply the framework developed in [44, 82]. In particular, we show that Assumptions (A-2) and (A-3) in [82] are satisfied. For Assumption (A-2), use (7.8) and (7.9):

$$B_0(0,t) = t + V_1(t) + V_2(t) - Krt - A_2(0,t) - V_1(0) - V_2(0).$$

Because the system is stable, both $(V_1(t)+V_2(t))/t \to 0$ and $(V_1(0)+V_2(0))/t \to 0$ when $t \to \infty$. Moreover, since $A_2(0,t)/t \to \rho$ for $t \to \infty$, we have

$$\lim_{t\to\infty} \frac{B_0(0,t)}{t} = 1 - \rho - Kr,$$

giving Assumption (A-2). For Assumption (A-3), note that $B_0(0,t) \geq \int_0^t 1/(K+1+N_{(K+1)}(u))\mathrm{d}u$. Thus, from the proof of [82, Theorem 3] (take $f(n) = \frac{1}{K+1+n}$), it follows that there exists a finite constant $D > 0$, such that

$$\mathbb{P}\{B_0(0,t) \leq Dt\} \leq \mathbb{P}\left\{\int_0^t \frac{1}{K+1+N_{(K+1)}(u)}\mathrm{d}u \leq Dt\right\} = o(\mathbb{P}\{B > x\}).$$

Since Assumptions (A-1)-(A-3) are satisfied, we may apply [82, Theorem 1] to obtain (7.39).

Lower bound for (7.38). Let $B^-_{2,\leq\kappa t}(s,t)$ $(B^-_{2,>\kappa t}(s,t))$ be the amount of service received by class-2 users with initial service requirements smaller than (larger than) κt, excluding the persistent class-2 user. Also, let $s_t := \sup\{0 \leq u \leq t : V_1(u) = 0\}$ be the last epoch before time t that the class-1 workload was zero. Recall that $V_2^c(t) = \sup_{0\leq s\leq t}\{A_2(s,t) - c(t-s)\}$. Using (7.8) and (7.9) in addition to Observation 7.7.1, we deduce

$$\begin{aligned}
& B_0(s_t,t) + B_1(s_t,t) + B^-_{2,>\kappa t}(s_t,t) \\
= \;& t - s_t - B^-_{2,\leq\kappa t}(s_t,t) \\
\geq \;& t - s_t - A_{2,\leq\kappa t}(s_t,t) - V^-_{2,\leq\kappa t}(s_t) \\
\geq \;& (1-\rho-\epsilon)(t-s_t) + (\rho+\epsilon)(t-s_t) - A_2(s_t,t) - V^-_{2,\leq\kappa t}(s_t) \\
\geq \;& (1-\rho-\epsilon)(t-s_t) - V_2^{\rho+\epsilon}(t) - M\kappa t,
\end{aligned}$$

where $A_{2,\leq\kappa t}(s,t)$ is the amount of "small" class-2 traffic generated during $(s,t]$ (see also Subsection 7.6.2), and $V^-_{2,\leq\kappa t}(s)$ is the workload at time s associated with "small" class-2 users, excluding the persistent user. Because class 1 uses the total available capacity during $(s_t,t]$, we have $B_1(s_t,t) = KB_0(s_t,t)$. Also, $V_1(s_t) = 0$ implies $B_1(0,s_t) \geq Krs_t$. Combining the above, and taking $\epsilon > 0$ sufficiently small, yields

$$\begin{aligned}
B_1(0,t) + B^-_{2,>\kappa t}(0,t) \;&\geq\; Krs_t + \frac{K}{K+1}[(1-\rho-\epsilon)(t-s_t) - V_2^{\rho+\epsilon}(t) - M\kappa t] \\
&\geq\; \frac{K}{K+1}[(1-\rho-\epsilon)t - V_2^{\rho+\epsilon}(t) - M\kappa t]. \qquad (7.41)
\end{aligned}$$

Now, applying (7.8) and (7.9), we obtain

$$\begin{aligned}
B_0(0,t) \;&\leq\; t - B_1(0,t) - B^-_{2,>\kappa t}(0,t) - B^-_{2,\leq\kappa t}(0,t) \\
&\leq\; (1-\rho+\epsilon)t - \frac{K}{K+1}[(1-\rho-\epsilon)t - V_2^{\rho+\epsilon}(t) - M\kappa t] \\
&\quad +V^-_{2,\leq\kappa t}(t) + (\rho-\epsilon)t - A_{2,\leq\kappa t}(0,t).
\end{aligned}$$

Moreover, observe that $V^-_{2,\leq\kappa t}(t) \leq V_{(K+1)}(t)$. Using these sample-path arguments and (7.40) yields

$$\begin{aligned}
& \mathbb{P}\{S_0 > t\} \\
\geq \;& \mathbb{P}\bigg\{B_0 > \frac{1-\rho+(2K+1)\epsilon}{K+1}t + \frac{K}{K+1}[V_2^{\rho+\epsilon}(t) + M\kappa t] \\
& \qquad\qquad +V^-_{2,\leq\kappa t}(t) + (\rho-\epsilon)t - A_{2,\leq\kappa t}(0,t) \leq \epsilon t\bigg\} \\
\geq \;& \mathbb{P}\left\{B_0 > \frac{1-\rho+(4K+2)\epsilon+KM\kappa}{K+1}t\right\} \qquad (7.42) \\
& \times\mathbb{P}\left\{\frac{K}{K+1}V_2^{\rho+\epsilon}(t) + V_{(K+1)}(t) + (\rho-\epsilon)t - A_{2,\leq\kappa t}(0,t) \leq \frac{2K+1}{K+1}\epsilon t\right\}.
\end{aligned}$$

Note that $V_2^{\rho+\epsilon}(t)$, $V_{(K+1)}(t)$, and $A_{2,\leq\kappa t}(0,t)$ are not independent. However, the second probability in (7.42) can be bounded from below by

$$\mathbb{P}\{A_{2,\leq\kappa t}(0,t) \geq (\rho-\epsilon)t\} - \mathbb{P}\{V_2^{\rho+\epsilon} \geq \epsilon t\} - \mathbb{P}\{V_{(K+1)}(t) \geq \epsilon t\}. \qquad (7.43)$$

The first term of (7.43) may be treated as in Subsection 7.6.2, which gives $\mathbb{P}\left\{A_{2,\leq\kappa t}(0,t) \geq (\rho-\epsilon)t\right\} \to 1$ as $t \to \infty$. For the second term, we note that $V_2^{\rho+\epsilon}$ has a non-defective distribution. Moreover, we use the fact that a system with $K+1$ permanent customers also has a proper limiting distribution to handle the third probability in (7.43) (see also Subsection 7.6.1).

Now, use the fact that $B(\cdot) \in \mathcal{R}_{-\nu}$ and let $\epsilon, \kappa \downarrow 0$ to obtain the lower bound for (7.39), which completes the proof. □

7.D Proof of (7.12) for variable-rate streaming traffic

We now extend Proposition 7.A.1 (relation (7.12)) to variable-rate streaming traffic. In fact, a slightly stronger result is needed in the proof of the lower bound of Theorem 7.8.1. However, $\mathbb{P}\left\{V_1^{\mathrm{var}} + V_2^{\mathrm{var}} > x\right\} \geq (1+o(1))\mathbb{P}\left\{V_2^{1-Kr} > x\right\}$ is a direct consequence of the proof, and the following proposition may be of independent interest. It also shows that the asymptotic equivalence holds under non-critical load if the system is work-conserving.

Proposition 7.D.1 *Suppose that the process $\{A_1(-t,0), t \geq 0\}$ satisfies Assumption 7.8.1 and $\rho + Kr < 1$. If $B(\cdot) \in \mathcal{R}_{-\nu}$ and one of the two following conditions is satisfied*

(i) the system is critically loaded, i.e., $1-\rho < (K+1)r$;

(ii) the system is work-conserving, i.e., $C_2(s,t) \equiv t - s - B_1(s,t)$;

then

$$\mathbb{P}\left\{V_1^{\mathrm{var}} + V_2^{\mathrm{var}} > x\right\} \sim \mathbb{P}\left\{V_2^{1-Kr} > x\right\}.$$

This asymptotic relation also holds when $V_1 + V_2$ and V_2^{1-Kr} represent the workloads embedded at class-2 arrival epochs rather than at arbitrary instants.

Proof The proofs again involve lower and upper bounds which asymptotically coincide. The lower bound is the same for both cases (i) and (ii).

(Lower bound) In fact, we will prove a slightly stronger result. Define $\overleftarrow{U}_1^c(0) := \sup_{t\geq 0}\{ct - A_1(-t,0)\}$ and recall that $\overrightarrow{U}_1^c(0) = \sup_{t\geq 0}\{ct - A_1(0,t)\}$. We show that

$$\liminf_{x\to\infty} \frac{\mathbb{P}\left\{V_1^{\mathrm{var}}(0) + V_2^{\mathrm{var}}(0) > x; \overrightarrow{U}_1^{K(r-\psi)}(0) \leq \phi x\right\}}{\mathbb{P}\left\{V_2^{1-Kr} > x\right\}} \geq 1. \tag{7.44}$$

Using the work-conserving scenario as a lower bound in addition to (7.10), we have for any $\xi > 0$,

$$\begin{aligned}
&\mathbb{P}\left\{V_1^{\mathrm{var}}(0) + V_2^{\mathrm{var}}(0) > x; \overrightarrow{U}_1^{K(r-\psi)}(0) \leq \phi x\right\} \\
\geq\ & \mathbb{P}\Bigg\{\sup_{t\geq 0}\{A_1(-t,0) - K(r-\psi)t + A_2(-t,0) - (1-K(r-\psi))t\} > x; \\
& \qquad\qquad\qquad \overrightarrow{U}_1^{K(r-\psi)}(0) \leq \phi x\Bigg\} \\
\geq\ & \mathbb{P}\left\{V_2^{1-K(r-\psi)}(0) - \overleftarrow{U}_1^{K(r-\psi)}(0) > x; \overrightarrow{U}_1^{K(r-\psi)}(0) \leq \phi x\right\} \\
\geq\ & \mathbb{P}\left\{V_2^{1-K(r-\psi)}(0) \geq (1+\xi)x\right\} \mathbb{P}\left\{\overleftarrow{U}_1^{K(r-\psi)}(0) \leq \xi x; \overrightarrow{U}_1^{K(r-\psi)}(0) \leq \phi x\right\}.
\end{aligned}$$

Note that

$$\begin{aligned}
&\mathbb{P}\left\{\overleftarrow{U}_1^{K(r-\psi)}(0) \leq \xi x; \overrightarrow{U}_1^{K(r-\psi)}(0) \leq \phi x\right\} \\
&\qquad \geq\ \mathbb{P}\left\{\overleftarrow{U}_1^{K(r-\psi)}(0) \leq \xi x\right\} - \mathbb{P}\left\{\overrightarrow{U}_1^{K(r-\psi)}(0) \geq \phi x\right\}.
\end{aligned}$$

Because both $\overleftarrow{U}_1^{K(r-\psi)}(0)$ and $\overrightarrow{U}_1^{K(r-\psi)}(0)$ have a proper distribution, it holds that $\mathbb{P}\left\{\overleftarrow{U}_1^{K(r-\psi)}(0) \leq \xi x\right\} \to 1$ and $\mathbb{P}\left\{\overrightarrow{U}_1^{K(r-\psi)}(0) \geq \phi x\right\} \to 0$ as $x \to \infty$ (see also (7.31)). Hence, we have

$$\liminf_{x\to\infty} \frac{\mathbb{P}\left\{V_1^{\mathrm{var}}(0) + V_2^{\mathrm{var}}(0) > x; \overrightarrow{U}_1^{K(r-\psi)}(0) \leq \phi x\right\}}{\mathbb{P}\left\{V_2^{1-K(r-\psi)} > (1+\xi)x\right\}} \geq 1.$$

Finally, let $\xi, \psi, \phi \downarrow 0$ and use Theorem 7.5.2 and the fact that $B^r(\cdot) \in \mathcal{R}_{1-\nu}$ to obtain (7.44). The lower bound is a direct consequence.

(Upper bound for part (i)) We now show that for a critically loaded system

$$\limsup_{x\to\infty} \frac{\mathbb{P}\left\{V_1^{\mathrm{var}} + V_2^{\mathrm{var}} > x\right\}}{\mathbb{P}\left\{V_2^{1-Kr} > x\right\}} \leq 1. \tag{7.45}$$

To do so, we apply the leaky-bucket idea of Section 7.8. Recall that in the reference system, class 1 generates traffic at constant rate $K(r+\psi)$, and class 2 receives service at rate $N_{(K)}(t)/(K + N_{(K)}(t))$, independently of class 1. Note that $V_2^{\mathrm{var}}(t) \leq V_{(K)}(t) = V_2^{\mathrm{cst},\psi}(t)$, with $V_{(K)}(t)$ the workload at time t in an isolated queue fed by class 2 with K permanent customers, and $V_2^{\mathrm{cst},\psi}(t)$ the class-2 workload at time t in the reference system. Thus, combining the above with (7.34) yields

$$V_1^{\mathrm{var}}(t) + V_2^{\mathrm{var}}(t) \leq V_1^{K(r+\psi)}(t) + V_1^{\mathrm{cst},\psi}(t) + V_2^{\mathrm{cst},\psi}(t).$$

Converting this sample-path relation into a probabilistic upper bound gives

$$\mathbb{P}\left\{V_1^{\mathrm{var}} + V_2^{\mathrm{var}} > x\right\} \leq \mathbb{P}\left\{V_1^{K(r+\psi)} > \xi x\right\} + \mathbb{P}\left\{V_1^{\mathrm{cst},\psi} + V_2^{\mathrm{cst},\psi} > (1-\xi)x\right\}.$$

Again, the first term can be controlled by Assumption 7.8.1. For the second term, apply Proposition 7.A.1 and Theorem 7.5.2, use the fact that $B^r(\cdot) \in \mathcal{R}_{1-\nu}$, and then let $\psi, \xi \downarrow 0$.

(Upper bound for part (ii)) It remains to be shown that (7.45) holds if the system is work-conserving. Using sample-path arguments, we have that $V_1^{\rm var}(t) + V_2^{\rm var}(t) \leq V_1^{K(r+\psi)}(t) + V_2^{1-K(r+\psi)}(t)$, so that, for any $\phi \in (0,1)$,

$$\mathbb{P}\left\{V_1^{\rm var} + V_2^{\rm var} > x\right\} \leq \mathbb{P}\left\{V_1^{K(r+\psi)} > \phi x\right\} + \mathbb{P}\left\{V_2^{1-K(r+\psi)} > (1-\phi)x\right\}.$$

It follows from Assumption 7.8.1, Theorem 7.5.2, and the fact that $B^r(\cdot) \in \mathcal{R}_{1-\nu}$ that

$$\mathbb{P}\left\{V_1^{K(r+\psi)} > \phi x\right\} = o(\mathbb{P}\left\{V_2^{1-K(r+\psi)} > (1-\phi)x\right\}),$$

as $x \to \infty$. Thus,

$$\limsup_{x\to\infty} \frac{\mathbb{P}\left\{V_1^{\rm var} + V_2^{\rm var} > x\right\}}{\mathbb{P}\left\{V_2^{1-K(r+\psi)} > (1-\phi)x\right\}} \leq 1.$$

Finally, let ψ, $\phi \downarrow 0$ and use Theorem 7.5.2 and the fact that $B^r(\cdot) \in \mathcal{R}_{1-\nu}$ to obtain (7.45).

Note that the above proof applies regardless of whether 0 is an arbitrary instant or a class-2 arrival epoch. □

References

[1] S. Aalto and W.R.W. Scheinhardt (2000). Tandem fluid queues fed by homogeneous on-off sources. *Oper. Res. Lett.* **27**, 73–82.

[2] I.J.B.F. Adan, E.A. van Doorn, J.A.C. Resing, and W.R.W. Scheinhardt (1998). Analysis of a single-server queue interacting with a fluid reservoir. *Queueing Syst.* **29**, 313–336.

[3] R. Agrawal, A.M. Makowski, and Ph. Nain (1999). On a reduced load equivalence for fluid queues under subexponentiality. *Queueing Syst.* **33**, 5–41.

[4] E. Altman, K.E. Avrachenkov, C. Barakat, and R. Núñez-Queija (2001). State-dependent M/G/1 type queueing analysis for congestion control in data networks. In *Proc. IEEE Infocom*, 1350–1359, Anchorage, AK.

[5] E. Altman, K.E. Avrachenkov, A.A. Kherani, and B.J. Prabhu (2004). Performance analysis and stochastic stability of congestion control protocols. RR-5262, INRIA Sophia Antipolis, France.

[6] D. Anick, D. Mitra, and M.M. Sondhi (1982). Stochastic theory of a data-handling system with multiple sources. *Bell System Tech. J.* **61**, 1871–1894.

[7] S. Asmussen (1998). Extreme value theory for queues via cycle maxima. *Extremes* **1**, 137–168.

[8] S. Asmussen (1998). Subexponential asymptotics for stochastic processes: extremal behavior, stationary distributions and first passage probabilities. *Ann. Appl. Probab.* **8**, 354–374.

[9] S. Asmussen (2000). *Ruin Probabilities.* World Scientific Publishing Co. Inc., New York.

[10] S. Asmussen (2003). *Applied Probability and Queues.* Springer-Verlag, New York.

[11] S. Asmussen and O. Kella (1996). Rate modulation in dams and ruin problems. *J. Appl. Probab.* **33**, 523–535.

[12] S. Asmussen and D. Perry (1992). On cycle maxima, first passage problems and extreme value theory for queues. *Comm. Statist. Stochastic Models* **8**, 421–458.

[13] S. Asmussen and S. Schock Petersen (1988). Ruin probabilities expressed in terms of storage processes. *Adv. Appl. Probab.* **20**, 913–916.

[14] S. Asmussen, H. Schmidli, and V. Schmidt (1999). Tail probabilities for non-standard risk and queueing processes with subexponential jumps. *Adv. Appl. Probab.* **31**, 422–447.

[15] S. Asmussen and K. Sigman (1996). Monotone stochastic recursions and their duals. *Probab. Engrg. Inform. Sci.* **10**, 1–20.

[16] F. Baccelli and P. Brémaud (2003). *Elements of Queueing Theory.* Springer-Verlag, Berlin.

[17] J. Bae, S. Kim, and E.Y. Lee (2003). Average cost under the $P^M_{\lambda,\tau}$ policy in a finite dam with compound Poisson inputs. *J. Appl. Probab.* **40**, 519–526.

[18] F. Baskett, K.M. Chandy, R.R. Muntz, and F.G. Palacios (1975). Open, closed, and mixed networks of queues with different classes of customers. *J. Assoc. Comput. Mach.* **22**, 248–260.

[19] R. Bekker (2004). Finite-buffer queues with workload-dependent service and arrival rates. SPOR-Report 2004-01, Eindhoven University of Technology, The Netherlands.

[20] R. Bekker (2005). Finite-buffer queues with workload-dependent service and arrival rates. *Queueing Syst.* **50**, 231–253.

[21] R. Bekker and S.C. Borst (2005). Optimal admission control in queues with workload-dependent service rates. PNA-E0514, CWI, The Netherlands.

[22] R. Bekker, S.C. Borst, O.J. Boxma, and O. Kella (2004). Queues with workload-dependent arrival and service rates. *Queueing Syst.* **46**, 537–556.

[23] R. Bekker, S.C. Borst, and R. Núñez-Queija (2004). Integration of TCP-friendly streaming sessions and heavy-tailed elastic flows. *Perf. Eval. Rev.* **32**, 41–43.

[24] R. Bekker, S.C. Borst, and R. Núñez-Queija (2005). Performance of TCP-friendly streaming sessions in the presence of heavy-tailed elastic flows. PNA-R0504, CWI, The Netherlands.

[25] R. Bekker, S.C. Borst, and R. Núñez-Queija (2005). Performance of TCP-friendly streaming sessions in the presence of heavy-tailed elastic flows. *Perf. Eval.* **61**, 143–162.

[26] R. Bekker and O.J. Boxma (2005). An M/G/1 queue with adaptable service speed. SPOR-Report 2005-09, Eindhoven University of Technology, The Netherlands.

[27] R. Bekker and O.J. Boxma (2005). Queues with adaptable speed. In B.D. Choi, editor, *Korea-Netherlands Joint Conference on Queueing Theory and its Applications to Telecommunication Systems*, 91–100.

[28] R. Bekker and A.P. Zwart (2005). On an equivalence between loss rates and cycle maxima in queues and dams. *Probab. Engrg. Inform. Sci.* **19**, 241–255.

[29] S. Ben Fredj, T. Bonald, A. Proutière, G. Régnié, and J.W. Roberts (2001). Statistical bandwidth sharing: a study of congestion at the flow level. In *Proc. ACM SIGCOMM*, 111–122.

[30] J.W.M. Bertrand and H.P.G. van Ooijen (2002). Workload based order release and productivity: a missing link. *Production Planning & Control* **13**, 665–678.

[31] N.H. Bingham, C.M. Goldie, and J.L. Teugels (1987). *Regular Variation.* Cambridge University Press, Cambridge.

[32] T. Bonald and A. Proutière (2004). On performance bounds for the integration of elastic and adaptive streaming flows. In *Proc. ACM Sigmetrics/Performance*, 235–245.

[33] N.K. Boots and H.C. Tijms (1999). A multiserver queueing system with impatient customers. *Management Sci.* **45**, 444–448.

[34] S.C. Borst, O.J. Boxma, R. Núñez-Queija, and A.P. Zwart (2003). The impact of the service discipline on delay asymptotics. *Perf. Eval.* **54**, 175–206.

[35] S.C. Borst, R. Núñez-Queija, and M.J.G. van Uitert (2002). User-level performance of elastic traffic in a differentiated-services environment. *Perf. Eval.* **49**, 507–519.

[36] S.C. Borst and A.P. Zwart (2005). Fluid queues with heavy-tailed M/G/∞ input. *Math. Oper. Res.*, to appear.

[37] O.J. Boxma (1996). Fluid queues and regular variation. *Perf. Eval.* **27 & 28**, 699–712.

[38] O.J. Boxma and V. Dumas (1998). Fluid queues with long-tailed activity period distributions. *Computer Communications* **21**, 509–529.

[39] O.J. Boxma, S.G. Foss, J.-M. Lasgouttes, and R. Núñez-Queija (2004). Waiting time asymptotics in the single server queue with service in random order. *Queueing Syst.* **46**, 35–73.

[40] O.J. Boxma, H. Kaspi, O. Kella, and D. Perry (2005). On/off storage systems with state-dependent input, output, and switching rates. *Probab. Engrg. Inform. Sci.* **19**, 1–14.

[41] O.J. Boxma, O. Kella, and D. Perry (2001). An intermittent fluid system with exponential on-times and semi-Markov input rates. *Probab. Engrg. Inform. Sci.* **15**, 189–198.

[42] O.J. Boxma, D. Perry, and W. Stadje (2001). Clearing models for M/G/1 queues. *Queueing Syst.* **38**, 287–306.

[43] O.J. Boxma, D. Perry, and F.A. van der Duyn Schouten (1999). Fluid queues and mountain processes. *Probab. Engrg. Inform. Sci.* **13**, 407–427.

[44] J. Boyer, F. Guillemin, Ph. Robert, and A.P. Zwart (2003). Heavy tailed M/G/1-PS queues with impatience and admission control in packet networks. In *Proc. IEEE Infocom*, San Francisco, USA.

[45] P.J. Brockwell, S.I. Resnick, and R.L. Tweedie (1982). Storage processes with general release rule and additive inputs. *Adv. Appl. Probab.* **14**, 392–433.

[46] S. Browne and K. Sigman (1992). Work-modulated queues with applications to storage processes. *J. Appl. Probab.* **29**, 699–712.

[47] E. Çinlar and M. Pinsky (1971). A stochastic integral in storage theory. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **17**, 227–240.

[48] E. Çinlar and M. Pinsky (1972). On dams with additive inputs and a general release rule. *J. Appl. Probab.* **9**, 422–429.

[49] J.W. Cohen (1969). Single server queues with restricted accessibility. *J. Engrg. Math.* **3**, 265–284.

[50] J.W. Cohen (1973). Some results on regular variation for distributions in queueing and fluctuation theory. *J. Appl. Probab.* **10**, 343–353.

[51] J.W. Cohen (1974). Superimposed renewal processes and storage with gradual input. *Stochastic Processes Appl.* **2**, 31–57.

[52] J.W. Cohen (1976). *On Regenerative Processes in Queueing Theory.* Springer-Verlag, Berlin.

[53] J.W. Cohen (1976). On the optimal switching level for an M/G/1 queueing system. *Stochastic Processes Appl.* **4**, 297–316.

[54] J.W. Cohen (1977). On up- and downcrossings. *J. Appl. Probab.* **14**, 405–410.

[55] J.W. Cohen (1979). The multiple phase service network with generalized processor sharing. *Acta Inform.* **12**, 245–284.

[56] J.W. Cohen (1982). *The Single Server Queue.* North-Holland, Amsterdam.

[57] J.W. Cohen and O.J. Boxma (1985). A survey of the evolution of queueing theory. *Statist. Neerlandica* **39**, 143–158.

[58] J.W. Cohen and M. Rubinovitch (1977). On level crossings and cycles in dam processes. *Math. Oper. Res.* **2**, 297–310.

[59] R.B. Cooper and S.C. Niu (1986). Beneš's formula for M/G/1-FIFO "explained" by preemptive-resume LIFO. *J. Appl. Probab.* **23**, 550–554.

[60] M. Crovella and A. Bestavros (1996). Self-similarity in world wide web traffic: evidence and possible causes. In *Proc. ACM Sigmetrics*, 160–169.

[61] D. J. Daley (1965). General customer impatience in the queue GI/G/1. *J. Appl. Probab.* **2**, 186–205.

[62] A.G. de Kok and H.C. Tijms (1985). A queueing system with impatient customers. *J. Appl. Probab.* **22**, 688–696.

[63] K. Dębicki, M.R.H. Mandjes, and M.J.G. van Uitert (2005). A tandem queue with Levy input: a new representation of the downstream queue length. Technical Report No. 10, Mittag-Leffler, Sweden.

[64] D. Denisov (2005). A note on the asymptotics for the maximum on a random time interval of a random walk. *Markov Process. Related Fields* **11**, 165–169.

[65] H. Dette, J.A. Fill, J. Pitman, and W.J. Studden (1997). Wall and Siegmund duality relations for birth and death chains with reflecting barrier. *J. Theoret. Probab.* **10**, 349–374.

[66] B.T. Doshi (1974). *Continuous Time Control of Markov Processes on an Arbitrary State Space.* PhD thesis, Cornell University, Ithaca, USA.

[67] B.T. Doshi (1977). Continuous time control of the arrival process in an M/G/1 queue. *Stochastic Processes Appl.* **5**, 265–284.

[68] B.T. Doshi (1992). Level-crossing analysis of queues. In U.N. Bhat and I.V. Basawa, editors, *Queueing and Related Models*, 3–33. Oxford Univ. Press, New York.

[69] V. Dumas and A. Simonian (2000). Asymptotic bounds for the fluid queue fed by sub-exponential On/Off sources. *Adv. Appl. Probab.* **32**, 244–255.

[70] I. Eliazar and J. Klafter (2003). Lévy-driven Langevin systems: targeted stochasticity. *J. Statist. Phys.* **111**, 739–768.

[71] A.I. Elwalid and D. Mitra (1991). Analysis and design of rate-based congestion control of high speed networks, I: stochastic fluid models, access regulation. *Queueing Syst.* **9**, 29–64.

[72] A.I. Elwalid and D. Mitra (1992). Fluid models for the analysis and design of statistical multiplexing with loss priorities on multiple classes of bursty traffic. In *Proc. IEEE Infocom*, 415–425.

[73] A.I. Elwalid and D. Mitra (1994). Statistical multiplexing with loss priorities in rate-based congestion control of high-speed networks. *IEEE Trans. Commun.* **42**, 2989–3002.

[74] P. Embrechts, C. Klüppelberg, and T. Mikosch (1997). *Modelling Extremal Events.* Springer-Verlag, Berlin.

[75] E.A. Feinberg and O. Kella (2002). Optimality of D-policies for an M/G/1 queue with a removable server. *Queueing Syst.* **42**, 355–376.

[76] W. Feller (1971). *An Introduction to Probability Theory and its Applications. Vol. II.* John Wiley & Sons Inc., New York.

[77] S. Floyd, M. Handley, J. Padhey, and J. Widmer (2000). Equation-based congestion control for unicast applications. In *Proc. ACM SIGCOMM*, 43–54.

[78] J. Gani (1955). Some problems in the theory of provisioning and of dams. *Biometrika* **42**, 179–200.

[79] J. Gani (1957). Problems in the probability theory of storage systems. *J. Roy. Statist. Soc. Ser. B.* **19**, 181–206; discussion 212–233.

[80] D.P. Gaver and R.G. Miller (1962). Limiting distributions for some storage problems. In K.J. Arrow, S. Karlin, and H. Scarf, editors, *Studies in Applied Probability and Management Science*, 110–126. Stanford Univ. Press, Stanford, Calif.

[81] B. Gavish and P.J. Schweitzer (1977). The Markovian queue with bounded waiting time. *Management Sci.* **23**, 1349–1357.

[82] F. Guillemin, Ph. Robert, and A.P. Zwart (2004). Tail asymptotics for processor-sharing queues. *Adv. Appl. Probab.* **36**, 525–543.

[83] J.M. Harrison and S.I. Resnick (1976). The stationary distribution and first exit probabilities of a storage process with general release rule. *Math. Oper. Res.* **1**, 347–358.

[84] A.M. Hasofer (1963). On the integrability, continuity and differentiability of a family of functions introduced by L. Takács. *Ann. Math. Statist.* **34**, 1045–1049.

[85] G. Hooghiemstra (1987). A path construction for the virtual waiting time of an M/G/1 queue. *Statist. Neerlandica* **41**, 175–181.

[86] D.L. Iglehart (1972). Extreme values in the GI/G/1 queue. *Ann. Math. Statist.* **43**, 627–635.

[87] V. Jacobson (1988). Congestion avoidance and control. In *Proc. ACM SIGCOMM*, 314–329.

[88] D. Jagerman (1985). Certain Volterra integral equations arising in queueing. *Comm. Statist. Stochastic Models* **1**, 239–256.

[89] P.R. Jelenković and A.A. Lazar (1999). Asymptotic results for multiplexing subexponential on-off processes. *Adv. Appl. Probab.* **31**, 394–421.

[90] P.R. Jelenković and P. Momčilović (2003). Large deviation analysis of subexponential waiting times in a processor-sharing queue. *Math. Oper. Res.* **28**, 587–608.

[91] P.R. Jelenković, P. Momčilović, and A.P. Zwart (2004). Reduced load equivalence under subexponentiality. *Queueing Syst.* **46**, 97–112.

[92] S. Karlin and H.M. Taylor (1981). *A Second Course in Stochastic Processes.* Academic Press Inc., New York.

[93] H. Kaspi, O. Kella, and D. Perry (1996). Dam processes with state dependent batch sizes and intermittent production processes with state dependent rates. *Queueing Syst.* **24**, 37–57.

[94] H. Kaspi and D. Perry (1989). On a duality between a non-Markovian storage/production process and a Markovian dam process with state-dependent input and output. *J. Appl. Probab.* **26**, 835–844.

[95] J. Keilson and N.D. Mermin (1959). The second-order distribution of integrand shot noise. *IRE Trans.* **IT-5**, 75–77.

[96] O. Kella (1993). Parallel and tandem fluid networks with dependent Lévy inputs. *Ann. Appl. Probab.* **3**, 682–695.

[97] O. Kella (2001). Markov-modulated feedforward fluid networks. *Queueing Syst.* **37**, 141–161.

[98] O. Kella and W. Stadje (2002). Exact results for a fluid model with state-dependent flow rates. *Probab. Engrg. Inform. Sci.* **16**, 389–402.

[99] O. Kella and W. Stadje (2002). Markov-modulated linear fluid networks with Markov additive input. *J. Appl. Probab.* **39**, 413–420.

[100] O. Kella and W. Whitt (1992). A storage model with a two-state random environment. *Oper. Res.* **40**, S257–S262.

[101] O. Kella and W. Whitt (1992). A tandem fluid network with Lévy input. In U.N. Bhat and I.V. Basawa, editors, *Queueing and Related Models*, 112–128. Oxford Univ. Press, New York.

[102] O. Kella and W. Whitt (1999). Linear stochastic fluid networks. *J. Appl. Probab.* **36**, 244–260.

[103] F.P. Kelly (1976). Networks of queues. *Adv. Appl. Probab.* **8**, 416–432.

[104] D.G. Kendall (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *Ann. Math. Statistics* **24**, 338–354.

[105] D.G. Kendall (1957). Some problems in the theory of dams. *J. Roy. Statist. Soc. Ser. B.* **19**, 207–212; discussion 212–233.

[106] P.B. Key, L. Massoulié, A. Bain, and F.P. Kelly (2003). A network flow model for mixtures of file transfers and streaming traffic. In *Proc. ITC-18*, 1021–1030, Berlin, Germany.

[107] L. Kleinrock (1975). *Queueing Systems.* Wiley, New York.

[108] L. Kleinrock (1976). *Queueing Systems, Vol. II: Computer Applications.* Wiley, New York.

[109] L. Kosten (1974). Stochastic theory of a multi-entry buffer. I. *Delft Progress Rep.* **1**, 10–18.

[110] L. Kosten (1984). Stochastic theory of data handling systems, with groups of multiple sources. In H. Rudin and W. Bux, editors, *Performance of Computer-Communication Systems*, 321–331. Elsevier Science Publishers B.V., North-Holland.

[111] D.P. Kroese and W.R.W. Scheinhardt (2001). Joint distributions for interacting fluid queues. *Queueing Syst.* **37**, 99–139.

[112] V.G. Kulkarni (1997). Fluid models for single buffer systems. In J.H. Dshalalow, editor, *Frontiers in Queueing*, 321–338. CRC, Boca Raton, FL.

[113] J.F. Kurose and K.W. Ross (2003). *Computer Networking.* Addison Wesley, second edition.

[114] J. Lee and J. Kim (2005). A workload-dependent M/G/1 queue under a two-stage service policy. *Oper. Res. Lett.*, to appear.

[115] D.V. Lindley (1959). Discussion of a paper by C.B. Winsten. *Cambridge Philosophical Society* **48**, 277–289.

[116] P. Linz (1985). *Analytical and Numerical Methods for Volterra Equations.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

[117] J. Loris-Teghem (1972). On the waiting time distribution in a generalized queueing system with uniformly bounded sojourn times. *J. Appl. Probab.* **9**, 642–649.

[118] R.M. Loynes (1965). On a property of the random walks describing simple queues and dams. *J. Roy. Statist. Soc. Ser. B* **27**, 125–129.

[119] M.R.H. Mandjes, D. Mitra, and W.R.W. Scheinhardt (2002). A simple model of network access: feedback adaptation of rates and admission control. In *Proc. IEEE Infocom*, 3–12, New York, USA.

[120] M.R.H. Mandjes, D. Mitra, and W.R.W. Scheinhardt (2003). Models of network access using feedback fluid queues. *Queueing Syst.* **44**, 365–398.

[121] L. Massoulié and J.W. Roberts (1999). Bandwidth sharing: Objectives and algorithms. In *Proc. IEEE Infocom*, 1395–1403.

[122] M. Mathis, J. Semke, J. Mahdavi, and T.J. Ott (1997). The macroscopic behavior of the TCP congestion avoidance algorithm. *Comp. Commun. Rev.* **27**, 67–82.

[123] S.G. Mikhlin (1957). *Integral Equations and their Applications to Certain Problems in Mechanics, Mathematical Physics and Technology.* Pergamon Press, New York.

[124] V. Misra, W.-B. Gong, and D. Towsley (1999). Stochastic differential equation modeling and analysis of TCP-windowsize behavior. In *Proc. Performance*, Istanbul, Turkey.

[125] M. Miyazawa (1994). Time-dependent rate conservation laws for a process defined with a stationary marked point process and their applications. *J. Appl. Probab.* **31**, 114–129.

[126] P.A.P. Moran (1954). A probability theory of dams and storage systems. *Austral. J. Appl. Sci.* **5**, 116–124.

[127] P.A.P. Moran (1955). A probability theory of dams and storage systems: modifications of the release rules. *Austral. J. Appl. Sci.* **6**, 117–130.

[128] P.A.P. Moran (1959). *The Theory of Storage.* Methuen & Co. Ltd., London.

[129] P.A.P. Moran (1969). A theory of dams with continuous input and a general release rule. *J. Appl. Probab.* **6**, 88–98.

[130] R. Núñez-Queija (2000). *Processor-Sharing Models for Integrated-Services Networks.* PhD thesis, Eindhoven University of Technology, Eindhoven, The Netherlands.

[131] R. Núñez-Queija (2001). Note on the GI/GI/1 queue with LCFS-PR observed at arbitrary times. *Probab. Engrg. Inform. Sci.* **15**, 179–187.

[132] R. Núñez-Queija (2002). Queues with equally heavy sojourn time and service requirement distributions. *Ann. Oper. Res.* **113**, 101–117.

[133] J. Padhye, V. Firoiu, D. Towsley, and J. Kurose (2000). Modeling TCP Reno performance: a simple model and its empirical validation. *IEEE/ACM Trans. Netw.* **8**, 133–145.

[134] J. Padhye, J. Kurose, D. Towsley, and R. Koodli (1999). A model-based TCP-friendly rate control protocol. In *Proc. IEEE NOSSDAV.*

[135] A.G. Pakes (1975). On the tails of waiting-time distributions. *J. Appl. Probab.* **12**, 555–564.

[136] J. Paulsen and H.K. Gjessing (1997). Ruin theory with stochastic return on investments. *Adv. Appl. Probab.* **29**, 965–985.

[137] D. Perry and S. Asmussen (1995). Rejection rules in the M/G/1 queue. *Queueing Syst.* **19**, 105–130.

[138] D. Perry and W. Stadje (2003). Duality of dams via mountain processes. *Oper. Res. Lett.* **31**, 451–458.

[139] N.U. Prabhu (1965). *Queues and Inventories.* John Wiley & Sons Inc., New York.

[140] K. Ramanan and A. Weiss (1997). Sharing bandwidth in ATM. In *Proc. Allerton Conference*, 732–740.

[141] E. Reich (1958). On the integrodifferential equation of Takács I. *Ann. Math. Statist* **29**, 563–570.

[142] R. Rejaie, M. Handley, and D. Estrin (1999). RAP: an end-to-end rate-based congestion control mechanism for real-time streams in the internet. In *Proc. IEEE Infocom*, 1337–1346.

[143] S. Resnick and G. Samorodnitsky (1999). Activity periods of an infinite server queue and performance of certain heavy tailed fluid queues. *Queueing Syst.* **33**, 43–71.

[144] S.O. Rice (1954). Mathematical analysis of random noise. In N. Wax, editor, *Selected Papers on Noise and Stochastic Processes*, 133–294. Dover, New York.

[145] R.K. Ritt and L.I. Sennott (1992). Optimal stationary policies in general state space Markov decision chains with finite action sets. *Math. Oper. Res.* **17**, 901–909.

[146] T. Rolski, S. Schlegel, and V. Schmidt (1999). Asymptotics of Palm-stationary buffer content distributions in fluid flow queues. *Adv. Appl. Probab.* **31**, 235–253.

[147] S.M. Ross (1968). Arbitrary state Markovian decision processes. *Ann. Math. Statist.* **39**, 2118–2122.

[148] S.M. Ross (1970). Average cost semi-Markov decision processes. *J. Appl. Probab.* **7**, 649–656.

[149] S.M. Ross (1996). *Stochastic Processes.* John Wiley & Sons Inc., New York, second edition.

[150] M. Schäl (1993). Average optimality in dynamic programming with general state space. *Math. Oper. Res.* **18**, 163–172.

[151] W.R.W. Scheinhardt (1998). *Markov-Modulated and Feedback Fluid Queues.* PhD thesis, University of Twente, Enschede, The Netherlands.

[152] W.R.W. Scheinhardt, N.D. van Foreest, and M.R.H. Mandjes (2005). Continuous feedback fluid queues. *Oper. Res. Lett.* **33**, 551–559.

[153] W.R.W. Scheinhardt and A.P. Zwart (2002). A tandem fluid queue with gradual input. *Probab. Engrg. Inform. Sci.* **16**, 29–45.

[154] R.W. Schmenner (1988). The merit of making things fast. *Sloan Management Review* **30**, 11–17.

[155] D. Siegmund (1976). The equivalence of absorbing and reflecting barrier problems for stochastically monotone Markov processes. *Ann. Probab.* **4**, 914–924.

[156] K. Sigman (1995). *Stationary Marked Point Processes.* Chapman & Hall, New York.

[157] W.L. Smith (1953). On the distribution of queueing times. *Proc. Cambridge Philos. Soc.* **49**, 449–461.

[158] A.J. Stam (1973). Regular variation of the tail of a subordinated probability distribution. *Adv. Appl. Probab.* **5**, 308–327.

[159] L. Takács (1955). Investigation of waiting time problems by reduction to Markov processes. *Acta Math. Acad. Sci. Hungar.* **6**, 101–129.

[160] L. Takács (1965). Application of ballot theorems in the theory of queues. In W.L. Smith and W.E. Wilkinson, editors, *Proc. Sympos. Congestion Theory*, 337–398. Univ. North Carolina Press, Chapel Hill, N.C.

[161] L. Takács (1967). *Combinatorial Methods in the Theory of Stochastic Processes.* John Wiley & Sons Inc., New York.

[162] H.C. Tijms (1976). Optimal control of the workload in an M/G/1 queueing system with removable server. *Math. Operationsforsch. Statist.* **7**, 933–943.

[163] H.C. Tijms (2003). *A First Course in Stochastic Models.* John Wiley & Sons Inc., Chichester.

[164] H.C. Tijms and F.A. van der Duyn Schouten (1978). Inventory control with two switch-over levels for a class of M/G/1 queueing systems with variable arrival and service rate. *Stochastic Processes Appl.* **6**, 213–222.

[165] F.G. Tricomi (1957). *Integral Equations.* Interscience Publishers, New York.

[166] J.L. van den Berg and O.J. Boxma (1991). The M/G/1 queue with processor sharing and its relation to a feedback queue. *Queueing Syst.* **9**, 365–401.

[167] N.D. van Foreest (2004). *Queues with Congestion-Dependent Feedback.* PhD thesis, University of Twente, Enschede, The Netherlands.

[168] J.C.W. van Ommeren and A.G. de Kok (1987). Asymptotic results for buffer systems under heavy load. *Probab. Engrg. Inform. Sci.* **1**, 327–348.

[169] H.P.G. van Ooijen and J.W.M. Bertrand (2003). The effects of a simple arrival rate control policy on throughput and work-in-process in production systems with workload dependent processing rates. *International Journal of Production Economics* **85**, 61–68.

[170] M.J.G. van Uitert (2003). *Generalized Processor Sharing Queues.* PhD thesis, Eindhoven University of Technology, Eindhoven, The Netherlands.

[171] G.H. Weiss (1965). A survey of some recent research in road traffic. In W.L. Smith and W.E. Wilkinson, editors, *Proc. Sympos. Congestion Theory*, 253–288. Univ. North Carolina Press, Chapel Hill, N.C.

[172] C.D. Wickens and J.G. Hollands (1999). *Engineering Psychology and Human Performance.* Prentice Hall, New Jersey, third edition.

[173] S.F. Yashkov (1987). Processor-sharing queues: some progress in analysis. *Queueing Syst.* **2**, 1–17.

[174] S.F. Yashkov (1992). Mathematical problems in the theory of processor-sharing queueing systems. *J. Soviet Math.* **58**, 101–147.

[175] P.P. Zabreĭko, A.I. Koshelev, and M.A. Krasnosel′skiy (1975). *Integral Equations: a Reference Text.* Noordhoff, Leiden.

[176] A.P. Zwart (2000). A fluid queue with a finite buffer and subexponential input. *Adv. Appl. Probab.* **32**, 221–243.

[177] A.P. Zwart (2001). *Queueing Systems with Heavy Tails.* PhD thesis, Eindhoven University of Technology, Eindhoven, The Netherlands.

[178] A.P. Zwart (2003). Loss asymptotics for the M/G/1 queue with complete rejection. SPOR-Report 2003-27, Eindhoven University of Technology, Netherlands.

[179] A.P. Zwart, S.C. Borst, and M.R.H. Mandjes (2004). Exact asymptotics for fluid queues fed by multiple heavy-tailed on-off flows. *Ann. Appl. Probab.* **14**, 903–957.

[180] A.P. Zwart and O.J. Boxma (2000). Sojourn time asymptotics in the M/G/1 processor sharing queue. *Queueing Syst.* **35**, 141–166.

Samenvatting (Summary)

In klassieke wachtrijmodellen wordt aangenomen dat de bediende op constante snelheid werkt zolang er werk in het systeem aanwezig is. Er zijn echter tal van situaties waar deze aanname niet opgaat, zoals in productiesystemen, waterreservoirs of communicatienetwerken. Bovendien kan de aankomstintensiteit van nieuwe klanten worden beïnvloed door de mate van congestie in het systeem. In dit proefschrift concentreren we ons daarom specifiek op wachtrijen met toestandsafhankelijke snelheden.

We onderscheiden in dit proefschrift drie belangrijke toepassingsgebieden. Als eerste noemen we productiesystemen waarbij de productiviteit van het personeel afhangt van de aanwezige hoeveelheid werk. In de psychologie wordt de relatie tussen werkdruk en productiviteit beschreven door de Yerkes Dodson wet: In eerste instantie leidt een hogere werkdruk tot een verbeterde productiviteit, maar bij een aanhoudende stijging van de aanwezige hoeveelheid werk krijgen stressfactoren de overhand, resulterend in een scherpe productiviteitsdaling. Een tweede toepassingsgebied van modellen met toestandsafhankelijke snelheden zijn communicatienetwerken waarbij het verzendingsprotocol reageert op drukte in het netwerk. Een duidelijk voorbeeld hiervan is het veel gebruikte Transmission Control Protocol (TCP), waarbij informatie over netwerkcongestie de basis vormt voor de bepaling van de verzendingssnelheid van Internetverkeer. In hoofdstuk 7 richten we ons specifiek op de integratie van verkeersstromen van verschillende aard met uiteenlopende kwaliteitseisen. Als derde toepassing noemen we de studie van waterreservoirs, en opslagmodellen in het algemeen. Instromend water als gevolg van hevige regenval wordt opgevangen in een reservoir, terwijl de uitstroomsnelheid afhangt van de watervoorraad achter de dam. Deze toepassing is van een meer wiskundige aard en is met name vanuit een historisch perspectief van groot belang.

In hoofdstuk 1 geven we verdere achtergrondinformatie over de bovengenoemde drie toepassingsgebieden en bespreken we de relatie tot wachtrijmodellen met toestandsafhankelijke snelheden. Verder demonstreren we verschillende methoden uit het proefschrift aan de hand van de klassieke M/G/1 rij. De daaruit voortvloeiende bekende M/G/1 resultaten kunnen tevens worden gebruikt als referentie voor resultaten in latere hoofdstukken. De voornaamste prestatiemaat in dit proefschrift is de verdeling van de hoeveelheid werk (ook wel *werklast* genoemd) in de evenwichtssituatie.

In hoofdstuk 2 bestuderen we allereerst de M/G/1 wachtrij met werklastafhankelijke aankomst- en bedieningssnelheden. Dit model hangt nauw samen met het hierboven besproken waterreservoir waarbij de uitstroomsnelheid afhangt van de inhoud van het reservoir. Het belangrijkste resultaat is de relatie tussen grootheden, zoals de werklastverdeling, in twee verwante M/G/1 wachtrijen. Daarnaast werken we enkele speciale gevallen verder uit. Vervolgens beschouwen we het algemenere G/G/1 model en geven we relaties tussen de werklast op verschillende momenten.

In hoofdstuk 3 breiden we het M/G/1 model van hoofdstuk 2 uit door verschillende begrenzingen (of toelatingseisen) op de werklast toe te staan. We kijken daarbij opnieuw naar de relatie tussen grootheden in verwante M/G/1 wachtrijen en laten verder zien dat de werklastverdeling voor een aantal M/G/1 rijen met beperkte toelating proportioneel is aan de werklastverdeling van het model *zonder* toelatingsrestrictie. Tevens beschouwen we de verdeling van een andere prestatiemaat, het *cycle maximum*. Het cycle maximum is de maximale hoeveelheid werk gedurende een busy cycle (de periode dat de bediende onafgebroken werkt). We besluiten het hoofdstuk door een aantal speciale gevallen uit te werken.

Het cycle maximum speelt ook een centrale rol in hoofdstuk 4. We bestuderen daar een G/G/1 rij met eindige buffer en analyseren de relatie tussen de kans dat een klant niet volledig wordt geaccepteerd (de verlieskans) en de staartkans van het cycle maximum in de daarbij behorende rij met oneindige buffer. Voor het klassieke G/G/1 model laten we zien dat deze twee kansen identiek zijn. In het model waarbij de bedieningssnelheid afhangt van de aanwezige hoeveelheid werk zijn de staartkans van het cycle maximum en de verlieskans op een iets ingewikkeldere manier gerelateerd. Tenslotte passen we deze relaties toe om resultaten te verkrijgen voor de verlieskans in modellen waar de verdeling van het cycle maximum bekend is en vice versa.

Hoofdstuk 5 betreft opnieuw een M/G/1 rij met werklastafhankelijke bedieningssnelheden. Mede geïnspireerd door de hierboven beschreven productiviteitspatronen in, bijvoorbeeld, productiesystemen, richten we ons specifiek op bedieningssnelheden die eerst stijgen en dan dalen als functie van de aanwezige hoeveelheid werk. Besturing van het systeem vindt plaats door het al dan niet toelaten van klanten afhankelijk van de werklast bij aankomst, met als doel de lange-termijn gemiddelde hoeveelheid afgehandeld werk (ofwel de *throughput*) te maximaliseren. We laten zien dat, onder bepaalde voorwaarden, een drempelwaarde strategie voor het accepteren van klanten optimaal is. We geven ook een karakterisering van de optimale drempelwaarde, waarvan de berekening in bepaalde gevallen reduceert tot de oplossing van een betrekkelijk eenvoudige vergelijking.

In de bovengenoemde hoofdstukken 2–5 hebben we steeds verondersteld dat de bedieningssnelheid op elk moment (en continu door de tijd) kan worden aangepast. In verschillende praktische situaties kan het echter voorkomen dat niet op elk moment informatie over de toestand van het systeem aanwezig is, of dat er hoge kosten gepaard gaan met het continu aanpassen van de bedieningssnelheid. In hoofdstuk 6 nemen we daarom aan dat de snelheid van bediening alleen op momenten direct na een aankomst kan worden gewijzigd, terwijl deze constant wordt gehouden tussen aankomsten van klanten in. Voor het geval van bedieningsdisciplines met één of meer drempelwaarden vinden we de verdeling en de getransformeerde van de hoeveelheid werk in het systeem op verschillende momenten.

Tenslotte richten we ons in hoofdstuk 7 op een toepassing op het gebied van communicatienetwerken. We beschouwen twee typen verkeer, *stromend* en *elastisch*, die capaciteit delen volgens de Processor Sharing (PS) discipline. De PS discipline is een natuurlijke manier om het delen van capaciteit tussen TCP en *TCP-friendly* gestuurd Internetverkeer te modelleren. Bovendien nemen we aan dat het verkeer van de elastische klasse *zwaarstaartig* is en dat de verbinding kritiek belast is. Het belangrijkste resultaat betreft de werklast asymptotiek van de stromende klasse. Deze prestatiemaat is met name interessant omdat de werklast kan worden geïnterpreteerd als een bedieningstekort ten opzichte van een ideaal scenario. Verder geven we ook verschillende resultaten voor de elastische klasse. We merken op dat het model van dit hoofdstuk ook opgevat kan worden als een waterreservoir of vloeistofmodel in een zwaarstaartige stochastische omgeving. Die omgeving bestaat dan uit de elastische klanten, terwijl de bedieningssnelheid, of uitstroomsnelheid, van de dam gelijk is aan die van een permanent aanwezige klant in een G/G/1 rij bediend volgens de PS discipline.

About the author

René Bekker was born in Leiderdorp (the Netherlands) on August 17, 1978. He completed Grammar School at the Aquino College, Leiden, in June 1996. In September 2001, he received his master's degree in Econometrics from the Vrije Universiteit in Amsterdam. Subsequently, he became a PhD student at Eindhoven University of Technology, in a project funded by Philips Research. Since then he has also been affiliated with CWI (Center for Mathematics and Computer Science, Amsterdam). René defends his thesis on December 12, 2005. The first three months of 2006, he intends to work at CWI and visit INRIA in Sophia Antipolis, France.