Markov-modulated and feedback fluid queues

Werner Scheinhardt
Faculty of Mathematical Sciences
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

ISBN 90-3651248-4

MARKOV-MODULATED AND FEEDBACK FLUID QUEUES

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof. dr. F.A. van Vught,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op vrijdag 4 december 1998 te 15.00 uur.

door

Willem Richard Werner Scheinhardt

geboren op 24 februari 1969
te Santiago

Dit proefschrift is goedgekeurd door de promotor en de assistent promotor,

prof. dr. ir. J.H.A. de Smit
dr. ir. E.A. van Doorn

Voorwoord

Aan het eind van dit proefschrift gekomen, rest nog het schrijven van het begin ervan, het voorwoord. Gebruikelijk is om daarin allen te bedanken die op één of andere wijze aan de totstandkoming van het proefschrift hebben bijgedragen. Graag houd ik deze traditie in ere, en wel omdat de volgende personen dit ten volle verdienen.

Uiteraard wil ik beginnen met mijn dagelijks begeleider Erik van Doorn hartelijk te bedanken voor zijn inzet en enthousiasme. Van vóór het eerste sollicitatiegesprek tot na het laatste □-teken was hij intensief betrokken bij mijn doen en laten. Een belangrijk deel van het onderzoek in dit proefschrift is in samenwerking met hem tot stand gekomen.

Ook de samenwerking met Dick Kroese was aangenaam en productief. Hij wist vaak antwoorden die ik zelf niet zou hebben gevonden, en leerde mij omgaan met het symbolisch manipulatie-pakket "Mathematica", dat heel wat "tiresome analysis" heeft uitgevoerd.

De overige twee personen die direct betrokken waren bij het hier beschreven onderzoek zijn Jacques Resing en Ivo Adan van de Technische Universiteit Eindhoven; ook met hen was het een eer en genoegen om samen te werken.

Vervolgens gaat natuurlijk mijn dank uit naar de leden van de promotiecommissie, met op de eerste plaats mijn promotor, Jos de Smit. Hoewel hij het vaak druk had, toonde hij zich altijd geïnteresseerd en betrokken. Ook de andere commissieleden (Hans van den Berg, Onno Boxma, Wim Nawijn, Ignas Niemegeers en Henk Zijm) ben ik dank verschuldigd, zowel voor de uitvoering van de hen toevertrouwde formele taken als voor het aangename en nuttige contact daarbuiten.

De nog niet genoemde (oud-)collega's van de afdeling DOS worden allen hartelijk bedankt voor hun interesse en gezelschap. Helaas is het niet mogelijk om hen allen hier bij naam te noemen, dus wil ik het laten bij deze twee mensen: Ineke van Eijkern-Moraal, onze secretaresse die op praktisch gebied altijd de beste oplossing, het juiste formulier, en niet te vergeten het nieuwste nieuwtje kon bieden. En natuurlijk Rieske Hadianti, met wie ik bijna 3 jaar lang de kamer met plezier heb gedeeld, dus: terima kasih sekali, Rieske!

Tijs Huisman, Sindo Núñez Queija en Simone Sassen wil ik bedanken voor de stimulerende gesprekken en leuke vakanties in binnen- en buitenland. John Hoff dank ik voor de illustratie op de voorkant van dit proefschrift.

Ten slotte bedank ik hier graag van harte de vier mensen die deze vier jaar zonder twijfel het meest hebben meegeleefd, namelijk mijn ouders, Karen, en natuurlijk Sonja.

Werner Scheinhardt — Enschede, november 1998

Contents

Chapter 1

Introduction

1.1 Introduction

1.1.1 Fluid models for queueing systems

In traditional queueing theory the object of interest is usually a system of one or more *servers* at which *customers* arrive, who want to receive some kind of service. Since there is uncertainty about the actual arrival times of the customers and/or their service requirements, they may have to wait for service in a *queue*. In the course of this century many interesting and insightful results have been found for a wide range of variants of this model. The basic assumptions in any of these models can be summarized in the notation $A/B/n/m - S$, which is an extension of Kendall's characterisation of a queueing system. Here, A and B indicate the distributions of interarrival times and service times (that may or may not be independent), n is the number of servers, m the maximum number of customers that can be accommodated in the system, and S specifies the service discipline. In their search for more advanced ways to model practical queueing situations, researchers have added a wide variety of features to this basic model, such as various types of feedback, multi-class customers, priorities with or without preemption, server vacations, polling systems, state-dependent arrivals and services, impatient customers, batch arrivals, batch services, and many more. Indeed, the only thing that seemed to stand the test of time was that in all of these models *individual* customers arrive to receive some kind of service, reflecting the discrete nature of the modeled phenomena.

In the last ten to fifteen years, this last remaining pillar has been torn down as well, allowing for models in which some continuous entity, referred to as *fluid*, takes the role of the individual customers. In these models, fluid flows into a *fluid reservoir* according to some stochastic process. The server may be thought of as a tap at the bottom of the reservoir, allowing fluid to flow out. The rate at which this happens is often constant, but may also be stochastic, possibly including zero. Since the fluid reservoir takes the role of the traditional customer queue, it is often referred to as *fluid queue*. A third term which is often encountered is *fluid buffer*, stressing the fact that the storage of fluid is temporary, to reduce loss of fluid at times when the arrival rate exceeds the service rate.

In the last decade, the literature on queueing theory has paid considerable attention to *Markov-modulated fluid models.* In these models, a fluid buffer is either filled or depleted, or both, at rates which are determined by the current state of a *background Markov process*, also called *Markovian random environment.* Partly, this thesis will engage in analysing this type of model, as should be clear from its title. The second type of models we encounter is new, but closely related. We shall call them *feedback fluid models.* We refer to Section 1.5.1 for a more detailed description, but we mention already that the term feedback has a different meaning here than in the classical queueing literature. Rather it signifies that the state of the buffer content influences the behaviour of the regulating background process.

1.1.2 Outline of this chapter

The remainder of this chapter is organised as follows. In Section 1.2 we will introduce the traditional Markov-modulated fluid model more precisely (but without making many assumptions) and give a flavour of how this type of model can be solved. In doing so we will fix the basic notation and terminology that is used throughout this thesis.

Section 1.3 will present an overview of current literature on the subject of Markov-modulated fluid models. Since there are many topics which are of interest, we will often refer the reader to other sources where more in-depth overviews may be found. The same holds for Section 1.4 where some literature is discussed about other types of fluid models in queueing theory. We will conclude this chapter with Section 1.5, in which we give an overview of the contributions made in the remaining chapters in this thesis. This includes a more elaborate explanation of what we mean by feedback fluid models.

1.2 The traditional Markov-modulated fluid model

This section deals with the fairly general Markov-modulated fluid model (henceforth abbreviated to MMFM, one of the few acronyms in this thesis). We will sketch the traditional solution procedure and at the same time fix some notation.

Consider a fluid reservoir. Let C_t denote the amount of fluid at time t in this reservoir; the symbol C is chosen as a mnemonic for content. Furthermore, let (X_t) be a continuous-time Markov process. (X_t) will be said to evolve "in the background". In this section we will assume that (X_t) has a finite state space $\mathcal{N}$. The same holds for any MMFM discussed in this chapter, unless otherwise mentioned. In particular we assume here that $\mathcal{N} = \{1, 2, \ldots, N\}$.

The content of the reservoir is *regulated* (or *driven*) by (X_t) in such a way that the *net input rate* into the reservoir (i.e. the rate of change of its content) is r_i at times when (X_t) is in state $i \in \mathcal{N}$, unless this is not physically possible. Hence we have,

$$\frac{dC_t}{dt} = \begin{cases} 0 & \text{if } C_t = 0 \text{ and } r_{X_t} < 0 \\ r_{X_t} & \text{else.} \end{cases} \tag{1.1}$$

In other words, C_t is the reflected version,

$$C_t = A_t - \min_{0 \le s \le t} A_s, \tag{1.2}$$

of the *potential net input process*,

$$A_t = \int_0^t r_{X_s} ds. \tag{1.3}$$

The above holds for buffers with infinite capacity. When the buffer has a finite size K, we have

$$\frac{dC_t}{dt} = \begin{cases} 0 & \text{if } C_t = 0 \text{ and } r_{X_t} < 0, \quad \text{or if } C_t = K \text{ and } r_{X_t} > 0, \\ r_{X_t} & \text{else.} \end{cases} \tag{1.4}$$

It is assumed that at least one of the parameters $r_i, i \in \mathcal{N}$ be strictly positive, in order for the model to be meaningful. When the buffer is infinitely large, another assumption must be made in order to ensure stability of the buffer content. This stability condition is given by

$$\sum_{i \in \mathcal{N}} p_i r_i < 0,$$

where p_i is the stationary probability that (X_t) is in state $i \in \mathcal{N}$. When this condition is satisfied, a stochastic vector (X, C) exists to which the process (X_t, C_t) converges in distribution as $t \to \infty$. Hence, the stationary joint distribution of (X_t, C_t) exists and can be written as

$$F_i(y) = P[X = i, C \le y], \qquad i \in \mathcal{N}, \ y \ge 0.$$

The generator of the Markov process (X_t) is denoted by Q, thus $Q = [q_{ij}]$ where $q_{ij}, i, j \in \mathcal{N}, i \ne j$ is the transition rate from state i to state j, and $q_{ii} = -\sum_{j \ne i} q_{ij}, i \in \mathcal{N}$. Furthermore, we define the diagonal matrix R as

$$R = \text{diag}\ (r_1, \ldots, r_N).$$

It can be shown that the vector

$$\mathbf{F}(y) = [F_1(y), F_2(y), \ldots, F_N(y)]^T$$

satisfies the differential equation

$$R\mathbf{F}'(y) = Q^T \mathbf{F}(y), \tag{1.5}$$

where prime denotes differentiation and superscript T denotes transpose. By assuming that R is non-singular, i.e. $r_i \ne 0$ for $i \in \mathcal{N}$, we arrive at

$$\mathbf{F}'(y) = R^{-1} Q^T \mathbf{F}(y).$$

In case the eigenvalues are simple, it follows that

$$\mathbf{F}(y) = \sum_{j=1}^{N} c_j e^{\xi_j y} \mathbf{v}^{(j)}, \tag{1.6}$$

or, equivalently,

$$F_i(y) = \sum_{j=1}^{N} c_j e^{\xi_j y} v_i^{(j)}, \tag{1.7}$$

where the $(\xi_j, \mathbf{v^{(j)}})$ are the eigenvalue-eigenvector pairs of the matrix $R^{-1}Q^T$ and c_j are constants that can be determined by boundary conditions. In particular, when the buffer is infinitely large, we must have that $c_j = 0$ when $\mathrm{Re}(\xi_j) > 0$. This observation, together with the following remarkable property that holds for the eigenvalues of the matrix $R^{-1}Q^T$, will lead us to the conclusion of this short outline. We define

$$\mathcal{N}^+ \equiv \{i \in \mathcal{N} \mid r_i > 0\}, \quad \mathcal{N}^- \equiv \{i \in \mathcal{N} \mid r_i < 0\}, \tag{1.8}$$

and

$$N_+ \equiv |\mathcal{N}^+|, \quad N_- \equiv |\mathcal{N}^-|. \tag{1.9}$$

Then it turns out that when the stability condition is satisfied, the number of eigenvalues with negative real part equals N_+, the number of eigenvalues with positive real part equals $N_- - 1$, and the last eigenvalue equals zero (notice that the total number of eigenvalues is $N_+ + N_- = N$). For the case of an infinitely large reservoir, this means that the sum in (1.6) can be reduced to a sum with $N_+ + 1$ terms. Notice that one of these terms, namely the one corresponding to the eigenvalue 0, can be interpreted as the stationary distribution of the regulating process. The coefficients c_j that appear in the other $\mathcal{N}_+$ terms can be found using the boundary conditions

$$F_i(0) = 0, \quad i \in \mathcal{N}^+, \tag{1.10}$$

which must hold since the content of the reservoir is increasing whenever $X_t \in \mathcal{N}^+$.

1.3 Literature on Markov-modulated fluid models

1.3.1 Introduction

The amount of papers in which fluid queues play a role is enormous. In this section we give an overview of the main references for Markov-modulated fluid models (MMFMs). A short outline of other types of fluid models will be given in Section 1.4.

The main reason why the class of MMFMs has attracted so much attention is that they are relevant for modelling certain phenomena in telecommunication networks. We will concentrate on papers which have contributed to the theoretical development of these models, rather than looking at the practical contexts in which they have been proposed.

In the literature, relatively much attention is paid to models in which the buffer is infinitely large, because this case is easier to analyse. Furthermore, for most practical situations in telecommunications the infinite buffer case is a good approximation for the finite buffer case, since overflow of the buffer is supposed to be extremely rare. The overflow probability for a finite buffer (typically in the range of 10^{-8} to 10^{-12}) can then be approximated by the *overshoot* probability $P[C > y]$ in the corresponding infinite buffer model.

In the following, various aspects of Markov-modulated fluid models will be discussed. In Section 1.3.2 we explain how the use of fluid models for queueing situations that are essentially discrete has been motivated. The next section shows early developments in methods that are aimed at finding the stationary distribution of the process (X_t, C_t) in the basic model of the previous section. Then, in Section 1.3.4, more recent extensions, as well as other aspects receive some attention, such as transient behaviour and output characterisation of the basic model, and simple Markov-modulated networks of fluid queues.

Throughout, we will stick to the notation introduced in the previous section as much as possible, regardless of the notation used in the papers which are described. A final remark concerns the fact that some papers have a more general setup than described here, or consider other models as well. Although we realise that this may not do the authors justice, we will not pay attention to these aspects.

1.3.2 Justification for the use of fluid models

In most queueing situations of interest, discrete entities have to be serviced, e.g. customers, data cells, etc. Therefore the question arises how it can be justified that fluid models, in which a continuous entity plays the central role, can be used to describe such situations.

The central point here is that random phenomena may play a role at various *time scales*. When the variations on the smaller time scale have less impact than those on the larger time scale, the use of fluid models can be justified.

This intuition behind the use of fluid models is often emphatically present in telecommunication networks, where *bursts* of data are usually transmitted in many smaller-sized data packets or *cells*. Here, the use of fluid models is particularly useful, since the variations on the cell level are almost neglegible compared to those on the more important burst level. In many papers, this observation was made. Several papers were written in which MMFMs were favourably compared with other candidates for modelling traffic in telecommunication networks, or otherwise shown to give good approximations for the actual behaviour of network traffic. Therefore they seemed to make more complicated models superfluous.

A more theoretical basis for the intuitive idea of time scales above has been offered in [71], where the authors look at the behaviour of a multiplexing queue. The "approximating" fluid queue is not regarded as an approximation of the real queue, but rather as one of its components, accounting for the long-term correlations in the arrival process (in other words, describing the burst-scale behaviour). The second, cell-scale component must be added to this, to account for the local fluctuations of the cell arrival rate around the fluid average. We note that the fluid component may be a Markov-modulated process, but also a fluid input process with a more general density, (see Section 1.4). The cell-scale component is related to a $\sum D_i/D/1$ queue.

A more general justification for the use of fluid models is offered in [21], where a $G/G/s$ queue in a random environment (X_t) is analysed. The time scale at which the random environment changes states is much larger than that at which arrivals and service completions take place. Furthermore, in some of the environment states the traffic intensity exceeds 1, leading to growing queue length and workload. By normalizing appropriately, these processes can be shown to converge to fluid processes, irrespective of the particular choice for the distributions of interarrival and service times (in other words only the first moments are important). It is also shown that the limit can be refined such that the limiting process is a diffusion process. An important and obvious conclusion is that it is natural to study the more tractable approximating fluid (and diffusion) processes than the real queueing systems.

1.3.3 Finding the stationary distribution

In this section we give an overview of the “early” literature regarding the *stationary behaviour* of MMFMs. The reason why many papers are concerned with the stationary distribution of the process (C_t) or (X_t, C_t), is that these provide much information that may be of interest for practical applications, such as tail probabilities, expected buffer content, expected delay and sojourn time and traffic intensity of in- and outgoing traffic. Loss probabilities can be found from finite-buffer models or approximated using models in which the buffer is infinitely large.

An early model which can be viewed as a MMFM is described in [98]. In the context of production facilities, a finite fluid buffer is considered which is fed and emptied by two unreliable production units with exponential up- and down-times. In other words, it is a MMFM driven by a four-state Markov process for which the stationary distribution is found.

Another early paper on MMFM is [39] which looks into a fluid approximation for a system in which transmission capacity that is not used by voice calls is used for the transmission of data packets. Both voice calls and data packets arrive according to Poisson processes and have exponential holding times, however they operate at different time-scales. Thus, the data can be approximated as fluid, with a constant input rate into the buffer, while the output varies, being regulated by the number of active voice calls.

Almost half a year later, the paper which would become *the* main reference for work in the area of MMFM was published by Anick, Mitra and Sondhi [6]. They give an elegant analysis of a simple fluid model (which, by the way, had been considered many years before in [78] according to [26]). The model describes an infinitely large fluid reservoir which is fed by n identical, exponential on-off sources and emptied by an output channel with constant capacity. Thus, the net input is regulated by a specific *birth-death process* (X_t) with state space $\mathcal{N} = \{0, \ldots, N\}$ and $r_i = a\,i - c$. The differential equation for the stationary joint distribution of (X_t, C_t) is derived, and the *spectrum* of the key matrix $R^{-1}Q^T$ is analysed. Boundary conditions are given and used to find the constants in the solution. Thus, the procedure in Section 1.2 results for this special case, leading to $F_i(y)$.

In a way, the model in [6] was a generalisation of the earlier model by Kosten [52], where a limiting case is considered. In particular, both the average off period and the number of sources (N) approach infinity, in such a way that the total traffic intensity remains finite. Since the state space $\mathcal{N}$ of the regulating process X_t is infinitely large in this model, we will discuss [52] in more detail in Section 1.3.4 together with other papers that consider models with this property.

In [54], it was Kosten's turn to generalize the model in [6]. A fluid buffer is analysed which is fed by a number of groups of i.i.d. exponential sources. By decomposing the output capacity, the eigenvalues and -vectors can be found in principle, which makes it possible to find the stationary buffer content. As an aside, we note that later, in [16] it is argued that the determination of the coefficients c_i (see Section 1.2) still constitutes a considerable problem. Therefore, by fitting certain characteristics, they approximate the system by a birth-death fluid queue.

In the last paper by Kosten on a MMFM, [55], he is once more ahead of his time. Based on the observation that for general MMFMs it is difficult to determine the full stationary distribution, he focuses on the *decay rate* of the buffer, i.e. the largest negative eigenvalue of $R^{-1}Q^T$. After all, for the analysis of loss-probabilities the asymptotic behaviour of $P[C > y]$ for large y is of particular importance, and from (1.7) it is easily seen that $P[C > y] \sim A\,exp(\xi_1 y)$ as $y \to \infty$, where ξ_1 is the decay rate and A is some constant. However, since the subject of this section is stationary behaviour – rather than asymptotic behaviour – we will quickly continue the course of our story.

In [31] we find another way in which the model in [6] has been generalized. They retain a birth-death structure for the regulating process, but allow general transition probabilities. The only extra condition imposed is that there is a state $k \in \mathcal{N}$ such that $r_i < 0$ for $i \leq k$ and $r_i > 0$ for $k < i \leq N$. Using the theory of orthogonal polynomials an explicit representation for the stationary joint distribution was found.

Apparently unaware of [31], a slightly less general birth-death fluid model is considered in [22]; however they mainly pay attention to numerical aspects.

Meanwhile, Mitra [69] had been working in another direction. Not only does he show how to deal with states $i \in \mathcal{N}$ for which $r_i = 0$, he also generalizes the model of [6] in two ways. Again, he considers a buffer which receives input from N i.i.d. exponential sources, here called producers. However now also the output is assumed to be Markov-modulated, namely by M i.i.d. exponential consumers. Thus, the net input is regulated by a (time-reversible) Markov process on the product space of the state spaces of the input and output processes.

A second generalisation is that he also considers the case of a finite reservoir. (In the same year [96] solves a particular model with finite reservoir). Among other things, Mitra states a theorem for the eigenvalues of the equation $RF'(y) = QF(y)$ for the general case of a fluid reservoir regulated by a *reversible* Markov process. In particular he shows that all eigenvalues are real and that the number of negative ones equals N_+, the number of positive diagonal elements in the matrix R. This had been proven ad hoc in other papers, for instance in [31] for the particular case of a birth-death fluid queue, where it is also proven that the eigenvalues are simple. Also in [92], the eigenvalue-structure receives attention. In this manuscript it is shown that, for the more general case in which the regulating process need not be reversible, the number of eigenvalues with negative real part is equal to N_+, and that the eigenvalue with the smallest negative real part is simple and real, while others may coincide and be non-real.

In [93] a fluid queue is driven by a *separable* Markov process, which means that the rate process can be seen as a superposition of independent, not necessarily identical, reversible Markov processes. Hence, this paper presents a generalisation of [54]. The state space explosion problem is solved by a decomposition for the equations for the equilibrium probabilities, similar to the one in [54]. The complexity of the problem is thus reduced from $\prod N_k^3$ to $\sum N_k^3$, where the constants N_k are the sizes of the state spaces of the independent Markov processes which generate the total net input. For a special case of this model, namely the superposition of two types of sources, approximations were developed in [41].

A next step is the analysis of fluid queues which are regulated by a Markov process with a *nearly completely decomposable* state space. These models arise in situations where some sources change states on very large time scales, while others operate on smaller time scales. In [51] the stationary distribution is approximated by solving the subsystems that result from a decomposition, and solving an "aggregative" system.

1.3.4 Extensions and other aspects

Alternative solution procedures

We mention two alternatives to the classical, i.e. spectral analysis, method for solving the stationary distribution of a MMFM that have been considered in the literature. In [9], a probabilistic analysis yields that the joint distribution of (X, C) is of phase type; its parameters can be found by an iterative procedure and the relation between these parameters and the solution in the form of (1.7) is given.

Another approach is to look first at the embedded content process at epochs when (X_t) changes states. For this discrete time process the stationary distribution can be found using *Wiener-Hopf factorization.* From this distribution the continuous-time stationary distribution can be found. See further [81, Chapter 4], [84], [85] and [74].

Infinite-state regulating processes

In all papers described until now, the state space $\mathcal{N}$ of the regulating process is finite, apart from [52]. Although this is not a recent reference, we will discuss it in this section, since it fits well with the other papers here. The model in [52] is that of a fluid reservoir that receives fluid from messages that arrive according to a Poisson process. Upon arrival, the content of a message flows into the reservoir at rate 1; the holding time of each message is exponentially distributed. Thus, the model can be characterised as a MMFM driven by a process (X_t), where X_t is the number of messages in an $M/M/\infty$ system and $r_i = a\,i - c$. The analysis, which reminds the reader of that in the well-known and more recent paper [6], leads to a system of equations from which $\mathbf{F}(0)$ can be found; $\mathbf{F}(y)$ can then be computed numerically. Also the moments of the buffer content distribution can be found, and the asymptotical behaviour is studied. The second and third parts of this paper, published separately in [53] and [56], deal with generalisations of the model in which the holding times of messages are assumed to have a (generalised) Erlang distribution and a hyperexponential distribution, respectively.

Many years passed before the same model as in [52] was investigated again in [73]. Here, an exact representation for the transform of $\mathbf{F}$ is given. Independently, the model was studied in [76], where it is called the $M/M/1$ model with gradual input. After a moment's thought it becomes clear why this is also a good characterisation for the same $M/M/\infty$-driven system (for a short introduction to models with gradual input we refer to Section 1.4). Its main result is an explicit expression for the transform of the stationary buffer content distribution.

Another model in wich $|\mathcal{N}| = \infty$ was introduced by Virtamo and Norros [97]. Here, X_t is the number of customers in an $M/M/1$ queue. The net rate into the fluid reservoir is $r > 0$ when $X_t > 0$ and -1 when $X_t = 0$. The stationary distribution is found via the spectral analysis of the matrix $R^{-1}\tilde{Q}^T$, where $\tilde{Q}$ is the symmetrized version of the matrix Q. A simpler approach to obtain the same result is given in [4], where the relation with the classical $M/G/1$ model is exploited in a similar way as in [46].

The model of [97] will be revisited in Chapter 2, where we present procedures which can be applied to various Markov-modulated fluid queues that are driven by (infinite-state) birth-death processes.

Fluid networks

In contrast to the vast body of literature on Jackson networks, the amount of literature on networks of fluid queues is small, one reason being that the stationary joint distribution of the contents of the reservoirs is not of product form. Also in the class of MMFMs there is hardly any literature on models involving more than one reservoir. In fact all three papers on this topic consider the same model. This model was first introduced in [100]. Two types of arriving traffic are buffered, each in a separate reservoir. A constant capacity server (output channel) serves both buffers such that buffer 1 has priority over buffer 2; hence buffer 2 is served with the rate which is not used for buffer 1. Thus, if we let (C_t^i) be the content process of buffer i, and (X_t) the background Markov process which drives both input processes, it is clear that (C_t^2) is regulated by (X_t, C_t^1). Hence, in fact we have a MMFM where the regulating process has an infinite, nondenumerable state space, and is in fact a MMFM itself. An implicit expression is found for the double Laplace transform of the stationary joint distribution of (C_t^1) and $(C_t^1 + C_t^2)$. For the case of a two-state Markov process the expression can be made explicit, while in the general case first and second moments can be found. For the general case, an explicit expression was found in

[20] for the same joint distribution (now Laplace-transformed for the second variable only) in terms of the eigenvalues and -vectors of a certain key matrix. For the two-state driving Markov process, the latter are expressed explicitly in terms of the parameters of the model, leading to the same result as in [100].

The same model is treated in [36], where the output process of buffer 1 is approximated by a Markov-modulated fluid process, based on the assumption that the length of a period during which $C_t^1 > 0$ is exponentially distributed. Since a model in which buffer 1 feeds into buffer 2 (together with the low priority fluid input) is equivalent to the original model, this leads to a fast and simple way to find a robust approximation for the stationary distribution of the content of buffer 2.

In Chapter 4 we will look into some Markov-modulated two-buffer fluid models.

Output characterisation and transient behaviour

The characterisation of the output of a fluid reservoir can be interesting for various reasons, e.g. because the output may be the input for a next queueing station or to decide how the burstiness of the traffic is affected (for the latter subject see also [94]).

The main references we like to mention are [1] and [2], where much insight is given into the subject, as well as other references.

An explicit expression for the Laplace transform of the busy period (although computationally inattractive) is found in [8], as well as a procedure to find the mean busy period. However, the mean is more easily obtained using the expression in [10]. (For the MMFM driven by the infinite-state $M/M/\infty$ system, the mean busy period is approximated in [29]). First passage times to empty buffer are studied in [42], also leading to expressions for the distributions of busy and idle periods.

It may be useful to note that in general, busy and idle times only give a very rough impression of the output process. For systems with constant output capacity, the output rate may be unknown during idle times of the buffer, while for other systems the same is true both for idle and busy periods.

In [34], the output of a fluid model for a so-called *dual leaky bucket* is considered. We refer to Section 3.6 for more information on leaky bucket mechanisms, including a short discussion of [34].

One important reference regarding transient behaviour is [95], where the same system as in [93] is considered. The system of partial differential equations for $F_i(t,y) = P[X_t = i, C_t \le y]$ is derived and a representation is found for the Laplace transformation of the solution. Earlier work in this area can be found in [49] and [99].

State-dependent input

The overview paper [61] cites [35], where a model is considered in wich the inflow of fluid not only depends on the state of the regulating Markov process, but also on the content of the fluid reservoir. In particular, a number of thresholds $0 = B_0 < \cdots < B_m,\ m > 0$, is assumed in the reservoir (where B_m is the size of the reservoir, possibly infinity), such that the regulation of C_t by X_t is determined via the rate matrix R_j at times t when $B_{j-1} \leq C_t < B_j$, $j = 1, 2, \ldots, m$, instead of via a constant matrix R as is usually the case. The system of differential equations, as well as the corresponding boundary conditions are derived and solved step by step. We will come back to this model in Section 1.5.1.

1.4 Connection with other types of fluid models

In this section we will go over some other types of fluid models. However, it is not our intent to give a comprehensive overview; rather, we will focus on the connection with Markov-modulated fluid models.

Lévy Processes

A well-known type of input, used in the context of dam models, is a Lévy process, which can be seen as the continuous-time analogue of a random walk. To explain this, let the amount of fluid arriving at a buffer during the time interval $[0, t]$ be denoted by A_t. It is then said that (A_t) is a Lévy process when: *i.* (A_t) has stationary and independent increments; *ii.* (A_t) is continuous in probability, and its sample paths are right-continuous with left limits. (see [80, Chapter 3]). In general (A_t) is the independent sum of a deterministic linear drift with rate r (say), a Brownian motion with variance σ^2; and a pure jump process with Lévy measure $\nu(dy)$ (see [7, Section III.8]). If $\lambda = ||\nu|| < \infty$, the jump process is a compound Poisson process with rate λ and jump size distribution given by ν/λ.

Clearly, the process (A_t) can be decreasing at times. Hence it can be useful to describe the behaviour of the content of a fluid buffer or dam, if we prevent negative values by adding a reflecting boundary at zero. However, if we want (A_t) to describe an input process, it seems natural to demand that its sample paths are non-decreasing. Thus, we arrive at the general *non-decreasing* Lévy process, which can be written as the sum of a linear drift and a compound Poisson process with positive jumps.

We note that, although (A_t) is a Markov process, a fluid model with input (A_t) is not Markov-modulated in general, in the sense that there is no background Markov process which determines the instantaneous arrival rate of fluid. This is in fact *only* the case when the jump component is zero, leading to continuous linear (deterministic) input.

Markov Additive Process

A generalisation of the Lévy process is the so-called *Markov Additive Process* (MAP). We will restrict ourselves to MAPs with an underlying finite-state Markov process (X_t), see e.g. [11, page 441]. At times when $X_t = i$, let the process (A_t) evolve like a Lévy process with linear drift at rate r_i, Brownian motion component with variance σ_i^2 and pure jump component given by the Lévy measure $\nu_i(dy)$. Finally, when $X(t)$ makes a transition from state i to state j, we allow (A_t) to have a jump with probability p_{ij}, the size of which has distribution function B_{ij}. This defines a Markov additive process (A_t) with underlying Markov jump process (X_t), or, as it is also described in the literature, a Markov additive process (A_t, X_t) with Markov component (X_t) and additive component (A_t). We note that (A_t, X_t) is a Markov process, but in general (A_t) is not.

Clearly, The MAP is a proces with a very general structure. To be useful as input process, [75] defines a *MAP of arrivals* by demanding that $A_t > 0$, implying $r_i > 0$, $\sigma_i = 0$ and allowing only nonnegative jumps. It has many input processes as special cases, e.g. the Markov-modulated Poisson process (MMPP), Neuts' N-process (also known as Markovian arrival process, which unfortunately is also abbreviated to MAP), batch Markovian arrival process (BMAP) and of course the nonnegative Lévy process and its special cases (Poisson process, compound poisson process). Last but not least, we mention Markov-modulated fluid input.

In fact, the buffer content process (C_t) in a Markov-modulated fluid model can be seen as (the additive component of) a MAP (A_t), reflected at zero. To see this we set $\sigma_i = 0$, and allow no jumps to occur (while r_i can be positive and negative, as before). As a consequence, the sample paths of (A_t) are *piecewise linear*, so that we obtain the standard MMFM of Section 1.2. (For a survey of processes with piecewise linear sample paths we refer to [27].) We notice that MMFMs in which $\mathcal{N}$ is infinitely large have piecewise linear sample paths only when the set $\{r_i \mid i \in \mathcal{N}\}$ is denumerable (assuming that (X_t) is nonexplosive), which is the case for all work in this thesis. In the next subsection we will see an example of a model where this is not the case.

Input processes with a continuous density

In this section we briefly discuss models in which the input process (A_t) has a density, i.e. we can write $A_t = \int_0^t X_s ds$. The instantaneous net input rate, or density, is assumed to be a stochastic process (X_t). Many input processes have this property, including standard MMFMs (as in Section 1.2 with $\mathcal{N}$ finite), when we allow the process (X_t) to have jumps. In the references we consider here, the process (X_t) varies continuously, its state space typically given by $\mathbb{R}$. It follows that when (X_t) is a Markov process, the model formally belongs to the category of MMFMs, which is easily seen by taking $\mathcal{N} = \mathbb{R}$ and $r_i = i$ in Section 1.2. Specific examples of these models can be found in [90], [91] and [62], see also [61], where (X_t) is an Ornstein-Uhlenbeck process. This model can be seen as a limiting case of the model in [6] in heavy traffic when we let $N \to \infty$. More general (Gaussian) input processes are considered in [50] and [28].

Two-state random environment

When we think of a fluid analogue of the standard $G/G/1$ model, we can do so in two ways; these will receive some attention in this and the following subsection. The first type of model is characterized by alternating on- and off periods. During on-periods fluid flows into the buffer at a constant rate, while during off-periods no fluid arrives. Thus, assuming that the output rate is constant, we again have a piecewise linear process.

One way of looking at these models is by considering the buffer to be regulated by an *on-off source*. This view is often encountered in the telecommunications literature, where the distribution of the on-times is nowadays often chosen such that it has a *heavy tail.* Another view, inspired by manufacturing systems, is presented in [18]. In this paper the on- and off-periods are called down- and up-periods respectively, as they describe the state of the output channel; the down-periods are also called *random disruptions.* They find the Laplace transform of the stationary distribution without any assumptions on the up- and down-times.

Notice that when both on- and off-times (or down- and up-times) are i.i.d. sequences of random variables with phase-type distributions, this model belongs to the class of MMFMs. When both times are exponentially distributed, we obtain a fluid model driven by the well-known *exponential on-off source.*

In [46] the concept of the (not necessarily Markovian) two-state random environment is also used, but the deterministic linear flows during up- and down-times are generalized to deterministic, nonlinear flows and to stochastic flows.

Queueing models with gradual input

The other generalisation of the $G/G/1$ model is more natural and has been introduced as early as 1974 by Cohen in [24]. Here the arriving traffic is characterised by (t_n, S_n), where t_n is the time at which the n-th *burst* starts being active, and S_n is the lenght of the time period it remains active. Each active burst produces fluid at rate 1, while the output rate of the reservoir is also 1. Note in particular that in this context more than one burst can be active simultaneously, giving rise to inflows at possible rates $0, 1, 2, \ldots$. In [24] several results are obtained by exploiting the connection with traditional queueing models.

In [76] the situation was studied in which both the interarrival times $(t_{n+1} - t_n)$ and the workloads (S_n) are exponentially distributed. As a consequence, this model can be seen as a MMFM, and has as such received some attention in Section 1.3.4. The generalization to the $G/G/1$ burst model came in [48], where a formula is derived for the Laplace-Stieltjes transform of the workload. See also [89], where first moments were found based on a pathwise comparison with the regular $G/G/1$ queue.

1.5 Contributions in this thesis

Before giving an overview of the subsequent chapters, we will introduce the main contributions of this work. In Section 1.5.1 we introduce the notion of feedback, while Section 1.5.2 concentrates on the state space $\mathcal{N}$ of the regulating process (X_t) in the various models.

1.5.1 Feedback

In this section we introduce the notion of feedback for fluid queues. We start off by mentioning that it is totally different from the well-known type of feedback that we encounter in traditional queueing systems. In fact, assimilating this type of feedback in the context of fluid models would by its very nature be meaningless, since it could well be described by an ordinary fluid queue fed by a different input process.

The feedback fluid models we will investigate have much in common with Markov-modulated fluid models. In particular we have that the rate of change of the content C_t of the fluid reservoir is determined by the current state of a stochastic process (X_t) evolving in the background. However, the evolution of this regulating process is no longer autonomous (let alone Markovian), but depends on the current state of the fluid reservoir. In other words, the processes (X_t) and (C_t) now *interact*, since the dependence works both ways.

To narrow down the variety of feedback fluid models that one can think of, and to hold out a prospect of success in analysing them, we confine ourselves to models with the following property. We will assume that the process (X_t) behaves like a Markov process $(X_t^{(0)})$ when the fluid reservoir is empty, and that it behaves as another Markov process $(X_t^{(1)})$ on the same state space $\mathcal{N}$ otherwise. Furthermore, the process (X_t) does not change states with positive probability at times when the reservoir becomes empty. (Note that it *does* change states when the reservoir starts filling up after an idle period, since the net input into the reservoir can only change from negative to positive due to a state transition of the process (X_t)). The main consequence is that, although the process (X_t) does not constitute a Markov process, the joint process (X_t, C_t) *does*.

As far as we know, the models at hand are the first to be analysed in which the dependence works both ways. Hence, an important motivation for studying these models has been to investigate whether such behaviour yields to analysis. In this thesis it is shown that this appears to be the case.

We like to mention one other model in which feedback is distinctly present. In [35] a MMFM with state-dependent input was considered; we refer to the description of this model in Section 1.3.4 and adopt the notation thereof. The reason why we come back to this model in the present context is the following. If we take the rate matrix R constant, but instead make *the Q-matrix* of the process (X_t) depend on the state of the fluid reservoir, we obtain a feedback fluid model. In the resulting model we have that (X_t) evolves as a Markov process $(X_t^{(k)})$ with generator Q_k at times t when $B_{k-1} \leq C_t < B_k$. Although we did not examine this thoroughly, it seems that this model can be solved, at least in principle, in the same way as the model in [35].

As an aside we mention that, at least formally, the model in [35] itself can also be described as a feedback fluid model. The, somewhat unnatural, way to do this, is to view the model as one in which a fluid reservoir is being driven via a rate matrix $\operatorname{diag}(R_1, \ldots, R_m)$ by a process (X_t, Y_t) with state space $\mathcal{N} \times \mathcal{M}$ and generator $\Delta^{(k)} \oplus Q$ at times when $B_{k-1} \leq C_t < B_k$; here $\mathcal{M} = \{1, 2, \ldots, m\}$, $\oplus$ denotes the Kronecker sum, and $\Delta^{(k)}$ is the $m \times m$ matrix with elements $\Delta_{ij}^{(k)} = \delta_{jk} - \delta_{ij}$, $i, j, k \in \mathcal{M}$, where δ_{ij} denotes Kronecker's delta. As a consequence $Y_t = k$ at times when $B_{k-1} \leq C_t < B_k$.

1.5.2 Infinite state space $\mathcal{N}$

In almost all Markov-modulated fluid models in the literature, the state space $\mathcal{N}$ of the regulating process (X_t) is finite, three exceptions being mentioned in Section 1.3.4. In the models we will encounter, $\mathcal{N}$ is mostly infinitely large, namely either countably infinite, in Chapters 2 and 3 (apart from Sections 2.3 and 3.2), or nondenumerable, in Chapters 4 and 5.

Although little attention has been paid to fluid models driven by infinite-state Markov processes, it appears that their analysis need not always be much more complicated than that of the more standard models. In fact we can easily distinguish between relatively "simple" and more "difficult" models. To do so, we recall the subdivision in $\mathcal{N}$ that was made in Section 1.2 in $\mathcal{N}_-$ and $\mathcal{N}_+$. It will turn out, roughly speaking, that we can call a model simple when $\mathcal{N}_+$ is finite and the fluid reservoir is infinitely large.

For a schematic overview of various possibilities for fluid models which are driven by a Markov process with a not necessarily finite state space, we refer to Figure 1.1. The reader should keep in mind the three assumptions made in this figure, namely that $r_i \neq 0$ for any $i \in \mathcal{N}$, that the buffer is infinitely large (otherwise there is no essential difference between the north-east and the south-west parts of the figure), and that the set $\{r_i \mid i \in \mathcal{N}\}$ is denumerable (so that the process (C_t) has piecewise linear sample paths).

Some relevant references are indicated in Figure 1.1, while the section numbers show how the Markov-modulated models in this thesis fit into the framework. This can be helpful when reading the following overview; the same is true for Table 1.1, where the connection between Chapters 2 – 5 is clarified.

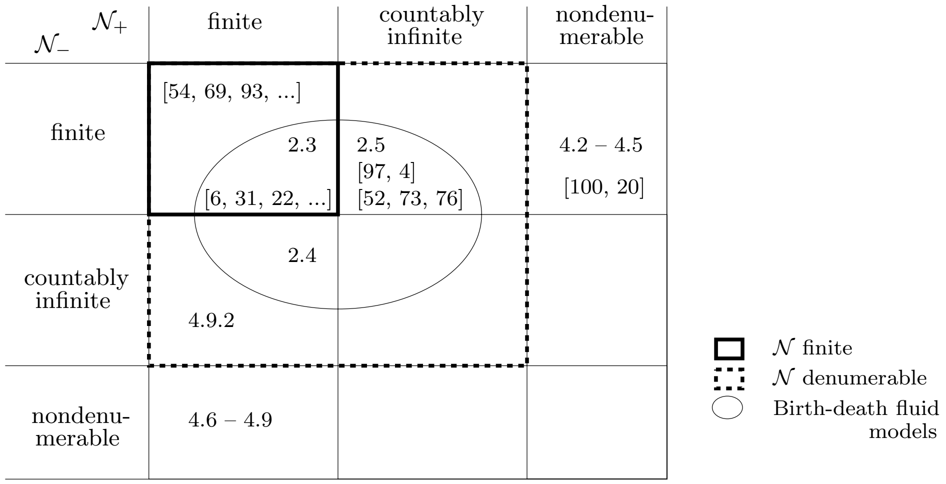


Figure 1.1: Various possibilities for $\mathcal{N}$

State space $\mathcal{N}$	Without feedback	With feedback
Countably infinite (mostly)	Some birth-death fluid models (Chapter 2)	A birth-death fluid model with feedback (Chapter 3)
Non-denumerable	Some two-buffer fluid models (Chapter 4)	A two-buffer fluid model with feedback (Chapter 5)

Table 1.1: The various models in this thesis

1.5.3 Overview

Beforehand, we note that in all models in this thesis we concentrate heavily on finding the stationary distributions of the processes involved.

In Chapter 2 we do so for some MMFMs in which the fluid reservoir is infinitely large and the regulating process (X_t) has a birth-death structure. In view of Figure 1.1 we have four possibilities, since both N_+ and N_- can be either finite or infinite. The situation in which $N_+ = N_- = \infty$ does not seem to yield to analysis and is therefore not studied. In the other three cases we present representation formulas for $\mathbf{F}$, using the theory of orthogonal polynomials. For each of the remaining two cases in which $\mathcal{N}$ is infinitely large we give an example in which explicit results can be found. In these examples, X_t can be interpreted as the number of customers in an $M/M/1$ queueing system, while the respective rate structures (characterised by the matrix R) are dual in some sense. The exemplary model for the case $N_- < \infty$ has been analysed before in [97] and [4]. Chapter 2 is based on [32] and [33].

In Chapter 3 we leave the world of MMFMs and concentrate on a particular feedback fluid model. This model is closely related to the first exemplary model in Chapter 2, the essential difference being that the service rate of the $M/M/1$ system now has a different value at times when $C_t = 0$. It turns out that a similar solution procedure works well here. However, in this chapter we also look into the situation where the fluid reservoir has a finite capacity. Since an analytic solution does not seem feasible in this case, we find an approximating procedure for computing the quantities of interest, involving a discretisation of the fluid reservoir. Chapter 3 is based on [3].

In Chapter 4 we analyse two particular models in which two infinitely large buffers play a role. The content of the first buffer, D_t, is regulated by a two state Markov process (M_t), while the content of the second buffer, C_t is regulated by the first one. For the latter regulation we consider two possibilities, one of which describes a tandem fluid queue. The two models can be considered dual in the same sense as the examples in Chapter 2. In fact they are "almost" generalisations of these examples. In particular, they can be categorized in the class of MMFMs. This can be understood by taking $(X_t) = (M_t, D_t)$ as the regulating Markov process for the second buffer. Notice that the state space of this process is non-denumerable. The stationary marginal distribution of the process (C_t) in both models is found by establishing a connection with the classical $M/G/1$ and $G/M/1$ models respectively. The joint distributions are found via a Laplace approach, that appears to be a powerful alternative to the classical spectral approach . For one of the models we also show how this spectral approach can be employed, illustrating why the solutions to both models are so different. Chapter 4 is based on [60] and [59].

In Chapter 5 we consider another two-buffer fluid model, this time in the presence of feedback. Thus, again the process (C_t) is regulated by $(X_t) = (M_t, D_t)$, but the net input rates for the first reservoir adopt different values at times when $C_t = 0$. Another difference is that the second buffer is assumed to be finite (the infinite case being included as a special case). Also it is shown how the results can be applied to the case where the first buffer is finite, provided that it is not too small. The results obtained include the stationary joint distribution of the process (M_t, D_t, C_t), which is obtained by combining a regenerative approach with the results in Chapter 4. Chapter 5 is based on [87]

Chapter 2

Some birth-death fluid models

2.1 Introduction

In this chapter we shall study the stationary behaviour of the content of a fluid reservoir which receives and releases fluid flows at rates which are determined by the actual state of an ergodic birth-death process evolving in the background. The reservoir is assumed to be infinitely large, which implies that for the stationary distribution of the content of the reservoir to exist it is necessary that some stability condition be satisfied.

The state space of the background birth-death process will be denoted by $\mathcal{N}$ and may be finite or infinite; in the former case $\mathcal{N} = \{0, 1, \ldots, N\}$ for some natural number $N \geq 1$, in the latter case $\mathcal{N}$ is the set of nonnegative integers. We shall denote the state of the background process at time t by X_t and the content of the reservoir at time t by C_t. The obvious approach to obtaining the stationary distribution of the process (C_t) is by analysing the two-dimensional process (X_t, C_t), which is Markovian.

Our assumption that the flow rates of fluid into and out of the reservoir are determined by the current state of the background process, entails that for each $i \in \mathcal{N}$ there is a real number r_i, the *drift* in state i, such that r_i is the slope of (C_t) when the birth-death process is in state i, as long as this is physically possible. That is, the rate of change of the content of the reservoir (or the *net input rate*) at time t is r_{X_t}, provided $r_{X_t} \geq 0$, or $r_{X_t} < 0$ and $C_t > 0$; if the reservoir has emptied at time t it stays empty as long as the drift remains negative. We shall assume throughout that $r_i \neq 0$ for all states $i \in \mathcal{N}$. We shall also assume that $r_i > 0$ for at least one $i \in \mathcal{N}$, since otherwise the reservoir is always empty.

When $\mathcal{N} = \{0, 1, \ldots, N\}$ for some natural number N, the model at hand is a generalization of the fluid flow models studied in [6], [39], [31] and [22], see Section 1.3. Although the latter two allow (X_t) to be an arbitrary birth-death process, they require the drift matrix R to have a particular sign structure. In Section 2.3 we generalize their results, the only remaining condition being that R should be non-singular, i.e. $r_i \neq 0$, $i \in \mathcal{N}$.

Relatively few results are available in the literature dealing with (variants of) our model when $\mathcal{N}$ is infinite, see Section 1.3.4. In Section 2.4, we take the approach of letting N tend to infinity in the expressions obtained for the truncated model in which $\mathcal{N} = \{0, 1, \ldots, N\}$.

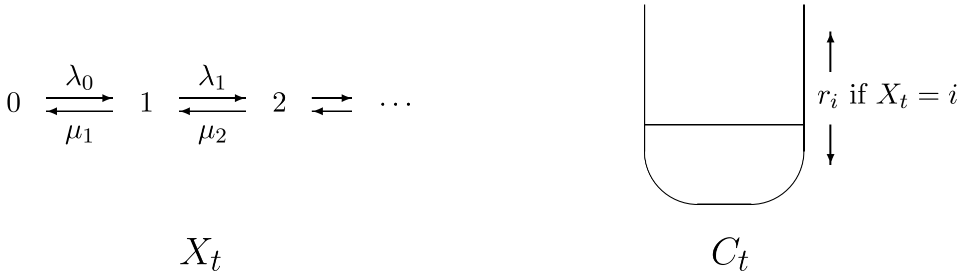


Figure 2.1: Regulation of the process (C_t) by (X_t)

It appears that this is a viable procedure whenever N_+, the number of positive components of the drift vector $(r_0, r_1, \ldots)$ is finite.

In Section 2.5 we deal with the opposite case, namely in which N_-, the number of negative components of the drift vector is finite. We shall present a general procedure for solving models with this property and show that the procedure works by analysing a particular model that has been analysed before in [97] and [4], where entirely different approaches were chosen.

For the sake of completeness we notice that a fourth class of models is conceivable that fall within the category of birth-death fluid queues. In these models both N_+ and N_- are infinite. However, suchlike models are not described here, since their analysis, if at all possible, appears to be more complicated.

The various analyses in this chapter amount to solving a finite (in Section 2.3) or infinite (in Sections 2.4 and 2.5) system of differential equations under certain boundary conditions. The derivation of this system of differential equations will be outlined in Section 2.2, where also the notation will be introduced. Much of this section is similar to Section 1.2, but we include it for the sake of completeness.

2.2 Preliminaries

We shall let λ_i denote the birth rate and μ_i the death rate in state i, $i \in \mathcal{N}$, of the birth-death process (X_t) with state space $\mathcal{N}$ which regulates the content of the reservoir. We shall assume that the birth and death rates are positive with the exception of the death rate μ_0 in the lowest state and, if $\mathcal{N} = \{0, 1, \ldots, N\}$, the birth rate λ_N in the highest state, which are zero. Also, it will be convenient to interpret λ_i and μ_i as zero if $i \notin \mathcal{N}$. Now that we have introduced all parameters of the model, we refer to Figure 2.1 where the behaviour of the processes (X_t) and (C_t) is illustrated.

Focussing on (X_t) for a while, we let

$$\pi_i \equiv \prod_{j=0}^{i-1} \frac{\lambda_j}{\mu_{j+1}}, \quad i \in \mathcal{N}, \tag{2.1}$$

where the empty product should be interpreted as unity. The stationary state probabilities $p_i,\ i \in \mathcal{N}$, of the birth-death process can then be represented as

$$p_i = \frac{\pi_i}{\sum_{j\in\mathcal{N}} \pi_j}, \quad i \in \mathcal{N}. \tag{2.2}$$

When $\mathcal{N}$ is infinite we shall always assume that the stationary distribution of the birth-death process exists, that is, $\sum_{i\in\mathcal{N}} \pi_i$ is finite. In order that a stationary distribution for (C_t), the content of the reservoir at time t, exists, the mean drift should evidently be negative, that is, $\sum_{i\in\mathcal{N}} p_i r_i < 0$, or, equivalently,

$$\sum_{i\in\mathcal{N}} \pi_i r_i < 0. \tag{2.3}$$

We shall assume throughout this chapter that this stability condition is satisfied.

As before we let

$$\mathcal{N}^+ \equiv \{i \in \mathcal{N} |\ r_i > 0\}, \quad \mathcal{N}^- \equiv \{i \in \mathcal{N} |\ r_i < 0\}, \tag{2.4}$$

and

$$N_+ \equiv |\mathcal{N}^+|, \quad N_- \equiv |\mathcal{N}^-|. \tag{2.5}$$

Obviously, $\mathcal{N}^+ \cup \mathcal{N}^- = \mathcal{N}$, since we have assumed that the drift in each state is nonzero. Also, when $\mathcal{N}$ is infinite at least one of N_+ or N_- is infinity.

Putting

$$F_i(t,y) \equiv P[X_t = i,\ C_t \le y], \quad t \ge 0,\ y \ge 0,\ i \in \mathcal{N},$$

and $F_i(t,y) \equiv 0$ if $i \notin \mathcal{N}$, it is not difficult to show that the Kolmogorov forward equations for the Markov process (X_t, C_t) are given by

$$\begin{aligned} \frac{\partial F_i(t,y)}{\partial t} \ &= \ -r_i \frac{\partial F_i(t,y)}{\partial y} \ - \ (\lambda_i + \mu_i) F_i(t,y) \\ &\qquad + \ \lambda_{i-1} F_{i-1}(t,y) \ + \ \mu_{i+1} F_{i+1}(t,y), \quad i \in \mathcal{N}. \end{aligned} \tag{2.6}$$

But since we will assume that the process is in equilibrium, we may set $F_i(t,y) \equiv F_i(y)$ and $\partial F_i(t,y)/\partial t \equiv 0$ and, hence, obtain the system

$$r_i F_i^{'}(y) = \lambda_{i-1} F_{i-1}(y) - (\lambda_i + \mu_i) F_i(y) + \mu_{i+1} F_{i+1}(y), \quad i \in \mathcal{N}, \tag{2.7}$$

where $F_i(y)$ denotes the equilibrium probability that the birth-death process is in state i and the content of the reservoir does not exceed y, again with the convention $F_i(y) \equiv 0$ if $i \notin \mathcal{N}$.

Since the content of the reservoir is increasing whenever the drift is positive, the solution to (2.7) must satisfy the boundary conditions

$$F_i(0) = 0, \quad i \in \mathcal{N}^+. \tag{2.8}$$

Also, we must obviously have

$$F_i(\infty) \equiv \lim_{y\to\infty} F_i(y) = p_i, \quad i \in \mathcal{N}. \tag{2.9}$$

2.3 Finite state space

In this section we will describe the procedure for solving the differential equations (2.7), subject to the boundary conditions (2.8) and (2.9), assuming that $\mathcal{N} = \{0, 1, \dots, N\}$ with $N \geq 1$ and that condition (2.3) is satisfied.

It will be convenient to write the homogeneous system (2.7) in matrix form as

$$\mathbf{F}'(y) = R^{-1}Q^T\mathbf{F}(y), \tag{2.10}$$

where

$$\mathbf{F}(y) \equiv (F_0(y), F_1(y), \dots, F_N(y))^T,$$

$$R \equiv \mathrm{diag}(r_0, r_1, \dots, r_N),$$

and the $(N+1) \times (N+1)$ matrix Q is the generator of the modulating birth-death process, that is,

$$Q \equiv \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & \cdots & \\ \mu_1 & -(\lambda_1 + \mu_1) & \lambda_1 & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & 0 & \mu_{N-1} & -(\lambda_{N-1} + \mu_{N-1}) & \lambda_{N-1} \\ & \cdots & 0 & \mu_N & -\mu_N \end{pmatrix}. \tag{2.11}$$

We start off by deriving a representation formula for the *characteristic polynomial* of the matrix $R^{-1}Q^T$. To this end we define the sequence of polynomials $\{\Delta_i^*(x)\}_{i=0}^N$ by the recurrence relations

$$\begin{aligned} &\Delta_0^*(x) = 1, \quad \Delta_1^*(x) = x + \frac{\lambda_0}{r_0} + \frac{\mu_1}{r_1}, \\ &\Delta_i^*(x) = \left(x + \frac{\lambda_{i-1}}{r_{i-1}} + \frac{\mu_i}{r_i}\right)\Delta_{i-1}^*(x) - \frac{\lambda_{i-1}\mu_{i-1}}{r_{i-1}^2}\Delta_{i-2}^*(x), \quad 2 \leq i \leq N, \end{aligned} \tag{2.12}$$

and observe the following, where I denotes the $(N+1) \times (N+1)$ identity matrix.

Lemma 2.1 *The characteristic polynomial* $\det\left[xI - R^{-1}Q^T\right]$ *of the matrix* $R^{-1}Q^T$ *can be represented as* $x\Delta_N^*(x)$.

Proof. It is easy to see that the statement of the lemma is true if $N = 0$, so in the remainder of the proof we will assume $N > 0$. We define another sequence of polynomials $\{\Delta_i(x)\}_{i=0}^{N+1}$ by the recurrence relations

$$\begin{aligned} &\Delta_0(x) = 1, \quad \Delta_1(x) = x + \frac{\lambda_0}{r_0}, \\ &\Delta_i(x) = \left(x + \frac{\lambda_{i-1} + \mu_{i-1}}{r_{i-1}}\right)\Delta_{i-1}(x) - \frac{\lambda_{i-2}\mu_{i-1}}{r_{i-2}r_{i-1}}\Delta_{i-2}(x), \quad 2 \leq i \leq N. \end{aligned} \tag{2.13}$$

The polynomial $\Delta_i(x)$ can be interpreted as the characteristic polynomial of the $i \times i$ north-west corner truncation of $R^{-1}Q^T$. Upon expanding $\det\left[xI - R^{-1}Q^T\right]$ by its last row we now obtain

$$\det\left[xI - R^{-1}Q^T\right] = \left(x + \frac{\mu_N}{r_N}\right)\Delta_N(x) - \frac{\lambda_{N-1}\mu_N}{r_{N-1}r_N}\Delta_{N-1}(x). \tag{2.14}$$

It can readily be established by induction, however, that

$$x\Delta_i^*(x) = \left(x + \frac{\mu_i}{r_i}\right)\Delta_i(x) - \frac{\lambda_{i-1}\mu_i}{r_{i-1}r_i}\Delta_{i-1}(x), \qquad 0 \le i \le N,$$

which proves the lemma. □

By Favard's Theorem, see e.g. [19, Theorem I.4.4], the polynomials $\Delta_i^*(x), i = 0, 1, \ldots, N$ constitute the first $N+1$ elements of a sequence of *orthogonal polynomials*. It follows, see [19, Theorem I.5.2], that the zeros of these polynomials, and the zeros of $\Delta_N^*(x)$ in particular, are real and simple. We can therefore conclude from the above lemma that the eigenvalues of $R^{-1}Q^T$ are real and simple, with the possible exception of the eigenvalue 0. Since it has been shown, in a more general setting, in [69] and [93], that the matrix $R^{-1}Q^T$ must have N_+ negative eigenvalues, $N_- - 1$ positive eigenvalues and one eigenvalue 0, we can conclude the following.

Lemma 2.2 *The eigenvalues $\xi_j,\ j \in \mathcal{N}$, of $R^{-1}Q^T$ are all real and simple; ordering them in increasing magnitude we have $\xi_j < 0,\ j = 0, \ldots, N_+ - 1,\ \xi_{N_+} = 0,\ \xi_j > 0,\ j = N_+ + 1, \ldots, N$.*

Knowing that all eigenvalues are simple it is straightforward to verify that the solution of (2.10) must be of the form

$$\mathbf{F}(y) = \sum_{j=0}^{N} c_j \exp\{\xi_j y\}\mathbf{v}^{(j)}, \quad y \ge 0, \tag{2.15}$$

where, for each $j \in \mathcal{N}$, the vector $\mathbf{v}^{(j)} \equiv \left(v_0^{(j)}, v_1^{(j)}, \ldots, v_N^{(j)}\right)$ is the suitably normalized eigenvector corresponding to the eigenvalue ξ_j, and c_j is a constant. However, since boundary condition (2.9) must be satisfied, we find that the coefficients c_j corresponding to positive eigenvalues must vanish — that is, $c_j = 0$ for $j = N_+ + 1, \ldots, N$, by Lemma 2.2 — and that $c_{N_+}\mathbf{v}^{(N_+)} = \mathbf{p}$, where $\mathbf{p} \equiv (p_0, p_1, \ldots, p_N)^T$ and p_i is given in (2.2). Consequently, (2.15) reduces to

$$\mathbf{F}(y) = \mathbf{p} + \sum_{j=0}^{N_+-1} c_j \exp\{\xi_j y\}\mathbf{v}^{(j)}, \quad y \ge 0. \tag{2.16}$$

The N_+ negative eigenvalues ξ_j in (2.16) can be found by determining the negative zeros of the polynomial $\Delta_N^*(x)$ of Lemma 2.1. Since $\Delta_N^*(x)$ is an element of a sequence of orthogonal polynomials, very efficient methods exist for finding these zeros, which can be interpreted as eigenvalues of a *symmetric* tridiagonal matrix, see, e.g., [64] and [57]. Since $R^{-1}Q^T$ is a tridiagonal matrix, the eigenvectors $\mathbf{v}^{(0)}, \ldots, \mathbf{v}^{(N_+-1)}$ have nonzero first components. Hence, for $j = 0, \ldots, N_+ - 1$, we can normalize $\mathbf{v}^{(j)}$ to have $v_0^{(j)} = 1$ and subsequently find the remaining components by solving the recurrence relations

$$\begin{aligned} v_0^{(j)} &= 1, \qquad \mu_1 v_1^{(j)} = r_0\xi_j + \lambda_0 \\ \mu_i v_i^{(j)} &= (r_{i-1}\xi_j + \lambda_{i-1} + \mu_{i-1})v_{i-1}^{(j)} - \lambda_{i-2}v_{i-2}^{(j)}, \quad i = 2, 3, \ldots, N. \end{aligned} \tag{2.17}$$

Finally, the constants $c_0, \ldots, c_{N_+-1}$ must be determined by the boundary conditions (2.8), which translate into

$$p_i + \sum_{j=0}^{N_+-1} c_j v_i^{(j)} = 0, \quad i \in \mathcal{N}^+. \tag{2.18}$$

As an aside we note that the system (2.18) can be solved explicitly when the drift vector has a particular sign structure, see [22] and [31].

The above is summarized in the following theorem.

Theorem 2.3 *The stationary joint distribution $F_i(y) \equiv P[X_t = i,\ C_t \le y]$, $i \in \mathcal{N} = \{0, 1, \ldots, N\}$, $y \ge 0$, of the process (X_t, C_t) is given by*

$$F_i(y) = p_i + \sum_{j=0}^{N_+-1} c_j v_i^{(j)} \exp\{\xi_j y\}, \tag{2.19}$$

where ξ_j, $j = 0, \ldots, N_+ - 1$, are the negative eigenvalues of $R^{-1}Q^T$, or, equivalently, the negative zeros of the polynomial $\Delta_N^*(x)$ defined in (2.12), and the constants p_i, $v_i^{(j)}$ and c_j, are determined by (2.2), (2.17) and (2.18), respectively.

2.4 Infinite state space with $N_+ < \infty$

2.4.1 Analysis

In this section our goal is to obtain the solution of the differential equations (2.7), subject to the boundary conditions (2.8) and (2.9), assuming $\mathcal{N} = \{0, 1, \ldots\}$ and $N_+ \equiv |\mathcal{N}^+| < \infty$. Our approach involves truncation of the state space of the birth-death process to the set $\{0, 1, \ldots, N\}$ for some sufficiently large N and letting N tend to infinity in the expressions found for the ensuing finite model by the procedure of the previous section. As we shall see, the viability of this approach hinges on the fact that N_+ is finite.

Concretely, we choose N such that $N > \max \mathcal{N}^+$ and

$$\sum_{i=0}^{n} \pi_i r_i < 0, \quad \text{for all } n \ge N, \tag{2.20}$$

which is always possible since stability condition (2.3) is assumed to be satisfied. Next we truncate the state space of the birth-death process to $\{0, 1, \ldots, N\}$ and make state N reflecting by setting $\lambda_N = 0$. Theorem 2.3 then tell us that for the truncated system, which is stable because of (2.20), the stationary probability that the birth-death process is in state i and the content of the reservoir does not exceed y is given by

$$F_i^{(N)}(y) = p_i^{(N)} + \sum_{j=0}^{N_+-1} c_j^{(N)} v_i^{(N,j)} \exp\{\xi_j^{(N)} y\}, \quad y \geq 0, \ i = 0, 1, \ldots, N, \tag{2.21}$$

where we have indicated dependence on N. Of crucial importance is the fact that the number of terms in the summation appearing in (2.21) equals $N_+ < \infty$ independent of N, which allows us to interchange limit and summation when we let N tend to infinity in (2.21). Before doing so, however, we must determine the limiting behaviour as $N \to \infty$ of the quantities $p_i^{(N)}$, $\xi_j^{(N)}$, $v_i^{(N,j)}$ and $c_j^{(N)}$.

First, it is obvious from (2.2) that

$$p_i^{(\infty)} \equiv \lim_{N\to\infty} p_i^{(N)} = p_i, \quad i \in \mathcal{N}. \tag{2.22}$$

Subsequently turning to the eigenvalues $\xi_j^{(N)}$, $j = 0, 1, \ldots, N_+ - 1$, we can show the following.

Lemma 2.4 *The limits*

$$\xi_j^{(\infty)} \equiv \lim_{N\to\infty} \xi_j^{(N)}, \quad j = 0, 1, \ldots, N_+ - 1,$$

exist and satisfy $-\infty < \xi_0^{(\infty)} < \xi_1^{(\infty)} < \cdots < \xi_{N_+-1}^{(\infty)} < 0$.

Proof. We recall that the polynomial $\Delta_N^*(x)$ defined in (2.12) has negative zeros $\xi_j^{(N)}$, $j = 0, 1, \ldots, N_+ - 1$, while its other zeros are positive. By lifting the restriction $i \leq N$ in (2.12) we make $\Delta_N^*(x)$ element of an infinite sequence $\{\Delta_i^*(x)\}_{i=0}^{\infty}$ which, by Favard's Theorem, constitutes a sequence of orthogonal polynomials, see [19]. The lemma can now be established with the help of two results about zeros of orthogonal polynomials, see [19, Theorems I.5.3 and II.4.6]. Letting x_{ij} denote the jth zero in ascending order (counting from $j = 1$ to $j = i$) of the ith polynomial in an orthogonal polynomial sequence, the first result says that for any fixed j the sequence $\{x_{ij}\}_{i=j}^{\infty}$ is decreasing, so that its limit exists (possibly $-\infty$). Letting $x_j \equiv \lim_{i\to\infty} x_{ij}$, $j = 1, \ldots, i$, and $x_0 = -\infty$, the second result says that if $x_j = x_{j+1}$ for some $j \geq 0$, then $x_j = x_{j+k}$ for all $k = 1, 2, \ldots$. Considering that $\xi_{N_++1}^{(N)}$, the $(N_+ + 1)$st zero of $\Delta_N^*(x)$, is positive for all N, the validity of the lemma is now evident. □

Remark 2.1 Interestingly, the sequence $\{\Delta_i^*(x)\}_{i=0}^{\infty}$ is orthogonal with respect to a positive measure which – in the current setting where $\mathcal{N}_+ < \infty$ – has point masses precisely at the points $\xi_0^{(\infty)}, \xi_1^{(\infty)}, \ldots, \xi_{N_+-1}^{(\infty)}$, while the rest of its mass lies on the positive axis and in 0.

Having established the existence of the limits $\xi_j^{(\infty)}$ we can obviously let N tend to infinity in the recurrence relations (2.17) for $v_i^{(N,j)}$, $i = 0, 1, \ldots, N$, by which we get the infinite system

$$\begin{aligned} v_0^{(j)} &= 1, \qquad \mu_1 v_1^{(j)} = r_0 \xi_j^{(\infty)} + \lambda_0 \\ \mu_i v_i^{(j)} &= (r_{i-1}\xi_j^{(\infty)} + \lambda_{i-1} + \mu_{i-1}) v_{i-1}^{(j)} - \lambda_{i-2} v_{i-2}^{(j)}, \quad i \in \mathcal{N} \backslash \{0, 1\}, \end{aligned} \tag{2.23}$$

where, for convenience, we have written $v_i^{(j)} \equiv \lim_{N\to\infty} v_i^{(N,j)}$.

Next, we turn to the N_+ equations (2.18) for the constants $c_j^{(N)}$, $j = 0, 1, \ldots, N_+ - 1$. It is clear that the constants $c_j^{(\infty)} \equiv \lim_{N\to\infty} c_j^{(N)}$, $j = 0, 1, \ldots, N_+ - 1$, exist and form the unique solution to the N_+ equations (2.18), where $v_i^{(j)}$ must now satisfy (2.23).

Finally, we can let N tend to infinity in the right-hand side of (2.21) and check that the resulting expressions indeed represent the solution to (2.7) – (2.9). Summarizing we have the following.

Theorem 2.5 *When N_+ is finite, the stationary joint distribution $F_i(y) \equiv P[X_t = i,\ C_t \le y]$, $i \in \mathcal{N} = \{0, 1, \ldots\}$, $y \ge 0$, of the process (X_t, C_t) is given by*

$$F_i(y) = p_i + \sum_{j=0}^{N_+-1} c_j v_i^{(j)} \exp\{\xi_j^{(\infty)} y\}, \tag{2.24}$$

where $\xi_0^{(\infty)}, \xi_1^{(\infty)}, \ldots, \xi_{N_+-1}^{(\infty)}$ are the limits in Lemma 2.4 and the constants p_i, $v_i^{(j)}$ and c_j are determined by (2.2), (2.23), and (2.18).

As in the finite case, it is evident that we cannot find explicit expressions for the quantities ξ_j, p_i, $v_i^{(j)}$ and c_j in general. However, in special cases explicit results can be obtained. The following is an example.

2.4.2 Example

We consider a simple model in which

$$r_0 \equiv r_+, \quad r_i \equiv -r_- < 0,\ i = 1, 2, \ldots,$$

so that

$$\mathcal{N}^+ \equiv \{0\}, \quad \mathcal{N}^- \equiv \{1, 2, \ldots\}.$$

Furthermore, the birth and death rates are constant, viz.,

$$\lambda_i \equiv \lambda \ \text{ and } \ \mu_{i+1} \equiv \mu, \quad i \in \mathcal{N}.$$

In the remainder of this section it will be convenient to define $\rho = \lambda/\mu$ and $\sigma = r_+/(r_+ + r_-)$. Since $\pi_i = \rho^i$, stability of the system is ensured if

$$\sigma < \rho < 1, \tag{2.25}$$

which we shall assume in the remainder of this example.

Our main problem is to find $\xi_0^{(\infty)} \equiv \lim_{N\to\infty} \xi_0^{(N)}$, where $\xi_0^{(N)}$ is the smallest zero of $\Delta_N^*(x)$ defined in (2.12). This problem is solved in the following lemma, the proof of which hinges on the fact that the sequence $\{\Delta_i^*(x)\}_{i=0}^\infty$ can, after appropriate renormalization, be recognized as a sequence of *perturbed Chebysev polynomials*, see [19] or [86].

Lemma 2.6 *The sequence* $\{\xi_0^{(N)}\}_{N=1}^\infty$ *constitutes a strictly decreasing sequence with limit*

$$\xi_0^{(\infty)} \equiv \lim_{N\to\infty} \xi_0^{(N)} = -\frac{\lambda}{r_+} + \frac{\mu}{r_+ + r_-} = -\frac{\mu}{r_+}(\rho - \sigma) \tag{2.26}$$

Proof. When we write

$$T_i(x) \equiv \left(\frac{r_-}{\sqrt{\lambda\mu}}\right)^i \Delta_i^* \left(\frac{2x\sqrt{\lambda\mu} + \lambda + \mu}{r_-}\right), \tag{2.27}$$

we see that

$$\begin{aligned} &T_0(x) = 1, \quad T_1(x) = 2x + \frac{\sqrt{\rho}}{\sigma}, \\ &T_i(x) = 2xT_{i-1}(x) - T_{i-2}(x), \quad i = 2, 3, \ldots, \end{aligned} \tag{2.28}$$

so that $\{T_i(x)\}$ constitutes a sequence of perturbed Chebysev polynomials. Since by (2.25) we have $T_1(0) > 1$ we find from [19, Section II.4 and page 205] that the sequence $\{\zeta_i\}_{i=1}^\infty$, where ζ_i is the smallest zero of $T_i(x)$, constitutes a strictly decreasing sequence which converges as $i \to \infty$ to

$$-\frac{1}{2}\left\{\frac{\sqrt{\rho}}{\sigma} + \frac{\sigma}{\sqrt{\rho}}\right\}.$$

Translating this result in terms of $\Delta_i^*(x)$ completes the proof. □

Writing $v_i \equiv v_i^{(0)}$ the recurrence relations (2.23) reduce to

$$\begin{aligned} v_0 &= 1, \qquad v_1 = \sigma \\ \mu v_i &= \left(\frac{\lambda}{\sigma} + \mu\sigma\right) v_{i-1} - \lambda v_{i-2}, \quad i \in \mathcal{N}\backslash\{0,1\}, \end{aligned}$$

which immediately yields

$$v_i = \sigma^i, \quad i \in \mathcal{N}. \tag{2.29}$$

Since (2.18) becomes

$$p_0 + c_0 v_0 = 0,$$

so that $c_0 = -p_0$, and evidently

$$p_i = (1-\rho)\rho^i, \quad i \in \mathcal{N}, \tag{2.30}$$

we finally obtain, for $y \geq 0$ and $i \in \mathcal{N}$,

$$F_i(y) = (1-\rho)\left(\rho^i - \sigma^i \exp\{-\frac{\mu}{r_+}(\rho-\sigma)y\}\right). \tag{2.31}$$

In particular, the stationary marginal distribution of the buffer content process (C_t) is given by

$$P[C > y] = \frac{1-\rho}{1-\sigma}\exp\{-\frac{\mu}{r_+}(\rho-\sigma)y\}, \quad y \geq 0. \tag{2.32}$$

2.5 Infinite state space with $N_- < \infty$

2.5.1 Analysis

We finally consider the case in which $\mathcal{N} = \{0, 1, \ldots\}$ and $N_+ \equiv |\mathcal{N}^+| = \infty$, but $N_- \equiv |\mathcal{N}^-| < \infty$. As announced we shall present an approach to obtain the equilibrium distribution of the content of the reservoir under these circumstances.

As a starting point we take the (infinite) system of differential equations (2.7) again, but, for the time being, we forget about the boundary conditions (2.8) and (2.9). Instead, we shall try to obtain, for each $j \in \mathcal{N}$, the solution of (2.7) under the initial conditions

$$F_i(0) = \delta_{ij}, \quad i \in \mathcal{N}, \tag{2.33}$$

where δ_{ij} is Kronecker's delta. We shall assume that, for each $j \in \mathcal{N}$, this solution is unique and denote it by $\{F_i^{(j)}(y),\ i \in \mathcal{N}\}$. Later on we shall try to find a linear combination of solutions of this type which fits our original boundary conditions.

We now form the infinite matrices $\mathcal{F}(y),\ y \geq 0$, with elements

$$(\mathcal{F}(y))_{ij} \equiv F_i^{(j)}(y), \quad i, j \in \mathcal{N}, \tag{2.34}$$

and note that the system of differential equations and initial conditions satisfied by the functions $F_i^{(j)}(y),\ i, j \in \mathcal{N}$, may be represented by

$$\mathcal{F}'(y) = R^{-1}Q^T\mathcal{F}(y) \tag{2.35}$$

and

$$\mathcal{F}(0) = I, \tag{2.36}$$

respectively, where I denotes the infinite identity matrix, $R \equiv \mathrm{diag}(r_0, r_1, \ldots)$ and Q is the generator of the birth-death process, that is,

$$Q \equiv \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & \cdots & & \\ \mu_1 & -(\lambda_1+\mu_1) & \lambda_1 & 0 & \cdots & \\ 0 & \mu_2 & -(\lambda_2+\mu_2) & \lambda_2 & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}. \tag{2.37}$$

Writing for convenience

$$A \equiv R^{-1}Q^T, \tag{2.38}$$

it now follows, formally at least, that

$$\mathcal{F}(y) = \exp(yA) = \sum_{n=0}^{\infty} A^n \frac{y^n}{n!}, \tag{2.39}$$

and hence that

$$F_i^{(j)}(y) = \sum_{n=0}^{\infty} (A^n)_{ij} \frac{y^n}{n!}, \quad i, j \in \mathcal{N}. \tag{2.40}$$

To obtain an alternative expression for $F_i^{(j)}(y)$, $i, j \in \mathcal{N}$, we next consider the polynomials $P_i(x)$, $i \in \mathcal{N}$, recurrently defined by $P_0(x) = 1$ and

$$xP_i(x) = \sum_{k \in \mathcal{N}} (A)_{ki} P_k(x), \quad i \in \mathcal{N}, \tag{2.41}$$

which is equivalent to

$$\begin{aligned} &P_0(x) = 1, \qquad \frac{\lambda_0}{r_1} P_1(x) = x + \frac{\lambda_0}{r_0}, \\ &\frac{\lambda_{i-1}}{r_i} P_i(x) = \left(x + \frac{\lambda_{i-1} + \mu_{i-1}}{r_{i-1}} \right) P_{i-1}(x) - \frac{\mu_{i-1}}{r_{i-2}} P_{i-2}(x), \quad i \in \mathcal{N} \backslash \{1, 2\}. \end{aligned} \tag{2.42}$$

It is not difficult to see by induction that we also have

$$x^n P_i(x) = \sum_{k \in \mathcal{N}} (A^n)_{ki} \, P_k(x), \quad i \in \mathcal{N}, \tag{2.43}$$

for all $n = 0, 1, \ldots$, and as a consequence we can write

$$e^{xy} P_i(x) = \sum_{n=0}^{\infty} x^n \frac{y^n}{n!} P_i(x) = \sum_{n=0}^{\infty} \sum_{k \in \mathcal{N}} (A^n)_{ki} \frac{y^n}{n!} P_k(x), \quad i \in \mathcal{N}, \tag{2.44}$$

which, after interchanging summation signs and substituting (2.40), reduces to

$$e^{xy} P_i(x) = \sum_{k \in \mathcal{N}} F_k^{(i)}(y) P_k(x) \ , \quad i \in \mathcal{N}. \tag{2.45}$$

Now, if the sequence of polynomials $\{P_i(x)\}_{i=0}^{\infty}$, would be orthogonal with respect to some inner product $(\ . \ , \ . \ )$, then the previous result would imply

$$(e^{xy} P_j(x), P_i(x)) = F_i^{(j)}(y) (P_i(x), P_i(x)), \quad i, j \in \mathcal{N}, \tag{2.46}$$

that is,

$$F_i^{(j)}(y) = \frac{(e^{xy}P_j(x), P_i(x))}{(P_i(x), P_i(x))}, \quad i, j \in \mathcal{N}. \tag{2.47}$$

Fortunately, the sequence $\{P_i(x)\}_{i=0}^{\infty}$ can be shown to constitute a system of so-called *chain-sequence polynomials* , see [30]. It follows that a sequence of associated *kernel polynomials* can be found that is orthogonal with respect to the inner product defined by

$$(f, g) = \int_{-\infty}^{\infty} f(x)g(x)\psi^*(dx), \tag{2.48}$$

where ψ^* is some positive measure on $\mathbb{R}$. In fact, the corresponding kernel polynomials are, apart from normalization, the polynomials $\Delta_i^*(x)$ that are defined in (2.12), with the convention $N = \infty$. The following lemma states that, although the polynomials $P_i(x)$ are not orthogonal with respect to an inner product in the classical sense, an equally valuable relation holds as a result of these considerations.

Lemma 2.7 *If the sequence $\{\Delta_i^*(x)\}_{i=0}^{\infty}$ is orthogonal with respect to a unique positive measure ψ^* with finite moment of order -1 (in the sense that $\psi^*(\{0\}) = 0$ and the integrals $\int_{-\infty}^{0-} x^{-1}\psi^*(dx)$ and $\int_{0+}^{\infty} x^{-1}\psi^*(dx)$ converge), then there exists a signed measure ψ of total mass 1 such that*

$$\int_{-\infty}^{\infty} P_i(x)P_j(x)\psi(dx) = \frac{r_i}{r_0\pi_i}\delta_{ij}, \quad i, j \in \mathcal{N}. \tag{2.49}$$

If $r_0 < 0$, the mass on the positive axis is positive and the mass on the negative axis is negative, while the reverse holds true if $r_0 > 0$.

Proof. It is easily verified that the monic polynomials $\Delta_i(x)$ given in (2.13) (again with $N = \infty$) satisfy

$$\Delta_i(x) = \prod_{k=1}^{i} \frac{\lambda_{k-1}}{r_k} \; P_i(x). \tag{2.50}$$

From [30, Theorem 3 and (9)] we see that $\{\Delta_i(x)\}_{i=0}^{\infty}$, and hence $\{P_i(x)\}_{i=0}^{\infty}$, constitutes a system of chain-sequence polynomials for which the associated system $\{\Delta_i^*(x)\}_{i=0}^{\infty}$ of kernel polynomials satisfies the recurrence relations in (2.12). In particular, these polynomials are orthogonal with respect to a positive measure ψ^*. When $r_0 < 0$, we normalize ψ^* such that its total mass equals $-\lambda_0/r_0$, so that, under the required conditions, it now follows from [30, Theorem 5] that a measure ψ exists, signed and normalized as indicated, with respect to which the polynomials $\Delta_i(x)$, and hence the $P_i(x)$, are orthogonal. When $r_0 > 0$, the same normalization of ψ^* implies this to be a negative measure. However, the validity of Theorem 5 in [30] can easily be extended to this case, the only difference being the sign of ψ. Relation (2.49) can now be found with the help of (2.50), since for the monic polynomials $\Delta_i(x)$ it is known that

$$\int_{-\infty}^{\infty} \Delta_i(x)\Delta_j(x)\psi(dx) = \delta_{ij} \prod_{k=1}^{i} \frac{\lambda_{k-1}\mu_k}{r_{k-1}r_k}, \quad i, j \in \mathcal{N}. \tag{2.51}$$

see [19, Theorem I.4.2]. □

As a consequence of this lemma, $F_i^{(j)}(y)$ can be represented as

$$F_i^{(j)}(y) = \frac{r_0\pi_i}{r_i} \int_{-\infty}^{\infty} e^{xy} P_i(x)P_j(x)\psi(dx), \quad y \geq 0,\ i, j \in \mathcal{N}. \tag{2.52}$$

As announced our next step is to assume that the solution of the system (2.7) with boundary conditions (2.8) and (2.9) is a linear combination of the solutions $\{F_i^{(j)}(y),\ i \in \mathcal{N}\}$, that is, we assume that there are constants $a_j,\ j \in \mathcal{N}$, such that

$$F_i(y) = \sum_{j\in\mathcal{N}} a_j F_i^{(j)}(y) = \frac{r_0\pi_i}{r_i} \sum_{j\in\mathcal{N}} a_j \int_{-\infty}^{\infty} e^{xy} P_i(x)P_j(x)\psi(dx), \quad y \geq 0,\ i \in \mathcal{N}, \tag{2.53}$$

and our next task is to use the boundary conditions to determine these constants, which, since $F_i^{(j)}(0) = \delta_{ij}$, have the interpretation

$$a_j = F_j(0), \quad j \in \mathcal{N}. \tag{2.54}$$

At this point our assumption $N_- \equiv |\mathcal{N}^-| < \infty$ starts playing its crucial role. Indeed, it follows from boundary condition (2.8) that

$$F_i(0) = a_i = 0, \quad i \in \mathcal{N}^+, \tag{2.55}$$

so that the summation in (2.53) is essentially finite, and hence

$$F_i(y) = \frac{r_0\pi_i}{r_i} \int_{-\infty}^{\infty} e^{xy} P_i(x) \left(\sum_{j\in\mathcal{N}^-} a_j P_j(x) \right) \psi(dx), \quad y \geq 0,\ i \in \mathcal{N}. \tag{2.56}$$

The following lemma is helpful in determining the constants a_j.

Lemma 2.8 *The part of the measure ψ which is concentrated on the positive axis consists of $N_- - 1$ isolated point masses.*

Proof. In Chihara's book [19] we find the relationship between the zeros of a sequence of monic orthogonal polynomials and the *support* of the positive measure with respect to which these polynomials are orthogonal. We therefore start off by analysing the zeros of the polynomials $\Delta_i^*(x)$, as this will give us information about the measure ψ^* which is closely related to the measure ψ. Let $i \geq \max\mathcal{N}_-$ and let x_{ij} (x_{ij}^*) denote the jth zero,

$j = 1, \ldots, i$, arranged in increasing order, of the polynomial $\Delta_i(x)$ $(\Delta_i^*(x))$. Then [30, Theorem 12] tells us that the number of positive zeros of $\Delta_i(x)$ equals N_-, and that

$$x_{i1} < x_{i2} < \cdots < x_{i,i-N_-} < 0 < x_{i,i-N_-+1} < \cdots < x_{ii}, \tag{2.57}$$

while by [30, Theorem 13] we have that

$$x_{i,j-1} < x_{ij}^* < x_{ij}, \qquad j = 1, \ldots, i, \tag{2.58}$$

where $x_{i0} = -\infty$. To find the position of $x_{i,i-N_-+1}^*$ relative to 0, we make two observations. The first is that the monicity of $\Delta_i^*(x)$ implies $\lim_{x\to\infty} \Delta_i^*(x) = \infty$, and hence $x_{i,i-N_-+1}^* < 0$ if and only if $\mathrm{sign}(\Delta_i^*(0)) = (-1)^{N_- -1}$. The second observation is that $\mathrm{sign}(\Delta_i^*(0)) = (-1)^{N_-}\mathrm{sign}(\sum_{k=0}^{i} \pi_k r_k)$, see [30, (15)]. By (2.3) we therefore find for sufficiently large i,

$$x_{i1}^* < x_{i2}^* < \cdots < x_{i,i-N_-+1}^* < 0 < x_{i,i-N_-+2}^* < \cdots < x_{ii}^*, \tag{2.59}$$

so that, in particular, $\Delta_i^*(x)$ has $N_- - 1$ strictly positive zeros. To determine the consequences for the support of ψ^*, denoted by $\mathrm{supp}(\psi^*)$, we follow [19, Section II.4]. We define $\zeta_{N_-} = \infty$ and the limits $\zeta_j = \lim_{i\to\infty} x_{i,i-N_-+j+1}^*$, $j = 1, \ldots, N_- - 1$, which exist in $\mathbb{R}^+ \cup \{\infty\}$, since for any $k \geq 1$, the sequence $\{x_{i,i-k+1}^*\}_{i=k}^{\infty}$ is increasing. In fact we have

$$0 < \zeta_1 < \zeta_2 < \cdots < \zeta_{N_- -1} < \zeta_{N_-} = \infty, \tag{2.60}$$

since $\zeta_k = \zeta_{k+1}$ would imply that also $\lim_{i\to\infty} x_{i,i-N_-+1}^* = \zeta_k$, which cannot be the case, see (2.59). It now follows, still from [19], that $\mathrm{supp}(\psi^*) \cap \mathbb{R}^+ \setminus \{\zeta_1, \ldots, \zeta_{N_- -1}\} = \varnothing$ and that at least one supporting point of ψ^* lies in every open interval $(x_{ij}^*, x_{i,j+1}^*)$, $j = 1, \ldots, i$, with $x_{i,i+1}^* = \infty$, so that we may conclude that $\mathrm{supp}(\psi^*) \cap \mathbb{R}^+ = \{\zeta_1, \ldots, \zeta_{N_- -1}\}$. The proof is completed by noting that the support of ψ coincides with that of ψ^*, possibly supplemented with $\{0\}$, see [30, Theorem 5]. □

As in the proof of the lemma, we let the point masses on the positive axis be located at the points $\zeta_1, \zeta_2, \ldots, \zeta_{N_- -1}$. Considering that $F_i(y)$ is a probability, and hence uniformly bounded, it follows that the constants a_j, $j \in \mathcal{N}^-$, must be such that

$$\sum_{j\in\mathcal{N}^-} a_j P_j(\zeta_k) = 0, \quad k = 1, 2, \ldots, N_- - 1, \tag{2.61}$$

and that the interval of integration in (2.56) may be reduced to $(-\infty, 0]$. The missing equation for the constants a_j, $j \in \mathcal{N}^-$, comes from the observation that the average amount of fluid flowing into the buffer should balance the average amount of fluid flowing out, that is,

$$\sum_{j\in\mathcal{N}^+} p_j r_j = -\sum_{j\in\mathcal{N}^-} (p_j - a_j) r_j, \tag{2.62}$$

or, with (2.2),

$$\sum_{j\in\mathcal{N}^-} a_j r_j = \frac{\sum_{j\in\mathcal{N}} \pi_j r_j}{\sum_{j\in\mathcal{N}} \pi_j}. \tag{2.63}$$

We observe that this equation can be rewritten as

$$r_0 \sum_{j\in\mathcal{N}^-} a_j P_j(0) = \frac{\sum_{j\in\mathcal{N}} \pi_j r_j}{\sum_{j\in\mathcal{N}} \pi_j}. \tag{2.64}$$

since $P_i(0) = r_i/r_0,\ i \in \mathcal{N}$, as can easily be verified.

We are now ready to state the main theorem of this section.

Theorem 2.9 *When N_- is finite, and the condition in Lemma 2.7 is satisfied, the stationary joint distribution $F_i(y) \equiv P[X_t = i,\ C_t \le y],\ i \in \mathcal{N} = \{0, 1, \ldots\},\ y \ge 0$, of the process (X_t, C_t) can be represented as*

$$F_i(y) = p_i + \frac{\pi_i}{r_i} \int_{-\infty}^{0-} e^{xy} P_i(x) R(x) \psi(dx). \tag{2.65}$$

Here, $P_i(x),\ i \in \mathcal{N}$, are the polynomials defined in (2.41), and ψ is the signed measure of Lemma 2.7 with respect to which these polynomials are orthogonal. Furthermore, $R(x)$ is a polynomial defined by

$$R(x) \equiv r_0 \sum_{j\in\mathcal{N}^-} a_j P_j(x), \tag{2.66}$$

with constants $a_j,\ j \in \mathcal{N}^-$, such that

$$R(0) = \frac{\sum_{j\in\mathcal{N}} \pi_j r_j}{\sum_{j\in\mathcal{N}} \pi_j} \tag{2.67}$$

and

$$R(\zeta_j) = 0, \quad j = 1, 2, \ldots, N_- - 1, \tag{2.68}$$

where $\zeta_1, \zeta_2, \ldots, \zeta_{N_--1}$ are the $N_- - 1$ supporting points of the measure ψ on the positive axis.

Proof. By substitution and using (2.1), (2.42), and Lemmas 2.7 and 2.8, it is not difficult to check that $F_i(y) = (\pi_i/r_i) \int_{-\infty}^{0+} e^{xy} P_i(x) R(x)\, \psi(dx)$ satisfies the differential equations (2.7) and the boundary conditions (2.8). It remains to check that condition (2.9) is satisfied. Assuming that ψ^* is normalized as before, its total mass being $-\lambda_0/r_0$, we use that $\psi(\{0\}) = 1 - \int_{-\infty}^{\infty} x^{-1}\psi^*(dx)$, see [30, Theorem 5] again, and that $\int_{-\infty}^{\infty} x^{-1}\psi^*(dx) = 1 - r_0/\sum_{j=0}^{\infty} r_j\pi_j$; the proof of the latter equation is deferred to Section 2.5.3 since it is rather technical. Since we now find

$$\begin{aligned}
\lim_{y\to\infty} F_i(y) &= \frac{\pi_i}{r_i} P_i(0) R(0) \psi(\{0\}) \\
&= \frac{\pi_i}{r_i} \frac{r_i}{r_0} \frac{\sum_{j=0}^{\infty} r_j \pi_j}{\sum_{j=0}^{\infty} \pi_j} \frac{r_0}{\sum_{j=0}^{\infty} r_j \pi_j} = p_i,
\end{aligned}$$

we conclude that (2.9) is satisfied. □

Corollary 2.10 *When the sum $S(x)$, given by*

$$S(x) \equiv \sum_{i \in \mathcal{N}} \frac{\pi_i}{r_i} P_i(x), \tag{2.69}$$

converges uniformly on an interval containing the negative part of the support of the measure ψ, the stationary distribution of the buffer content process (C_t) can be represented as

$$P[C > y] = -\int_{-\infty}^{0-} e^{xy} S(x) R(x) \; \psi(dx), \quad y \geq 0. \tag{2.70}$$

Evidently, the main problem in concrete examples is to find the signed measure ψ with respect to which the polynomials $P_i(x)$ are orthogonal. In the proofs of both lemmas in this subsection, we already touched on the technique described in [30] by which this problem is transformed into the easier problem of finding the (positive) orthogonalizing measure ψ^* for the associated sequence of polynomials $\{\Delta_i^*(x)\}_{i=0}^{\infty}$. At least in some cases, such as the model studied in [4] and [97], this technique enables us to find the measure explicitly, as we show in the next section.

2.5.2 Example

We look at the case where we have

$$r_0 \equiv -r_- \text{ and } r_i \equiv r_+ > 0, \quad i = 1, 2, \ldots,$$

so that

$$\mathcal{N}^+ \equiv \{1, 2, \ldots\}, \quad \mathcal{N}^- \equiv \{0\},$$

and assume that the birth and death rates are constant and given by

$$\lambda_i \equiv \lambda \text{ and } \mu_{i+1} \equiv \mu, \quad i \in \mathcal{N}.$$

As in Section 2.4.2 we find it convenient to set $\rho \equiv \lambda/\mu$; furthermore we now let $\sigma \equiv r_-/(r_+ + r_-)$. Since $\pi_i = \rho^i$, it follows from (2.3) that we must have

$$\rho < \sigma, \tag{2.71}$$

for the system to be stable. We shall assume in what follows that this condition is satisfied.

Since $\mathcal{N}^- \equiv \{0\}$ and $P_0(x) = 1$, the polynomial $R(x)$ in Theorem 2.9 is in fact a constant, which, from (2.67), is readily seen to be

$$R(x) = R(0) = -r_-(1 - \rho/\sigma). \tag{2.72}$$

The polynomials $P_i(x)$, $i \in \mathcal{N}$, satisfy the recurrence relations

$$\begin{aligned} P_0(x) &= 1, \quad \lambda P_1(x) = r_+ x - \lambda r_+/r_-, \\ \lambda P_2(x) &= (r_+ x + \lambda + \mu) P_1(x) + \mu r_+/r_-, \\ \lambda P_i(x) &= (r_+ x + \lambda + \mu) P_{i-1}(x) - \mu P_{i-2}(x), \quad i = 3, 4, \ldots . \end{aligned} \tag{2.73}$$

Following the procedure outlined in [30] and using results on perturbed Chebysev polynomials in Chihara's book [19], the required orthogonalizing measure ψ for these polynomials can be found as follows. First, we return to the sequence of monic chain-sequence polynomials $\{\Delta_i(x)\}_{i=0}^{\infty}$ of (2.13) and the associated system $\{\Delta_i^*(x)\}_{i=0}^{\infty}$ of kernel polynomials. The latter sequence satisfies the recurrence relations in (2.12), which in this example are given by

$$\begin{aligned}
\Delta_0^*(x) &= 1, \quad \Delta_1^*(x) = x - \frac{\lambda}{r_-} + \frac{\mu}{r_+}, \\
\Delta_i^*(x) &= (x + \frac{\lambda+\mu}{r_+})\Delta_{i-1}^*(x) - \frac{\lambda\mu}{r_+^2}\Delta_{i-2}^*(x), \quad i = 2, 3, \ldots . \qquad (2.74)
\end{aligned}$$

By the transformation

$$T_i(x) \equiv \left(\frac{r_+}{\sqrt{\lambda\mu}}\right)^i \Delta_i^*\left(\frac{2x\sqrt{\lambda\mu} - \lambda - \mu}{r_+}\right),$$

we once more find a sequence of perturbed Chebysev polynomials as in (2.28), but now with $T_1(x) = 2x - \sqrt{\rho}/\sigma$, where σ is defined as in this section. The corresponding (positive) orthogonalizing measure with total mass 1 is given in [19, page 205], whence we find,

$$\begin{aligned}
&\frac{2}{\pi}\int_{-1}^{1} T_i(x)T_j(x)\,\frac{\sqrt{1-x^2}}{1+\rho/\sigma^2 - 2x\sqrt{\rho}/\sigma}\,dx + \\
&\qquad \left(1 - \frac{\sigma^2}{\rho}\right) T_i\left(\frac{\sqrt{\rho}}{2\sigma} + \frac{\sigma}{2\sqrt{\rho}}\right) T_j\left(\frac{\sqrt{\rho}}{2\sigma} + \frac{\sigma}{2\sqrt{\rho}}\right) \; = \; \delta_{ij}. \qquad (2.75)
\end{aligned}$$

By appropriately transforming the latter result we can find the positive measure ψ^* with respect to which the polynomials $\Delta_i^*(x)$, $i = 0, 1, \ldots,$ are orthogonal. After normalizing ψ^* to have total mass λ/r_-, it follows from [30, Theorem 5] that the signed measure ψ, corresponding to the sequence $\{\Delta_i(x)\}_{i=0}^{\infty}$ (and hence also to $\{P_i(x)\}_{i=0}^{\infty}$), is given by

$$\psi(dx) = \frac{\psi^*(dx)}{x}, \quad x \neq 0, \qquad (2.76)$$

while ψ has an atom in zero of size

$$\psi(\{0\}) = 1 - \int_{-\infty}^{\infty} x^{-1}\psi^*(dx). \qquad (2.77)$$

Hence, after some straightforward calculations we find that the required orthogonalizing measure ψ for the polynomials $P_i(x)$ is given by

$$\psi(dx) = \frac{1}{2\pi x}\frac{\sqrt{4\rho - (r_+x/\mu + \rho + 1)^2}}{(\rho - r_-x/\mu)/\sigma - 1}dx, \qquad (2.78)$$

for x in the interval

$$-\frac{\mu(1+\sqrt{\rho})^2}{r_+} \leq x \leq -\frac{\mu(1-\sqrt{\rho})^2}{r_+}, \tag{2.79}$$

while ψ has no mass outside this interval, with the exception of an atom at zero (the size of which need not concern us), and, if $\rho > \sigma^2$, an atom at $-\frac{\mu}{r_-}(\sigma-\rho)$ of size

$$\psi\left(\left\{-\frac{\mu}{r_-}(\sigma-\rho)\right\}\right) = -\frac{\rho-\sigma^2}{\sigma-\rho}. \tag{2.80}$$

We note that ψ has no mass on the positive axis, which is in accordance with Lemma 2.8, since $N_- = 1$ in the model at hand.

Finally, we turn to $S(x)$ of (2.69) by writing down the generating function

$$S(x,z) = \sum_{i\in\mathcal{N}} \frac{\pi_i}{r_i} P_i(x)\ z^i, \tag{2.81}$$

which, by using the recurrence relations (2.73), can be shown to be equal to

$$S(x,z) = \frac{1}{r_-}\,\frac{-1+((r_+ + r_-)x/\mu+1)z}{\rho z^2-(r_+x/\mu+\rho+1)z+1}. \tag{2.82}$$

When we interpret $S(x,z)$ as a function of z for fixed x, it turns out that it is analytic for $|z| \leq 1$ if and only if $-2\mu(1+\rho)/r_+ < x < 0$ (when x lies inside the interval (2.79), the numerator of S has no real roots, otherwise the roots lie outside the interval $(-1,1)$). Therefore we conclude that the series $S(x)$ converges on the interval $-2\mu(1+\rho)/r_+ < x < 0$ to the constant limit

$$S(x) = -\frac{1}{r_+} - \frac{1}{r_-}. \tag{2.83}$$

As a consequence, $S(x)$ is uniformly convergent in an interval that contains the support of the measure ψ, which justifies the interchange of summation and integration signs when summing $F_i(y)$ of (2.65) over all $i \in \mathcal{N}$.

Summarizing and slightly rewriting our results we conclude that the stationary distribution of the buffer content can be represented as

$$P[C > y] = \frac{1-\rho/\sigma}{2\pi}\int_{-(1+\sqrt{\rho})^2}^{-(1-\sqrt{\rho})^2} \frac{\sqrt{4\rho-(x+\rho+1)^2}}{r_-x^2+r_+(\sigma-\rho)x}\, e^{\frac{\mu}{r_+}xy}\, dx \\ + \frac{r_-}{r_+}\left(\frac{\rho}{\sigma^2}-1\right) e^{-\frac{\mu}{r_-}(\sigma-\rho)y}\mathbf{1}_{\{\rho>\sigma^2\}}, \quad y \geq 0, \tag{2.84}$$

where $\mathbf{1}_A$ denotes the indicator function of the event A. It is not difficult to see that (2.84) is in accordance with the expression given in [97], apart from the erroneous minus sign there that has already been noted in [4].

Remark 2.2 If $\rho > \sigma^2$ the second term in (2.84) is the dominating one as $y \to \infty$. When $\rho \leq \sigma^2$ it is interesting to know the asymptotic behaviour of the first term, which is then given by

$$P[C > y] \;\sim\; \frac{1-\rho/\sigma}{2\sqrt{\pi}} \, \frac{\rho^{1/4}}{r_-(1-\sqrt{\rho})^4 - r_+(\sigma-\rho)(1-\sqrt{\rho})^2} \times \left(\frac{\mu}{r_+}\, y\right)^{-3/2} \exp\{-(1-\sqrt{\rho})^2 \, \frac{\mu}{r_+}\, y\}, \tag{2.85}$$

where $f(y) \sim g(y)$ means $\lim_{y\to\infty} f(y)/g(y) = 1$. This result is due to [101], where a proof is given based on Laplace's method, see [72, page 80].

Remark 2.3 The models in this section and in Section 2.4.2 are *dual*, in the sense that the only difference between the two models is the interchange of $\mathcal{N}_+$ and $\mathcal{N}_-$. Thus, if we identify r_+ (r_-) from Section 2.4.2 with r_- (r_+), the model of the current section follows from the model in Section 2.4.2 by adding a minus-sign to the matrix R. Minor as it may seem at first sight, this difference appears to have considerable consequences when we compare the expressions for the stationary distributions in (2.32) and (2.84). The key to understanding the difference in complexity of these solutions clearly lies in Lemma 2.2. If we truncate the state space to $\{0, 1, \ldots, N\}$ and then let $N \to \infty$ for the case $N_- < \infty$ — as we did in Section 2.4 for the case $N_+ < \infty$ — the number of relevant (i.e. negative) eigenvalues tends to infinity. In the limit we obtain a continuum as well as one or two single points, together forming the support of the measure ψ.

Something similar happens in the dual case. The measure is again the orthogonalizing measure for the polynomials $\Delta_i(x)$, $i = 0, 1, \ldots$, which was already mentioned in Remark 2.1 for the general (dual) case. Due to the minus-sign in front of the matrix R, this measure is basically equal to $-\psi$, with ψ as in this section. In particular, the continuous part of its support lies on the positive axis, and therefore does not play a role in the solution. The only real asymmetry, namely the fact that the atom outside 0 is on the *negative* axis, is due to the different stability condition: for the dual case we have $\sigma - \rho < 0$.

2.5.3 Completion of proof of Theorem 2.9

In this section we will prove that, under the conditions of Theorem 2.9,

$$\int_{-\infty}^{\infty} \frac{\psi^*(dx)}{x} = 1 - \frac{r_0}{\sum_{j=0}^{\infty} r_j \pi_j}, \tag{2.86}$$

where, as before, ψ^* is the measure with respect to which the sequence $\{\Delta_i^*(x)\}_{i=0}^{\infty}$ is orthogonal. Again, we will assume that ψ^* is normalized such that its total mass is $-\lambda_0/r_0$.

We first introduce some notation. Suppose p_i and q_{i+1}, $i = 0, 1, \ldots$, are real constants such that $p_i q_i > 0$ for $i = 1, 2, \ldots$. Let the infinite-dimensional matrix A be such that its (i, j)th element, $i, j = 0, 1, \ldots$, is p_{i+1} if $j = i + 1$, $-(p_i + q_{i+1})$ if $j = i$, q_i if $j = i - 1$ and 0 if $|j - i| > 1$. We define the monic polynomial $Q_i(x)$ as the characteristic polynomial of the $i \times i$ north-west corner truncation of the matrix A. Hence, $Q_i(x)$ is given by the recurrence formula,

$$\begin{aligned} &Q_0(x) = 1, \quad Q_1(x) = x + p_0 + q_1, \\ &Q_i(x) = (x + p_{i-1} + q_i)Q_{i-1}(x) - p_{i-1}q_{i-1}Q_{i-2}(x), \quad i = 2, 3, \dots . \end{aligned} \tag{2.87}$$

We assume that there is a unique positive measure with total mass 1 with respect to which these polynomials are orthogonal, and denote it by $\phi(dx)$. Notice that when we let $p_i = \lambda_i/r_i$ and $q_i = \mu_i/r_i$, we find that $Q_i(x) = \Delta_i^*(x)$ and $\phi(dx) = -r_0/\lambda_0\, \psi^*(dx)$.

For $n = 1, 2, \dots$, we construct a discrete measure ϕ_n such that its kth moment equals the kth moment of ϕ, $k = 0, \dots, 2n-1$, see e.g. [12]. In particular we define ϕ_n as consisting of point masses $1/\sum_{j=0}^{n-1} \tilde{Q}_j^2(x_{nk})$ in x_{nk}, $k = 1, \dots, n$, where $\{\tilde{Q}_j(x)\}_{j=0}^{\infty}$ is the orthonormal version of the sequence $\{Q_j(x)\}_{j=0}^{\infty}$ and x_{nk} is the kth zero of $Q_n(x)$. The "method of moments" tells us that ϕ_n converges weakly to ϕ as $n \to \infty$, see [12] again. As a result we can conclude for $s \notin \operatorname{supp}(\phi)$ that

$$\int_{-\infty}^{\infty} \frac{\phi(dx)}{s - x} = \lim_{n\to\infty} \int_{-\infty}^{\infty} \frac{\phi_n(dx)}{s - x}. \tag{2.88}$$

This relation is useful because by [19, Theorem III.4.3] we have

$$\int_{-\infty}^{\infty} \frac{\phi_n(dx)}{s - x} = \frac{\tilde{Q}_{n-1}^{(1)}(s)}{\tilde{Q}_n(s)} = \frac{Q_{n-1}^{(1)}(s)}{Q_n(s)},$$

where $\{Q_n^{(1)}(x)\}_{n=0}^{\infty}$ and $\{\tilde{Q}_n^{(1)}(x)\}_{n=0}^{\infty}$ are the *numerator polynomials* corresponding to the sequences $\{Q_n(x)\}_{n=0}^{\infty}$ and $\{\tilde{Q}_n(x)\}_{n=0}^{\infty}$ respectively. As a consequence, if $0 \notin \operatorname{supp}(\psi^*)$, we can take $s = 0$ and find after some calculations using [30, equations (15) and (17)] that (2.86) holds.

In the case where 0 *is* a part of $\operatorname{supp}(\psi^*)$, it must be the largest point of accumulation of $\operatorname{supp}(\psi^*)$, see Lemmas 2.7 and 2.8. In the remainder of this section we will therefore assume that the support of ϕ has at least one accumulation point. We will denote the largest by τ and assume furthermore that ϕ has only finitely many supporting points to the right of τ, since ψ^* has this property. It is our objective to extend the validity of (2.88) to the case $s \geq \tau$ (although irrelevant for our current purpose, we mention that when $s \in \operatorname{supp}(\phi)$ and $s > \tau$, both sides of (2.88) may be interpreted as infinity). In this way we can show that (2.86) also holds when $0 \in \operatorname{supp}(\psi^*)$ by simply taking $s = 0\,(= \tau)$.

Notice that an immediate proof of (2.88) using weak convergence is not possible, since the function $x \mapsto (s - x)^{-1}$ is not bounded on $\operatorname{supp}(\phi)$ for $s \in \operatorname{supp}(\phi)$. We will follow the approach in the proof of [43, Lemma 6] to handle this problem. First we define for $t \geq 0$,

$$f_n(t) \equiv \int_{-\infty}^{\infty} e^{xt}\phi_n(dx), \qquad n = 1, 2, \dots, \tag{2.89}$$

and

$$f(t) \equiv \int_{-\infty}^{\infty} e^{xt}\phi(dx). \tag{2.90}$$

Since $x \mapsto e^{xt}$ is bounded on $\operatorname{supp}(\phi)$, $f(t)$ is the pointwise limit of $f_n(t)$ as $n \to \infty$ by weak convergence. The reason why we introduce these functions is that by interchanging the order of integration we easily find for s sufficiently large that

$$\int_0^\infty e^{-st} f(t)dt = \int_{-\infty}^\infty \frac{\phi(dx)}{s-x}, \tag{2.91}$$

while a similar relation holds when $f(t)$ and $\phi(dx)$ are replaced by $f_n(t)$ and $\phi_n(dx)$ respectively.

We will now show that $f_n(t)$ is increasing in n, so that, by monotone convergence, we will be justified in concluding that $\int_0^\infty e^{-st} f(t)dt$ exists and equals $\lim_{n\to\infty} \int_0^\infty e^{-st} f_n(t)dt$. In order to show the monotonicity of $f_n(t)$ in n we prove two lemmas that give some properties for related functions. Before defining these functions we introduce the $n \times n$ matrix $A(n,q)$, which is obtained from the matrix A by truncation and a small perturbation, replacing q_n by q. Concretely, we define

$$A(n,q) \equiv \begin{pmatrix} -(p_0+q_1) & p_1 & 0 & \cdots & \\ q_1 & -(p_1+q_2) & p_2 & 0 & \cdots \\ 0 & \cdots & \cdots & \cdots & 0 \\ \cdots & 0 & q_{n-2} & -(p_{n-2}+q_{n-1}) & p_{n-1} \\ & \cdots & 0 & q_{n-1} & -(p_{n-1}+q) \end{pmatrix}. \tag{2.92}$$

It is not difficult to see that the characteristic polynomial $Q^{(n,q)}(x)$ of this matrix satisfies $Q^{(n,q)}(x) = Q_n(x) + (q - q_n)Q_{n-1}(x)$. We will denote the zeros of $Q^{(n,q)}(x)$ by $x_k^{(n,q)}$, $k = 1, \ldots, n$ and define the discrete measure $\phi^{(n,q)}(x)$ as consisting of point masses $1/\sum_{j=0}^{n-1} \tilde{Q}_j^2(x_k^{(n,q)})$ at the points $x_k^{(n,q)}$, $k = 1, \ldots, n$. We note that for $k = 0, \ldots, 2n-2$, the kth moment of $\phi^{(n,q)}$ does not depend on q, since it coincides with the kth moment of ϕ, see [19, equation (II.5.4)].

The next step is to define for $t \geq 0$,

$$f_{i,j}^{(n,q)}(t) = \frac{1}{\prod_{k=1}^j p_k \prod_{k=1}^i q_k} \int_{-\infty}^\infty e^{xt} Q_i(x) Q_j(x) \phi^{(n,q)}(dx), \qquad i,j = 0, \ldots, n-1.$$

It can be shown by substitution that the $n \times n$ matrix $F^{(n,q)}(t)$ with elements $f_{i,j}^{(n,q)}(t)$ is the unique solution to the system of differential equations $(d/dt)F^{(n,q)}(t) = A^T(n,q)F^{(n,q)}(t)$ with initial condition $F^{(n,q)}(0) = I$; hence $F^{(n,q)}(t) = e^{tA^T(n,q)}$.

We are now ready to prove the announced lemmas.

Lemma 2.11 *$f_{i,j}^{(n,q)}(t)$ is decreasing in q for each $t > 0$ and $i,j = 0, \ldots, n-1$.*

Proof. Let $c_k(q) = \int_{-\infty}^\infty x^k \phi^{(n,q)}(dx)$, $k = 0, 1, \ldots$. The first moment that truly depends on q, namely $c_{2n-1}(q)$, can be found from

$$0 = \int x^{n-1} Q^{(n,q)}(x) \phi^{(n,q)}(dx) = c_{2n-1}(q) + qc_{2n-2} + \sum_{k=0}^{2n-3} a_k c_k,$$

where the constants a_k are independent of q. It follows that $c_{2n-1}(q)$ is a decreasing linear function of q. Putting $q_1 < q_2$ and using a Taylor expansion, we now find for small t that

$$\begin{aligned}
f_{i,j}^{(n,q_1)}(t) - f_{i,j}^{(n,q_2)}(t) &= \\
&= \frac{1}{\prod_{k=1}^{j} p_k \prod_{k=1}^{i} q_k} \int_{-\infty}^{\infty} e^{xt} Q_i(x) Q_j(x) \left(\phi^{(n,q_1)}(dx) - \phi^{(n,q_2)}(dx) \right) \\
&= \frac{1}{\prod_{k=1}^{j} p_k \prod_{k=1}^{i} q_k} \frac{t^{2n-1-i-j}}{(2n-1-i-j)!} \left(c_{2n-1}(q_1) - c_{2n-1}(q_2) + o(t^{2n-1-i-j}) \right) \\
&> 0. \qquad (2.93)
\end{aligned}$$

Finally, when (2.93) holds for t_1 and t_2, we also have $F^{(n,q_1)}(t_1+t_2) = F^{(n,q_1)}(t_1)F^{(n,q_1)}(t_2) > F^{(n,q_2)}(t_1)F^{(n,q_2)}(t_2) = F^{(n,q_2)}(t_1+t_2)$, where the inequality holds componentwise. The lemma now follows immediately. □

Lemma 2.12 *We have for $t \geq 0$,*

$$\lim_{q\to\infty} f_{i,j}^{(n,q)}(t) = f_{i,j}^{(n-1,q_{n-1})}(t), \qquad i,j = 0,\ldots,n-1.$$

Proof. As $q \to \infty$, one zero of $Q^{(n,q)}(x)$, $x_n^{(n,q)}$ say, tends to $-\infty$, while for the other zeros we find $x_k^{(n,q)} \to x_{n-1,k}$, that is, they converge to the zeros of $Q_{n-1}(x)$. Furthermore we find that $\phi^{(n,q)}(x_k^{(n,q)}) \to \phi_{n-1}(x_{n-1,k})$, $k = 1,\ldots,n-1$, and as a consequence $\phi^{(n,q)}(x_n^{(n,q)}) \to 0$. Hence $\phi^{(n,q)}$ converges to $\phi_{n-1} = \phi^{(n-1,q_{n-1})}$ and the lemma follows. □

The fact that $f_n(t) = f_{0,0}^{(n,q_n)}(t)$ increases in n is now immediate from the lemmas above.

As announced we conclude that $\int_0^\infty e^{-st} f(t)dt = \lim_{n\to\infty} \int_0^\infty e^{-st} f_n(t)dt$, or, in other words, that (2.88) holds. However, this does not help us to show (2.88) for $s \geq \tau$, since (2.91) holds true only for sufficiently large s, namely for $s > \max \operatorname{supp}(\phi)$. We resolve this problem by defining for $t \geq 0$,

$$f_n^\tau(t) \equiv \int_{-\infty}^{\tau} e^{xt} \phi_n(dx), \qquad n = 1,2,\ldots,$$

and

$$f^\tau(t) \equiv \int_{-\infty}^{\tau} e^{xt} \phi(dx),$$

where, as before, the latter one is the pointwise limit of the first as $n \to \infty$. We can now write *for $s \geq \tau$*

$$\int_{-\infty}^{\tau} \frac{\phi(dx)}{s-x} = \int_0^\infty e^{-st} f^\tau(t)dt = \lim_{n\to\infty} \int_0^\infty e^{-st} f_n^\tau(t)dt = \lim_{n\to\infty} \int_{-\infty}^{\tau} \frac{\phi_n(dx)}{s-x}, \qquad (2.94)$$

where the second equality is justified by Lebesgue's dominating convergence theorem, using the fact that $f_n^\tau(t) \leq f_n(t) \leq f(t)$. Finally, since interchanging a limit and a finite summation is allowed, and hence

$$\int_{\tau}^{\infty} \frac{\phi(dx)}{s-x} = \lim_{n\to\infty} \int_{\tau}^{\infty} \frac{\phi_n(dx)}{s-x}, \qquad (2.95)$$

we can conclude the following.

Proposition 2.13 *Let ϕ and ϕ_n be the corresponding measures to a sequence of orthogonal polynomials $\{Q_i(x)\}_{i=0}^{\infty}$, as before, with a finite set of supporting points to the right of τ, the largest point of accumulation of* $\mathrm{supp}(\phi)$*. Then, for $s \geq \tau$,*

$$\int_{-\infty}^{\infty} \frac{\phi(dx)}{s-x} = \lim_{n\to\infty} \int_{-\infty}^{\infty} \frac{\phi_n(dx)}{s-x} = \lim_{n\to\infty} \frac{Q_{n-1}^{(1)}(s)}{Q_n(s)} \tag{2.96}$$

Applying this in the setting of Section 2.5.1 with $s = \tau = 0$, we find that (2.86) holds under the conditions in Theorem 2.9 (or Lemma 2.7), also when $0 \in \mathrm{supp}(\psi^*)$.

Chapter 3

A birth-death fluid model with feedback

3.1 Introduction

We consider a single-server queueing system at which customers arrive according to a Poisson process with rate λ. Customers arriving while the server is busy wait for their turn if there is a free waiting position and are lost otherwise.

During idle periods of the server a fluid commodity which we shall designate as *credit* accumulates in a reservoir at a constant rate $r_+ > 0$. The credit reservoir depletes during busy periods of the server at a constant rate $r_- > 0$ as long as the reservoir is nonempty. It may be helpful to think of credit as energy which the server gathers when idle and consumes when busy.

In the following it will become clear that the current model involves *feedback* in the sense that the current state of the fluid reservoir influences the behaviour of the regulating queueing system. The amounts of service which customers require *in the presence of credit* are independent and exponentially distributed random variables with mean $1/\mu_1$, which results in a departure rate of μ_1 as long as there are customers in the system *and* the credit reservoir is nonempty. When the credit reservoir becomes empty, however, the server slows down and the departure rate drops to $\mu_2 \leq \mu_1$.

Notice that for $\mu_1 = \mu_2 = \mu$ we have a (Markov-modulated) fluid model without feedback. In fact this is the same model as that in Section 2.4.2, where the regulating process can be interpreted as the number of customers in an $M/M/1$ queueing system with parameters λ and μ.

We shall let X_t denote the number of customers in the system and C_t the content of the credit reservoir at time t. The interaction between the processes (X_t) and (C_t) is summarized schematically in Figure 3.1. Obviously, the two-dimensional process (X_t, C_t) constitutes a Markov process which, under a suitable stability condition, possesses a unique stationary distribution. Our foremost aim is to obtain this distribution.

An important motivation for studying the present model is to investigate whether

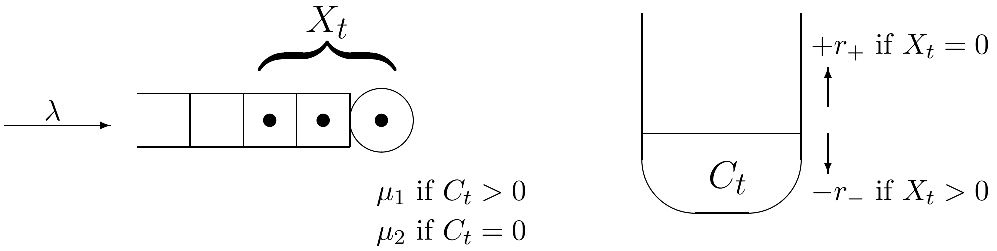


Figure 3.1: Interaction between the processes (X_t) and (C_t)

models with feedback yield to analysis. As it turns out, the particular type of feedback we consider (different behaviour of (X_t) when $C_t = 0$) does not influence the analysis heavily when we compare it to Section 2.4.2.

Another motivation for studying the present model is that it may serve as a model for a traffic regulation mechanism in an ATM network. This application is described in Section 3.6.

In the following we will first discuss the case of an infinite credit reservoir. Although primarily interested in the model with infinite waiting room, we shall start off, in Section 3.2 by studying the case in which the number of customers in the system is bounded by some number N. Subsequently, in Section 3.3, we let $N \to \infty$ in the expressions found to obtain the stationary distribution of (X_t, C_t) (and related quantities of interest) when the waiting room is unbounded. The reason that we try this approach, which worked so well in Section 2.4, is that the analysis amounts to solving a system of differential-difference equations that resembles the system of equations appearing in the corresponding fluid model without feedback.

An approach similar to that for the infinite credit reservoir does not seem to lead to tractable results when the reservoir has finite capacity. By way of introduction to our approach to this case, we will present, in Section 3.4, an alternative analysis for the model with infinite credit reservoir and infinite waiting room, in which we keep track of the amount of credit by observing the *number* of suitably defined credit quanta rather than the actual *volume* of credit in the reservoir. Subsequently, in Section 3.5, we use an extension of this discretization technique to obtain an approximative solution for the finite-reservoir model. We also validate this approximation by simulation. We believe that the discretization technique of Sections 3.4 and 3.5 has much wider applicability (see, e.g., [5]), and consider its presentation as one of the main contributions of this chapter.

Finally, in Section 3.6 we turn to the application we have in mind, that of a traffic regulation mechanism operating on a very bursty on-off source. We present some numerical results that have been obtained by the methods of Section 3.5 and show the trade-off between extra delay incurred by the regulation mechanism on the one hand and burstiness reduction of the regulated traffic stream on the other.

3.2 Finite waiting room and infinite credit reservoir

3.2.1 Preliminaries

We assume in this section that the waiting room is bounded and has $N-1$ waiting positions, so that the state space $\mathcal{N}$ of the process (X_t) is given by $\mathcal{N} = \{0, 1, \dots, N\}$. Clearly, to have stability of the process (X_t, C_t), it is necessary and sufficient that the expected net rate of credit into the reservoir, *conditional* on the reservoir being non-empty, be negative. Since

$$\left\{1 + \frac{\lambda}{\mu_1} + \left(\frac{\lambda}{\mu_1}\right)^2 + \cdots + \left(\frac{\lambda}{\mu_1}\right)^N\right\}^{-1}$$

is the stationary probability of the server being idle in an $M/M/1/N$ system with arrival rate λ and service rate μ_1, this condition translates into

$$\frac{\lambda}{\mu_1} > \sigma_N, \tag{3.1}$$

where $\sigma_N \equiv \sigma_N(r_+, r_-)$ denotes the unique positive solution of the equation

$$x + x^2 + \cdots + x^N = \frac{r_+}{r_-}. \tag{3.2}$$

In what follows we shall assume that condition (3.1) is satisfied.

Letting

$$F_i(t, y) \equiv P[X_t = i,\ C_t \le y], \quad t \ge 0,\ y \ge 0,\ i \in \mathcal{N}, \tag{3.3}$$

it is not difficult to show that the Kolmogorov forward equations for the process (X_t, C_t) are here given by

$$\begin{aligned}
\frac{\partial F_0(t,y)}{\partial t} + r_+\frac{\partial F_0(t,y)}{\partial y} &= -\lambda F_0(t,y) + \mu_1 F_1(t,y) - (\mu_1 - \mu_2)F_1(t,0) \\
\frac{\partial F_i(t,y)}{\partial t} - r_-\frac{\partial F_i(t,y)}{\partial y} &= \lambda F_{i-1}(t,y) - (\lambda + \mu_1)F_i(t,y) + \mu_1 F_{i+1}(t,y) \\
&\quad + (\mu_1 - \mu_2)(F_i(t,0) - F_{i+1}(t,0)), \quad i \in \mathcal{N}\backslash\{0, N\} \\
\frac{\partial F_N(t,y)}{\partial t} - r_-\frac{\partial F_N(t,y)}{\partial y} &= \lambda F_{N-1}(t,y) - \mu_1 F_N(t,y) + (\mu_1 - \mu_2)F_N(t,0).
\end{aligned} \tag{3.4}$$

Assuming as before, that the process is in equilibrium, we may set $F_i(t,y) \equiv F_i(y)$ and also $\frac{\partial}{\partial t}F_i(t,y) \equiv 0$ for all $i \in \mathcal{N}$ in (3.4) and, hence, obtain the system

$$\begin{aligned}
r_+F_0'(y) &= -\lambda F_0(y) + \mu_1 F_1(y) - (\mu_1 - \mu_2)F_1(0) \\
-r_-F_i'(y) &= \lambda F_{i-1}(y) - (\lambda + \mu_1)F_i(y) + \mu_1 F_{i+1}(y) \\
&\quad + (\mu_1 - \mu_2)(F_i(0) - F_{i+1}(0)), \quad i \in \mathcal{N}\backslash\{0, N\} \\
-r_-F_N'(y) &= \lambda F_{N-1}(y) - \mu_1 F_N(y) + (\mu_1 - \mu_2)F_N(0).
\end{aligned} \tag{3.5}$$

Notice that this system of differential equations is non-homogeneous, as opposed to situations in which no feedback is present.

Since credit accumulates whenever the server is idle, the solution to (3.5) must satisfy the boundary condition

$$F_0(0) = 0. \tag{3.6}$$

Also, letting

$$p_i \equiv \lim_{y\to\infty} F_i(y) = \lim_{t\to\infty} P[X_t = i], \quad i \in \mathcal{N}, \tag{3.7}$$

the limiting distribution of the (non-Markov) process (X_t), we must obviously have

$$\sum_{i\in\mathcal{N}} p_i = 1. \tag{3.8}$$

Finally, the solution to (3.5) should satisfy the rate balance equations

$$\lambda p_i = \mu_1 \left(p_{i+1} - F_{i+1}(0)\right) + \mu_2 F_{i+1}(0), \quad i \in \mathcal{N}\backslash\{N\}. \tag{3.9}$$

We note that, by letting $y \to \infty$ in (3.5), these balance equations are readily seen to be equivalent to

$$\lim_{y\to\infty} F_i'(y) = 0, \quad i \in \mathcal{N}. \tag{3.10}$$

3.2.2 Stationary joint distribution

By differentiating (3.5) we obtain a homogeneous system of differential equations for the derivatives

$$f_i(y) \equiv F_i'(y)\ , \quad i \in \mathcal{N}, \tag{3.11}$$

which is conveniently written down in matrix notation as

$$\mathbf{f}'(y) = R^{-1}Q^T\mathbf{f}(y). \tag{3.12}$$

Here,

$$\mathbf{f}(y) \equiv \left(f_0(y), f_1(y), \ldots, f_N(y)\right)^T,$$

and R and Q are the $(N+1)\times(N+1)$ matrices

$$R \equiv \operatorname{diag}(r_+, \overbrace{-r_-, -r_-, \ldots, -r_-}^{N}), \tag{3.13}$$

and

$$Q \equiv \begin{pmatrix} -\lambda & \lambda & 0 & \cdots & \\ \mu_1 & -(\lambda+\mu_1) & \lambda & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & 0 & \mu_1 & -(\lambda+\mu_1) & \lambda \\ & \cdots & 0 & \mu_1 & -\mu_1 \end{pmatrix}. \tag{3.14}$$

As in Section 2.3 we first turn to the eigenvalues of the matrix $R^{-1}Q^T$. This can now be done simply by invoking Lemma 2.2 for the current situation. We find

Lemma 3.1 *The eigenvalues $\xi_j,\ j \in \mathcal{N}$, of $R^{-1}Q^T$ are all real and simple; ordering them in increasing magnitude we have $\xi_0 < 0,\ \xi_1 = 0$, and $\xi_j > 0$ for $j = 2, \ldots, N$.*

Since the eigenvalues of $R^{-1}Q^T$ are all distinct, the general solution of (3.12) is of the form

$$\mathbf{f}(y) = \sum_{j=0}^{N} c_j \exp\{\xi_j y\}\, \mathbf{v}^{(j)},$$

where as before for every $j \in \mathcal{N}$, the vector $\mathbf{v}^{(j)}$ is the suitably normalized eigenvector corresponding to the eigenvalue ξ_j and c_j is a constant. However, the boundary conditions (3.10) and the above lemma are readily seen to imply that with the exception of c_0 all constants c_i must vanish. As a consequence we must have

$$\mathbf{f}(y) = c_0 \exp\{\xi_0 y\}\, \mathbf{v}^{(0)}$$

for some constant c_0. Upon integrating this result and writing $\mathbf{v}$ for $-c_0\mathbf{v}^{(0)}/\xi_0$, it now follows immediately that

$$\mathbf{F}(y) = \mathbf{p} - \exp\{\xi_0 y\}\, \mathbf{v}, \tag{3.15}$$

where $\mathbf{p} \equiv (p_0, p_1, \ldots, p_N)^T$ with $p_i,\ i \in \mathcal{N}$, as defined in (3.7).

Since $\mathbf{v}$ is, apart from normalization, the unique eigenvector of $R^{-1}Q^T$ corresponding to the single negative eigenvalue ξ_0, its components $v_i,\ i \in \mathcal{N}$, satisfy the relations

$$\begin{aligned} \mu_1 v_1 &= (\lambda + r_+ \xi_0) v_0 \\ \mu_1 v_{i+1} &= (\lambda + \mu_1 - r_- \xi_0) v_i - \lambda v_{i-1}, \quad i \in \mathcal{N} \backslash \{0, N\}. \end{aligned} \tag{3.16}$$

As for the components of $\mathbf{p}$, it is clear that these cannot be determined in a straightforward manner as in (2.2), since (X_t) is not Markovian due to the presence of feedback. However, the boundary conditions (3.6) and (3.9) entail that the components of $\mathbf{p}$ should satisfy the recurrence relations

$$\begin{aligned} p_0 &= v_0 \\ \mu_2 p_{i+1} &= \lambda p_i - (\mu_1 - \mu_2) v_{i+1}, \quad i \in \mathcal{N} \backslash \{N\}, \end{aligned} \tag{3.17}$$

besides the normalization condition (3.8). Thus we have a total of $2N+2$ linearly independent equations for the $2N+2$ quantities v_i and $p_i,\ i \in \mathcal{N}$.

Summarizing, we have found the following.

Theorem 3.2 *The stationary joint distribution $F_i(y) \equiv P[X_t = i,\ C_t \leq y],\ i \in \mathcal{N} = \{0, 1, \ldots, N\},\ y \geq 0$, of the process (X_t, C_t) is given by*

$$F_i(y) = p_i - v_i \exp\{\xi_0 y\}, \tag{3.18}$$

where ξ_0 is the smallest eigenvalue of $R^{-1}Q^T$, and the quantities p_i and $v_i,\ i \in \mathcal{N}$, are determined by (3.8) and the recurrence relations (3.16) and (3.17).

It is now a simple exercise to determine the stationary distribution of the process (X_t, C_t), once ξ_0 is known. To obtain more information on this smallest eigenvalue we turn back to the sequence of polynomials $\{\Delta_i^*\}_{i=0}^{\infty}$ in (2.12), which are here given by

$$\begin{aligned}
&\Delta_0^*(x) = 1, \quad \Delta_1^*(x) = x + \frac{\lambda}{r_+} - \frac{\mu_1}{r_-}, \\
&\Delta_i^*(x) = \left(x - \frac{\lambda + \mu_1}{r_-}\right)\Delta_{i-1}^*(x) - \frac{\lambda\mu_1}{{r_-}^2}\Delta_{i-2}^*(x), \quad i = 2, 3, \ldots.
\end{aligned} \tag{3.19}$$

Letting

$$\sigma \equiv \lim_{N\to\infty} \sigma_N = \frac{r_+}{r_+ + r_-}, \tag{3.20}$$

and indicating dependence of R, Q and ξ_0 on N we can state the following result, which is immediate from Lemmas 2.1 and 2.6.

Theorem 3.3 *The smallest eigenvalue $\xi_0^{(N)}$ of the matrix $R_N^{-1}Q_N^T$ is the unique negative zero of the polynomial $\Delta_N^*(x)$, $N = 1, 2, \ldots$, and $\{\xi_0^{(N)}\}_{N=1}^{\infty}$ constitutes a strictly decreasing sequence with limit*

$$\xi_0^{(\infty)} \equiv \lim_{N\to\infty} \xi_0^{(N)} = -\frac{\lambda - \mu_1\sigma}{r_+}. \tag{3.21}$$

This theorem tells us that the eigenvalue $\xi_0^{(N)}$ is the unique zero of $\Delta_N^*(x)$ in the interval $(\xi_0^{(\infty)}, 0)$. This knowledge enables us to use a very stable and efficient bisection algorithm to compute $\xi_0^{(N)}$ for any particular value of $N \geq 1$.

3.2.3 Examples

First looking into the case $N = 1$ we have $\sigma_1 = r_+/r_-$, and, by Theorem 3.3,

$$\xi_0^{(1)} = -\frac{\lambda}{r_+} + \frac{\mu_1}{r_-}. \tag{3.22}$$

The relations (3.16), (3.17) and (3.8) reduce to

$$\begin{aligned}
v_0 &= p_0 \\
v_1 &= v_0\, r_+/r_- \\
\mu_2 p_1 &= \lambda p_0 - (\mu_1 - \mu_2)v_1 \\
p_0 + p_1 &= 1.
\end{aligned}$$

It follows that

$$\begin{aligned}
v_0 \;=\; p_0 \;&=\; \frac{\mu_2}{\lambda + \mu_2 - (\mu_1 - \mu_2)\sigma_1} \\
v_1 \;&=\; \frac{\mu_2\sigma_1}{\lambda + \mu_2 - (\mu_1 - \mu_2)\sigma_1} \\
p_1 \;&=\; \frac{\lambda - (\mu_1 - \mu_2)\sigma_1}{\lambda + \mu_2 - (\mu_1 - \mu_2)\sigma_1}
\end{aligned}$$

Hence,

$$\begin{aligned} F_0(y) &= \frac{\mu_2}{\lambda+\mu_2-(\mu_1-\mu_2)\sigma_1}\left(1-\exp\left\{-\left(\frac{\lambda}{r_+}-\frac{\mu_1}{r_-}\right)y\right\}\right) \ , &(3.23)\\ F_1(y) &= \frac{1}{\lambda+\mu_2-(\mu_1-\mu_2)\sigma_1}\times \\ &\qquad \left(\lambda-(\mu_1-\mu_2)\sigma_1-\mu_2\sigma_1\exp\left\{-\left(\frac{\lambda}{r_+}-\frac{\mu_1}{r_-}\right)y\right\}\right) \ , &(3.24) \end{aligned}$$

and, in particular,

$$P[C>y]=\frac{(1+\sigma_1)\mu_2}{\lambda+\mu_2-(\mu_1-\mu_2)\sigma_1}\exp\left\{-\left(\frac{\lambda}{r_+}-\frac{\mu_1}{r_-}\right)y\right\}. \tag{3.25}$$

Next letting $N=2$ we have $\sigma_2=-\frac{1}{2}+\frac{1}{2}\sqrt{1+4r_+/r_-}$ and

$$\xi_0^{(2)}=\frac{1}{2}\left\{-\frac{\lambda}{r_+}+\frac{\lambda+2\mu_1}{r_-}-\sqrt{\left(\frac{\lambda}{r_+}+\frac{\lambda}{r_-}\right)^2+\frac{4\lambda\mu_1}{r_-^2}}\right\}. \tag{3.26}$$

The relations (3.16), (3.17) and (3.8) reduce to the system

$$\begin{aligned} v_0 &= p_0\\ \mu_1 v_1 &= (\lambda+r_+\xi_0^{(2)})v_0\\ \mu_1 v_2 &= (\lambda+\mu_1-r_-\xi_0^{(2)})v_1-\lambda v_0\\ \mu_2 p_1 &= \lambda p_0-(\mu_1-\mu_2)v_1\\ \mu_2 p_2 &= \lambda p_1-(\mu_1-\mu_2)v_2\\ p_0+p_1+p_2 &= 1 \ , \end{aligned}$$

which can easily be solved numerically.

3.3 Infinite waiting room and infinite credit reservoir

3.3.1 Preliminaries

We will use the results of the previous section to analyse the system in which both waiting room and credit reservoir are unbounded. Throughout this section $\mathcal{N}=\{0,1,\ldots\}$, but otherwise the notation and terminology of the previous section are maintained.

Obviously, stability of the service system (apart from the credit reservoir) now requires that $\lambda<\mu_2$. To have stability of the content of the credit reservoir we must impose

$$\left(1-\frac{\lambda}{\mu_1}\right)r_+-\frac{\lambda}{\mu_1}r_-<0,$$

since λ/μ_1 is the probability of the server being busy in an $M/M/1/\infty$ system with arrival rate λ and service rate μ_1. Thus, in order to have stability of the entire system (and $\mu_1 \geq \mu_2$), we will assume

$$\sigma < \frac{\lambda}{\mu_1} \leq \frac{\lambda}{\mu_2} < 1, \tag{3.27}$$

with σ given by (3.20).

As an aside we note that the assumption $\mu_1 \geq \mu_2$ is motivated by the application we have in mind, see Section 3.6, but the model yields to analysis without it. Actually, the ensuing analysis remains valid under the stability condition $\mu_1\sigma < \lambda < \mu_2$, apart from the case $\lambda = \mu_2\sigma$, which requires some additional work.

Defining $F_i(t, y)$ and $F_i(y)$ as in the previous section, our task is now to solve the system

$$\begin{aligned} r_+F_0'(y) &= -\lambda F_0(y) + \mu_1 F_1(y) - (\mu_1 - \mu_2)F_1(0) \\ -r_-F_i'(y) &= \lambda F_{i-1}(y) - (\lambda + \mu_1)F_i(y) + \mu_1 F_{i+1}(y) \\ &\qquad + (\mu_1 - \mu_2)(F_i(0) - F_{i+1}(0)), \quad i \in \mathcal{N}\backslash\{0\}, \end{aligned} \tag{3.28}$$

with boundary conditions (3.6), (3.8) and (3.10). As in Section 2.4, our approach is to let $N \to \infty$ in the solution for finite N, and subsequently check whether the resulting expressions satisfy the required conditions.

3.3.2 Stationary joint distribution

In view of Theorems 3.2 and 3.3 we obtain

$$F_i(y) = p_i - v_i \exp\left\{-(\lambda - \mu_1\sigma)\frac{y}{r_+}\right\}, \quad i \in \mathcal{N}, \tag{3.29}$$

if we let $N \to \infty$. Also, the recurrence relation (3.16) reduces to

$$\begin{aligned} v_1 &= \sigma v_0 \\ v_{i+1} &= \left(\frac{\lambda}{\mu_1\sigma} + \sigma\right) v_i - \frac{\lambda}{\mu_1} v_{i-1}, \quad i \in \mathcal{N}\backslash\{0\}, \end{aligned} \tag{3.30}$$

from which we immediately obtain

$$v_i = \sigma^i v_0, \quad i \in \mathcal{N}. \tag{3.31}$$

With this result the relation (3.17) becomes

$$\begin{aligned} p_0 &= v_0 \\ \mu_2 p_{i+1} &= \lambda p_i - (\mu_1 - \mu_2)\sigma^{i+1} v_0, \quad i \in \mathcal{N}. \end{aligned} \tag{3.32}$$

Writing

$$P(x) \equiv \sum_{i=0}^{\infty} p_i x^i,$$

we subsequently get

$$(\lambda x - \mu_2)P(x) = \frac{\mu_1 \sigma x - \mu_2}{1 - \sigma x} v_0, \tag{3.33}$$

that is,

$$P(x) = \frac{v_0}{\lambda - \mu_2 \sigma} \left\{ \frac{(\mu_1 - \mu_2)\sigma}{1 - \sigma x} + \frac{\lambda - \mu_1 \sigma}{1 - (\lambda/\mu_2)x} \right\}. \tag{3.34}$$

It follows that

$$p_i = \frac{v_0}{\lambda - \mu_2 \sigma} \left\{ (\mu_1 - \mu_2)\sigma^{i+1} + (\lambda - \mu_1 \sigma) \left(\frac{\lambda}{\mu_2} \right)^i \right\}, \quad i \in \mathcal{N}. \tag{3.35}$$

It remains to determine v_0, but (3.8) tells us that $P(1) = 1$, so that (3.33) yields

$$v_0 = \frac{(\mu_2 - \lambda)(1 - \sigma)}{\mu_2 - \mu_1 \sigma}. \tag{3.36}$$

Substitution of (3.31), (3.35) and (3.36) in (3.29) finally gives us the expression for $F_i(y)$ given below in Theorem 3.4, which is readily checked to satisfy the required conditions.

Theorem 3.4 *The stationary joint distribution* $F_i(y) \equiv P[X_t = i,\ C_t \leq y],\ i \in \mathcal{N} = \{0, 1, \ldots\},\ y \geq 0,$ *of the process* (X_t, C_t) *is given by*

$$\begin{aligned} F_i(y) &= \frac{(\mu_2 - \lambda)(1 - \sigma)}{\mu_2 - \mu_1 \sigma} \\ &\times \left\{ \frac{\lambda - \mu_1 \sigma}{\lambda - \mu_2 \sigma} \left(\frac{\lambda}{\mu_2} \right)^i + \sigma^i \left(\frac{(\mu_1 - \mu_2)\sigma}{\lambda - \mu_2 \sigma} - \exp\left\{ -(\lambda - \mu_1 \sigma) \frac{y}{r_+} \right\} \right) \right\}. \end{aligned}$$

Corollary 3.5 *The stationary marginal distributions of the number of customers in the system and of the content of the credit reservoir are given by*

$$P[X = i] = \frac{(\mu_2 - \lambda)(1 - \sigma)}{(\lambda - \mu_2 \sigma)(\mu_2 - \mu_1 \sigma)} \left\{ (\lambda - \mu_1 \sigma) \left(\frac{\lambda}{\mu_2} \right)^i + (\mu_1 - \mu_2)\sigma^{i+1} \right\}, i \in \mathcal{N}. \tag{3.37}$$

and

$$P[C > y] = \frac{\mu_2 - \lambda}{\mu_2 - \mu_1 \sigma} \exp\left\{ -(\lambda - \mu_1 \sigma) \frac{y}{r_+} \right\}, \quad y \geq 0. \tag{3.38}$$

When we remove the feedback by setting $\mu_1 = \mu_2 = \mu$ the service system behaves as an independent $M/M/1$ system, and so the distribution of the number of customers in the system must be geometric, as indeed is indicated by (3.37). Moreover, the expression given in (3.38) reduces to (2.32) with $\rho = \lambda/\mu$, which is not surprising, since in Section 2.4.2 the model without feedback is discussed.

3.3.3 Stationary sojourn and waiting time distributions

We can obtain the distribution of the stationary sojourn time S of an arbitrary customer by conditioning on the state of the Markov process (X_t, C_t) on arrival of the customer. Indeed, invoking PASTA we easily get

$$\begin{aligned} P[S > s] &= \sum_{i=0}^{\infty} \left\{ P\left[E_{i+1}(\mu_1) > \frac{\mu_2}{\mu_1} s\right] F_i(0) \right. \\ &\quad + \int_0^{sr_-} P\left[E_{i+1}(\mu_1) > \frac{y}{r_-} + \frac{\mu_2}{\mu_1}\left(s - \frac{y}{r_-}\right)\right] dF_i(y) \\ &\quad \left. + \int_{sr_-}^{\infty} P[E_{i+1}(\mu_1) > s] dF_i(y) \right\}, \end{aligned}$$

where $E_{i+1}(\mu_1)$ denotes an Erlang-distributed random variable with parameters $i+1$ and μ_1, representing the amount of work in the system (the time required to serve all customers present at rate μ_1) immediately *after* the arrival of a customer, given that this customer finds i customers in the system. Subsequent substitution of the result of Theorem 3.4 yields after tedious but straightforward calculations an explicit expression for the distribution of S. The distribution of the stationary waiting time W may be obtained by a similar calculation. The results are summarized in the next theorem.

Theorem 3.6 *The distribution of the stationary sojourn time S is given by*

$$P[S > s] = \zeta \exp\left\{-\lambda\sigma^{-1}(1-\sigma)s\right\} + (1-\zeta)\exp\left\{-(\mu_2 - \lambda)s\right\}, \quad s \geq 0, \tag{3.39}$$

while the distribution of the stationary waiting time W satisfies

$$P[W > s] = \zeta\sigma \exp\left\{-\lambda\sigma^{-1}(1-\sigma)s\right\} + \eta \exp\left\{-(\mu_2 - \lambda)s\right\}, \quad s \geq 0, \tag{3.40}$$

where

$$\zeta \equiv \frac{(\mu_1 - \mu_2)(\mu_2 - \lambda)\sigma}{(\lambda - \mu_2\sigma)(\mu_2 - \mu_1\sigma)} < 1 \quad \text{and} \quad \eta \equiv \frac{\lambda(\lambda - \mu_1\sigma)(1-\sigma)}{(\lambda - \mu_2\sigma)(\mu_2 - \mu_1\sigma)} < 1 - \zeta\sigma.$$

Clearly, when $\mu_1 > \mu_2$, then $\zeta > 0$, so that S has a hyperexponential distribution. When $\mu_1 = \mu_2$ the sojourn time of a customer is not influenced by credit and therefore its distribution is simply the sojourn time distribution in an $M/M/1$-queue, and hence exponential. Indeed, $\zeta = 0$ in this case.

3.4 Alternative analysis via discretization

The model in which both waiting room and credit reservoir are bounded can in principle be analysed in a way similar to that of Section 3.2, with the complication that all N nonzero eigenvalues of the matrix $R^{-1}Q^T$, rather than only one, play a role in the solution. As a result of this complicating aspect it does not seem possible to obtain the solution of the model with infinite waiting room and finite credit reservoir by letting N tend to infinity as in Section 3.3.2. By way of introduction to our alternative (approximative) approach to this problem in the next section, we will now present an alternative analysis for the model in which both waiting room and credit reservoir are unbounded.

The basic idea behind the analysis is to discretize the state space for credit by observing *quanta* of credit rather than *volume* of credit. Indeed, we imagine that at the beginning of each idle period the reservoir receives a quantum of credit whose size is initially zero but increases at rate r_+ during the idle period, so that the size at the end of the idle period is exponentially distributed with parameter λ/r_+. We also imagine that during busy periods these quanta of credit are drained at rate r_- in their order of arrival, and disposed of as soon as their sizes are reduced to zero. We shall denote the number of credit quanta in the reservoir at time t by $\tilde{C}_t$. In other words, $\tilde{C}_t$ can be viewed as the state of a counter at time t, which increases by one at the beginning of each idle period and decreases by one each time a quantum of credit has been disposed of.

Our interest now focuses on the two-dimensional process $(X_t, \tilde{C}_t)$, for which we want to compute the stationary distribution. Once we know this distribution we shall be able to calculate the stationary distribution of the original process (X_t, C_t).

The process $(X_t, \tilde{C}_t)$ does not constitute a Markov process, since the length of an idle period of the server determines the size of a credit quantum, and, hence, influences future behaviour of the process after the idle period. However, if we disregard periods of time corresponding to idle periods of the server and consider the process only during busy periods, then we are dealing with a process which *is* a Markov process, since the sizes of the credit quanta currently present (including the one that is being drained, if any) are independent of the past. Note in particular that the *reduction in size* of the quantum that is currently being drained (if any) is determined by the past of the new process, but, since idle periods and hence credit quanta before being drained are exponentially distributed (with parameter λ/r_+), the past of the new process does not provide any information about the current *size* of the quantum, which is still exponentially distributed.

Letting $q(i,j)$ denote the stationary probability that the new process is in state (i,j), the balance equations for this process are readily seen (see also Figure 3.2) to be given by

$$(\lambda+\mu_2)q(i,0) = \lambda q(i-1,0) + (\lambda r_-/r_+)q(i,1) + \mu_2 q(i+1,0), \quad i \in \mathcal{N}\backslash\{0\}, \tag{3.41}$$

with the convention $q(0,0) \equiv 0$, and

$$\begin{aligned}
(\lambda(1+r_-/r_+)+\mu_1)q(1,1) &= \mu_2 q(1,0) + (\lambda r_-/r_+)q(1,2) + \mu_1 q(2,1) && (3.42)\\
(\lambda(1+r_-/r_+)+\mu_1)q(1,j) &= \mu_1 q(1,j-1) + (\lambda r_-/r_+)q(1,j+1) + \mu_1 q(2,j), \\
&\qquad\qquad j \in \mathcal{N}\backslash\{0,1\}, && (3.43)\\
(\lambda(1+r_-/r_+)+\mu_1)q(i,j) &= \lambda q(i-1,j) + (\lambda r_-/r_+)q(i,j+1) + \mu_1 q(i+1,j), \\
&\qquad\qquad i \in \mathcal{N}\backslash\{0,1\},\ j \in \mathcal{N}\backslash\{0\}. && (3.44)
\end{aligned}$$

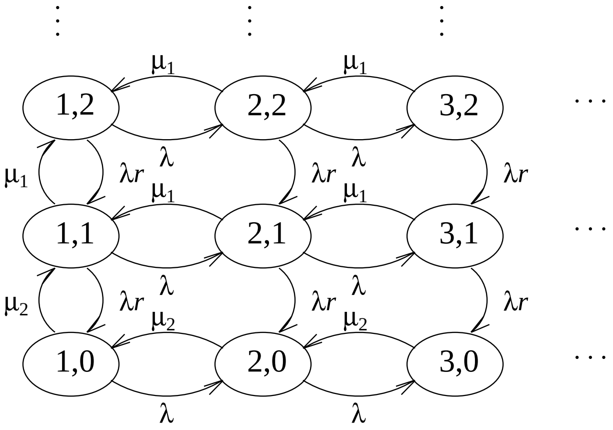


Figure 3.2: Flow diagram of the Markov process with $r = r_-/r_+$

The solution of the system (3.41) – (3.44) is given in the next lemma, which can easily be verified by substitution.

Lemma 3.7 *The stationary probabilities* $q(i, j)$, $i \in \mathcal{N}\backslash\{0\}$, $j \in \mathcal{N}$, *satisfy*

$$q(i, j) = b\sigma^i \left(\frac{\mu_1\sigma}{\lambda}\right)^j, \quad i, j \in \mathcal{N}\backslash\{0\}, \tag{3.45}$$

and

$$q(i, 0) = \frac{b\mu_1\sigma}{\lambda - \mu_2\sigma}\left(\left(\frac{\lambda}{\mu_2}\right)^i - \sigma^i\right), \quad i \in \mathcal{N}\backslash\{0\}, \tag{3.46}$$

with σ *as defined in (3.20) and* b *being a normalizing constant.*

As an aside we note that a deductive proof of this lemma may be based on the fact that for $j \geq 1$ the probabilities $q(i, j)$ must be of the form

$$q(i, j) = b\alpha^i\beta^j,$$

which is revealed by a careful analysis of the transition structure.

We next look at the complete process $(X_t, \tilde{C}_t)$ and let $p(i, j)$ denote the stationary probability that the process is in state (i, j) (state $(0, j)$ now corresponds to an idle server and j quanta of credit in the reservoir, including the one in development). We recall that the process $(X_t, \tilde{C}_t)$ is not a Markov process, because the sojourn time of the process in state $(0, j)$ equals the amount of credit that is added during that period, which, in turn, influences future behaviour of the process. However, it is clear that $p(i, j)/q(i, j)$ equals the stationary probability that the server is busy, and hence is constant for all $i \in \mathcal{N}\backslash\{0\}$ and $j \in \mathcal{N}$. Moreover, it is intuitively obvious that the rate balance equations

$$\begin{aligned} \lambda p(0, 1) &= \mu_2 p(1, 0) \\ \lambda p(0, j) &= \mu_1 p(1, j - 1), \quad j \in \mathcal{N}\backslash\{0, 1\}, \end{aligned} \tag{3.47}$$

must hold true, a result that can be formally justified by Miyazawa's *rate conservation law* (see [70], in particular the arguments on page 17). Interestingly, substitution of the preceding results in (3.41) – (3.44) leads to equations for the probabilities $p(i,j)$ which are precisely the balance equations that we would be justified in writing down directly if $(X_t, \tilde{C}_t)$ *were* a Markov process. We can now conclude the following.

Theorem 3.8 *The stationary distribution* $p(i,j) \equiv P[X_t = i,\ \tilde{C}_t = j],\ i,\ j \in \mathcal{N}$, *of the process* $(X_t, \tilde{C}_t)$ *is given by* $p(0,0) = 0$,

$$p(i,0) = \frac{c\mu_1\sigma}{\lambda - \mu_2\sigma}\left(\left(\frac{\lambda}{\mu_2}\right)^i - \sigma^i\right), \quad i \in \mathcal{N}\backslash\{0\}, \tag{3.48}$$

and

$$p(i,j) = c\sigma^i\left(\frac{\mu_1\sigma}{\lambda}\right)^j, \quad i \in \mathcal{N},\ j \in \mathcal{N}\backslash\{0\}, \tag{3.49}$$

where c is a normalization constant given by

$$c \equiv \frac{(\mu_2 - \lambda)(\lambda - \mu_1\sigma)(1-\sigma)}{\mu_1(\mu_2 - \mu_1\sigma)\sigma}. \tag{3.50}$$

We can now easily recover the result of Theorem 3.4. Indeed, given the number of credit quanta at an arbitrary point in time, their sizes must be independent and identically distributed according to an exponential distribution with parameter λ/r_+. Hence, with $E_j(\lambda/r_+)$ denoting an Erlang-distributed random variable with parameters j and λ/r_+, we have

$$F_i(y) = p(i,0) + \sum_{j=1}^{\infty} p(i,j)P[E_j(\lambda/r_+) \le y]. \tag{3.51}$$

Substitution of (3.48) – (3.50) and a little algebra subsequently lead to the required result.

3.5 Infinite waiting room and finite credit reservoir

3.5.1 Preliminaries

We will now assume that the credit reservoir has finite capacity K, but otherwise maintain the notation and assumptions of the previous section. Evidently, the stability condition for this model is $\lambda/\mu_2 < 1$. We will not analyse the model directly but rather apply a modification of the approach of Section 3.4 to a model which approximates the model at hand. As a result we obtain an approximation for the stationary distribution of (X_t, C_t), which, however, can be made arbitrarily accurate at the cost of increasing computing time.

The approximative model differs from the model at hand in the way credit is collected and spent. As in Section 3.4 , we let credit be composed of quanta, of which a maximum number M, say, is now allowed in the credit reservoir. Instead of collecting one credit quantum in the reservoir during an idle period, as in Section 3.4, we now collect a random number of credit quanta in the following way. At the beginning of each idle period the reservoir, if it is not full, receives a credit quantum whose size is initially zero but increases at rate r_+ until *either* the idle period *or* an exponentially distributed spell of mean $1/\nu$ has ended, whichever happens first. In the latter case, and if there is still room, a second credit quantum is added whose size again grows at rate r_+ until either the remaining idle period or the length of a new, exponentially distributed spell of mean $1/\nu$, has ended. If the latter happens first, a third quantum is added to the reservoir if possible, and so on, until either the complete idle period has come to an end or the total number of credit quanta has reached level M, whichever happens first. Note that letting $\nu = 0$ amounts to creating one credit quantum per idle period, as in Section 3.4.

Clearly, the size of each credit quantum is exponentially distributed with mean $r_+/(\lambda + \nu)$. Moreover, if there were no restriction on the number of credit quanta, the total volume of credit added during an idle period would be exponentially distributed with mean r_+/λ. As in Section 3.4, we imagine that during busy periods of the server the credit quanta are drained at rate r_- in their order of arrival, and disposed of as soon as their sizes are reduced to zero.

It is intuitively clear that the approximative model will resemble the original model closer and closer by letting $1/\nu$, and hence the mean size of a credit quantum, tend to zero and simultaneously letting M, the maximum number of credit quanta in the reservoir, tend to infinity in such a way that

$$\frac{Mr_+}{\lambda + \nu} = K, \tag{3.52}$$

with K being the maximum volume of credit in the reservoir in the original model. This intuition is validated by the numerical results of Section 3.5.3.

In the next subsection we will show how to perform an exact analysis of the approximative model. We let $\tilde{C}_t$ again denote the number of credit quanta in the reservoir at time t, and our aim is to compute the stationary distribution of the process $(X_t, \tilde{C}_t)$. It will be convenient to let

$$\gamma_n \equiv \left(1 - \frac{\lambda}{\lambda + \nu}\right)^{n-1} \frac{\lambda}{\lambda + \nu}, \quad \bar{\gamma}_n \equiv \sum_{m=n}^{\infty} \gamma_m = \left(1 - \frac{\lambda}{\lambda + \nu}\right)^{n-1}, \quad n = 1, 2, \ldots,$$

so that γ_n is the probability that n credit quanta are added to the reservoir during an idle period, conditional on the reservoir having sufficient capacity. Evidently, the stability conditions for the approximative and the original model are identical.

3.5.2 Analysis of the approximative model

As in Section 3.4, we first disregard periods of time corresponding to idle periods of the server and consider the process $(X_t, \tilde{C}_t)$ only during busy periods, as a result of which we are dealing with a two-dimensional Markov process with state space $\{(i, j),\ i \in \mathcal{N} \backslash \{0\},\ j \in$

$\mathcal{M}\}$, where $\mathcal{M} \equiv \{0, 1, \ldots, M\}$. Letting $q(i,j)$ denote the stationary probability that this new process is in state (i,j), and writing

$$\tau \equiv (\lambda + \nu)\frac{r_-}{r_+}, \tag{3.53}$$

the balance equations for the new process are given by

$$(\lambda + \mu_2)q(i,0) = \lambda q(i-1,0) + \tau q(i,1) + \mu_2 q(i+1,0), \quad i \in \mathcal{N}\backslash\{0\}, \tag{3.54}$$

with the convention $q(0,0) \equiv 0$, and

$$\begin{aligned}
(\lambda + \mu_1 + \tau)q(1,j) &= \mu_2\gamma_j q(1,0) + \mu_1 \sum_{k=1}^{j-1} \gamma_{j-k} q(1,k) + \tau q(1,j+1) + \mu_1 q(2,j), \\
&\qquad\qquad\qquad\qquad\qquad\qquad j \in \mathcal{M}\backslash\{0, M\}, \qquad (3.55) \\
(\lambda + \tau)q(1,M) &= \mu_2\bar{\gamma}_M q(1,0) + \mu_1 \sum_{k=1}^{M-1} \bar{\gamma}_{M-k} q(1,k) + \mu_1 q(2,M), \qquad (3.56) \\
(\lambda + \mu_1 + \tau)q(i,j) &= \lambda q(i-1,j) + \tau q(i,j+1) + \mu_1 q(i+1,j), \\
&\qquad\qquad\qquad i \in \mathcal{N}\backslash\{0,1\},\ j \in \mathcal{M}\backslash\{0\}, \qquad (3.57)
\end{aligned}$$

with the convention $q(i, M+1) \equiv 0$. Evidently, one of these equations is redundant; in what follows we shall not use (3.56).

To solve these equations we first note that taking $j = M$ in equation (3.57) yields the difference equation

$$(\lambda + \mu_1 + \tau)q(i,M) = \lambda q(i-1,M) + \mu_1 q(i+1,M), \quad i \in \mathcal{N}\backslash\{0,1\}. \tag{3.58}$$

The most general solution of this difference equation gives $q(i,M)$ as a linear combination of ith powers of the roots of the equation

$$(\lambda + \mu_1 + \tau)x = \lambda + \mu_1 x^2. \tag{3.59}$$

But one of the roots being larger than 1, its weight in the linear combination must be zero, and so

$$q(i,M) = A_{0,0}\omega^i, \quad i \in \mathcal{N}\backslash\{0\}, \tag{3.60}$$

where $A_{0,0}$ is some constant and ω is the smallest root of the equation (3.59), that is,

$$\omega \equiv \frac{\lambda + \mu_1 + \tau - \sqrt{(\lambda + \mu_1 + \tau)^2 - 4\lambda\mu_1}}{2\mu_1}. \tag{3.61}$$

Subsequently substituting (3.60) into equation (3.57) for $j = M - 1$, we obtain an inhomogeneous difference equation for the probabilities $q(i, M-1)$. It follows that

$$q(i, M-1) = A\omega^i + A_{1,1} i\omega^i, \quad i \in \mathcal{N}\backslash\{0\}, \tag{3.62}$$

for some constant A and

$$A_{1,1} \equiv \frac{\tau\omega A_{0,0}}{\lambda - \mu_1\omega^2},$$

since the first term in (3.62) constitutes the (summable) solution of the homogeneous equation, while the second term is a particular solution of the inhomogeneous equation. It will be convenient to represent $q(i, M-1)$ as

$$q(i, M-1) = \omega^i \sum_{k=0}^{1} A_{1,k} \binom{1+i}{k}, \quad i \in \mathcal{N}\backslash\{0\}, \tag{3.63}$$

with a constant $A_{1,0}$ which is yet to be determined.

Our next step is to substitute (3.63) into (3.57) for $j = M-2$ and solve the resulting difference equation for the probabilities $q(i, M-2)$. Thus proceeding, we can work our way back from $q(i, M)$ to $q(i, 1)$ and find after some simple calculations that

$$q(i, M-j) = \omega^i \sum_{k=0}^{j} A_{j,k} \binom{j+i}{k}, \quad j \in \mathcal{M}\backslash\{M\},\ i \in \mathcal{N}\backslash\{0\}, \tag{3.64}$$

with constants $A_{j,k}$ satisfying

$$A_{j,k} = \frac{\mu_1\omega^2 A_{j,k+1} + \tau\omega A_{j-1,k-1}}{\lambda - \mu_1\omega^2}, \quad j \in \mathcal{M}\backslash\{0, M\},\ k = 1, 2, \ldots, j, \tag{3.65}$$

where $A_{j,j+1} \equiv 0$.

Upon substitution in equation (3.54) of the expression we have thus found for $q(i, 1)$ we subsequently obtain an inhomogeneous difference equation for the probabilities $q(i, 0)$. Solving this equation under the conditions $q(0, 0) = 0$ and $\sum_i q(i, 0) < \infty$, yields after some algebra

$$q(i, 0) = \sum_{k=0}^{M} A_{M,k} \left\{ \omega^i \binom{M+i}{k} - \left(\frac{\lambda}{\mu_2}\right)^i \binom{M}{k} \right\}, \quad i \in \mathcal{N}, \tag{3.66}$$

with constants $A_{M,k}$ satisfying

$$A_{M,k} = \frac{\omega((\lambda + \mu_2 - 2\mu_2\omega)A_{M,k+1} - \mu_2\omega A_{M,k+2} - \tau A_{M-1,k})}{\lambda - (\lambda + \mu_2)\omega + \mu_2\omega^2}, \quad k \in \mathcal{M}\backslash\{M\}, \tag{3.67}$$

where $A_{M,M} = A_{M,M+1} \equiv 0$.

At this point it is convenient to express the stationary state probabilities $p(i, j)$ of the process $(X_t, \tilde{C}_t)$ in terms of the probabilities $q(i, j)$ in a way similar to that of Section 3.4. Indeed, it is clear that $p(i, j)/q(i, j)$ must be equal to the stationary probability that the server is busy, and, hence, must be constant for $i \in \mathcal{N}\backslash\{0\}$ and $j \in \mathcal{M}$. Moreover, the rate balance equations

$$\begin{aligned} (\lambda + \nu)p(0, 1) &= \mu_2 p(1, 0) \\ (\lambda + \nu)p(0, j) &= \nu p(0, j-1) + \mu_1 p(1, j-1), \quad j = 2, 3, \ldots, M-1 \\ \lambda p(0, M) &= \nu p(0, M-1) + \mu_1\{p(1, M-1) + p(1, M)\}, \end{aligned} \tag{3.68}$$

must hold true, by Miyazawa's rate conservation law (see again [70]). It follows in particular that the equations (3.55), which we have not used yet, may be rewritten as

$$(\lambda + \mu_1 + \tau)p(1,j) = \lambda p(0,j) + \tau p(1,j+1) + \mu_1 p(2,j), \quad j \in \mathcal{M}\backslash\{0,M\}, \tag{3.69}$$

precisely the equation balancing probability flow in and out of state $(1,j)$ which we would have written down directly if the process $\{(X_t, \tilde{C}_t),\ t \geq 0\}$ were a Markov process.

After a little algebra we can now conclude the following.

Theorem 3.9 *The stationary distribution $p(i,j) \equiv P[X_t = i,\ \tilde{C}_t = j]$, $i \in \mathcal{N}$, $j \in \mathcal{M}$, of the process $(X_t, \tilde{C}_t)$ is given by $p(0,0) = 0$,*

$$p(0,j) = \frac{c}{\nu} \sum_{k=1}^{j} \left(\frac{\nu}{\lambda+\nu}\right)^{k} \{\mu_1 + \delta_{kj}(\mu_2 - \mu_1)\}\ q(1, j-k), \quad j \in \mathcal{M}\backslash\{0,M\}, \tag{3.70}$$

$$p(0,M) = \frac{c}{\lambda} \sum_{k=0}^{M} \left(\frac{\nu}{\lambda+\nu}\right)^{k-1+\delta_{k0}} \{\mu_1 + \delta_{kM}(\mu_2 - \mu_1)\}\ q(1, M-k), \tag{3.71}$$

and

$$p(i,j) = cq(i,j), \quad i \in \mathcal{N}\backslash\{0\},\ j \in \mathcal{M}, \tag{3.72}$$

where c is a normalization constant, δ_{ij} is Kronecker's delta as before, the $q(i,j)$ are given by (3.64) and (3.66), and the $(M+1)(M+2)/2$ constants $A_{j,k}$, $j \in \mathcal{M}$, $k = 0, 1, \ldots, j$, are determined (apart from normalization) by $A_{M,M} = 0$, and the linear equations (3.65), (3.67), and (3.69).

It is now a matter of routine to calculate performance characteristics such as the mean number of customers in the system and hence, by applying Little's law, the mean sojourn time of a customer.

3.5.3 Validation of the approximative model

To investigate how well the discretization technique works we compare the mean sojourn time of a customer in the original model with the mean sojourn time of a customer in the approximative model. The results for the original model have been obtained by simulation, while the results for the approximative model have been calculated via the procedure outlined in the previous subsection.

In the last column of Table 3.1 we have listed the values of $E[S_K]$, the mean sojourn time of a customer in the original model when the credit reservoir is bounded by K, for several values of K and six sets of values of the other parameters. Throughout we have chosen $\lambda = 1$ and $r_+ = 1$. We have also indicated a 95% confidence interval for the values of $E[S_K]$. In each case these confidence intervals were obtained from 40 runs of 10^7 arrivals.

μ_1	μ_2	r_-	K	$E[S_{K,M}]$						$E[S_K]$
			M	2	4	6	8	10	20	
2	1.5	0.5	1	1.617	1.603	1.598	1.596	1.594	1.591	1.587 ± 0.003
			2	1.436	1.410	1.400	1.395	1.392	1.385	1.378 ± 0.003
			5			1.151	1.143	1.138	1.128	1.117 ± 0.002
2	1.5	1	1	1.776	1.777	1.777	1.777	1.778	1.778	1.777 ± 0.003
			2	1.644	1.646	1.647	1.648	1.648	1.648	1.648 ± 0.003
			5			1.436	1.437	1.437	1.437	1.437 ± 0.003
2	1.5	2	1	1.879	1.884	1.886	1.886	1.887	1.888	1.887 ± 0.003
			2	1.805	1.817	1.821	1.823	1.824	1.826	1.826 ± 0.003
			5			1.732	1.735	1.737	1.741	1.744 ± 0.003
1.5	1.1	0.5	2	8.865	8.870	8.871	8.872	8.872	8.873	8.855 ± 0.066
			3		8.436	8.438	8.439	8.440	8.441	8.438 ± 0.060
			5			7.726	7.727	7.728	7.730	7.695 ± 0.065
1.5	1.1	1	2	9.455	9.488	9.499	9.504	9.507	9.513	9.507 ± 0.069
			3		9.322	9.339	9.347	9.352	9.361	9.373 ± 0.072
			5			9.120	9.133	9.140	9.155	9.162 ± 0.074
1.5	1.1	2	2	9.750	9.774	9.781	9.784	9.786	9.790	9.789 ± 0.071
			3		9.720	9.729	9.734	9.736	9.741	9.746 ± 0.082
			5			9.681	9.685	9.688	9.693	9.693 ± 0.064

Table 3.1: Convergence of $E[S_{K,M}]$ to its limit $E[S_K]$ ($\lambda = 1, r_+ = 1$)

Other columns in Table 3.1 list the corresponding values of $E[S_{K,M}]$, the mean sojourn time in the approximative model when the number of credit quanta is bounded by M, for various values of M. The parameter ν in the approximative model has always been chosen such that (3.52) is satisfied. Computing the quantities $E[S_{K,M}]$ requires a fraction of a second, which is negligible compared to the effort required to obtain $E[S_K]$. We can conclude from the results of Table 3.1 that $E[S_{K,M}]$ is a good approximation for $E[S_K]$ already for small values of M.

3.6 Two-level traffic shaper

3.6.1 Introduction

We will show that the system with finite credit reservoir can serve as a model for a *traffic regulation* mechanism operating on a very bursty traffic source in an ATM network. However, we will first give some information on traffic regulation in ATM networks.

Probably the best-known techniques for regulating (or *shaping*) cell streams entering an ATM network are variants of the *leaky-bucket* mechanism, also known as *token-bank throttle* or *generic cell rate algorithm* (see, e.g., [13] and [88]). The mechanism has recently been included in the recommendations of the ITU and the ATM Forum; for more information we refer to Section 2.1 of [83] and the references there.

The basic operation of the leaky-bucket scheme is simple. Before entering the network cells are sent to a buffer. In order to get access to the network a cell at the head of the line in this buffer needs a token from a token bank. If no token is available the cell has to wait. Tokens arrive deterministically and evenly spaced to the token bank at a rate which equals the specified average arrival rate of the source (the *sustainable cell rate*). The capacity of the token bank is finite and tokens that arrive at a full bank are blocked and lost. The token-bank throttle guarantees that the long-term average rate at which cells enter the network never exceeds the sustainable cell rate. However, for some period of time, determined by the size of the token bank and the cell arrival stream, the scheme permits a higher rate, equalling, in fact, the actual cell arrival rate.

There have been proposals for extensions of the token-bank throttle, which behave as *two-level regulators*, see, e.g., [34, 40, 68, 77, 82], see also [83]. Such shapers have the additional feature that during periods in which the token bank is nonempty, the rate at which cells enter the network will not exceed a second specified rate (the *peak cell rate*). Again, this scheme allows a higher rate than the sustainable cell rate for some period of time, but the input rate will never exceed the peak cell rate. The size of the token bank determines the *maximum burst duration* (i.e., the maximum duration of a peak-rate period). Thus, a two-level shaper forces the traffic to conform to three (previously negotiated) traffic parameters: sustainable cell rate, peak cell rate and maximum burst duration.

Exact analyses of various versions of the token-bank throttle have appeared in many papers (see [13, 14, 15, 34, 66, 88], and the references mentioned therein). Crucial in these analyses is that at any moment in time either the cell buffer or the token bank is empty. Hence, the process describing both the number of cells waiting and the content of the token bank is essentially one-dimensional. This feature is not shared by a two-level traffic shaper and therefore its analysis is much more difficult. The behaviour of a two-level traffic shaper has been studied through simulations in [77].

We note that in [34] an exact analysis using a fluid approximation was carried out. However, this paper does not really describe a two-level traffic shaper as we have just introduced it, since in this model cells are delayed in the cell buffer only to enforce sustainable cell rate conformation, and not to enforce peak rate conformation. Rather, cells that violate the peak rate constraint are marked and then transmitted into the network anyway, to be discarded later within the network if necessary. As a consequence, this model *does* share the above-mentioned feature, along with the relative ease of the ensuing performance analysis.

We will now picture a setting involving a two-level traffic shaper which is well described by the model of Section 3.1. Indeed, consider a cell stream generated by an on-off source with exponentially distributed on-times and off-times, for which the on-times are short and the arrival rate during on-times is high. Ignoring the duration of the on-times and the discrete nature of the cells, the stream may be looked upon as a Poisson process in which an event is the generation of a burst of information (corresponding to a batch of cells) whose total size is exponentially distributed with mean θ^{-1}, say.

Next, suppose that the cell stream is sent to a buffer at the entrance of a network, access to which is regulated by a two-level traffic shaper. We ignore the discrete nature of tokens (as we did for the cells) and regard them as a fluid commodity, which, following [63], we may call *credit*. This credit then flows continuously into a reservoir (corresponding to the token bank) as long as it is not completely filled, at a constant rate r_+, say. Credit is released from the reservoir only when the information buffer is nonempty, the output rate being equal to the input rate r_+ if the credit reservoir is empty, but to $r_+ + r_-$, say, if the reservoir is nonempty. Note that, as far as the content of the credit reservoir is concerned, this is equivalent to saying that the reservoir fills at rate r_+ (as long as it is not full) during idle periods of the server, and empties at rate r_- (as long as it is nonempty) during busy periods of the server. The information itself is released from the information buffer at rate $r_+ + r_-$ as long as there is credit, and at rate r_+ otherwise, implying that bursts of information leave the network at rate $\theta(r_+ + r_-)$ as long as there is credit, and at rate θr_+ otherwise.

It is not difficult to see that by choosing $\mu_1 \equiv \theta(r_+ + r_-)$ and $\mu_2 \equiv \theta r_+$ and interpreting customers as bursts of information the model of Section 3.1 matches the setting described above.

3.6.2 Numerical results

With the results of this chapter we can evaluate the behaviour of the two-level traffic shaper in the setting described above. To illustrate this, we look into the influence of K, the maximum amount of credit in the reservoir, on the mean sojourn time. Furthermore, we study the trade-off, as K decreases, between extra delay on the one hand and reduction of burstiness of the output stream on the other.

In Figure 3.3, we display the mean sojourn time as a function of the maximum amount of credit for two different parameter settings. When $K = 0$ we are dealing with a simple $M/M/1$ system in which the mean sojourn time equals $1/(\mu_2 - \lambda)$. For $K > 0$, the mean sojourn time has been calculated with the method of Section 3.5. Note that, as $K \to \infty$,

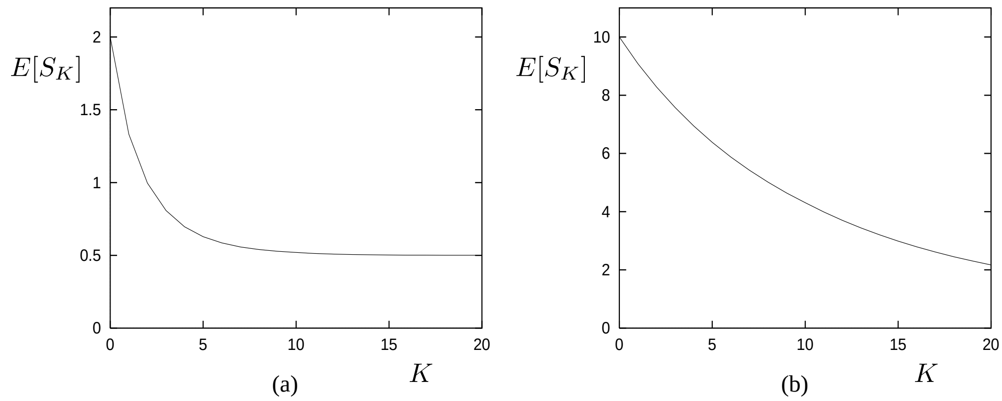


Figure 3.3: Behaviour of $E[S_K]$ as a function of K for the parameter settings
(a) $\lambda = 1$, $\mu_1 = 3$, $\mu_2 = 1.5$, $r_- = r_+ = 1$ and (b) $\lambda = 1$, $\mu_1 = 2.2$, $\mu_2 = 1.1$, $r_- = r_+ = 1$

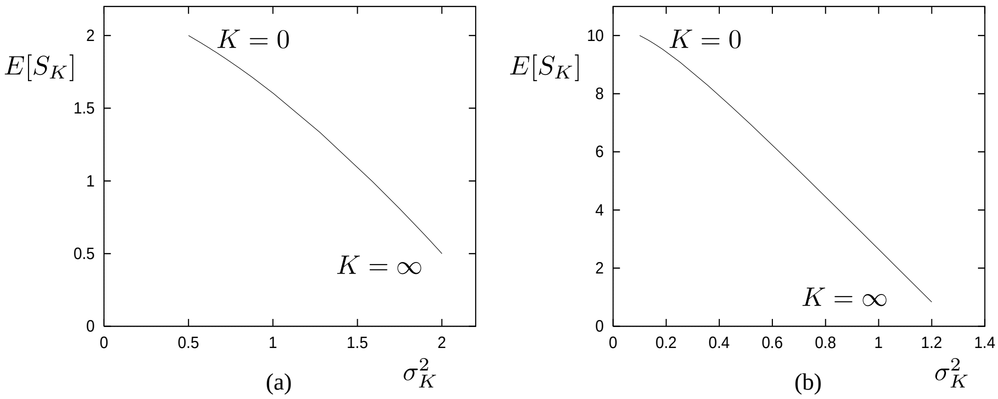


Figure 3.4: Behaviour of $E[S_K]$ and σ_K^2 when K runs from 0 to ∞ for the parameter settings (a) $\lambda = 1$, $\mu_1 = 3$, $\mu_2 = 1.5$, $r_- = r_+ = 1$ and (b) $\lambda = 1$, $\mu_1 = 2.2$, $\mu_2 = 1.1$, $r_- = r_+ = 1$

the mean sojourn time tends to $1/(\mu_1 - \lambda)$, the mean sojourn time in an $M/M/1$ system with service rate μ_1. The parameter settings of Figures 3.3(a) and 3.3(b) lead to mean sojourn times of 0.5 and 0.833, respectively, when $K \to \infty$. We observe that a large part of the possible reduction of the mean sojourn time is already achieved for small values of K.

In Figure 3.4 we show, for the same parameter settings as before, the relation between mean delay and burstiness of the output stream, when the maximum amount of credit in the reservoir varies from 0 to ∞. We quantify the burstiness of the output stream by the variance σ_K^2 of the output rate of the service system, that is,

$$\sigma_K^2 = p_0(0-\lambda)^2 + p_1(\mu_1 - \lambda)^2 + p_2(\mu_2 - \lambda)^2,$$

where p_0, p_1 and p_2 are the fractions of time the output rate equals 0, μ_1 and μ_2, respectively. Obviously, when $K = 0$ then $p_0 = 1-\lambda/\mu_2$, $p_1 = 0$ and $p_2 = \lambda/\mu_2$. For $K > 0$, the fractions have been calculated numerically using the method of Section 3.5. When $K = \infty$, then, clearly, $p_0 = 1 - \lambda/\mu_1$, $p_1 = \lambda/\mu_1$ and $p_2 = 0$. Graphs such as Figure 3.4 may be used to make the trade-off between the benefit of burstiness reduction and the drawback of extra delay.

Chapter 4

Some two-buffer fluid models

4.1 Introduction

In this chapter we consider two closely related *systems* of fluid queues. Both these systems consist of two infinitely large reservoirs. The first one receives its input from an exponential on-off source and is emptied at a constant rate. The second reservoir is driven by the first one.

For the way in which the regulation of this second reservoir takes place we will consider two possibilities, leading to the two systems just mentioned. In the first part of this chapter we will consider the situation in which the content of the second reservoir increases at times when the first reservoir is nonempty, while it decreases if this is not the case (unless also the second reservoir itself is empty). We will refer to this model as the *tandem model*, although it will be formulated slightly more general than the model in which the output of the first reservoir feeds into the second one.

In the second part of the chapter a similar model with two buffers is studied, which will be referred to as the *dual model*. The reason for this is that it may be regarded dual to the tandem model just described, in the same way as the models in Sections 2.4.2 and 2.5.2 are dual: in the dual model the content of the second reservoir increases when the first one is empty, and decreases otherwise.

In both models, (M_t) will denote the two-state Markov process that regulates the first reservoir. Furthermore, the contents of the first and second buffer at time t will be given by D_t and C_t respectively. One reason for choosing this "counter-alphabetical" notation is that it ensures that the relation of the current model(s) with those in the next chapter (where D and C are used as mnemonics for data and credit) will not become blurred by non-corresponding notation. Another, more significant, reason is that it clarifies how the models fit into the context of standard Markov-modulated fluid models in Section 1.2: a fluid reservoir with content C_t is driven by a Markov process (X_t) which in this case is given by $(X_t) \equiv (M_t, D_t)$. Note that M_t is included, since (D_t) is not a Markov process. Thus, in both models we have a fluid reservoir that is regulated by a Markov process with a nondenumerable state space $\mathcal{N} \equiv \{0,1\} \times [0,\infty)$.

Our aim is to derive the stationary marginal distributions of (D_t) and (C_t), as well as the stationary joint distribution of (M_t, D_t, C_t), both for the tandem model and its dual. We show how a variety of techniques from renewal theory, Laplace transformation, stochastic integration and standard queueing theory can be fruitfully used to achieve this goal. A starting point for the analyses is the relationship between the waiting time in a $G/G/1$-queue and the content of the second buffer. In fact, the duality between the two models considered is closely related to the duality between the $M/G/1$- and the $G/M/1$-queue.

As we already mentioned in Section 1.3.4, networks of fluid queues have not received much attention in the literature, since their solutions do not have a product form. In addition to the references on Markov-modulated fluid networks in that section, we mention here [47]. In this paper an n-node tandem fluid queue with increasing Lévy input and deterministic linear internal flows is analyzed, generalizing the dam model of [80]. It is one of the few examples where the (Laplace-transform of) the stationary joint distribution of the contents of more than one reservoir is actually derived. More recent work in this context can be found in [44] and [45] and references there.

The rest of this chapter is organized as follows. The tandem model is described in Section 4.2. In Section 4.3 we derive the stability conditions for this system. We obtain the marginal distributions of the buffers in stationarity in Section 4.4. Other performance measures such as the expected buffer content, the utilization and the decay rate of the buffers are also given. In Section 4.5 the first main result of this chapter is given: the stationary joint distribution of the process (M_t, D_t, C_t) for the tandem model.

In Section 4.6 the dual model is formulated. Again, we first determine the stability conditions, which is done in Section 4.7. As a by-product of this stability analysis we find the limiting distribution of the content of the second buffer *given* that the first one is empty. In Section 4.8 we use this information to derive the second main result of this chapter: the joint stationary distribution of the process (M_t, D_t, C_t) for the dual model. Finally, in 4.9 we present another method to find this result, based on the classical spectral expansion method, enabling us to compare both solution techniques.

Notation We note that several variables and parameters are used in both the tandem and the dual model. Although the *interpretation* of these variables and parameters remains the same throughout the chapter, the actual algebraic *form* may differ for the two models.

4.2 Tandem model

We consider a fluid system consisting of two infinitely large reservoirs, with contents D_t and C_t at time t respectively, and a continuous-time Markov process (M_t), wich is characterised by its state space $\{0, 1\}$ and its Q-matrix, which is given by

$$Q = \begin{pmatrix} -a & a \\ b & -b \end{pmatrix}. \tag{4.1}$$

The first reservoir is driven by (M_t) in the following manner. When (M_t) is in state 1, the content of the first buffer increases at constant rate d_+, otherwise it decreases at rate d_-,

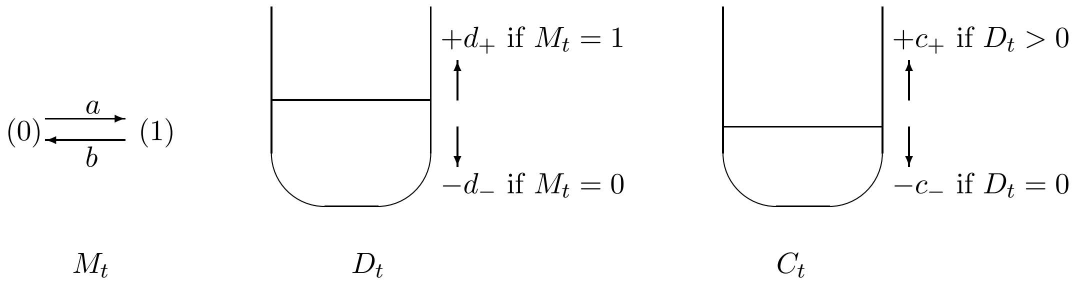


Figure 4.1: Interaction between the subsystems of the tandem system

provided that it is not empty. The second buffer is driven by the first one, in such a way that its content increases at rate c_+ when the first buffer is not empty, and else decreases at rate c_-, provided that the second buffer is not empty. We note that c_+, c_-, d_+ and d_- are *positive* numbers, and that the meaning of these symbols is reflected in the notation (d and c for the rates pertaining to the first and second buffer respectively)

A schematic overview of the behaviour of the interaction between the processes (M_t), (D_t) and (C_t) is given in Figure 4.1, while a realization of the processes (D_t) and (C_t) is given in Figure 4.2. The parameter values used here and in other figures pertaining to the tandem model are $a = 1$, $b = 2$, $d_+ = 2$, $d_- = 6$, $c_+ = 3$ and $c_- = 2.5$.

Observe that the stochastic process (M_t, D_t, C_t) is a Markov process with state space $\{0, 1\} \times S$, where

$$S = \{(x, y) \in \mathbb{R} \mid y \geq xc_+/d_+\}.$$

This model may be used to describe a fluid version of the classical tandem model: two fluid buffers with constant release rates are placed in series, the first buffer is fed by an

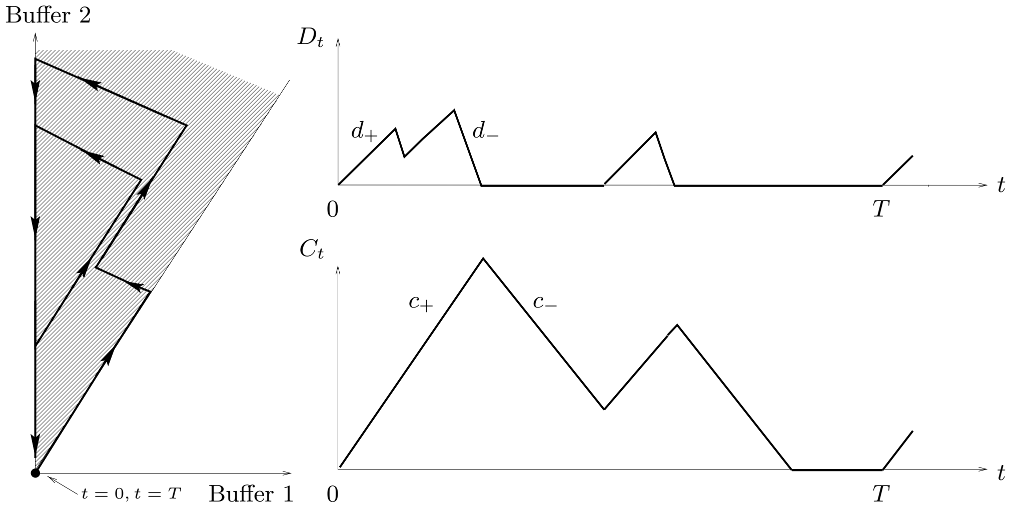


Figure 4.2: Realisation of the buffer content processes for the tandem model

exponential on-off source while the second one is fed by the output of the first. In this case $d_- = c_+ + c_-$; notice however, that our model can handle slightly more general scenarios.

The analysis of the model is now carried out through a study of the process (M_t, D_t, C_t). For simplicity, we assume that $M_0 = 1$ and $D_0 = C_0 = 0$.

4.3 Tandem model: stability

Before we study the limiting behaviour as $t \to \infty$ for this model, we need to find out under which conditions the process (M_t, D_t, C_t) has a limiting distribution. In doing so, we will need some preliminary results.

Let (V_t) be the potential net input process for the first buffer, defined as

$$V_t = \int_0^t r_{M_s} ds. \tag{4.2}$$

where the rates are $r(0) = -d_-$ and $r(1) = d_+$. The Markov process (V_t) is called a *random velocity process* in [79], where more information on these processes can be found. Analogous to the analysis on page 12 of this reference, we find the following result.

Proposition 4.1 *Let Y be the first entrance time of (V_t) into 0, and let $P_{i,x}$ denote the probability measure under which (M_t) starts in i and (V_t) in x, for $i = 0, 1$, $x \geq 0$. The corresponding expectation operator is denoted by $E_{i,x}$. We have,*

$$E_{0,x}\, e^{-sY} = e^{\lambda_1(s)x}, \tag{4.3}$$

$$E_{1,x}\, e^{-sY} = \frac{b\, e^{\lambda_1(s)x}}{s + b - \lambda_1(s)d_+}, \tag{4.4}$$

where

$$\lambda_1(s) = \frac{\eta(s) - \sqrt{\xi(s)}}{2d_-d_+}, \tag{4.5}$$

with

$$\eta(s) = bd_- - ad_+ + s(d_- - d_+),$$

and

$$\xi(s) = (bd_- - ad_+)^2 + 2s(d_- + d_+)(bd_- + ad_+) + s^2(d_- + d_+)^2.$$

Proof. For fixed $s \geq 0$, let $\phi_0(x)$ and $\phi_1(x)$ denote the left-hand sides of (4.3) and (4.4), respectively. By conditioning on the first transition epoch of (M_t), we obtain the following integral equations:

$$\phi_0(x) = e^{-(s+a)x/d_-} + a \int_0^{x/d_-} e^{-(s+a)u}\, \phi_1(x - ud_-)\, du, \tag{4.6}$$

$$\phi_1(x) = b \int_0^{\infty} e^{-(s+b)u}\, \phi_0(x + ud_+)\, du. \tag{4.7}$$

Differentiation with respect to x leads to the differential equation

$$\boldsymbol{\phi}'(x) = A\,\boldsymbol{\phi}(x),$$

where $\boldsymbol{\phi}(x) = (\phi_0(x), \phi_1(x))^T$ and

$$A = \begin{pmatrix} -(s+a)/d_- & a/d_- \\ -b/d_+ & (s+b)/d_+ \end{pmatrix}.$$

The eigenvalues of A are $\lambda_1(s)$ as in (4.5) and

$$\lambda_2(s) = \frac{\eta + \sqrt{\xi(s)}}{2d_-d_+}. \tag{4.8}$$

In particular, $\phi_0(x) = c_1\, e^{\lambda_1(s)x} + c_2\, e^{\lambda_2(s)x}$, for some constants c_1 and c_2. It is not difficult to see that $\lambda_1(s) \leq 0$ and $\lambda_2(s) \geq 0$, for $s \geq 0$. Since $\phi_0(x)$ must remain bounded as $x \to \infty$ and $\phi_0(0) = 1$, we have $c_2 = 0$ and $c_1 = 1$, which gives (4.3). Finally, if we insert (4.3) into (4.7), we find (4.4) after some algebra. □

Corollary 4.2 *The expectations $E_{0,x}Y$ and $E_{1,x}Y$ are finite if and only if*

$$bd_- - ad_+ > 0.$$

When this condition holds, we have for any $x \geq 0$,

$$E_{0,x}Y = \frac{(a+b)x}{bd_- - ad_+}, \tag{4.9}$$

$$E_{1,x}Y = \frac{d_- + d_+ + (a+b)x}{bd_- - ad_+}. \tag{4.10}$$

We are now ready to derive the stability conditions for the process (M_t, D_t, C_t). Clearly, this process is regenerative. As regeneration epochs we may, and henceforth will, choose the times when (M_t, D_t, C_t) is in state $(1, 0, 0)$, including time 0. The point at issue is under which conditions the expected length of a regeneration cycle is finite. This question is answered in the following theorem.

Theorem 4.3 *The process (M_t, D_t, C_t) is regenerative with regeneration cycles that have a non-lattice distribution with finite expectation, if and only if*

$$\frac{bd_-}{c_+d_- + c_-d_+ + c_+d_+} - \frac{a}{c_-} > 0. \tag{4.11}$$

Proof. Let $B_0, B_1, \ldots$ and $I_0, I_1, \ldots$ denote respectively the lengths of the busy periods and the idle periods of (D_t), and let B (I) be a generic busy (idle) period.

We first consider the process at embedded points in time. Specifically, let Z_i be the content of the second buffer at the beginning of the ith busy period, $i = 0, 1, 2, \ldots$. We have $Z_0 = 0$ and

$$Z_{i+1} = [Z_i + c_+B_i - c_-I_i]^+,\ i = 0, 1, 2 \ldots, \tag{4.12}$$

where $[x]^+$ denotes the maximum of x and 0. In other words, Z_i has the same distribution as the waiting time of the ith customer in an $M/G/1$-queue with interarrival times distributed as $c_- I$ and service times distributed as $c_+ B$.

Obviously, (Z_i) is a regenerative process, the regeneration epochs being the (discrete) times i where $Z_i = 0$. By e.g. [7], Proposition VIII.1.3, the expected length of a regeneration cycle is finite if and only if

$$c_- \, EI > c_+ \, EB. \tag{4.13}$$

Let τ and T denote the first strictly positive regeneration epoch of (Z_i) and (M_t, D_t, C_t), respectively. We have

$$T = B_0 + I_0 + \cdots + B_{\tau-1} + I_{\tau-1},$$

and consequently by Wald's Lemma

$$ET = E\tau \, E(B + I) = E\tau \, (EB + \frac{1}{a}).$$

Thus, it remains to show that both $E\tau$ and EB are finite if and only if (4.11) holds. From Corollary 4.2 we know that a necessary and sufficient condition for EB to be finite, is

$$bd_- - ad_+ > 0. \tag{4.14}$$

In that case,

$$EB = \frac{d_- + d_+}{bd_- - ad_+}, \tag{4.15}$$

and hence, in view of (4.13), $E\tau$ is finite if and only if (4.11) holds. Sufficiency and necessity of (4.11) for $ET < \infty$ is now immediate, since (4.14) is implied by (4.11).

Finally, T must have a non-lattice distribution, since $P[T \in [x, x + h]] > 0$ for any choice of $x \geq 0$ and $h > 0$. This can be established by choosing the first transition epochs of (M_t) appropriately. □

Remark 4.1 The relationship between the buffer content and the waiting time distribution of an associated $G/G/1$-queue is a well-known property of fluid models in a "two-state random environment" or "with random disruptions". In such models, the buffer content is driven by an i.i.d. sequence $\{(D_i, U_i)\}$ of down- and up-times, such that the content increases at down-times and decreases at up-times, see e.g. [18] and [46]. Indeed, for the second buffer, the two-state environment consists of the i.i.d. sequence $\{(B_i, I_i)\}$ of busy and idle periods of the first buffer.

Corollary 4.4 *If (4.11) holds, a random vector (M, D, C) exists, to which (M_t, D_t, C_t) converges in distribution as $t \to \infty$.*

We will henceforth assume (4.11) to be satisfied and interpret (M, D, C) as the state of the system in stationarity. Its distribution will be denoted by $\mathbf{F} = (F_0(dx, dy), F_1(dx, dy))$, where

$$\begin{aligned} F_i(dx, dy) &= P[M = i,\ D \in dx, C \in dy] \\ &= \lim_{t\to\infty} P[M_t = i,\ D_t \in dx, C_t \in dy], \qquad i \in \{0, 1\}. \end{aligned} \tag{4.16}$$

Before trying to find this distribution we will first concentrate on the marginal distributions of D and C.

4.4 Tandem model: stationary marginal distributions

The stationary distribution of the first buffer is well-known (see e.g. Theorem 2.3). We include it for completeness.

Proposition 4.5 *The stationary distribution of the process (M_t, D_t) is for $x \geq 0$ given by*

$$P[M = 0, D \leq x] = \frac{b}{a+b} - \frac{a}{a+b}\frac{d_+}{d_-}e^{-\alpha x}, \tag{4.17}$$

$$P[M = 1, D \leq x] = \frac{a}{a+b} - \frac{a}{a+b}e^{-\alpha x}, \tag{4.18}$$

where

$$\alpha = \frac{b}{d_+} - \frac{a}{d_-}. \tag{4.19}$$

Note that $\alpha > 0$ is equivalent to (4.14), and hence implied by (4.11).

Corollary 4.6 *The stationary distribution of the content of the first buffer decreases exponentially with* decay rate α. *Moreover, the* utilization ρ_d *of the first buffer is given by*

$$\rho_d = P[D > 0] = \frac{a}{a+b}\frac{d_- + d_+}{d_-}. \tag{4.20}$$

The expected stationary buffer content is

$$ED = \frac{d_+}{a+b}\frac{\rho_d}{1-\rho_d}.$$

Next, we consider the stationary distribution of the content of the second buffer. In view of Remark 4.1, we may view the second buffer as a system with random disruptions. We will therefore proceed as in [46] and [4].

Let B and I be a generic busy and idle period of the first buffer, as before. By Proposition 4.1, the Laplace-Stieltjes transform (L_B, say) of B is given by

$$L_B(s) = \frac{b}{s + b - d_+\lambda_1(s)}, \tag{4.21}$$

with $\lambda_1(s)$ as in (4.5).

Let the stochastic variable Z have the limiting distribution of the Lindley process in (4.12). Since Z can be interpreted as the actual waiting time of an arbitrary customer in an $M/G/1$-queue, its Laplace-Stieltjes transform L_Z satisfies the Pollaczek-Khintchine formula, which leads to

$$L_Z(s) = \frac{s\,(c_-EI - c_+EB)}{c_-sEI - 1 + L_B(c_+s)}, \tag{4.22}$$

where EB is given in (4.15) and $EI = 1/a$.

The following lemma gives another interpretation of the distribution of Z.

Lemma 4.7 *The conditional distribution of* $(C \,|\, D = 0)$ *is the same as the distribution of* Z.

Proof. Assuming an equilibrium situation, we compare the content of the second buffer at two instants, namely at an arbitrary instant in an idle period of the first reservoir and at the end of such an idle period. Clearly, the age of the idle period is exponentially distributed with parameter a in both cases. Therefore, we may conclude that also the amount of fluid that has flown into the second buffer during the present idle period has the same distribution for both cases, and hence our statement follows. □

As a consequence we find by standard renewal theory that the distribution of C is given by the following mixture of two distributions,

$$C \overset{d}{=} \begin{cases} Z & \text{w.p. } 1 - \rho_d, \\ Z + c_+B^* & \text{w.p. } \rho_d, \end{cases} \tag{4.23}$$

where B^* is independent of Z and distributed as the residual lifetime of B. The Laplace-Stieltjes transform L_C of C therefore satisfies

$$L_C(s) = L_Z(s)\left(\rho_d\,\frac{1 - L_B(c_+s)}{c_+sEB} + (1 - \rho_d)\right). \tag{4.24}$$

The following theorem asserts that it is possible to invert this expression analytically.

Theorem 4.8 *The stationary marginal distribution of the process* (C_t) *is given by*

$$\begin{aligned} P[C = 0] &= 1 - \rho_c, \\ P[C \in dy] &= (1 - \rho_c)\,\frac{c_- + c_+}{c_+}\;e^{-\beta y}\;\times \\ &\qquad \left(\frac{a}{c_-} - \frac{c_+\nu\omega}{2}\int_0^y e^{-(\theta-\beta)u}\,H_0(0,u)\,du\right)\,dy, \qquad y > 0. \end{aligned}$$

Here,

$$H_0(0,y) = \frac{I_1(y\sqrt{\omega})}{y\sqrt{\omega}}, \tag{4.25}$$

where I_1 is the modified Bessel function of the first kind of order 1 (see also (4.58)), and

$$\begin{aligned}
\beta &= \frac{bd_-}{c_+d_- + c_-d_+ + c_+d_+} - \frac{a}{c_-}, &(4.26)\\
\theta &= \frac{bd_- + ad_+}{c_+(d_- + d_+)}, &(4.27)\\
\rho_c &= \frac{a}{a+b}\,\frac{c_- + c_+}{c_-}\,\frac{d_- + d_+}{d_-}, &(4.28)\\
\nu &= \frac{d_- + d_+}{c_+d_- + c_-d_+ + c_+d_+}, &(4.29)\\
\omega &= \frac{4abd_-d_+}{c_+^2(d_- + d_+)^2}. &(4.30)\\
& &(4.31)
\end{aligned}$$

Proof. To find the inverse Laplace transform of (4.24), we apply the shift $s \to s - \theta$ to obtain

$$L_C(s-\theta) = (1-\rho_c)\left(1 + \frac{a(c_- + c_+)}{c_+c_-(s-(\theta-\beta))} - \frac{s-\sqrt{s^2-\omega}}{s-(\theta-\beta)}\,\frac{c_- + c_+\nu}{2}\right), \tag{4.32}$$

Using the fact that the inverse Laplace transform of the function $s \mapsto s - \sqrt{s^2-\omega}$ is the function $y \mapsto \omega H_0(0,y)$, see e.g. [37, page 235,(28)], the result follows. □

Notice that the stability condition (4.11) can also be written as $\beta > 0$. Theorem 4.8 gives a probabilistic interpretation for this parameter, as well as for ρ_c. Moreover, since the expectation of C can be derived straightforwardly from L_C, we find the following counterpart to Corollary 4.6.

Corollary 4.9 *The stationary distribution of the content of the second buffer has* decay rate *β, in the usual sense that $\lim_{y\to\infty} -y^{-1}\log P[C > y] = \beta$. Moreover, the* utilization *of the second buffer is given by ρ_c in (4.28). The expected stationary buffer content is*

$$EC = \frac{bc_+}{a(a+b)}\,\frac{\rho_c}{1-\rho_c}\,\frac{\rho_d}{1-\rho_d}.$$

Remark 4.2 The fact that ρ_d and ρ_c are of the form (4.20) and (4.28) has an easy interpretation: if the first buffer is stable, the average input rate must be equal to the average output rate, in other words,

$$(d_+ + d_-)\,\frac{a}{a+b} = d_-\,P[D > 0],$$

which leads to (4.20). For the second buffer, a similar inflow-outflow analysis leads to

$$(c_+ + c_-)\, P[D > 0] = c_-\, P[C > 0],$$

which gives (4.28). To make these arguments rigorous, observe that by conditioning on D and using Lemma 4.7, we have $P[C > 0] = (1 - \rho_d)P[Z > 0]$, while $P[Z > 0]$ follows from

$$P[Z > 0] = 1 - \lim_{s\to\infty} L_Z(s) = \frac{c_+ EB}{c_- EI} = \frac{ac_+(d_- + d_+)}{c_-(bd_- - ad_+)}, \tag{4.33}$$

where we used (4.22), (4.15) and the fact that $EI = 1/a$. Equation (4.28) now follows immediately.

Equation (3.8) of [84] provides yet another way to derive $P[C = 0]$. Let the potential net input process for the second reservoir be given by

$$W_t = \int_0^t \left(c_+ \mathbf{1}_{\{D_s > 0\}} - c_- \mathbf{1}_{\{D_s = 0\}}\right) ds,$$

and let $\underline{W}_t = \inf_{s \le t} W_s$. As before, $\mathbf{1}_A$ denotes the indicator function of the event A. Since $d\underline{W}_t/dt = -c_-\ \mathbf{1}_{\{C_t = 0\}}$, we have

$$\begin{aligned} P[C = 0] &= \lim_{t\to\infty} t^{-1} \int_0^t \mathbf{1}_{\{C_s = 0\}}\, ds \ = \ -\lim_{t\to\infty} \frac{\underline{W}_t}{c_- t} \ = \ -\lim_{t\to\infty} \frac{W_t}{c_- t} \\ &= -\frac{1}{c_-}\, \{c_+ P[D > 0] - c_- P[D = 0]\} \ = \ P[D = 0] - \frac{c_+}{c_-}\, P[D > 0], \end{aligned}$$

where $P[D > 0]$ is given in (4.20).

Remark 4.3 The *decay rate* of the stationary distribution of the content of a fluid buffer has received much attention in the telecommunication literature, since it gives an indication of the probability of buffer overflow for large buffers. Moreover, once the decay rate has been established, efficient simulation procedures (importance sampling) can be used to estimate the actual loss probability, see e.g. [17]. The evaluation of the decay rate may be carried out using the theory of *large deviations*. In this context, [67] considers a two-node tandem fluid model, in which two buffers are connected in series via a channel of capacity c_1 (we use the notation of [67]). The output capacity of the second buffer is c_2 and the input rate into the first one is r_i at times when a modulating n-state Markov process is in state i, $i \in \{1, \ldots, n\}$. Let $R = \text{diag}\{r_1, r_2, \ldots, r_n\}$, and let Λ be the infinitesimal generator of the driving Markov process. Finally, define $c(\theta)$ to be the largest eigenvalue of the matrix $R + \Lambda/\theta$. Then the decay rate of the second buffer is the unique positive solution to $c(\theta) = c_2$. (Actually, the decay rate is the solution to a slightly more complicated equation – see (2) of [67] – but the present formulation suffices for our purposes.)

For $n = 2$, the model of [67] is a special case of the present model, and we may take $\Lambda = Q$, $r_1 = 0$, $r_2 = d_+ + d_-$, $c_1 = c_- + c_+ = d_-$ and $c_2 = c_-$. For this parameter setting, the solution of $c(\theta) = c_2$ is given by $\theta = b/(c_+ + d_+) - a/c_-$, which is in accordance with the expression we found for the decay rate in (4.26), see also Corollary 4.9.

Since the distribution of the buffer content is given in Theorem 4.8, we do not need (fast) simulation to find the probability of buffer overflow in this particular model.

4.5 Tandem model: stationary joint distribution

In this section we derive the joint distribution $\mathbf{F}$ of the random vector (M, D, C). The form of the distribution is easily established (see also Figure 4.2). As a consequence of Theorem 4.3, the state $(0, 0, 0)$ is a positive recurrent state of the Markov process (M_t, D_t, C_t). This state is entered via the set $\{(0, 0, y) \mid y \geq 0\}$ and left via the set $\{(1, x, y) \mid x \geq 0, y = xc_+/d_+\}$. Moreover, the set $\{0, 1\} \times \{(x, y) \mid y < xc_+/d_+\}$ is never entered. These considerations suggest that $\mathbf{F}$ be of the following form,

$$F_0(\{0, 0\}) = 1 - \rho, \tag{4.34}$$

$$F_0(\{0\}, dy) = \sigma_0(y)\, dy,\ y > 0, \tag{4.35}$$

$$F_1(dx, c_+/d_+dx) = \sigma_1(x)\, dx,\ x > 0, \tag{4.36}$$

$$F_i(dx, dy) = f_i(x, y)\, dx\, dy\ ,\ x > 0,\ y > xc_+/d_+,\ i = 0, 1, \tag{4.37}$$

for some constant ρ and certain densities σ_0, σ_1, f_0 and f_1. From Theorem 4.8, we have $F_0(\{0, 0\}) = P[C = 0] = 1-\rho_c$, so that the constant ρ is immediately given by ρ_c in (4.28). We now set out to find the densities in three consecutive steps.

Density $\boldsymbol{\sigma_0}$

By Lemma 4.7 we have

$$E\, e^{-sC}\, \mathbf{1}_{\{D=0\}} = (1 - \rho_d)L_Z(s), \tag{4.38}$$

with L_Z given in (4.22). Applying the shift $s \mapsto s - \theta$, as in (4.32), yields after some algebra

$$E\, e^{-(s-\theta)C}\, \mathbf{1}_{\{D=0\}} = (1 - \rho_c) \left(1 + \frac{a}{c_-(s - (\theta - \beta))} - \frac{s - \sqrt{s^2 - \omega}}{s - (\theta - \beta)} \frac{c_+\nu}{2}\right),$$

so that σ_0 is found by inverse Laplace transformation in exactly the same way as $P[C \in dy]$ in the proof of Theorem 4.8. We have, for $y > 0$,

$$\sigma_0(y) = (1 - \rho_c)\, e^{-\beta y} \left(\frac{a}{c_-} - \frac{c_+\nu\omega}{2} \int_0^y e^{-(\theta-\beta)u} H_0(0, u)\, du\right). \tag{4.39}$$

Density $\boldsymbol{\sigma_1}$

The expected sojourn time of the process (D_t, C_t) in the set $\{(\hat{x}, \hat{y}) \mid c_+\hat{x} = d_+\hat{y}, \hat{x} \leq x\}$ during the first regeneration period can be found by conditioning on the time the process stays on the line $\{(x, y) \mid x \geq 0, y = xc_+/d_+\}$ after $t = 0$. Since this time is exponentially distributed with parameter b, it follows after some calculations and applying the theory of regenerative processes, that

$$P[c_+D = d_+C, D \leq x] = \frac{1 - e^{-bx/d_+}}{bET}.$$

Since we also have that

$$1 - \rho_c = P[C = 0] = \frac{1}{ET}\frac{1}{a},$$

we obtain

$$P[c_+ D = d_+ C, D \leq x] = \frac{a(1-\rho_c)}{b}\left(1 - e^{-bx/d_+}\right). \tag{4.40}$$

Finally, by differentiating with respect to x, we find

$$\sigma_1(x) = (1-\rho_c)\frac{a}{d_+}e^{-bx/d_+} \tag{4.41}$$

Densities $\boldsymbol{f_0}$ and $\boldsymbol{f_1}$

This last step is the most difficult one. Our approach is to determine the densities f_0 and f_1 via a Laplace-transformed version of the stationary Kolmogorov forward equations for the Markov process (M_t, D_t, C_t). Thereto, we define the joint Laplace transforms q_i by

$$q_i(p,s) = E\mathbf{1}_{\{M=i\}}\, e^{-pD-sC}, \ \ i \in \{0,1\}, \ \ p, s \geq 0. \tag{4.42}$$

We will write $\mathbf{q}(p,s)$ for the column vector with entries $q_0(p,s)$ and $q_1(p,s)$.

Lemma 4.10 *The vector* $\mathbf{q}(p,s)$ *satisfies:*

$$A(p,s)\ \mathbf{q}(p,s) = B(p,s)\begin{pmatrix} q_0(\infty, s) \\ q_0(\infty,\infty) \end{pmatrix}, \tag{4.43}$$

where

$$A(p,s) = \begin{pmatrix} -a + d_- p - c_+ s & b \\ a & -d_+ p - c_+ s - b \end{pmatrix},$$

and

$$B(p,s) = \begin{pmatrix} d_- p - c_+ s - c_- s & c_- s \\ 0 & 0 \end{pmatrix}$$

Proof. We only prove the first row of the matrix equation, the second row can be proved in a similar manner.

Consider the stochastic processes $(X_i(t),\, t \geq 0)$, $i \in \{0,1\}$, defined by

$$X_i(t) = e^{-pD_t - sC_t}\,\mathbf{1}_{\{M_t = i\}}.$$

Notice that both these processes are of bounded variation. We denote the continuous part of $(X_i(t))$ by $(X_i^c(t))$. In particular, we have for $t > 0$,

$$X_0(t) = X_0(0) + X_0^c(t) + \sum_{0<u\leq t}[X_0(u) - X_0(u-)]. \tag{4.44}$$

We now concentrate on $(X_0(t))$. The derivative of $(X_0^c(t))$ is easily found to be

$$\frac{d}{dt}X_0^c(t) = X_0(t)\left(d_-p\mathbf{1}_{\{D_t>0\}} - c_+s\mathbf{1}_{\{D_t>0\}} + c_-s\mathbf{1}_{\{D_t=0,C_t>0\}}\right), \; t > 0. \tag{4.45}$$

Moreover, the pure jump part of $(X_0(t))$ can be written in stochastic integral form,

$$\sum_{0<u\le t}[X_0(u) - X_0(u-)] = -\int_0^t X_0(u-)\,dA_u + \int_0^t X_1(u-)\,dB_u, \tag{4.46}$$

where (A_t) and (B_t) denote the counting processes that count the number of jumps of (M_t) from state 0 to 1 and from 1 to 0, respectively. The stochastic intensities at time t of (A_t) and (B_t) are given by $a\,\mathbf{1}_{\{M_t=0\}}$ and $b\,\mathbf{1}_{\{M_t=1\}}$, respectively. Because $(X_0(u-))$ is a left-continuous adapted process, we have by the theory of stochastic integration, (see e.g. [65]) that

$$E\int_0^t X_0(u-)\,dA_u = E\int_0^t X_0(u-)\,a\,\mathbf{1}_{\{M_u=0\}}\,du, \tag{4.47}$$

and a similar result holds for the other integral in (4.46). If we now take expectations in (4.44) and use (4.45)–(4.47), we arrive at

$$\begin{aligned} EX_0(t) &= EX_0(0) \\ &+ d_-p\int_0^t EX_0(u)\mathbf{1}_{\{D_u>0\}}\,du - c_+s\int_0^t EX_0(u)\mathbf{1}_{\{D_u>0\}}\,du \\ &+ c_-s\int_0^t EX_0(u)\mathbf{1}_{\{D_u=0,C_u>0\}}\,du \\ &- a\int_0^t EX_0(u)\,du + b\int_0^t EX_1(u)\,du. \end{aligned}$$

Now differentiate both sides of the previous equation with respect to t and let $t \to \infty$. By the continuity of Laplace transforms, we obtain

$$\begin{aligned} 0 &= d_-p\Big(q_0(p,s) - q_0(\infty,s)\Big) - c_+s\Big(q_0(p,s) - q_0(\infty,s)\Big) \\ &+ c_-s\Big(q_0(\infty,s) - q_0(\infty,\infty)\Big) - a\,q_0(p,s) + b\,q_1(p,s). \end{aligned}$$

The first row of (4.43) now follows immediately. □

Notice that the quantities in the right-hand side of (4.43) are known. In particular, using (4.42), (4.38) and (4.22), we have

$$q_0(\infty,s) = (1-\rho_d)\frac{c_-s - ac_+sEB}{c_-s - a + aL_B(c_+s)}, \tag{4.48}$$

where EB is given in (4.15) and L_B in (4.21). Furthermore, $q_0(\infty,\infty) = P[M=0, D = 0, C=0] = P[C=0] = 1-\rho_c$ with ρ_c given in (4.28).

Solving $\mathbf{q}(p,s)$ from equation (4.43) yields for all $p, s \geq 0$,

$$q_1(p,s) = a\,\frac{(-d_-p + c_-s + c_+s)\,q_0(\infty,s) - c_-s\,(1-\rho_c)}{\det A(p,s)} \tag{4.49}$$

and

$$q_0(p,s) = \frac{b + d_+p + c_+s}{a}\,q_1(p,s),$$

which, after some algebra, reduces to

$$q_0(p,s) = \frac{b + d_+p + c_+s}{d_+(p + \lambda_2(c_+s))}\,q_0(\infty,s) \tag{4.50}$$

and

$$q_1(p,s) = \frac{a}{d_+(p + \lambda_2(c_+s))}\,q_0(\infty,s), \tag{4.51}$$

where $\lambda_2(s)$ is given in (4.8), and $q_0(\infty,s)$ in (4.48).

Remark 4.4 It is possible to derive $q_0(\infty,s)$ directly from Lemma 4.10. For this, write

$$\det A(p,s) = -d_-d_+(p + \lambda_1(c_+s))(p + \lambda_2(c_+s)),$$

where $\lambda_1(s)$ and $\lambda_2(s)$ are given in (4.5) and (4.8); recall that $\lambda_1(s) \leq 0 \leq \lambda_2(s)$ for $s \geq 0$. Since for all $p, s \geq 0$, $\mathbf{q}(p,s)$ must remain bounded, in particular for $p = -\lambda_1(c_+s)$, the numerator in (4.49) must be zero on the set $\{(p,s) \mid s \geq 0,\, p = -\lambda_1(c_+s)\}$. This gives a linear equation in $q_0(\infty,s)$, from which (4.48) follows.

We are now ready to specify the complete distribution of (M,D,C).

Theorem 4.11 *The stationary joint distribution* $\mathbf{F}$ *of the process* (M_t, D_t, C_t) *is of the form (4.34) – (4.37), where*

$$\begin{aligned}
\sigma_0(y) &= (1-\rho_c)\,e^{-\beta y}\left(\frac{a}{c_-} - \frac{c_+\nu\omega}{2}\int_0^y e^{-(\theta-\beta)u} H_0(0,u)\,du\right), && (4.52)\\
\sigma_1(x) &= (1-\rho_c)\,\frac{a}{d_+}\,e^{-bx/d_+}, && (4.53)\\
f_0(x,y) &= (1-\rho_c)\,\frac{\nu b c_-}{d_- + d_+}\,e^{-\frac{b}{d_+}x} \times && (4.54)\\
&\quad \left(\frac{d_+\gamma\omega}{b}\quad e^{-\theta\left(y-\frac{c_+}{d_+}x\right)}\quad H_1(x, y - \frac{c_+}{d_+}x)\right.\\
&\quad + \frac{a}{c_-}\quad e^{-\beta\left(y-\frac{c_+}{d_+}x\right)}\quad \{1 + x\omega\gamma\int_0^{y-\frac{c_+}{d_+}x} e^{-(\theta-\beta)u} H_0(x,u)du\}\\
&\quad \left. - \frac{c_+\nu\omega}{2}\quad e^{-\beta\left(y-\frac{c_+}{d_+}x\right)}\quad \int_0^{y-\frac{c_+}{d_+}x} e^{-(\theta-\beta)u} H_1(x,u)du\right),
\end{aligned}$$

$$\begin{aligned} f_1(x,y) &= (1-\rho_c)\,\frac{a}{d_+}\,e^{-\frac{b}{d_+}x} \;\times \qquad (4.55)\\ &\Bigg(\omega\gamma\, x \quad e^{-\theta\left(y-\frac{c_+}{d_+}x\right)}\;\; H_0(x, y-\frac{c_+}{d_+}x)\\ &\quad +\frac{a}{c_-} \quad e^{-\beta\left(y-\frac{c_+}{d_+}x\right)} \;\; \{1+x\omega\gamma\int_0^{y-\frac{c_+}{d_+}x} e^{-(\theta-\beta)u}H_0(x,u)du\}\\ &\quad -\frac{c_+\nu\omega}{2} \quad e^{-\beta(y-\frac{c_+}{d_+}x)} \quad \int_0^{y-\frac{c_+}{d_+}x} e^{-(\theta-\beta)u}H_1(x,u)du\Bigg)\,. \end{aligned}$$

Here, the functions H_0 and H_1 are given by

$$H_0(x,y) = \frac{I_1\left(\sqrt{\omega(y^2+2xy\gamma)}\right)}{\sqrt{\omega(y^2+2xy\gamma)}}, \tag{4.56}$$

$$\begin{aligned} H_1(x,y) &= \frac{y^2+xy\gamma}{y^2+2xy\gamma}H_0(x,y)\\ &\quad +\frac{xy\gamma}{y^2+2xy\gamma}\;\frac{I_0\left(\sqrt{\omega(y^2+2xy\gamma)}\right)+I_2\left(\sqrt{\omega(y^2+2xy\gamma)}\right)}{2} \end{aligned} \tag{4.57}$$

where I_i is the modified Bessel function of the first kind of order i, i.e.,

$$I_i(z) = \left(\frac{z}{2}\right)^i \sum_{k=0}^{\infty} \frac{\left(\frac{z}{2}\right)^{2k}}{k!(k+i)!}. \tag{4.58}$$

Furthermore,

$$\gamma = \frac{c_+(d_-+d_+)}{2d_-d_+}, \tag{4.59}$$

while all other constants are the same as in Theorem 4.8.

Proof. It remains to be shown how $f_0(x,y)$ and $f_1(x,y)$ can be found from $q_0(p,s)$ and $q_1(p,s)$. First, inverse transformation of $q_0(p,s)$ and $q_1(p,s)$ with respect to p yields the functions

$$\begin{aligned} g_0(s) &= \left(\delta_0(x)+\frac{b-d_+\lambda_2(c_+s)+c_+s}{d_+}\,e^{-\lambda_2(c_+s)\,x}\right)q_0(\infty,s),\\ g_1(s) &= \frac{a}{d_+}\,e^{-\lambda_2(c_+s)\,x}\,q_0(\infty,s). \end{aligned}$$

where δ_0 denotes Dirac's delta function at 0. Since the distribution F_1 only has mass on S, we know that for fixed $x \geq 0$, g_1 must be the Laplace transform of a (generalized) function on the interval $[xc_+/d_+, \infty)$. Therefore, by multiplying $g_1(s)$ with $\exp(sxc_+/d_+)$ we obtain the Laplace transform $\tilde{h}_1(s) = \exp(sxc_+/d_+)g_1(s)$ of a function h_1 on $[0, \infty)$. After some calculations we find,

$$\begin{aligned}\tilde{h}_1(s-\theta) &= (1-\rho_c)\frac{a}{d_+}\, e^{-\frac{b}{d_+}x} e^{x\gamma(s-\sqrt{s^2-\omega})} \times \\ &\qquad \left(1 + \frac{a}{c_-(s-(\theta-\beta))} - \frac{c_+\nu}{2}\,\frac{s-\sqrt{s^2-\omega}}{s-(\theta-\beta)}\right).\end{aligned}$$

We can invert $\tilde{h}_1(s-\theta)$ straightforwardly (still for fixed $x \geq 0$) by using the following two facts. First, the function

$$y \mapsto H_0(x,y)x\omega\gamma,$$

is the inverse Laplace transform of

$$s \mapsto \exp(x\gamma(s-\sqrt{s^2-\omega})) - 1,$$

see e.g. [37, page 250, (41)]. Secondly, by differentiating H_0 with respect to x we see that

$$y \mapsto \omega H_1(x,y),$$

is the inverse Laplace transform of

$$s \mapsto (s-\sqrt{s^2-\omega})\exp(x\gamma(s-\sqrt{s^2-\omega})).$$

It follows that $h_1(y) = \delta_0(y)\,\sigma_1(x) + f_1(x, y + xc_+/d_+)$, with σ_1 and f_1 as in (4.53) and (4.55).

Similarly, for fixed $x > 0$, let $\tilde{h}_0(s) = \exp(sxc_+/d_+)g_0(s)$. We find

$$\begin{aligned}\tilde{h}_0(s-\theta) &= (1-\rho_c)\,\nu\, e^{-\frac{b}{d_+}x}\, e^{x\gamma(s-\sqrt{s^2-\omega})}\left(\frac{ab}{(d_-+d_+)(s-(\theta-\beta))}\right. \\ &+ \left.\frac{c_-c_+}{2d_-}(s-\sqrt{s^2-\omega}) - \frac{bc_-c_+\nu}{2(d_-+d_+)}\,\frac{s-\sqrt{s^2-\omega}}{s-(\theta-\beta)}\right).\end{aligned}$$

Notice that the term $\delta_0(x)\,q_0(\infty, s)$ in $g_0(s)$ does not play a role, since we assume x to be strictly positive. Inversion of $\tilde{h}_0$ finally yields $h_0(y) = f_0(x, y + xc_+/d_+)$. □

Remark 4.5 The current model can be seen as a refinement of the model in Section 2.5.2 if we interpret the driving process (X_t) in that model as the number of customers in an $M/M/1$ queueing system with parameters λ and μ. However, as far as the content of the fluid reservoir is concerned, it does not matter if we choose the regulating process to be the amount of work (D_t) in the system, since clearly $X_t = 0$ if and only if $D_t = 0$.

Now if we take in the current model $d_+ = b/\mu$, the net amount of fluid that flows into the buffer during an on-time is exponentially distributed with parameter μ. Thus, if we let $b \to \infty$ (and hence $d_+ \to \infty$) and take $a = \lambda$, $d_- = 1$, $c_- = r_-$, and $c_+ = r_+$, both models are identical, so that the joint distribution of the amount of work in the $M/M/1$

system and the content of the buffer is given by σ_0 and f_0 in Theorem 4.11 (notice that the set S is now given by $[0,\infty) \times [0,\infty)$, while the densities corresponding to $M = 1$ are identically 0). In particular, the marginal distribution of Theorem 4.8 can be shown to turn into the one in (2.84) by using the fact that $I_1(z) = (z/\pi)\int_{-1}^{1}\sqrt{1-x^2}\exp(zx)dx$ and that for $b \geq 1$, $\int_{-1}^{1}\sqrt{1-x^2}/(x-b)dx = -\pi(b-\sqrt{b^2-1})$, see [4]. As an aside we mention that we cannot easily find an expression for the joint distribution of (X_t, C_t), where X_t is the *number of customers* in the $M/M/1$-queue, which remained implicit in Section 2.5.2.

Clearly, it is not difficult to obtain numerical results from Theorem 4.11. In Figures 4.3 and 4.4 the various densities are shown for the parameter values given in Section 4.2. Notice that β, the decay rate of the second reservoir is rather small in this case, $\beta \approx 0.014$.

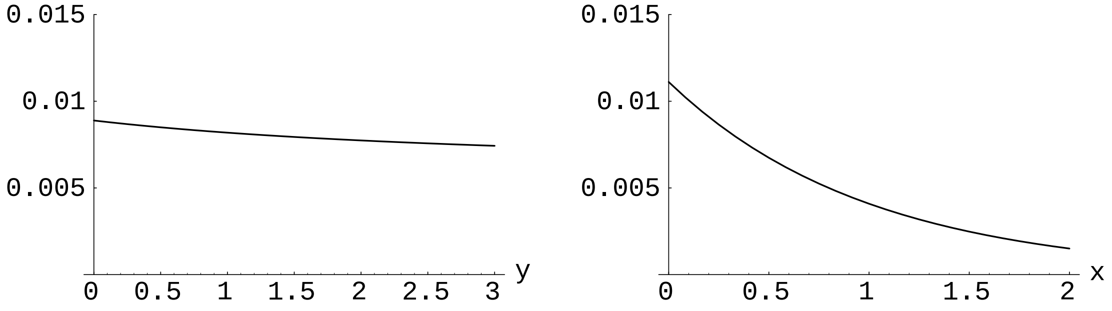


Figure 4.3: The densities σ_0 and σ_1 as functions of y and x, respectively

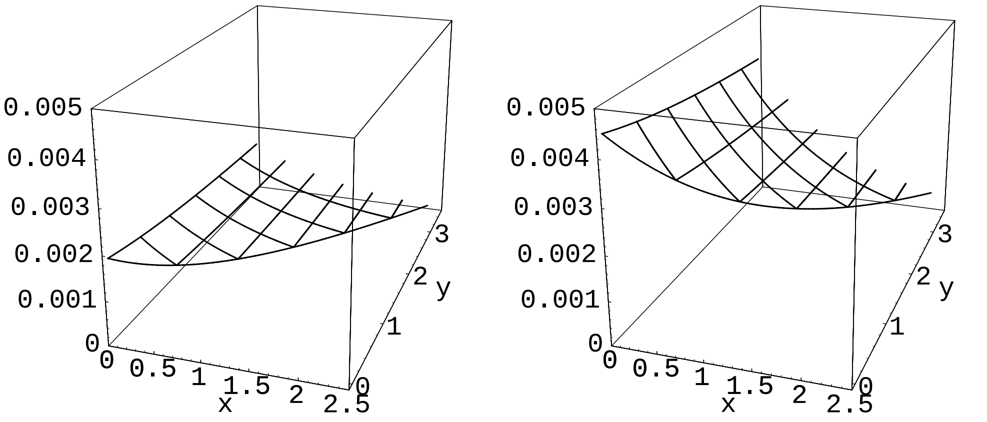


Figure 4.4: The densities f_0 and f_1 as functions of x and y

As announced in Section 4.1, the rest of this chapter deals with the dual model, which can be considered as a similar refinement to the model in Section 2.4.2.

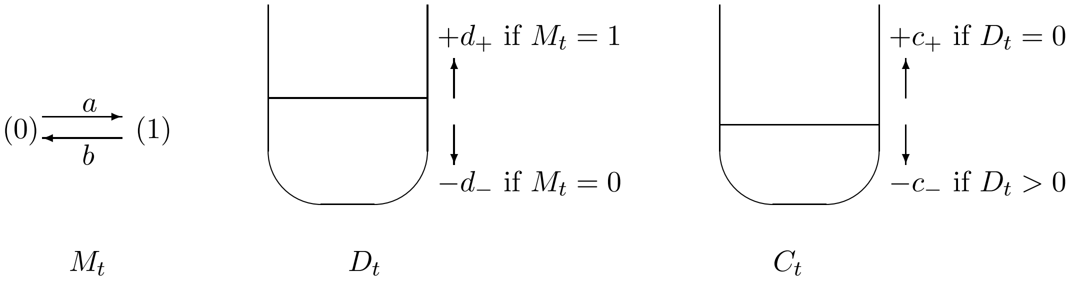


Figure 4.5: Interaction between the subsystems of the dual system

4.6 Dual model

As in the first part of this chapter we consider a fluid system consisting of two infinitely large reservoirs. The first one is regulated by a two-state continuous-time Markov process (M_t) in the same way as before; also the transition intensities of (M_t) are given by a (from 0 to 1) and b (from 1 to 0). The only difference with the tandem model is that the content of the second buffer increases at rate c_+ when the first buffer is *empty*, and decreases at rate c_- otherwise, provided that it is not empty.

As before, we let D_t and C_t denote the contents of the first and second buffer at time t, respectively. A schematic overview of the behaviour of the three subsystems is given in Figure 4.5, while a realisation of the processes (D_t) and (C_t) is given in Figure 4.6. This time we assume that $(M_0, D_0, C_0) = (0, 0, 0)$.

As for the tandem model, the stochastic process (M_t, D_t, C_t) is a Markov process. Its state space is simply given by $\{0, 1\} \times \mathbb{R}_+ \times \mathbb{R}_+$.

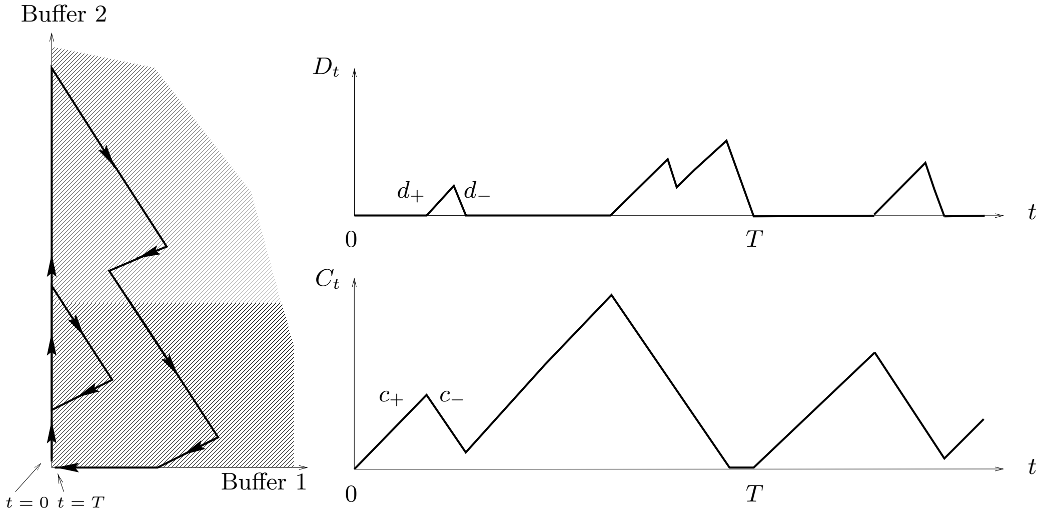


Figure 4.6: Realisation of the buffer content processes for the dual model

4.7 Dual model: stability

In analogy to Theorem 4.3, we find the conditions under which the limiting distribution of the process (M_t, D_t, C_t) exists. As regeneration epochs we choose the times (including 0) at which $(M_t, D_t, C_t) = (0, 0, 0)$.

Theorem 4.12 *The process (M_t, D_t, C_t) is regenerative with regeneration cycles that have a non-lattice distribution and finite expectation if and only if*

$$\frac{b}{d_+} - \frac{a}{d_-} > 0, \tag{4.60}$$

and

$$\frac{a}{c_+} - \frac{bd_-}{c_-d_- + c_-d_+ + c_+d_+} > 0. \tag{4.61}$$

Proof. Let $I_0, I_1, \ldots$ and $B_0, B_1, \ldots$ denote respectively the lengths of the idle periods and the busy periods of (D_t), and let I (B) be a generic idle (busy) period. Note that these periods alternate as $I_0, B_0, I_1, B_1, \ldots$, rather than as $B_0, I_0, B_1, I_1, \ldots$, which was the case in the tandem model. We consider an embedded process (Z_i) where Z_i is the content of the second buffer at the beginning of the ith *idle* period of the first buffer, $i = 0, 1, 2, \ldots$. Clearly, this leads to $Z_0 = 0$ and

$$Z_{i+1} = [Z_i + c_+I_i - c_-B_i]^+, \; i = 0, 1, 2 \ldots, \tag{4.62}$$

the Lindley equation for a $G/M/1$ queue with interarrival times distributed as c_-B and service times distributed as c_+I.

The rest of the proof can be copied from the proof of Theorem 4.3, apart from (4.13), which is replaced by

$$c_- \, EB > c_+ \, EI. \tag{4.63}$$

Note however that here (4.60) is not implied by (4.61). □

Remark 4.6 While for the tandem case, the content of the second buffer at the beginning of a busy period is the actual waiting time in an $M/G/1$-queue (with inter-arrival times distributed as c_-I and service times distributed as c_+B), we now find an embedded process that is related to the waiting time in a $G/M/1$-queue.

Corollary 4.13 *If (4.60) and (4.61) hold, a random vector (M, D, C) exists, to which the process (M_t, D_t, C_t) converges in distribution as $t \to \infty$.*

We will henceforth assume (4.60) and (4.61) to be satisfied. The interpretation of (M, D, C) and the definition of the limiting distribution $\mathbf{F}$ are the same as for the tandem model case, see (4.16).

The following lemma can be compared to Lemma 4.7.

Lemma 4.14 *The conditional distribution of* $(C \mid D=0)$ *is exponential with intensity*

$$\beta = \frac{a}{c_+} - \frac{bd_-}{c_-d_- + c_-d_+ + c_+d_+}. \tag{4.64}$$

Proof. Let (Z_i) be the Lindley process in the proof of Theorem 4.12. Under the stability conditions given in Theorem 4.12, the process (Z_i) converges in distribution to a random variable Z, say. By Theorems IX.1.2(b) and IX.1.3 of [7], we have

$$P[Z \le z] = 1 - (1 - \beta c_+/a)\, e^{-\beta z}.$$

Here β is the unique strictly positive solution of the equation $1 = E\, e^{\beta U}$, where U is distributed as $c_+I - c_-B$ and I and B are generic idle and busy periods of the first buffer respectively. It follows that β satisfies

$$1 = \frac{a}{a - \beta c_+} \frac{b}{\beta c_- + b - \lambda_1(\beta c_-) d_+},$$

which is readily solved to give (4.64). Furthermore, by standard regenerative processes theory,

$$(C \mid D=0) \overset{d}{=} Z + c_+ I^*,$$

where I^* denotes the residual lifetime of an idle period. In particular,

$$E(e^{-sC} \mid D=0) = \left(\frac{\beta c_+}{a} + (1 - \frac{\beta c_+}{a}) \frac{\beta}{\beta + s} \right) \frac{a}{a + c_+ s} = \frac{\beta}{\beta + s},$$

which had to be shown. □

Notice that $\beta > 0$ due to our assumption after Corollary 4.13. In the following section we will thankfully use Lemma 4.14.

4.8 Dual model: stationary joint distribution

In this section we find the limiting distribution $\mathbf{F}$ of the process (M_t, D_t, C_t). First, we derive a set of algebraic equations for the Laplace transform of $\mathbf{F}$, as in Lemma 4.10. Secondly, we use Lemma 4.14 to solve these equations.

Again, $\mathbf{q}(p,s)$ is a vector with components $q_0(p,s)$ and $q_1(p,s)$, given by

$$q_i(p,s) = E\mathbf{1}_{\{M=i\}}\, e^{-pD-sC}, \;\; i \in \{0,1\}, \;\; p,s \ge 0. \tag{4.65}$$

Lemma 4.15 *The vector* $\mathbf{q}(p,s)$ *satisfies:*

$$A(p,s)\,\mathbf{q}(p,s) = B(p,s) \begin{pmatrix} q_0(\infty, s) \\ q_0(p, \infty) \\ q_1(p, \infty) \end{pmatrix}, \tag{4.66}$$

with

$$A(p,s) = \begin{pmatrix} -a + d_-p + c_-s & b \\ a & -b - d_+p + c_-s \end{pmatrix},$$

and

$$B(p,s) = \begin{pmatrix} d_-p + c_+s + c_-s & c_-s & 0 \\ 0 & 0 & c_-s \end{pmatrix}.$$

Proof. Similar to the proof of Lemma 4.10; note that here $q_0(\infty,\infty) = 0$. □

From Lemma 4.15 we obtain for all $p, s \geq 0$,

$$\mathbf{q}(p,s) = \frac{H(p,s)}{\det A(p,s)} \begin{pmatrix} q_0(\infty,s) \\ q_0(p,\infty) \\ q_1(p,\infty) \end{pmatrix}, \tag{4.67}$$

where

$$H(p,s) = \begin{pmatrix} -b - d_+p + c_-s & -b \\ -a & -a + d_-p + c_-s \end{pmatrix} B(p,s).$$

Next, we use Lemma 4.14 and the first part of (4.20), by which we have

$$q_0(\infty,s) = (1-\rho_d)\frac{\beta}{s+\beta}. \tag{4.68}$$

It remains to determine $q_i(p,\infty), i \in \{0,1\}$, which we will do via an argument that is similar to the argument in Remark 4.4. Let $s_1(p)$ and $s_2(p)$ denote the two roots of the quadratic equation $\det A(p,s) = 0$, see Figure 4.7. We note that both roots are real and that for the smallest, s_1 say, we have $s_1(-\alpha) = s_1(0) = 0$, where α is given in (4.19).

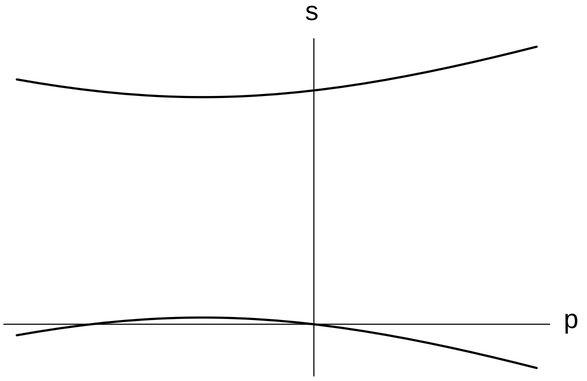


Figure 4.7: The roots s_1 and s_2 as functions of p

By writing out (4.67) we find that $q_0(p,s)$ is of the form

$$q_0(p,s) = \frac{c_3(p)s^3 + c_2(p)s^2 + c_1(p)s + c_0(p)}{(s-s_1(p))(s-s_2(p))(s+\beta)}, \tag{4.69}$$

where the c_i are unknown but analytic functions of p, at least for $p > -\alpha$ because $q_i(p,\infty) < Ee^{-pD}$ and α is the decay rate of the first reservoir. We now fix p such that $-\alpha < p < 0$. Because for $s > 0$ we have that $q_0(p,s) < Ee^{-pD}$ we can conclude that $q_0(p,s)$ must be bounded for $s > 0$. Moreover, since it is not difficult to show that $s_1(p) > 0$ and $s_2(p) > 0$ (see Figure 4.7), it follows that the numerator in (4.69) must be zero for $s = s_1(p)$ and for $s = s_2(p)$. This provides us with two linearly independent equations for $q_0(p,\infty)$ and $q_1(p,\infty)$. As an aside we note that taking $q_1(p,s)$ instead of $q_0(p,s)$ in the reasoning above leads to an equivalent set of equations. After quite a bit of algebra, the solution can be written as

$$\begin{aligned} q_0(p,\infty) &= (1-\rho_d)\frac{bc_+ + ac_- + c_+d_+p}{c_-d_- + c_+d_+}\;\frac{\zeta-\alpha}{(p+\alpha)(p+\zeta)}, & (4.70)\\ q_1(p,\infty) &= (1-\rho_d)\,\frac{a}{d_+}\;\frac{\zeta-\alpha}{(p+\alpha)(p+\zeta)}, & (4.71)\end{aligned}$$

where we have defined

$$\zeta = \alpha + \beta\frac{c_-d_- + c_+d_+}{d_-d_+} = \frac{ac_-}{c_+d_+} + \frac{bc_-}{c_-d_- + c_-d_+ + c_+d_+}. \tag{4.72}$$

The Laplace transforms q_0 and q_1 now follow from (4.67), (4.68), (4.70) and (4.71) and take, after some strenuous rewriting, the form

$$\begin{aligned} q_0(p,s) &= (1-\rho_d)\beta\;\frac{(p+\zeta)(p+b/d_+)\;+\;s(ac_- + bc_+ + c_+d_+p)/(d_+d_-)}{(p+\alpha)(p+\zeta)(s+\beta)}, & (4.73)\\ q_1(p,s) &= (1-\rho_d)\beta\;\frac{a}{d_+}\;\frac{p+\zeta+s(c_+d_+ + c_-d_-)/(d_+d_-)}{(p+\alpha)(p+\zeta)(s+\beta)}. & (4.74)\end{aligned}$$

We are now ready to state the main result for the dual model.

Theorem 4.16 *The stationary joint distribution* $\mathbf{F}$ *of the process* (M_t, D_t, C_t) *is of the form*

$$\begin{aligned} F_0(\{0\},dy) &= \sigma_0(y)\,dy, && y>0,\\ F_i(dx,\{0\}) &= \mu_i(x)\,dx, && x>0,\ i\in\{0,1\} \qquad (4.75)\\ F_i(dx,dy) &= f_i(x,y)\,dx\,dy, && x,y>0,\ i\in\{0,1\}\end{aligned}$$

where the densities σ_0, μ_i *and* f_i, $i \in \{0,1\}$ *are given by*

$$\begin{aligned}
\sigma_0(y) &= (1-\rho_d)\,\beta\, e^{-\beta y}, &(4.76)\\
\mu_0(x) &= (1-\rho_d)\left(\frac{a}{d_-}\,e^{-\alpha x} - \frac{bc_+}{c_-d_- + c_-d_+ + c_+d_+}e^{-\zeta x}\right), &(4.77)\\
\mu_1(x) &= (1-\rho_d)\,\frac{a}{d_+}\,(e^{-\alpha x} - e^{-\zeta x}), &(4.78)\\
f_0(x,y) &= (1-\rho_d)\frac{bc_+\beta}{c_-d_- + c_-d_+ + c_+d_+}\,e^{-\zeta x-\beta y}, &(4.79)\\
f_1(x,y) &= (1-\rho_d)\frac{a\beta}{d_+}\,e^{-\zeta x-\beta y}, &(4.80)
\end{aligned}$$

and the constants ρ_d, α, β *and* ζ *are given in (4.20), (4.19), (4.64) and (4.72) respectively.*

Proof. Lemma 4.14 gives (4.76), and inverse Laplace transformation of (4.70) and (4.71) yields (4.77) and (4.78). In order to obtain the densities f_i, we first rewrite $q_i(p,s)$ to a form in which we can recognize (the transforms of) the densities we just found. The result is given by

$$\begin{aligned}
q_0(p,s) &= (1-\rho_d)\,\beta\,\left\{\frac{ac_- + bc_+ + c_+d_+p}{d_+d_-(p+\alpha)(p+\zeta)} + \right.\\
&\qquad\left.\left(1+\frac{bc_+}{c_-d_- + c_-d_+ + c_+d_+}\,\frac{1}{p+\zeta}\right)\frac{1}{s+\beta}\right\}, &(4.81)\\
q_1(p,s) &= (1-\rho_d)\,\frac{a}{d_+}\left\{\frac{\zeta-\alpha}{(p+\alpha)(p+\zeta)} + \frac{\beta}{(p+\zeta)(s+\beta)}\right\}. &(4.82)
\end{aligned}$$

By inversion of these expressions, we now easily find (4.79) and (4.80). □

Corollary 4.17 *The stationary marginal distribution of the process* (C_t) *is given by*

$$P[C \le y] = 1 - \rho_c e^{-\beta y}, \qquad y \ge 0, \tag{4.83}$$

with

$$\rho_c = (1-\rho_d)\,\frac{c_- + c_+}{c_-}.$$

Therefore, as in the tandem case, the parameter β *as given in (4.64) may be interpreted as the* decay rate *for the second buffer, while the* utilization *of this buffer is given by* ρ_c. *The expected stationary content of the second buffer is given by*

$$EC = \frac{\rho_c}{\beta}.$$

Remark 4.7 As the tandem model offers a refinement to the model in Section 2.5.2, the dual model can be seen as an extension of the model in Section 2.4.2. The correspondance is the same as in Remark 4.5. In particular in the resulting model we have, as can be expected, that,

$$P[D = 0, C \leq y] = F_0(y),$$

and for $x > 0$,

$$P[D \leq x, C \leq y] = \sum_{i=1}^{\infty} F_i(y)\, P[E_i(\mu) \leq x],$$

where the distribution function F_i is given in (2.31) and $E_i(\mu)$ denotes an Erlang-distributed random variable with parameters i and μ.

The remarkable difference in complexity of the solutions for the tandem and the dual model can be explained as in Remark 2.3, by looking at the spectral expansion for the solution of each model. Therefore we conclude this chapter with the following section.

4.9 Dual model: spectral analysis

4.9.1 Introduction

In this section we show how the dual model can be solved using the spectral approach. First we will consider an approximative model, in which the second reservoir is being regulated by a Markov process with a countably infinite state space. Although this process has no birth-death structure, we will use similar techniques as in Section 2.4, to indicate how this model can be solved, without working out the details of the analysis. The second step is done in Section 4.9.3 where the distribution of (M_t, D_t, C_t) is obtained. We will end with some concluding remarks.

4.9.2 An approximative model

This section is devoted to an approximation of the model in Section 4.6, for which the state space of the modulating Markov process is denumerable. We prove a lemma about a system of difference equations and indicate how this can be used to solve the (approximative) model. More important however is that in Section 4.9.3 we can show that, by taking appropriate limits, the system of difference equations leads to a system of differential equations, which is used to find the solution to the original model.

Specifically, we approximate the process (M_t, D_t) by another Markov process $(M_t, \tilde{D}_t)$, in which $\tilde{D}_t$ now represents the number of *fluid quanta* in the first reservoir. Instead of a continuous inflow of fluid during on-periods, we suppose that fluid quanta of size $1/n$ arrive according to a Poisson process with rate nd_+. In order to avoid the possibility of an empty first reservoir during on-periods, we will suppose that a quantum is added as soon as an

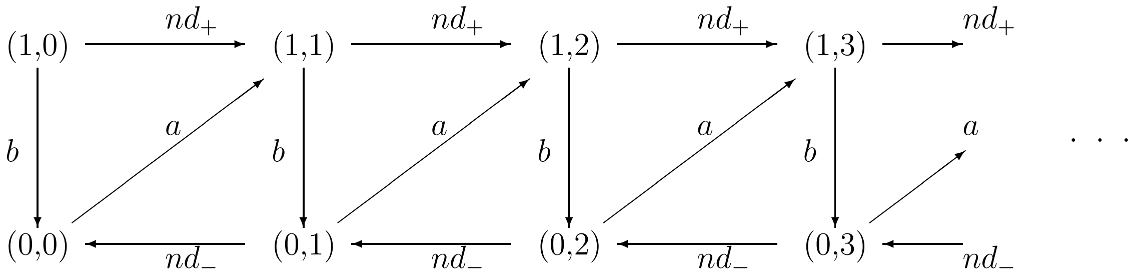


Figure 4.8: Transition diagram of the Markov process $(M_t, \tilde{D}_t)$

on-period starts. Also, we suppose that quanta are removed from the first reservoir during off periods according to a Poisson process with rate nd_-. Thus, the transition diagram of $(M_t, \tilde{D}_t)$ is given in Figure 4.8, and the corresponding (infinite-dimensional) generator has the following block-tridiagonal structure.

$$T \equiv \begin{pmatrix} H_0 & D_+ & & & \\ D_- & H & D_+ & & \\ & D_- & H & D_+ & \\ & & \cdots & \cdots & \cdots \end{pmatrix}, \tag{4.84}$$

where

$$H = \begin{pmatrix} -a - nd_- & 0 \\ b & -b - nd_+ \end{pmatrix}, \qquad H_0 = \begin{pmatrix} -a & 0 \\ b & -b - nd_+ \end{pmatrix},$$

$$D_- = \begin{pmatrix} nd_- & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad D_+ = \begin{pmatrix} 0 & a \\ 0 & nd_+ \end{pmatrix}$$

We assume that the Markov process $(M_t, \tilde{D}_t, C_t)$ is stationary, and define

$$\mathbb{G}(y) \equiv (\mathbf{G}_0(y), \mathbf{G}_1(y), \mathbf{G}_2(y), \ldots)^T, \qquad y \geq 0, \tag{4.85}$$

where

$$\mathbf{G}_j(y) = (G_{0,j}(y), G_{1,j}(y)), \quad j \in \mathbb{N},\ y \geq 0,$$

with

$$G_{i,j}(y) \equiv P[M_t = i, \tilde{D}_t = j, C_t \leq y], \quad i \in \{0,1\},\ j \in \mathbb{N},\ y \geq 0.$$

Notation. Throughout this section blackboard bold characters (such as $\mathbb{G}$, $\mathbb{P}$, etc.) will indicate vectors that can typically be partitioned into two-dimensional vectors (corresponding to the two possible values for M_t), which will be written in bold-faced type.

Note that the Markov process $(M_t, \tilde{D}_t)$ has been constructed such that state (1,0) is transient. We could therefore eliminate it from the state space as we are interested in stationary behaviour only. However, to maintain our notational convention we prefer not to do so, but rather let $G_{1,0}(y) = 0$ for all $y \geq 0$.

It can be shown by writing down the Kolmogorov forward equations for the system and then assuming equilibrium, that $\mathbb{G}$ must be a solution of the differential equation

$$\frac{d}{dy}\mathbb{G}(y) = R^{-1}T^T\mathbb{G}(y). \tag{4.86}$$

Here, T is the generator in (4.84), while the diagonal matrix R, containing the net input and output rates of fluid in the various states of the modulating proces $(M_t, \tilde{D}_t)$, is given by

$$R \equiv \operatorname{diag}(c_+, 1, -c_-, -c_-, \ldots). \tag{4.87}$$

Note that the choice of the second diagonal element is not relevant. Since the second reservoir cannot be empty while the first reservoir is empty, the solution must satisfy the boundary condition

$$G_{0,0}(0) = 0. \tag{4.88}$$

Moreover, we must impose

$$\lim_{y\to\infty} \sum_{j=0}^{\infty} (G_{0,j}(y) + G_{1,j}(y)) = 1. \tag{4.89}$$

To analyse (4.86), we apply a similar procedure as described in Section 2.4. First, we truncate the state space of the process $(M_t, \tilde{D}_t)$ to $\{0,1\} \times \{0,1,2,\ldots,m\}$, for some sufficiently large m. From this truncated state space we eliminate state $(1,0)$, so that the generator of the new process is the $(2m+1) \times (2m+1)$-matrix

$$T_m \equiv \begin{pmatrix} -a & \mathbf{d}_+ & & & \\ \mathbf{d}_- & H & D_+ & & \\ & \cdots & \cdots & \cdots & \\ & & D_- & H & D_+ \\ & & & D_- & H_1 \end{pmatrix}, \tag{4.90}$$

where H, D_+ and D_- are as before, and

$$\begin{aligned} H_1 &= \begin{pmatrix} -nd_- & 0 \\ b & -b \end{pmatrix}, \\ \mathbf{d}_+ &= (0, a), \\ \mathbf{d}_- &= (nd_-, 0)^T. \end{aligned}$$

When we define

$$R_m \equiv \operatorname{diag}(c_+, \overbrace{-c_-, -c_-, \ldots, -c_-}^{2m}), \tag{4.91}$$

it follows that the stationary distribution $\mathbb{G}_m$ of the truncated process is the solution of the differential system

$$\frac{d}{dy}\mathbb{G}_m(y) = R_m^{-1}T_m^T\mathbb{G}_m(y), \tag{4.92}$$

and is hence formally given by

$$\mathbb{G}_m(y) = e^{-R_m^{-1}T_m^T y}\mathbb{G}_m(0).$$

By [92], we know that the eigenvalues $\xi_k,\ k = 0, 1, \ldots, 2m$, of the matrix $R_m^{-1}T_m^T$, when ordered properly, satisfy $\xi_0 < 0 = \xi_1 < \mathrm{Re}(\xi_2) \leq \mathrm{Re}(\xi_3) \leq \cdots \leq \mathrm{Re}(\xi_{2m})$, where ξ_0 and ξ_1 are real. Since we are looking for a bounded solution, $\mathbb{G}_m$ must be of the form

$$\mathbb{G}_m(y) = c_0\mathbb{v}_0 e^{\xi_0 y} + c_1\mathbb{v}_1, \quad y \geq 0,$$

where $\mathbb{v}_k$ is a suitably normalized eigenvector corresponding to ξ_k and c_k is a constant, $k = 0, 1$. The values for c_0 and c_1 follow from boundary conditions, similar to the ones in (4.88) and (4.89). Therefore we may conclude that, *for any* m, the solution of (4.92) is of the form

$$\mathbb{G}_m(y) = \mathbb{p}_m - \mathbb{v}_m e^{-\beta_m y}, \quad y \geq 0.$$

Here, $\mathbb{p}_m$ denotes the stationary distribution of the truncated modulating Markov process, $-\beta_m$ is the unique negative eigenvalue of $R_m^{-1}T_m^T$, and $\mathbb{v}_m$ is the corresponding, suitably normalized eigenvector. Finally, letting $m \to \infty$, we get the same form for the original process,

$$\mathbb{G}(y) = \mathbb{p} - \mathbb{v}e^{-\beta y}, \quad y \geq 0, \tag{4.93}$$

where

$$\begin{array}{rcllrcll}
\beta &=& \lim_{m\to\infty}\beta_m, &&&&& \\
\mathbb{p} &=& (\mathbf{p}_0, \mathbf{p}_1, \ldots)^T, & \text{with} & \mathbf{p}_j &=& (p_{0,j}, p_{1,j}), & j = 0, 1, \ldots, \\
\mathbb{v} &=& (\mathbf{v}_0, \mathbf{v}_1, \ldots)^T, & \text{with} & \mathbf{v}_j &=& (v_{0,j}, v_{1,j}), & j = 0, 1, \ldots,
\end{array}$$

respectively. After determining β, $\mathbb{p}$ and $\mathbb{v}$, it can be shown by substitution that (4.93) indeed is a solution of (4.86). Note that the components of $\mathbb{p}$ and $\mathbb{v}$ have a probabilistic interpretation, namely

$$\begin{array}{rcll}
p_{i,j} &=& P[M_t = i, \tilde{D}_t = j], & i \in \{0, 1\},\ j = 0, 1, \ldots, \\
v_{i,j} &=& P[M_t = i, \tilde{D}_t = j, C_t > 0], & i \in \{0, 1\},\ j = 0, 1, \ldots,
\end{array}$$

so that we immediately have that $p_{1,0} = v_{1,0} = 0$.

To learn more about the relation between β and $\mathbb{v}$, we investigate the eigenvalue problem

$$T^T\mathbb{v} = -\beta R\mathbb{v}. \tag{4.94}$$

Writing out (4.94), where we leave out the equation concerning state $(1,0)$ and set $v_{1,0}=0$, we obtain

$$\begin{aligned}-av_{0,0}+nd_-v_{0,1} &= -\beta c_+v_{0,0}\\ av_{0,0}-(b+nd_+)v_{1,1} &= \beta c_-v_{1,1},\end{aligned}$$

and, for $j=1,2,\ldots,$

$$\begin{aligned}-(a+nd_-)v_{0,j}+bv_{1,j}+nd_-v_{0,j+1} &= \beta c_-v_{0,j}\\ av_{0,j}+nd_+v_{1,j}-(b+nd_+)v_{1,j+1} &= \beta c_-v_{1,j+1}.\end{aligned}$$

The following lemma is now immediate.

Lemma 4.18 *The components of the eigenvector* $\mathbb{V}$ *corresponding to eigenvalue* $-\beta$ *satisfy the difference equation*

$$\mathbf{v}_{j+1}=A\mathbf{v}_j \quad j=1,2,\ldots, \tag{4.95}$$

where

$$A=\begin{pmatrix}\dfrac{a+nd_-+\beta c_-}{nd_-} & -\dfrac{b}{nd_-}\\ \dfrac{a}{b+nd_++\beta c_-} & \dfrac{nd_+}{b+nd_++\beta c_-}\end{pmatrix},$$

with initial condition

$$\mathbf{v}_1=\begin{pmatrix}\dfrac{-\beta c_++a}{nd_-}\\ \dfrac{a}{b+nd_++\beta c_-}\end{pmatrix}v_{0,0}\ . \tag{4.96}$$

Since det $A>0$ and det $(A-I_2)<0$ for $\beta>0$, it turns out that for the eigenvalues $\zeta_1(\beta)$ and $\zeta_2(\beta)$ of A we have $0<\zeta_1(\beta)<1<\zeta_2(\beta)$. Therefore, in order for a bounded solution $\mathbb{V}$ to exist, $\mathbf{v}_1$ must be in the eigenspace of $\zeta_1(\beta)$, which gives an equation from which we can determine β. Also, since $\mathbb{P}$ is the eigenvector corresponding to eigenvalue 0, we can use Lemma 4.18 to find an expression for its components. Finally, applying conditions (4.88) and (4.89), it is possible to solve the model exactly. We will however proceed by showing how the results of this section can be used to obtain the solution of our original problem.

4.9.3 Taking the limit

In this section we will show how the stationary distribution $\mathbf{F}$ of the process (M_t,D_t,C_t) can be found, using the results for the approximative system of the previous section. We will first derive an analogue to Lemma 4.18 for our original model. Then, we proceed by carrying out an analysis similar to the one suggested after Lemma 4.18, which will lead to the same result as in Theorem 4.16.

We first consider a sequence of auxiliary systems, as described in the previous section. However, now we are interested in the *amount of fluid* in the first reservoir, rather than the *number of fluid quanta* (apart from the state of the on-off source and the content of the second reservoir). Concretely, let $M_t^{(n)}$, $D_t^{(n)}$ and $C_t^{(n)}$ denote the state of the on-off source, the first reservoir and the second reservoir at time t, respectively, for any $n \in \mathbb{N}\backslash\{0\}$, where $1/n$ is the size of a fluid quantum as before. In particular, $D_t^{(n)} = \tilde{D}_t/n$. Assuming stationarity of the Markov process $(M_t^{(n)}, D_t^{(n)}, C_t^{(n)})$, we define for any n,

$$F_i^{(n)}(x,y) \equiv P[M_t^{(n)} = i,\ D_t^{(n)} \leq x,\ C_t^{(n)} \leq y], \qquad x, y \geq 0,\ i \in \{0,1\},$$

and in vector notation,

$$\mathbf{F}^{(n)}(x,y) \equiv (F_0^{(n)}(x,y), F_1^{(n)}(x,y))^T, \qquad x, y \geq 0.$$

$\mathbf{F}^{(n)}(x,y)$ can be expressed easily in terms of the quantities in Section 4.9.2, namely

$$\mathbf{F}^{(n)}(x,y) = \sum_{j=0}^{\lfloor xn \rfloor} \mathbf{G}_j^{(n)}(y) = \sum_{j=0}^{\lfloor xn \rfloor} \mathbf{p}_j^{(n)} - \mathbf{v}_j^{(n)} e^{-\beta^{(n)} y} \qquad x, y \geq 0,$$

where we have indicated dependence of these quantities on n, and where $\lfloor x \rfloor$ is the largest integer $\leq x$.

Clearly, the stochastic process $(M_t^{(n)}, D_t^{(n)}, C_t^{(n)})$ is a good approximation for the original process (M_t, D_t, C_t) if n is large. It seems therefore plausible that the stationary distribution $\mathbf{F}$ of the latter satisfies

$$\mathbf{F}(x,y) = \lim_{n\to\infty} \mathbf{F}^{(n)}(x,y), \qquad x, y \geq 0.$$

We draw the conclusion that

$$\mathbf{F}(x,y) = \mathbf{P}(x) - \mathbf{V}(x) e^{-\beta y}, \qquad x, y \geq 0, \tag{4.97}$$

assuming that the following limits exist,

$$\begin{aligned} \mathbf{P}(x) &= \lim_{n\to\infty} \sum_{j=0}^{\lfloor xn \rfloor} \mathbf{p}_j^{(n)}, \qquad x \geq 0, & (4.98) \\ \mathbf{V}(x) &= \lim_{n\to\infty} \sum_{j=0}^{\lfloor xn \rfloor} \mathbf{v}_j^{(n)}, \qquad x \geq 0, & (4.99) \\ \beta &= \lim_{n\to\infty} \beta^{(n)}. & (4.100) \end{aligned}$$

We note that for $\mathbf{v}(x) \equiv \mathbf{V}'(x)$ we have

$$\mathbf{v}(x) = \lim_{n\to\infty} n \mathbf{v}_{\lfloor xn \rfloor + 1}^{(n)}, \qquad x \geq 0, \tag{4.101}$$

assuming the right hand side exists, since for $x \geq 0$,

$$\int_0^x \lim_{n\to\infty} n\mathbf{v}^{(n)}_{\lfloor tn\rfloor+1} dt = \lim_{n\to\infty} \int_0^x n\mathbf{v}^{(n)}_{\lfloor tn\rfloor+1} dt$$
$$= \lim_{n\to\infty} \left(\sum_{j=1}^{\lfloor xn\rfloor} \mathbf{v}^{(n)}_j + \mathbf{v}^{(n)}_{\lfloor xn\rfloor+1}(xn - \lfloor xn\rfloor) \right) = \mathbf{V}(x) - \mathbf{V}(0).$$

Similarly, we have

$$\mathbf{v}'(x) = \lim_{n\to\infty} n^2(\mathbf{v}^{(n)}_{\lfloor xn\rfloor+2} - \mathbf{v}^{(n)}_{\lfloor xn\rfloor+1}), \qquad x \geq 0. \tag{4.102}$$

We are now ready to prove the following lemma.

Lemma 4.19 *If it exists, the vector function* $\mathbf{v}(x)$ *satisfies the differential equation*

$$\mathbf{v}'(x) = B\mathbf{v}(x) \qquad x \geq 0, \tag{4.103}$$

where

$$B = \begin{pmatrix} \dfrac{a+\beta c_-}{d_-} & -\dfrac{b}{d_-} \\ \dfrac{a}{d_+} & -\dfrac{b+\beta c_-}{d_+} \end{pmatrix},$$

with initial condition

$$\mathbf{v}(0) = \begin{pmatrix} \dfrac{-\beta c_+ + a}{d_-} \\ \dfrac{a}{d_+} \end{pmatrix} V_0(0). \tag{4.104}$$

Proof. By (4.95), we have

$$\mathbf{v}^{(n)}_{\lfloor xn\rfloor+2} - \mathbf{v}^{(n)}_{\lfloor xn\rfloor+1} = (A^{(n)} - I)\mathbf{v}^{(n)}_{\lfloor xn\rfloor+1}\ , \qquad x \geq 0,$$

where we have indicated the dependence on n of matrix A in Lemma 4.18. Multiplying both sides of this equation by n^2 and taking the limit for $n \to \infty$ while applying (4.101) and (4.102), yields (4.103), where $B = \lim_{n\to\infty} n(A^{(n)} - I)$. Similarly, (4.104) follows from multiplying (4.96) by n and taking the limit for $n \to \infty$. □

Since for $-\beta < 0$ also $\det B < 0$, we have for the eigenvalues $\zeta_1(\beta)$ and $\zeta_2(\beta)$ of B that $\zeta_1(\beta) < 0 < \zeta_2(\beta)$. For $\mathbf{v}(x)$ to be bounded for $x \geq 0$, we must therefore have that $\mathbf{v}(0)$ is an eigenvector of $\zeta_1(\beta)$, or equivalently, $\mathbf{v}(0)$ must be orthogonal to the left eigenvector of $\zeta_2(\beta)$. After some calculus it follows that this is the case if and only if

$$\beta = \frac{a}{c_+} - \frac{bd_-}{c_-d_- + c_-d_+ + c_+d_+}. \tag{4.105}$$

Using (4.105), we now solve (4.103) – (4.104) and obtain

$$\mathbf{v}(x) = e^{-\zeta x}\mathbf{v}(0), \qquad x \geq 0,$$

where

$$\zeta = -\zeta_1(\beta) = \frac{ac_-}{c_+d_+} + \frac{bc_-}{c_-d_- + c_-d_+ + c_+d_+} \tag{4.106}$$

and

$$\mathbf{v}(0) = \begin{pmatrix} \dfrac{bc_+}{c_-d_- + c_-d_+ + c_+d_+} \\ \dfrac{a}{d_+} \end{pmatrix} V_0(0).$$

Integration yields for $x \geq 0$,

$$\begin{aligned} \mathbf{V}(x) &= \mathbf{V}(0) - \int_0^x \mathbf{v}(t)dt \\ &= \begin{pmatrix} V_0(0) \\ 0 \end{pmatrix} + \frac{V_0(0)}{\zeta} \begin{pmatrix} \dfrac{bc_+}{c_-d_- + c_-d_+ + c_+d_+} \\ \dfrac{a}{d_+} \end{pmatrix} \left(1 - e^{-\zeta x}\right). \end{aligned} \tag{4.107}$$

We could follow a similar procedure to obtain $\mathbf{P}(x)$ from $\mathbf{p}_j^{(n)}$ based on (4.98). However, since $\mathbf{P}(x) = \lim_{y\to\infty} \mathbf{F}(x, y)$, it is the solution to the standard fluid flow problem involving just the on-off source and first reservoir, which has already been given in Proposition 4.5. Finally, since the second reservoir cannot be empty while the first reservoir is empty, and hence $F_0(0, 0) = 0$, we have that

$$V_0(0) = P_0(0) = \frac{d_+\alpha}{a + b}. \tag{4.108}$$

Combining (4.97) with Proposition 4.5, (4.107) and (4.108), we can conclude that the stationary joint distribution of the Markov process (M_t, D_t, C_t) is given by

$$\mathbf{F}(x, y) = \mathbf{P}(x) - \mathbf{V}(x)e^{-\beta y}, \qquad x, y \geq 0,$$

with

$$\mathbf{P}(x) = \begin{pmatrix} \dfrac{b}{a+b} \\ \dfrac{a}{a+b} \end{pmatrix} - \begin{pmatrix} \dfrac{a}{a+b}\dfrac{d_+}{d_-} \\ \dfrac{a}{a+b} \end{pmatrix} e^{-\alpha x}, \qquad x \geq 0,$$

and

$$\mathbf{V}(x) = (1 - \rho_d)\begin{pmatrix} 1 \\ 0 \end{pmatrix} + \frac{1 - \rho_d}{\zeta} \begin{pmatrix} \dfrac{bc_+}{c_-d_- + c_-d_+ + c_+d_+} \\ \dfrac{a}{d_+} \end{pmatrix} \left(1 - e^{-\zeta x}\right), \qquad x \geq 0,$$

where the constants ρ_d, α, β and ζ are given in (4.20), (4.19), (4.105) and (4.106) respectively. It is not difficult to check that this is equivalent to the statement in Theorem 4.16.

Remark 4.8 In fact the discretization performed in this section is not essential. Instead of manipulating matrices for the approximative models, we could immediately find Lemma 4.19 by manipulating differential operators.

Remark 4.9 Now that we solved the dual model using the spectral analysis, we can explain the difference in complexity of the solutions for the tandem and the dual model. The reason that the dual model is amenable to spectral analysis is that there is only one state in the regulating process (M_t, D_t) for which the content of the second reservoir increases, namely (0,0). As a consequence, only one negative eigenvalue plays a role in the solution, namely $-\beta$.

When we try to solve the tandem model via this approach, we obtain an infinite, uncountable number of states in which the second buffer fills up, and a continuum of negative eigenvalues that play a role in the solution. This makes the analysis much harder, if not impossible, and explains the complexity of the solution in Theorem 4.11.

Chapter 5

A two-buffer fluid model with feedback

5.1 Introduction

In this final chapter we once more consider a system of two fluid reservoirs driven by a two-state (on and off) continuous-time Markov process. In fact it resembles the dual system of the previous chapter, since the net input rate into the first reservoir is positive (negative) when the on-off process is in the on-state (off-state), while the second reservoir accumulates fluid when the first reservoir is empty, and releases fluid otherwise, at constant rates. Thus, we may again view the second reservoir as being driven by the joint process of the content of the first reservoir and the on-off Markov process.

The main difference is that, unlike in the previous chapter, we incorporate the notion of feedback into our model, as described in Section 1.5.1. This feedback mechanism entails that the rates at which the first reservoir fills or depletes depend on the state (empty or nonempty) of the second reservoir.

The main motivation for studying this model is that it provides a more detailed description of a two-level traffic shaper than the model in Chapter 3. By this we mean the following (see also Remarks 4.5 and 4.7 for the relation between birth-death fluid models and two-buffer fluid models). In Section 3.6 we assumed that the cell stream that arrives at the traffic shaper is generated by an on-off source with exponentially distributed on-times and off-times, for which the on-times are short and the arrival rate during on-times is high. Ignoring the duration of the on-times and the discrete nature of the cells, we then looked upon the stream as a Poisson arrival process of exponentially distributed bursts of data cells. In the current chapter we do not ignore the duration of the on-times. The number of data cells in the buffer will not be described by the number of bursts in a queueing system as in Chapter 3, but by the amount of fluid in the first reservoir. This is why we will henceforth use the term *data buffer* for this reservoir, while the second reservoir will be called *credit buffer* as in Chapter 3.

In this context it is important that we let the credit buffer be finite. In most of this chapter we assume that this is the case and are able to find analytical results for this situation. Since we shall also consider the situation in which the credit buffer is infinitely large, we will find that the finiteness of the credit buffer complicates the form of the stationary distribution considerably.

Although the application mentioned above is the main motivation for our present study, the model is formulated in general terms to allow for other interpretations. Most notably, the model also captures a two-node fluid tandem queue with a finite second reservoir, driven by an on-off source. Another conceivable situation that can be described is that of such a tandem queue, where the flow rate from the first to the second reservoir is diminished when the latter is completely filled, to prevent or reduce loss of data.

We note that in [5] the same model as the one at hand is analysed in the context of two-level traffic shaping via an approximative discretization approach, along the lines of the discretization technique in Chapter 3.

The structure of this chapter is as follows. First we describe the model in Section 5.2. We also show here how the parameters should be chosen such that the two above-mentioned special cases of practical interest are obtained, viz. the two-level traffic shaper and the tandem queue with finite second buffer. Returning to the general model, we provide the condition under which the system is stable in Section 5.3.

In Section 5.4 we concentrate on the marginal distribution of the second reservoir. We employ two techniques, both of which make use of the distribution of an embedded Markov chain, which is derived in Section 5.4.2. The first of these techniques, described in Section 5.4.3 gives results for the two special cases of interest. The second technique, similar to the Laplace-transform techniques in Chapter 4, is described in Section 5.4.4 and yields results which are useful for any parameter values that satisfy the stability conditions.

The main result of this chapter is presented in Section 5.5, namely an explicit expression for the stationary joint distribution of the stochastic process that describes the state of the on-off source and the contents of both reservoirs. The proof of the theorem is given in three subsections, the first of which exploits a useful relationship between the current model and the tandem model of the previous chapter. This explains why the solution is fairly complicated, involving integrals of modified Bessel functions; however, as in Chapter 4, these can easily be evaluated numerically.

Section 5.6 shows how the stationary distribution simplifies for the special case in which both reservoirs are infinitely large. It is also shown that an extra stability condition must be satisfied in this situation.

Finally, in Section 5.7 we find the stationary distribution for the case in which both reservoirs are finite, provided that the first reservoir is not too small. When we apply these results to the earlier mentioned tandem situation, this actually gives us the distribution for a tandem fluid queue with two finite reservoirs, again when the first reservoir is not too small.

5.2 Model

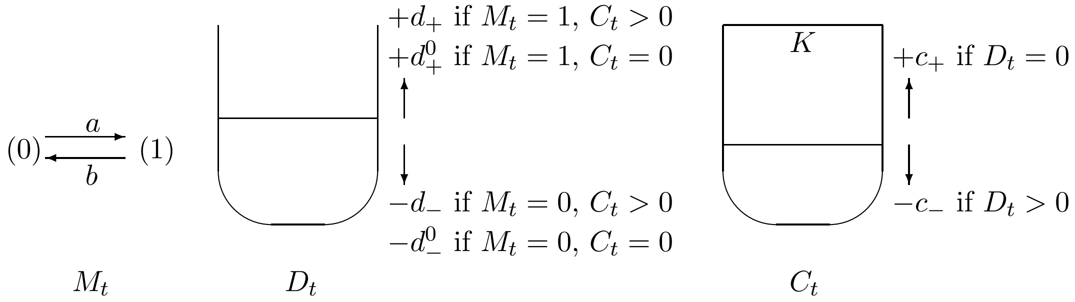
5.2.1 General model

We consider a fluid system consisting of two reservoirs, which we shall henceforth call buffers: an infinitely large *data buffer* and a finite *credit buffer* of size K. The amount of fluid that flows into and out of these buffers depends on the contents of both buffers and ultimately on a continuous-time Markov process (M_t), with state space $\{0, 1\}$ and transition intensities a (from 0 to 1) and b (from 1 to 0).

When the credit buffer is *not empty*, the content of the data buffer increases at rate d_+ when (M_t) is in state 1 and decreases at rate d_- when (M_t) is in state 0, provided that the data buffer is not empty. However, when the credit buffer *is* empty, the up and down rates are d_+^0 and d_-^0, instead of d_+ and d_- respectively.

Furthermore, the content of the credit buffer increases at rate c_+ when the data buffer is empty (provided that the credit buffer is not completely filled), and decreases at rate c_- otherwise (provided that the credit buffer is not empty) . Notice that $d_+, d_-, d_+^0, d_-^0, c_+$ and c_- are positive numbers, as in the previous chapter and that the meaning of the symbols is reflected in the notation (d for data, c for credit).

We let D_t and C_t denote the content of the data and credit buffer at time t, respectively, and observe that the stochastic process (M_t, D_t, C_t) is a Markov process. A schematic overview of the interaction between (M_t), (D_t) and (C_t) is given in Figure 5.1.



(0)
a
b
(1)
$+d_+$ if $M_t = 1,\ C_t > 0$
$+d_+^0$ if $M_t = 1,\ C_t = 0$
$-d_-$ if $M_t = 0,\ C_t > 0$
$-d_-^0$ if $M_t = 0,\ C_t = 0$
K
$+c_+$ if $D_t = 0$
$-c_-$ if $D_t > 0$
M_t
D_t
C_t

Figure 5.1: Interaction between the processes (M_t), (D_t) and (C_t)

A realization of the process (D_t, C_t) is given in Figure 5.2. The parameter values used here and in other figures in this chapter are $a = 1$, $b = 2$, $d_+ = 2$, $d_- = 6$, $d_+^0 = 4$, $d_-^0 = 3$, $c_+ = 2.5$, $c_- = 3$ and $K = 3$.

At first sight an obvious choice for the state space of (M_t, D_t, C_t) would be $\{0, 1\} \times [0, \infty) \times [0, K]$. However, a close inspection of the behaviour of the system shows that for any t we must have $D_t \le d_+(K - C_t)/c_-$, unless $C_t = 0$, see Figure 5.2. Therefore, we denote the state space of (M_t, D_t, C_t) by $\{0, 1\} \times S$ with

$$\begin{aligned} S &= S_1 \cup S_2, && (5.1)\\ S_1 &= \{(x, y) \mid 0 < y \le K,\ 0 \le x \le (K - y)d_+/c_-\}, && (5.2)\\ S_2 &= \{(x, y) \mid y = 0,\ x \ge 0\}. && (5.3) \end{aligned}$$

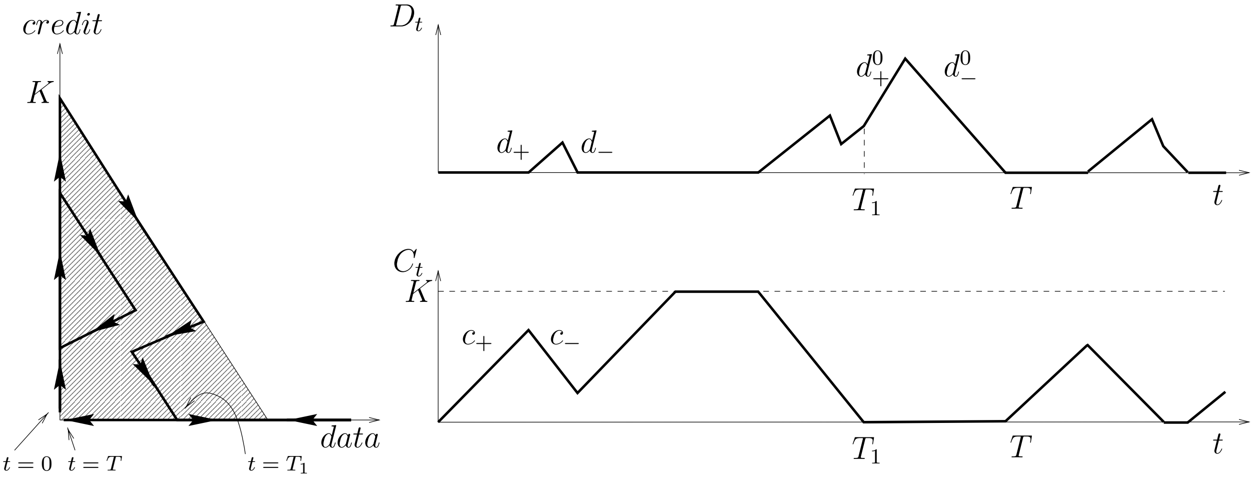


Figure 5.2: Realisation of the buffer content processes

In the remainder of this section we will discuss two possible applications of the model. In both cases the data buffer and credit buffer are closely related. The first application describes a two-level traffic shaper as in Section 3.6, controlling the traffic flow coming from an on-off source. The second model describes a fluid tandem queue as in the first part of Chapter 4, but with a finite second reservoir.

5.2.2 Modelling a two-level traffic shaper

Instead of six parameters $d_+, d_-, d^0_+, d^0_-, c_+, c_-$ for the behaviour of both buffers, we take three parameters v_0, v_1 and v_2 such that $v_0 > v_1 > v_2 > 0$ and choose

$$\begin{aligned} d_+ &= v_0 - v_1, & d_- &= v_1, \\ d^0_+ &= v_0 - v_2, & d^0_- &= v_2, \\ c_+ &= v_2, & c_- &= v_1 - v_2. \end{aligned} \tag{5.4}$$

The interpretation is the following. The data buffer only receives data when the on-off source is in the on-state, at rate v_0. The output rate is v_1 if credit is available and $v_2\,(< v_1)$ otherwise. We can think of v_2 as the long term average rate at which the data buffer is allowed to send. The rate v_1 is a higher rate that may be used for a limited period of time, namely as long as credit is available. The particular values of c_+ and c_- can be explained by arguing that whenever the data buffer is not sending (i.e., when it is empty), the “unused capacity” v_2 is saved up for later use in the form of credit, while this credit is consumed when the data buffer is sending at high rate; the “extra capacity” $v_1 - v_2$ that is used by the data buffer is taken from the credit buffer. Note that the above is equivalent to saying that the credit buffer is constantly filled at rate v_2, while it it is drained at the same rate as the data buffer (0, v_1 or v_2) at any time.

5.2.3 Modelling a tandem queue with finite second buffer

The second way in which the general model may be applied is given by the following choice of parameters. Again we have three parameters for the flow rates, v_0, v_1 and v_2, such that $v_0 > v_1 > v_2 > 0$, but now we take

$$\begin{array}{rclrcl} d_+ = d_+^0 & = & v_0 - v_1, & \qquad d_- = d_-^0 & = & v_1, \\ c_+ & = & v_2, & c_- & = & v_1 - v_2. \end{array} \tag{5.5}$$

Notice that the feedback has disappeared now, since $d_+ = d_+^0$ and $d_- = d_-^0$. Furthermore we define the process $(\bar{C}_t)$ by $\bar{C}_t \equiv K - C_t$. We can interpret $\bar{C}_t$ as the content of a buffer which receives fluid from the data buffer at rate v_1 whenever $D_t > 0$ and $\bar{C}_t < K$, while it releases fluid at rate v_2 when $\bar{C}_t > 0$. Hence the process $(M_t,\ D_t,\ \bar{C}_t)$ describes a fluid tandem queue with finite second buffer.

In Section 5.7 we extend the (general) model to the case where the data buffer is also finite, although it must in some sense be larger than the credit buffer. This provide us with results for the tandem model with two finite reservoirs, see Remark 5.2.

5.3 Stability

In this section we will derive the conditions under which the stochastic process (M_t, D_t, C_t) has a limiting distribution. It is clear that (M_t, D_t, C_t) is a regenerative process; as regeneration epochs we choose the times t when simultaneously $M_t = 0$, $D_t = 0$ and $C_t = 0$. We assume that $t = 0$ is a regeneration epoch and denote the next one by T, i.e.

$$T = \min\{t > 0 \mid M_t = 0,\ D_t = 0,\ C_t = 0\}. \tag{5.6}$$

We also define

$$T_1 = \min\{t > 0 \mid C_t = 0\}. \tag{5.7}$$

(See Figure 5.2 for a visualisation.) The following theorem provides the condition under which ET is finite.

Theorem 5.1 *The process (M_t, D_t, C_t) is regenerative with regeneration cycles that have a non-lattice distribution and finite expectation if and only if*

$$\alpha = \frac{b}{d_+^0} - \frac{a}{d_-^0} > 0. \tag{5.8}$$

Proof. It is clear that $C_t > 0$ for $t \in (0, T_1)$ and that $C_t = 0$ for $t \in [T_1, T]$. We will first show that ET_1 is finite; after this it remains to be shown when $E[T - T_1]$ is finite. Since we assume that $t = 0$ is a regeneration epoch, the process (D_t, C_t) stays on the y-axis after $t = 0$ for a while (see Figure 5.2). It will leave the y-axis after a time period of length t_1 that is exponentially distributed with parameter a. Then, after at most K/c_- time units, it either hits the x-axis or the y-axis. In the latter case, it again remains on the y-axis for an exponentially distributed period of time, t_2 say, and so on. We define N as the number of times that the process (D_t, C_t) enters the positive y-axis in the time interval $[0, T_1]$. Note that N takes values in $\{1, 2, \ldots\}$ and $EN < \infty$ since the probability that (D_t, C_t) will not reach the y-axis again (once it has left this) before T_1 is always greater than e^{-bK/c_-}. Since N is a stopping time for the i.i.d. sequence $\{t_1, t_2, \ldots\}$, and

$$T_1 \leq t_1 + \frac{K}{c_-} + \ldots + t_N + \frac{K}{c_-},$$

it follows by Wald's Lemma that

$$ET_1 \leq EN\left(\frac{1}{a} + \frac{K}{c_-}\right) < \infty.$$

To examine $E[T - T_1]$, we note that the conditional expectation $E[T - T_1 \mid M_{T_1} = i,\ D_{T_1} = x]$ is equal to the expected first entrance time into 0 of a Markov additive process (V_t) as in Section 4.3, with parameters a, b, d_+^0 and d_-^0, given that (V_t) starts in x and (M_t) in i. It now follows immediately from Corollary 4.2 that condition (5.8) is necessary and sufficient for $E[T - T_1 \mid M_{T_1} = i,\ D_{T_1} = x]$ to be finite, and hence for $E[T - T_1]$ to be finite since $D_{T_1} \leq Kd_+/c_- < \infty$.

Finally, by choosing the first three transition epochs of (M_t) appropriately, it is easily seen that T must have a non-lattice distribution. □

Corollary 5.2 *If (5.8) holds, a random vector (M, D, C) exists, to which the process (M_t, D_t, C_t) converges in distribution as $t \to \infty$.*

We will henceforth assume condition (5.8) to be satisfied. As in the previous chapter we will interpret (M, D, C) as the state of the system in stationarity. Its distribution $\mathbf{F}$ is given by $\mathbf{F}(dx, dy) = (F_0(dx, dy),\ F_1(dx, dy))$ with

$$\begin{aligned} F_i(dx, dy) &= P[M = i,\ D \in dx, C \in dy] \\ &= \lim_{t\to\infty} P[M_t = i,\ D_t \in dx, C_t \in dy], \qquad i \in \{0, 1\}. \end{aligned} \tag{5.9}$$

Our primary interest is in finding this distribution. However, before doing so in Section 5.5, we first analyse the marginal distribution of the content of the credit buffer.

5.4 Stationary marginal distribution of (C_t)

5.4.1 Introduction

In this section we concentrate on the stationary distribution of the content of the credit buffer. As in Chapter 4 it will be useful to look at an embedded process of (C_t), namely at points in time when an idle period of the data buffer is finished. In Section 5.4.2 we do so and derive the stationary distribution of this embedded process. We will use this information in a quite straightforward manner in Section 5.4.3 to find the distribution of C for the two special cases of our model that were introduced in Sections 5.2.2 and 5.2.3. Finally, in Section 5.4.4 we sketch a second approach that yields the distribution of C for the general case. This approach uses Laplace-transform techniques similar to those in Chapter 4, and also uses the results for the embedded process in Section 5.4.2.

5.4.2 An embedded process

Before introducing the embedded process, we will first give some preliminaries. Recall our assumption that $M_0 = 0$, $D_0 = 0$ and $C_0 = 0$, and let $I_0, I_1, \ldots$ and $B_0, B_1, \ldots$ denote respectively the lengths of the *idle* periods and the *busy* periods of (D_t). Note that $\{I_i\}$ is an i.i.d. sequence with generic idle period I that is exponentially distributed with parameter a, whereas the sequence $\{B_i\}$ is not i.i.d. Finally, let Y be a stochastic variable as in Proposition 4.1 starting in $x = 0$ and $i = 1$. Note that Y is distributed as a busy period of (D_t) when we forget the effect of an empty credit buffer. In particular, the Laplace-Stieltjes transform L_Y of Y is given by

$$L_Y(s) = \frac{b}{s + b - \lambda_1(s)d_+},$$

with λ_1 as in Proposition 4.1.

We will now analyse the embedded process (Z_k), where Z_k is the content of the credit buffer at the end of the kth idle period, $k = 0, 1, \ldots$. The behaviour of the process (Z_k) is given by $Z_0 = c_+ I_0$ and

$$Z_{k+1} = K - [K - c_+ I_{k+1} - [Z_k - c_- B_k]^+]^+, \qquad k = 0, 1 \ldots, \tag{5.10}$$

where $[x]^+$ denotes the maximum of x and 0. Direct analysis of (5.10) is problematic, because the variables B_k are not independent, and their distributions are unknown. Fortunately, the distribution of Z_k is the same as that of Z'_k when we define $Z'_0 = c_+ I_0$ and

$$Z'_{k+1} = K - [K - c_+ I_{k+1} - [Z'_k - c_- Y_k]^+]^+, \qquad k = 0, 1 \ldots, \tag{5.11}$$

where $\{I_k\}$ *and* $\{Y_k\}$ are independent i.i.d. sequences distributed as I and Y respectively. This identifies Z'_k as the virtual waiting time immediately after arrival of a customer in a $G/M/1$-queue with uniformly bounded virtual waiting time . Specifically, the capacity of the waiting room is K, the interarrival times are $c_- Y_0$, $c_- Y_1, \ldots$ and the service times $c_+ I_0$, $c_+ I_1, \ldots$.

The distribution of the stationary content immediately after an arrival, Z say, is given by $U(z)$ in (5.104) of [25, Part III] or in (6.10) of [23]. The fact that for certain parameter values $P[Y = \infty] > 0$ is of no consequence to the analysis, since (Z_k) still converges in distribution to a proper random variable Z. Thus we obtain

$$P[Z \le y] = \begin{cases} 1 - \dfrac{G(K-y)}{G(K)}, & y \in [0, K), \\ 1 & y \in [K, \infty), \end{cases} \tag{5.12}$$

where the function G is the inverse Laplace transform of the function

$$L_G(s) = \frac{1}{1 - sc_+/a - L_Y(sc_-)}. \tag{5.13}$$

It is now a matter of calculus to find the distribution of Z.

Proposition 5.3 *The stationary distribution of the process* (Z_k) *is given by*

$$\begin{aligned} P[Z = K] &= P_{ZK} && (5.14)\\ P[Z \in dy] &= f_Z(y)\,dy, \qquad y \in (0, K) && (5.15)\end{aligned}$$

where

$$P_{ZK} = \left(1 + \frac{a}{c_+\beta}(1 - e^{-\beta K}) + \frac{c_-\nu\omega}{2\beta}\int_0^K \left(e^{-\beta(K-u)} - 1\right) e^{-\theta u} H_0(0,u)\,du\right)^{-1}, \tag{5.16}$$

and

$$f_Z(y) = P_{ZK}\ e^{-\beta(K-y)}\left(\frac{a}{c_+} - \frac{c_-\nu\omega}{2}\int_0^{K-y} e^{-(\theta-\beta)u} H_0(0,u)\,du\right). \tag{5.17}$$

Here, the function H_0 *is the same as in Theorem 4.11, i.e.,*

$$H_0(0,u) = \frac{I_1\left(u\sqrt{\omega}\right)}{u\sqrt{\omega}}, \tag{5.18}$$

where I_1 *is the modified Bessel function of the first kind of order* 1*. The constants* β, θ, ν *and* ω *are given by*

$$\begin{aligned} \beta &= \frac{bd_-}{c_-d_- + c_-d_+ + c_+d_+} - \frac{a}{c_+}, && (5.19)\\ \theta &= \frac{bd_- + ad_+}{c_-(d_- + d_+)}, && (5.20)\\ \nu &= \frac{d_- + d_+}{c_-d_- + c_-d_+ + c_+d_+}, && (5.21)\\ \omega &= \frac{4abd_-d_+}{c_-^2(d_- + d_+)^2}. && (5.22)\end{aligned}$$

Proof. Applying a convenient shift we obtain from (5.13),

$$\begin{aligned} L_G(s-\theta) &= -\left(\frac{a^2}{c_+^2\beta} + \frac{a}{c_+}\right)\frac{1}{s-\theta} + \frac{a^2}{c_+^2\beta}\,\frac{1}{s-(\theta-\beta)} \\ &\quad + \frac{ac_-\nu}{2c_+}\,\frac{s - \sqrt{s^2-\omega}}{(s-\theta)(s-(\theta-\beta))}, \end{aligned}$$

where the constants β, θ, ν and ω are given in (5.19), (5.20), (5.21) and (5.22). This may be inverted directly to give

$$G(y) = -\frac{a}{c_+}\left(1 + \frac{a}{c_+\beta}(1 - e^{-\beta y}) + \frac{c_-\nu\omega}{2\beta}\int_0^y \left(e^{-\beta(y-u)} - 1\right) e^{-\theta u} H_0(0,u)\, du\right). \quad (5.23)$$

The result now follows from (5.12). □

The following observation is sufficiently important to state as a lemma. The proof is the same as for Lemma 4.7.

Lemma 5.4 *The conditional distribution of $(C \,|\, D = 0)$, is the same as the distribution of Z, and hence given in Proposition 5.3.*

It follows in particular, by conditioning on D, that

$$P[C = K] \;=\; P_{ZK}\, P[D = 0]. \quad (5.24)$$

5.4.3 Two special cases

We now return to the distribution of C. Clearly, it must have a density for $0 < y < K$, and probability masses in $y = 0$ and $y = K$. The actual form is given in the next proposition.

Proposition 5.5 *The stationary marginal distribution of the process (C_t) is given by*

$$\begin{aligned} P[C = 0\,] &= 1 + \frac{c_+}{c_-}P_{CK} - \left(1 + \frac{c_+}{c_-}\right)\frac{P_{CK}}{P_{ZK}} && (5.25)\\ P[C \in dy] &= f_C(y)\; dy, \qquad y \in (0, K) && (5.26)\\ P[C = K] &= P_{CK} && (5.27) \end{aligned}$$

where

$$f_C(y) = P_{CK}\left(1 + \frac{c_+}{c_-}\right) e^{-\beta(K-y)}\left(\frac{a}{c_+} - \frac{c_-\nu\omega}{2}\int_0^{K-y} e^{-(\theta-\beta)u} H_0(0,u)\, du\right), \quad (5.28)$$

H_0, P_{ZK}, β, θ, ν and ω are the same as in Proposition 5.3, and P_{CK} is a constant that is yet to be determined.

Proof. For the density $f_C(y)$ we have that

$$f_C(y) = \lim_{\varepsilon\to 0}\frac{EU(y,\varepsilon)}{\varepsilon\, ET} \quad (5.29)$$

where $U(y,\varepsilon)$ is defined as the total sojourn time of the process (C_t) in the set $(y, y+\varepsilon]$ during the first regeneration period. By defining $U_I(y,\varepsilon)$ and $U_B(y,\varepsilon)$ as the sojourn time of (D_t, C_t) in $\{0\} \times (y, y+\varepsilon]$ and $(0,\infty) \times (y, y+\varepsilon]$ respectively, we can make two observations. First, we have that $U(y,\varepsilon) = U_I(y,\varepsilon) + U_B(y,\varepsilon)$. Secondly, since the number of visits during one regeneration period of (D_t, C_t) to the sets $\{(0,y)\}$ and $(0,\infty) \times \{y\}$ is equal, we have that $c_+ U_I(y,\varepsilon) = c_- U_B(y,\varepsilon)$. Together, it follows that

$$U(y,\varepsilon) = U_I(y,\varepsilon)\left(1 + \frac{c_+}{c_-}\right). \tag{5.30}$$

Observing that

$$\lim_{\varepsilon \to 0} \frac{EU_I(y,\varepsilon)}{\varepsilon\, ET} = P[D=0] f_Z(y)$$

and combining this with (5.29), we arrive at

$$f_C(y) = \left(1 + \frac{c_+}{c_-}\right) P[D=0] f_Z(y), \tag{5.31}$$

from which we immediately have (5.28), using (5.17) and (5.24). Finally, from normalization it follows that

$$\begin{aligned} P[C=0] &= 1 - P_{CK} - \left(1 + \frac{c_+}{c_-}\right) P[D=0] \int_0^K f_Z(y) dy, \\ &= 1 - P_{CK} - \left(1 + \frac{c_+}{c_-}\right) \frac{P_{CK}}{P_{ZK}} (1 - P_{ZK}), \end{aligned}$$

which leads to (5.25). □

Remark 5.1 We note that, in accordance with Lemma 5.4, the result in Proposition 5.5 may be slightly refined to

$$\begin{aligned} P[D=0, C=K] &= P_{CK}, & (5.32) \\ P[D=0, C \in dy] &= \frac{c_-}{c_- + c_+} f_C(y)\, dy, & (5.33) \end{aligned}$$

and

$$\begin{aligned} P[D>0, C=\ 0\,] &= 1 + \frac{c_+}{c_-} P_{CK} - \left(1 + \frac{c_+}{c_-}\right) \frac{P_{CK}}{P_{ZK}} & (5.34) \\ P[D>0, C \in dy] &= \frac{c_+}{c_- + c_+} f_C(y)\, dy. & (5.35) \end{aligned}$$

Although the results obtained in this subsection so far are valid for the general case, they are only applicable to cases in which we can find the constant P_{CK}. In the remainder we show that we are able to do so in the two special cases of Sections 5.2.2 and 5.2.3, due to the close relationship between the data buffer and the credit buffer in both cases. We will not only give the resulting expression for P_{CK}, but also for other quantities of interest that can be obtained easily.

Two-level traffic shaper

We take the three parameters v_0, v_1 and v_2 as in (5.4) and show that simple expressions for $P_{CK} = P[C = K]$ and $P[D = 0]$ are easily obtained for this case. Balancing the long term input and output of the credit buffer yields

$$v_2\,(1 - P_{CK}) \;=\; v_1\,P[D > 0, C > 0] \;+\; v_2\,P[D > 0, C = 0], \tag{5.36}$$

while a similar balance for the data buffer gives

$$\frac{a}{a+b}\,v_0 \;=\; v_1\,P[D > 0, C > 0] \;+\; v_2\,P[D > 0, C = 0]. \tag{5.37}$$

As an aside we mention that (5.36) is equivalent to (5.25), which is easily seen when we use (5.24), and the equalities $P[D > 0, C > 0] = 1 - P[D = 0] - P[C = 0]$ and $P[D > 0, C = 0] = P[C = 0]$. From (5.36) and (5.37) it follows that

$$P_{CK} = 1 - \frac{a}{a+b}\,\frac{v_0}{v_2}. \tag{5.38}$$

Using (5.24) we can now also find a simple expression for $P[D = 0]$,

$$P[D = 0] = P_{ZK}^{-1}\left(1 - \frac{a}{a+b}\,\frac{v_0}{v_2}\right), \tag{5.39}$$

while from (5.36) or (5.37) we find

$$P[C = 0] = \frac{v_1}{v_1 - v_2}\left\{\left(1 - \frac{a}{a+b}\frac{v_0}{v_1}\right) - P_{ZK}^{-1}\left(1 - \frac{a}{a+b}\frac{v_0}{v_2}\right)\right\}. \tag{5.40}$$

The constant P_{ZK} in these expressions can clearly be expressed in the parameters of the model by combining (5.16) with (5.4).

The fact that P_{CK} is independent of K and v_1, may be surprising at first sight, but it can easily be understood by considering the process $(M_t, D_t - C_t + K)$. This process describes an elementary Markov-modulated fluid system in which an infinitely large fluid reservoir receives fluid at rate v_0 at times when $M_t = 1$, while there is a constant output rate v_2, as long as $D_t - C_t + K > 0$. Since the credit buffer can be completely filled only at times when the data buffer is empty, we have that $P[C = K] = P[D - C + K = 0]$. This leads to an alternative derivation of (5.38) in which the parameters K and v_1 clearly do not play any role. Also, this viewpoint gives us a means to find the expected data buffer occupancy, since we can derive that

$$E[D - C + K] = \frac{av_0}{a+b}\,\frac{v_0 - v_2}{bv_2 - a(v_0 - v_2)},$$

while EC follows immediately from Proposition 5.5.

Tandem queue with finite second buffer

A second way in which the general model may be applied is given by the parameter choice in (5.5). Since the process $(M_t,\, D_t)$ is not influenced by $(\bar{C}_t)$, it follows from (4.20) or directly from the balance equation for the data buffer, that

$$P[D=0] = 1 - \frac{a}{a+b}\,\frac{v_0}{v_1}. \tag{5.41}$$

As a consequence, we immediately find

$$P[\bar{C}=0] = P_{CK} = P_{ZK}\left(1 - \frac{a}{a+b}\,\frac{v_0}{v_1}\right), \tag{5.42}$$

and

$$P[\bar{C}=K] = P[C=0] = \frac{a}{a+b}\,\frac{v_0}{v_1} \; - \; \frac{v_2}{v_1 - v_2}\,(1 - P_{ZK})\left(1 - \frac{a}{a+b}\,\frac{v_0}{v_1}\right), \tag{5.43}$$

by using (5.24) and (5.25) respectively, where P_{ZK} can be found from (5.16) and (5.5).

5.4.4 General case

In this section we briefly sketch a way in which the distribution of C, the content of the credit buffer in stationarity, can be found for the general case. We will proceed along the lines of Chapter 4, which results in the Laplace-Stieltjes transform of C. In particular, we also find formulas for several probabilities of interest, including the constant P_{CK} that remained implicit for the general model in the previous section.

We start off by defining the functions $q_i(p,s)$ as

$$q_i(p,s) = E\mathbf{1}_{\{M=i\}}e^{-pD-sC}, \qquad i=0,1.$$

Notice immediately that $q_1(\infty,s)=0$, while Lemma 5.4 tells us that

$$q_0(\infty,s) = P[D=0]L_Z(s), \tag{5.44}$$

where L_Z, the Laplace-Stieltjes transform of Z, can be straightforwardly derived using Proposition 5.3, resulting in

$$\begin{aligned} L_Z(s) = E\,e^{-sZ} = P_{ZK}\Bigg\{e^{-sK} + \frac{1}{s-\beta}\Bigg\{&e^{-\beta K}\left(\frac{a}{c_+} - \frac{c_-\nu\omega}{2}\int_0^K e^{-(\theta-\beta)u}H_0(0,u)\,du\right) \\ &-e^{-sK}\left(\frac{a}{c_+} - \frac{c_-\nu\omega}{2}\int_0^K e^{(s-\theta)u}H_0(0,u)\,du\right)\Bigg\}\Bigg\}. \end{aligned} \tag{5.45}$$

More information on the functions q_i is given in the following lemma, where as before $\mathbf{q}(p,s) = (q_0(p,s), q_1(p,s))^T$.

Lemma 5.6 *The vector* $\mathbf{q}(p,s)$ *satisfies*

$$A(p,s)\,\mathbf{q}(p,s) = B(p,s)\begin{pmatrix} f(p,s) \\ q_0(p,\infty) \\ q_1(p,\infty) \end{pmatrix}, \tag{5.46}$$

with

$$A(p,s) = \begin{pmatrix} -a + d_-p + c_-s & b \\ a & -b - d_+p + c_-s \end{pmatrix},$$

$$B(p,s) = \begin{pmatrix} 1 & d_-p - d_-^0p + c_-s & 0 \\ 0 & 0 & -d_+p + d_+^0p + c_-s \end{pmatrix}$$

and

$$f(p,s) = (d_-p + c_+s + c_-s)q_0(\infty,s) - c_+se^{-sK}P_{CK}.$$

Proof. The derivation of (5.46) can be carried out analogous to the proof of Lemma 4.10 and is not much harder, despite the finiteness of the credit buffer and the presence of feedback. □

As a consequence of this lemma we have

$$\mathbf{q}(p,s) = \frac{H(p,s)}{\det A(p,s)}\begin{pmatrix} f(p,s) \\ q_0(p,\infty) \\ q_1(p,\infty) \end{pmatrix}, \tag{5.47}$$

where

$$H(p,s) = \begin{pmatrix} -b - d_+p + c_-s & -b \\ -a & -a + d_-p + c_-s \end{pmatrix} B(p,s).$$

We now denote the zeros of $\det A(p,s)$ for fixed $p \geq 0$ by $s_1(p)$ and $s_2(p)$ and observe that $s_1(p) \leq 0 \leq s_2(p)$. As in Remark 4.4 we can now use the fact that $\mathbf{q}(p,s)$ must remain bounded for all $p \geq 0$. Notice in particular that this must also be true when $s \leq 0$, since then $q_i(p,s) < E\,e^{-sC} < e^{-sK}$. Thus, we are able to express $q_0(p,\infty)$ and $q_1(p,\infty)$ in terms of $f_1(p) = f(p, s_1(p))$ and $f_2(p) = f(p, s_2(p))$, and find

$$\begin{aligned} q_0(p,\infty) &= \frac{(f_1(p) + f_2(p))(b + d_+^0p)g(p) + c_-(f_1(p) - f_2(p))g_0(p)}{2p(bd_-^0 - ad_+^0 + d_-^0d_+^0p)g(p)}, \\ q_1(p,\infty) &= a\,\frac{(f_1(p) + f_2(p))g(p) + c_-(f_1(p) - f_2(p))g_1(p)}{2p(bd_-^0 - ad_+^0 + d_-^0d_+^0p)g(p)}, \end{aligned}$$

where

$$g(p) = c_-\sqrt{(-a - b + d_-p - d_+p)^2 - 4p(-bd_- + ad_+ - d_-d_+p)},$$

$$g_0(p) = ab + b(d_- + d_+)p + d_+^0 p(b-a) + d_+^0 p^2 (d_- + d_+),$$

and

$$g_1(p) = a + b + d_- p - 2d_-^0 p + d_+ p.$$

Evaluating (5.47) for $p = 0$ and summing q_0 and q_1 now gives,

$$\begin{aligned} E\, e^{-sC} &= \frac{c_- + c_+}{c_-}\, q_0(\infty, s) - \frac{c_+}{c_-}\, P_{CK}\, e^{-sK} \\ &+ \frac{a(c_- d_- + c_- d_+ + c_+ d_+) - bc_+ d_-}{c_-(bd_-^0 - ad_+^0)}\, q_0(\infty, 0) \\ &+ \frac{c_+ P_{CK}}{c_-(bd_-^0 - ad_+^0)} \left(a\, e^{-(a+b)K/c_-} (d_- - d_-^0 + d_+ - d_+^0) + bd_- - ad_+\right) \\ &- \frac{a(c_- + c_+)(d_- - d_-^0 + d_+ - d_+^0)}{c_-(bd_-^0 - ad_+^0)}\, q_0(\infty, (a+b)/c_-). \end{aligned}$$

Finally, the following proposition follows, using (5.44) and (5.24).

Proposition 5.7 *The Laplace-Stieltjes transform of C is given by*

$$L_C(s) = E\, e^{-sC} \;=\; P[D=0] \left\{ \frac{c_- + c_+}{c_-}\, L_Z(s) - \frac{c_+}{c_-}\, P_{ZK}\, e^{-sK} + \frac{\chi}{c_-} \right\}, \tag{5.48}$$

where $L_Z(s)$, the Laplace-Stieltjes transform of Z, is given in (5.45), P_{ZK} is given in (5.16) and

$$\begin{aligned} \chi \;=\; & \Big\{ a(c_- d_- + c_- d_+ + c_+ d_+) - bc_+ d_- \\ & + c_+(bd_- - ad_+)\; P_{ZK} \\ & + ac_+(d_- - d_-^0 + d_+ - d_+^0)\; P_{ZK}\; e^{-(a+b)K/c_-} \\ & - a(c_- + c_+)(d_- - d_-^0 + d_+ - d_+^0)\, L_Z((a+b)/c_-) \Big\} / (bd_-^0 - ad_+^0). \end{aligned} \tag{5.49}$$

When we take the inverse Laplace transform we find the same result as in Proposition 5.5, apart from the expression for $P[C = 0] = L_C(\infty)$ in terms of P_{ZK} and $P[D = 0]$. This extra information makes it possible to find the following expressions, which are consistent with the ones for the special cases in the previous subsection.

Corollary 5.8 *The following equalities hold,*

$$P[D=0] \;=\; \frac{c_-}{c_- + c_+(1 - P_{ZK}) + \chi}, \tag{5.50}$$

$$P[C=0] \;=\; \frac{\chi}{c_- + c_+(1 - P_{ZK}) + \chi}, \tag{5.51}$$

$$P_{CK} \;=\; \frac{c_- P_{ZK}}{c_- + c_+(1 - P_{ZK}) + \chi}, \tag{5.52}$$

where P_{ZK} and χ are given in (5.16) and (5.49) respectively.

Proof. If we take $s = 0$ and $s \to \infty$ respectively in equation (5.48), we immediately obtain (5.50) and (5.51). P_{CK} is found by (5.24). □

In principle it should be possible to obtain expressions for $q_0(p, s)$ and $q_1(p, s)$, which could then lead to an explicit result for the joint distribution of the process (M_t, D_t, C_t) via inverse Laplace transformation. However, this analysis becomes too difficult to carry out. In the following section we will take another direction to find the joint distribution. The analysis will be completely independent from the current section.

5.5 Stationary joint distribution

We are interested in the stationary joint distribution $\mathbf{F}$ of the process (M_t, D_t, C_t), defined in (5.9). When we let $\overset{\circ}{S}$ denote the interior of S, we expect $\mathbf{F}$ to be of the following form,

$$\begin{aligned}
F_0(\{0\}, \{K\}) &= P_{CK}, && && (5.53)\\
F_i(dx, dy) &= f_i(x, y)\ dx\ dy, && (x, y) \in \overset{\circ}{S},\ i = 0, 1, && (5.54)\\
F_0(\{0\}, dy) &= \sigma_0(y)\ dy, && y \in [0, K], && (5.55)\\
F_1(dx, K - c_-/d_+\, dx) &= \sigma_1(x)\ dx, && x \in [0, Kd_+/c_-], && (5.56)\\
F_i(dx, \{0\}) &= \mu_i(x)\ dx, && x \in [0, \infty),\ i = 0, 1. && (5.57)
\end{aligned}$$

Observe that, as in the previous section, the notation P_{CK} for the probability mass in $(0, 0, K)$ is an abbreviation for $P[C = K]$. In Figure 5.3 the distribution $\mathbf{F}$ is rendered graphically.

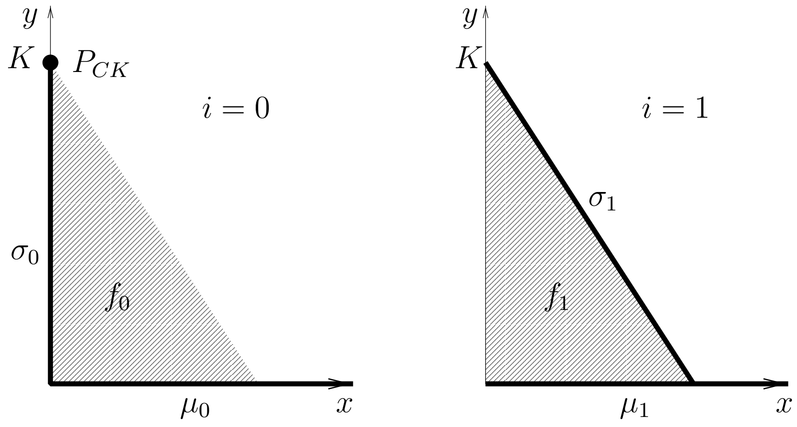


Figure 5.3: The stationary distribution

The following theorem states that the form above is correct and gives explicit expressions for the densities.

Theorem 5.9 *The stationary joint distribution* $\mathbf{F}$ *of the process* (M_t, D_t, C_t) *is of the form (5.53)–(5.57), where the various densities are given as follows.*

$$\sigma_0(y) \;=\; P_{CK}\; e^{-\beta(K-y)} \left(\frac{a}{c_+} - \frac{c_-\nu\omega}{2} \int_0^{K-y} e^{-(\theta-\beta)u} H_0(0,u)\, du \right), \tag{5.58}$$

$$\sigma_1(x) \;=\; P_{CK}\; \frac{a}{d_+} e^{-\frac{b}{d_+}x}, \tag{5.59}$$

$$\begin{aligned} f_0(x,y) \;=\; & P_{CK}\; \frac{\nu b c_+}{d_- + d_+}\; e^{-\frac{b}{d_+}x} \;\times \\ & \Big(\frac{d_+\gamma\omega}{b} \;\; e^{-\theta\left(K-y-\frac{c_-}{d_+}x\right)} \;\; H_1(x, K-y-\frac{c_-}{d_+}x) \\ & + \;\frac{a}{c_+} \;\; e^{-\beta\left(K-y-\frac{c_-}{d_+}x\right)} \;\; \{1 + x\omega\gamma \int_0^{K-y-\frac{c_-}{d_+}x} e^{-(\theta-\beta)u} H_0(x,u) du\} \\ & - \;\frac{c_-\nu\omega}{2} \;\; e^{-\beta\left(K-y-\frac{c_-}{d_+}x\right)} \;\; \int_0^{K-y-\frac{c_-}{d_+}x} e^{-(\theta-\beta)u} H_1(x,u) du \Big), \end{aligned} \tag{5.60}$$

$$\begin{aligned} f_1(x,y) \;=\; & P_{CK}\; \frac{a}{d_+}\; e^{-\frac{b}{d_+}x} \;\times \\ & \Big(\omega\gamma\; x \;\;\; e^{-\theta\left(K-y-\frac{c_-}{d_+}x\right)} \;\; H_0(x, K-y-\frac{c_-}{d_+}x) \\ & + \;\frac{a}{c_+} \;\; e^{-\beta\left(K-y-\frac{c_-}{d_+}x\right)} \;\; \{1 + x\omega\gamma \int_0^{K-y-\frac{c_-}{d_+}x} e^{-(\theta-\beta)u} H_0(x,u) du\} \\ & - \;\frac{c_-\nu\omega}{2} \;\; e^{-\beta\left(K-y-\frac{c_-}{d_+}x\right)} \;\; \int_0^{K-y-\frac{c_-}{d_+}x} e^{-(\theta-\beta)u} H_1(x,u) du \Big), \end{aligned} \tag{5.61}$$

$$\mu_0(x) \;=\; \frac{e^{-\alpha x}}{d_-^0} \{J_1(x \wedge Kd_+/c_-) \;+\; \eta_1(x \wedge Kd_+/c_-)\; J_2(x \wedge Kd_+/c_-)\} \tag{5.62}$$

$$\mu_1(x) \;=\; \frac{d_-^0}{d_+^0}\; \mu_0(x) \;-\; \frac{\mathbf{1}_{\{x<Kd_+/c_-\}}}{d_+^0}\; J_2(x). \tag{5.63}$$

Here, the constant P_{CK} *may be obtained by normalisation and the functions* H_0 *and* H_1 *are the same as in Theorem 4.11, that is, they are given by*

$$H_0(x,y) \;=\; \frac{I_1\left(\sqrt{\omega(y^2 + 2xy\gamma)}\right)}{\sqrt{\omega(y^2 + 2xy\gamma)}}, \tag{5.64}$$

$$\begin{aligned} H_1(x,y) \;=\; & \frac{y^2 + xy\gamma}{y^2 + 2xy\gamma} H_0(x,y) \\ & + \frac{xy\gamma}{y^2 + 2xy\gamma}\; \frac{I_0\left(\sqrt{\omega(y^2 + 2xy\gamma)}\right) + I_2\left(\sqrt{\omega(y^2 + 2xy\gamma)}\right)}{2} \end{aligned} \tag{5.65}$$

where I_i *is the modified Bessel function of the first kind of order* i*,*

$$I_i(z) \;=\; \left(\frac{z}{2}\right)^i \sum_{k=0}^{\infty} \frac{\left(\frac{z}{2}\right)^{2k}}{k!(k+i)!}.$$

Furthermore, $x \wedge Kd_+/c_- \equiv \min(x, Kd_+/c_-)$,

$$\eta_0(u) = \frac{a(e^{\alpha u}-1)}{d_-^0 \alpha}, \tag{5.66}$$

$$\eta_1(u) = \eta_0(u) + e^{\alpha u}, \tag{5.67}$$

$$J_1(x) = c_- \int_{u=0}^{x} \{\eta_0(u) f_0(u,0) + \eta_1(u) f_1(u,0)\}\, du, \tag{5.68}$$

$$J_2(x) = c_- \int_{u=x}^{Kd_+/c_-} \{f_0(u,0) + f_1(u,0)\}\, du \ + \ \sigma_1(Kd_+/c_-), \tag{5.69}$$

and finally,

$$\alpha = \frac{b}{d_+^0} - \frac{a}{d_-^0}, \tag{5.70}$$

$$\beta = \frac{bd_-}{c_-d_- + c_-d_+ + c_+d_+} - \frac{a}{c_+}, \tag{5.71}$$

$$\theta = \frac{bd_- + ad_+}{c_-(d_- + d_+)}, \tag{5.72}$$

$$\omega = \frac{4abd_-d_+}{c_-^2(d_- + d_+)^2}, \tag{5.73}$$

$$\gamma = \frac{c_-(d_- + d_+)}{2d_-d_+}, \tag{5.74}$$

$$\nu = \frac{d_- + d_+}{c_-d_- + c_-d_+ + c_+d_+}. \tag{5.75}$$

Notice that this result is in agreement with the results of the previous section. In particular we note that in correspondence with Lemma 5.4 we have that

$$\sigma_0(y) = P[D=0]\ f_Z(y), \tag{5.76}$$

where

$$P[D=0] = P_{CK} + \int_0^K \sigma_0(y)\, dy = \frac{P_{CK}}{P_{ZK}}. \tag{5.77}$$

To illustrate that calculation of the densities in Theorem 5.9 is numerically feasible, some graphs are shown in Figures 5.4 – 5.6, where the parameter values are the same as in Figure 5.2. The most difficult part of the numerical calculations is the normalization. For Figures 5.4–5.6 we used the explicit expression for P_{CK} that is given in (5.52).

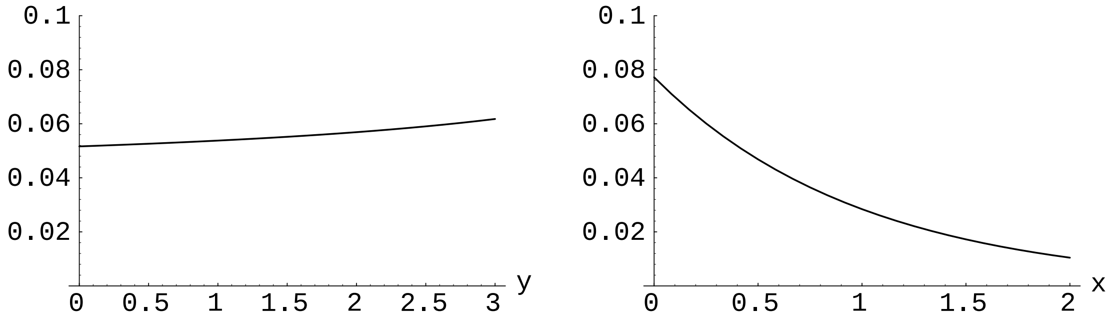


Figure 5.4: The densities σ_0 and σ_1 as functions of y and x, respectively

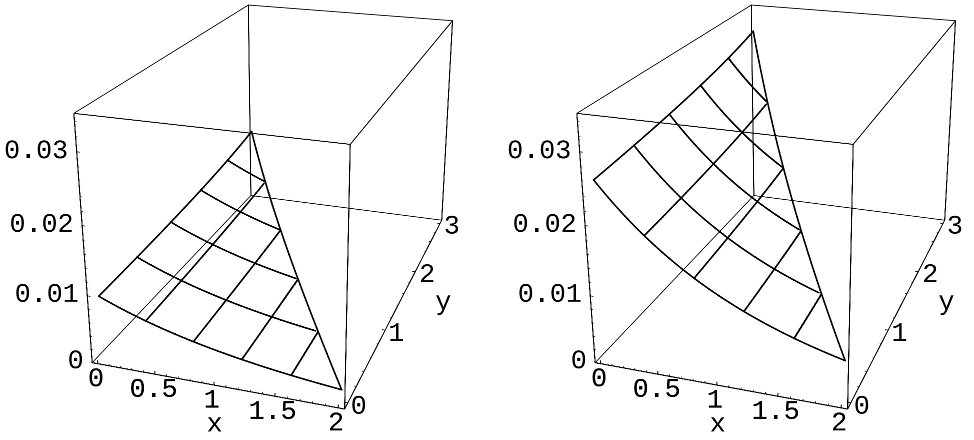


Figure 5.5: The densities f_0 and f_1 as functions of x and y

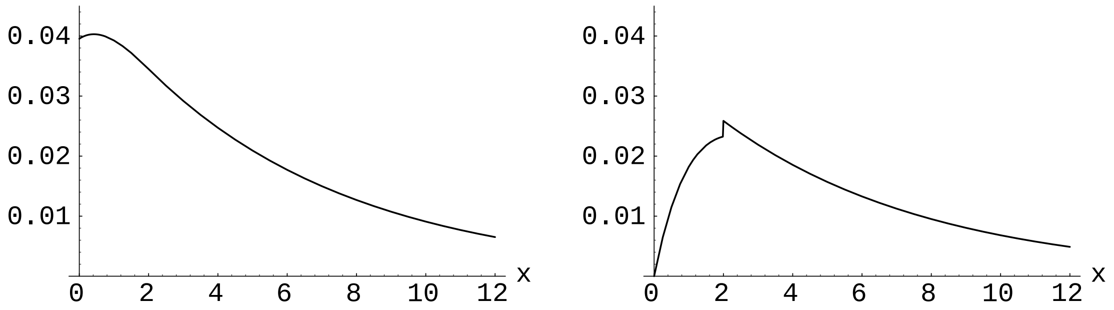


Figure 5.6: The densities μ_0 and μ_1 as functions of x

The proof of Theorem 5.9 requires that we split the state space $\{0,1\} \times S$ of the Markov process in two parts, namely $\{0,1\} \times S_1$ and $\{0,1\} \times S_2$, where S_1 and S_2 are defined in (5.2) and (5.3), see also Figure 5.7(a). The proof is presented in three steps. In the first step we will find $\mathbf{F}$ on the set $\{0,1\} \times S_1$ for the case $\beta > 0$ by relating it to the stationary distribution of a tandem fluid queue. In the second step, we find $\mathbf{F}$ on the set $\{0,1\} \times S_2$. Finally, in the third step we show that the results are also valid for parameter values for which $\beta \leq 0$.

Step 1: Densities $\boldsymbol{\sigma_0, \sigma_1, f_0}$ and $\boldsymbol{f_1}$

In this step we will establish a close relation between the model under consideration and the tandem fluid model in Chapter 4. Hereto, let $(M_t, D_t, \hat{C}_t)$ be the stochastic process that corresponds to the tandem fluid model with the following parameters. We identify the parameters a, b, d_+ and d_- with the parameters of the same name in the current model. Furthermore we will choose the parameters c_+ and c_- to be equal to the parameters c_- and c_+, respectively, of the current model, in other words the symbols are interchanged. In this and the following subsection we will assume that the stability condition for this tandem model holds; since this does not cover all parameter values for which the current model is stable, we will lift this restriction in Step 3. The condition can be found from (4.11) by interchanging the symbols c_+ and c_- and is given by

$$\frac{bd_-}{c_-d_- + c_-d_+ + c_+d_+} - \frac{a}{c_+} > 0, \tag{5.78}$$

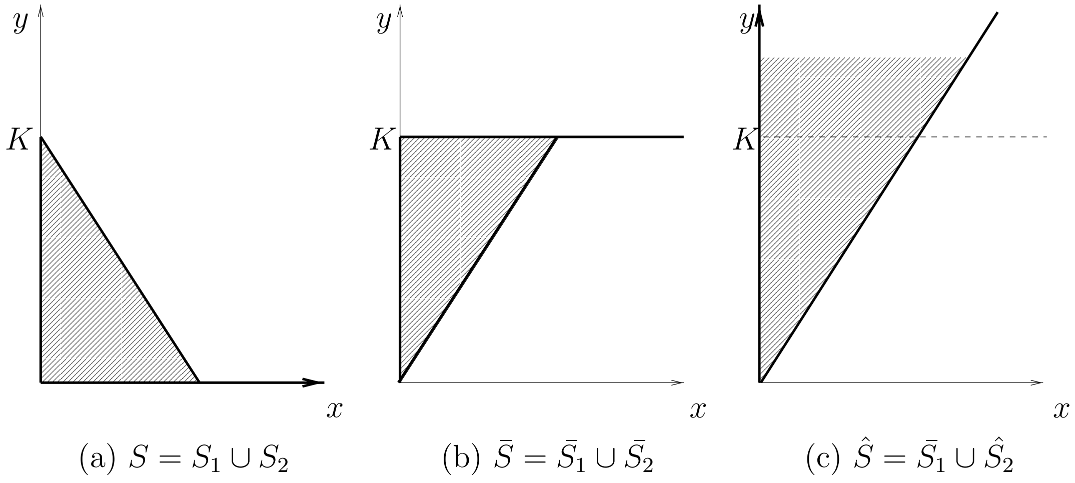
or, equivalently, $\beta > 0$, where β is given in (5.71). Corollary 4.4 now tells us that a stationary distribution for the process $(M_t, D_t, \hat{C}_t)$ exists. We will denote this distribution by $\hat{\mathbf{F}} = (\hat{F}_0(dx,dy), \hat{F}_1(dx,dy))$, where

$$\begin{aligned} \hat{F}_i(dx,dy) &= P[M = i,\, D \in dx, \hat{C} \in dy] \\ &= \lim_{t\to\infty} P[M_t = i,\, D_t \in dx, \hat{C}_t \in dy], \qquad i \in \{0,1\}. \end{aligned} \tag{5.79}$$

Clearly, $\hat{\mathbf{F}}$ can be found from Theorem 4.11, again by interchanging c_+ and c_-.

To find the announced relation between the processes (M_t, D_t, C_t) and $(M_t, D_t, \hat{C}_t)$, we consider yet another stochastic process $(\bar{C}_t)$, where $\bar{C}_t$ is the amount of free space in the credit buffer at time t. Hence, $\bar{C}_t = K - C_t$. In Figure 5.7 the respective state spaces of the processes (D_t, C_t), $(D_t, \bar{C}_t)$ and $(D_t, \hat{C}_t)$ are given.

We will now compare two processes. On the one hand we have the process $(M_t, D_t, \bar{C}_t)$, with state space $\{0,1\} \times (\bar{S}_1 \cup \bar{S}_2)$, where $\bar{S}_i \equiv \{(x,y) \mid (x, K-y) \in S_i\}$. On the other hand we have the process $(M_t, D_t, \hat{C}_t)$ with state space $\{0,1\} \times \hat{S}$ where $\hat{S} \equiv \{(x,y) \mid y \geq 0,\ 0 \leq x \leq yd_+/c_-\}$. It is clear that $\hat{S}$ can be written as $\hat{S} = \bar{S}_1 \cup \hat{S}_2$, with $\hat{S}_2 = \{(x,y) \mid y \geq K,\ 0 \leq x \leq yd_+/c_-\}$. Moreover, the behaviour of the two processes on $\{0,1\} \times \bar{S}_1$ is identical, and both processes enter this set in the same way if $\alpha > 0$ (namely via state $(0,0,K)$ with probability one). It is therefore possible to express the distribution



(a) $S = S_1 \cup S_2$ (b) $\bar{S} = \bar{S}_1 \cup \bar{S}_2$ (c) $\hat{S} = \bar{S}_1 \cup \hat{S}_2$

Figure 5.7: The sets S, $\bar{S}$ and $\hat{S}$

of $(M, D, \bar{C})$ on $\{0,1\} \times \bar{S}_1$ (and hence that of (M, D, C) on $\{0,1\} \times S_1$) in terms of $\hat{F}$, the stationary distribution of $(M_t, D_t, \hat{C}_t)$. This is done in the following proposition.

Proposition 5.10 *If $\alpha > 0$ and $\beta > 0$, the stationary joint distribution $\mathbf{F}$ of the process (M_t, D_t, C_t) on the set $\{0,1\} \times S_1$ is given by*

$$F_i(dx, dy) = k\ \hat{F}_i(dx, K - dy), \qquad (x, y) \in S_1,\ i = 0, 1. \tag{5.80}$$

The constant k is given by

$$k = \frac{P[\bar{C} < K]}{P[\hat{C} < K]} = \frac{E\hat{T}}{ET}, \tag{5.81}$$

where T ($\hat{T}$) is the length of a generic regeneration period of the process $(M_t, D_t, \bar{C}_t)$ (the process $(M_t, D_t, \hat{C}_t)$) if we choose state $(0, 0, K)$ as regeneration state.

Proof. We assume $\alpha, \beta > 0$ and consider Figures 5.7(b) and 5.7(c). The choice of $(0, 0, K)$ as regeneration state for the proces $(M_t, D_t, \bar{C}_t)$ entails that during any regeneration period this process first sojourns in $\{0,1\} \times \bar{S}_1$, for a time period that is distributed as T_1 (which was defined in (5.7)), while during the remainder of such a regeneration period it stays in $\{0,1\} \times \bar{S}_2$, with sojourn time distributed as $T - T_1$. A similar observation can be made for the process $(M_t, D_t, \hat{C}_t)$: first it resides in $\{0,1\} \times \bar{S}_1$, with sojourn time distributed as $\hat{T}_1$, say, after which it remains in $\{0,1\} \times \hat{S}_2$, for a time period distributed as $\hat{T} - \hat{T}_1$. Moreover, the pathwise behaviour of both processes in the time interval $(0, T_1)$ on $\{0,1\} \times \bar{S}_1$ is identical. Hence, we have for any $A \subset \{0,1\} \times \bar{S}_1$,

$$P[(M, D, \bar{C}) \in A \mid (D, \bar{C}) \in \bar{S}_1] \;=\; P[(M, D, \hat{C}) \in A \mid (D, \hat{C}) \in \bar{S}_1],$$

or

$$P[(M,D,\bar{C}) \in A] \;=\; \frac{P[(D,\bar{C}) \in \bar{S}_1]}{P[(D,\hat{C}) \in \bar{S}_1]}\, P[(M,D,\hat{C}) \in A]$$

$$= \frac{ET_1/ET}{E\hat{T}_1/E\hat{T}}\, P[(M,D,\hat{C}) \in A] \;=\; k\; P[(M,D,\hat{C}) \in A].$$

Finally, since

$$F_i(dx,dy) = P[M = i, D \in dx, \bar{C} \in K - dy], \qquad i = 0,1,$$

we easily find the stated results. □

It is now a matter of combining Proposition 5.10 and Theorem 4.11 (with the symbols c_+ and c_- interchanged), to find (5.53) – (5.56) and (5.58) – (5.61), when we take $P_{CK} \equiv F_0(\{0\},\{K\}) = k\ \hat{F}_0(\{0\},\{0\})$.

Step 2: Densities μ_0 and μ_1

Having found the distribution of (M_t, D_t, C_t) on $\{0,1\} \times S_1$ (apart from normalization) in the previous subsection, we proceed to derive the densities μ_0 and μ_1 in (5.57). To do so, we first need to prove two lemmas. The first one gives us the entrance distribution G of the process (M_t, D_t, C_t) into the set $\{0,1\} \times S_2$, that is,

$$G_i(dx) = P[M_{T_1} = i, D_{T_1} \in dx], \qquad 0 \le x \le Kd_+/c_-,\ i = 0,1.$$

with T_1 as in (5.7).

Lemma 5.11 *The joint distribution G of the stochastic variable (M_{T_1}, D_{T_1}) is given by*

$$G_0(dx) \;=\; ET\ c_-\ f_0(x,0)\ dx \tag{5.82}$$

$$G_1(dx) \;=\; ET\ \left\{c_-\ f_1(x,0)\ +\ \delta_{Kd_+/c_-}(x)\ \sigma_1(Kd_+/c_-)\right\} dx, \tag{5.83}$$

where δ_{Kd_+/c_-} denotes the Dirac measure at Kd_+/c_-, and σ_1, f_0 and f_1 are given in (5.59), (5.60), and (5.61).

Proof. We consider the set $\{i\} \times (0,x] \times (0,\varepsilon)$. The sojourn time $V_i(x,\varepsilon)$ of (M_t, D_t, C_t) in this set during the interval $[0,T]$ is equal to $\varepsilon/c_- + o(\varepsilon)$ if $\{M_{T_1} = i, D_{T_1} \le x\}$ occurs, and is $o(\varepsilon)$ otherwise. In other words, we have

$$V_i(x,\varepsilon) \;=\; \frac{\varepsilon}{c_-}\ \mathbf{1}_{\{M_{T_1}=i, D_{T_1} \le x\}} + o(\varepsilon).$$

If we take expectations, divide by ET and apply the Key Renewal Theorem, we obtain

$$P[M = i, D \le x, 0 < C < \varepsilon] = \frac{\varepsilon}{c_- ET} P[M_{T_1} = i, D_{T_1} \le x] + o(\varepsilon).$$

We now find for $x < Kd_+/c_-$,

$$G_i((0,x]) = c_-ET \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\int_0^{\varepsilon}\int_0^{x} f_i(u,v)du\ dv = c_-ET\int_0^x f_i(u,0)du,$$

while an extra term $ET\ \sigma_1(Kd_+/c_-)$ appears if $i=1$ and $x = Kd_+/c_-$. The result is now immediate. □

For the second lemma, we define $N_i(x)$ as the number of times that the process (M_t, D_t, C_t) visits $(i,x,0)$ before it reaches $(0,0,0)$ during the first regeneration period. Also, for $u \ge 0$ and $j = 0,1$, we let

$$P_{j,u}[\ \cdot\ ] \equiv P[\ \cdot \mid M_{T_1} = j, D_{T_1} = u],$$

and

$$E_{j,u}[\ \cdot\ ] \equiv E[\ \cdot \mid M_{T_1} = j, D_{T_1} = u].$$

Lemma 5.12 *The conditional expectations $E_{j,u}N_i(x)$ are given by*

$$E_{j,u}N_0(x) \;=\; \left\{\begin{array}{lcll} E_{j,u}N_1(x) &=& e^{-\alpha x}\ \eta_j(u), & u\le x,\ j=0,1,\\ E_{j,u}N_0(x) &=& e^{-\alpha x}\ \eta_1(x), & u> x,\ j=0,1,\\ E_{j,u}N_1(x) &=& e^{-\alpha x}\ \eta_0(x), & u> x,\ j=0,1,\end{array}\right. \tag{5.84}$$

where

$$\eta_0(u) \;=\; \frac{a(e^{\alpha u}-1)}{d_-^0\alpha}, \tag{5.85}$$

$$\eta_1(u) \;=\; \frac{bd_-^0e^{\alpha u} - ad_+^0}{d_-^0d_+^0\alpha} = \eta_0(u) + e^{\alpha u}, \tag{5.86}$$

and $\alpha = b/d_+^0 - a/d_-^0$.

Proof. First, we define

$$p_j(u,x) = P_{j,u}[D_t = x \text{ for some } t\in(T_1,T]\,]$$

By conditioning on the first transition epoch of the process (M_t), we obtain the following relations for $u \le x$,

$$\begin{aligned} p_0(u,x) &= \int_0^{u/d_-^0} p_1(u-d_-^0v,x)\ ae^{-av}\ dv,\\ p_1(u,x) &= \int_0^{(x-u)/d_+^0} p_0(u+d_+^0v,x)\ be^{-bv}\ dv\ +\ e^{-b(x-u)/d_+^0}, \end{aligned}$$

while for $u > x$ we have $p_j(u,x) = 1$.

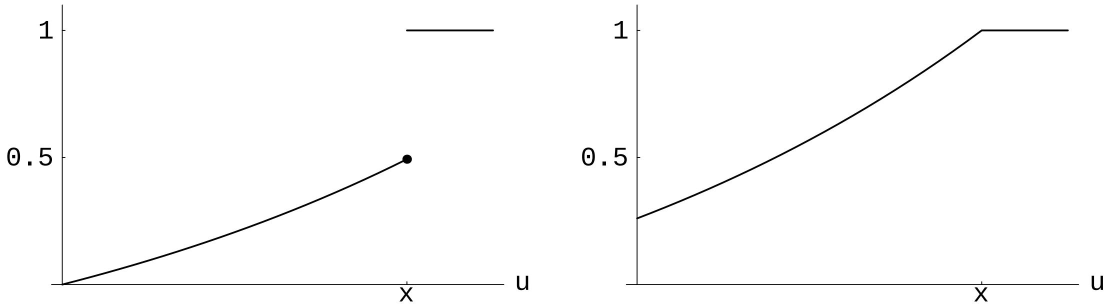


Figure 5.8: The probabilities $p_0(u,x)$ and $p_1(u,x)$ for fixed x

Using the transformations $v \mapsto u - d_-^0 v$ and $v \mapsto u + d_+^0 v$, respectively, and differentiating with respect to u gives the following differential equation for the vector $\mathbf{p}(u,x) = (p_0(u,x), p_1(u,x))^T$ in u,

$$\frac{\partial}{\partial u}\mathbf{p}(u,x) = \begin{pmatrix} -a/d_-^0 & a/d_-^0 \\ -b/d_+^0 & b/d_+^0 \end{pmatrix} \mathbf{p}(u,x), \qquad 0 \le u < x,$$

with boundary conditions $p_0(0,x) = 0$ and $p_1(x,x) = 1$. It follows that the probabilities $p_j(u,x)$ are given by

$$\begin{aligned} p_j(u,x) &= 1, & u > x,\ j = 0,1, \qquad (5.87)\\ p_j(u,x) &= \eta_j(u)/\eta_1(x), & u \le x,\ j = 0,1, \qquad (5.88) \end{aligned}$$

see Figure 5.8. Since the conditional distribution of $N_0(x)$ is given by

$$\begin{aligned} P_{j,u}[N_0(x) = 0] &= 1 - p_j(u,x) \\ P_{j,u}[N_0(x) = k] &= p_j(u,x)\ (1 - p_0(x,x))\ (p_0(x,x))^{k-1}, \qquad k = 1,2,\ldots, \end{aligned}$$

we have

$$E_{j,u}N_0(x) = \frac{p_j(u,x)}{1 - p_0(x,x)}.$$

Furthermore, we have

$$E_{j,u}N_1(x) = \begin{cases} E_{j,u}N_0(x), & \text{if } x > u, \\ E_{j,u}N_0(x) \ - \ 1, & \text{if } x < u. \end{cases}$$

The desired result now follows immediately using (5.87) and (5.88). □

We are now ready to specify the densities μ_i, $i = 0,1$.

Proposition 5.13 *If $\alpha > 0$ and $\beta > 0$, the stationary joint distribution $\mathbf{F}$ of the process (M_t, D_t, C_t) on the set $\{0,1\} \times S_2$ is given by*

$$F_i(dx, \{0\}) = \mu_i(x)\, dx, \qquad x > 0,\ i = 0,1. \tag{5.89}$$

The densities μ_i, $i = 0,1$, are given by

$$\begin{aligned} \mu_0(x) &= \frac{e^{-\alpha x}}{d_-^0}\{J_1(x \wedge Kd_+/c_-) + \eta_1(x \wedge Kd_+/c_-)J_2(x \wedge Kd_+/c_-)\} \\ \mu_1(x) &= \frac{d_-^0}{d_+^0}\mu_0(x) - \frac{\mathbf{1}_{\{x<Kd_+/c_-\}}}{d_+^0}J_2(x), \end{aligned}$$

where $x \wedge Kd_+/c_- \equiv \min(x, \frac{Kd_+}{c_-})$,

$$\begin{aligned} J_1(x) &= c_- \int_{u=0}^{x} \{\eta_0(u)f_0(u,0) + \eta_1(u)f_1(u,0)\}\, du, \\ J_2(x) &= c_- \int_{u=x}^{Kd_+/c_-} \{f_0(u,0) + f_1(u,0)\}\, du \ + \ \sigma_1(Kd_+/c_-), \end{aligned}$$

and η_0 and η_1 are given in (5.85) and (5.86).

Proof. First we denote the sojourn time of the process (M_t, D_t, C_t) in the set $\{i\} \times [x, x+\varepsilon] \times \{0\}$ by $W_i(x,\varepsilon)$, that is,

$$W_i(x,\varepsilon) = \int_{t=T_1}^{T} \mathbf{1}_{\{M_t=i, D_t \in [x,x+\varepsilon]\}} dt.$$

Clearly, we have

$$E_{j,u}W_0(x,\varepsilon) = \frac{E_{j,u}N_0(x)}{d_-^0}\varepsilon + o(\varepsilon), \tag{5.90}$$

$$E_{j,u}W_1(x,\varepsilon) = \frac{E_{j,u}N_1(x)}{d_+^0}\varepsilon + o(\varepsilon). \tag{5.91}$$

Combining

$$\mu_i(x) = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon\, ET} \sum_{j=0}^{1} \int_{u=0}^{Kd_+/c_-} E_{j,u}W_i(x,\varepsilon)G_j(du), \qquad x > 0,\ i = 0,1,$$

with (5.90) and (5.91) and then using Lemmas 5.11 and 5.12 leads to the result. □

It is not difficult to check that Propositions 5.10 and 5.13 together lead to the conclusion that the distribution $\mathbf{F}$ given in Theorem 5.9 indeed is the stationary distribution of the process (M_t, D_t, C_t) when $\alpha > 0$ and $\beta > 0$.

As a side result in this subsection, we find an expression for ET, namely $ET = 1/J_2(0)$. This can be found by normalisation of the distribution G in Lemma 5.11.

Step 3: $\boldsymbol{\beta \leq 0}$

In this last step it remains to be shown that the distribution in Theorem 5.9 not only represents the stationary distribution of the process (M_t, D_t, C_t) when $\alpha > 0$ and $\beta > 0$, as we showed in the previous steps, but also when $\alpha > 0$ and $\beta \leq 0$.

We fix the parameters b, d_+, d_-, d_+^0, d_-^0, c_+, c_- and K, and let a vary. Then we have $\alpha > 0$ if and only if $a < a_1 = bd_-^0/d_+^0$, while $\beta > 0$ is equivalent to $a < a_0 = (bd_-c_+)/(c_-d_- + c_-d_+ + c_+d_+)$, see Figure 5.9. We will assume that $a_0 < a_1$, otherwise $\alpha > 0$ would imply $\beta > 0$.

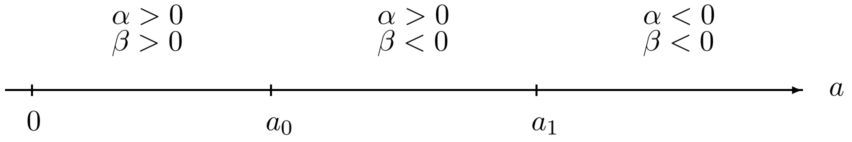


Figure 5.9: Behaviour of α and β as functions of a

In what follows we will need the infinitesimal generator $\mathcal{A}$ of the process (M_t, D_t, C_t), which is an operator mapping a function $\mathbf{h} : \mathbb{R}^2 \to \mathbb{R}^2$ to another function $\mathcal{A}\mathbf{h} : \mathbb{R}^2 \to \mathbb{R}^2$, with, for $x, y \geq 0$,

$$(\mathcal{A}\mathbf{h})(x,y) = \lim_{t \downarrow 0} t^{-1} \begin{pmatrix} E[h_{M_t}(D_t, C_t) - h_0(x,y) | M_0 = 0,\ D_0 = x,\ C_0 = y] \\ E[h_{M_t}(D_t, C_t) - h_1(x,y) | M_0 = 1,\ D_0 = x,\ C_0 = y] \end{pmatrix}.$$

It is not difficult to see that

$$\begin{aligned} (\mathcal{A}\mathbf{h})(x,y) &= Q\mathbf{h}(x,y) + (\mathcal{A}_0\mathbf{h})(x,y), & x > 0,\ 0 < y < K && (5.92) \\ (\mathcal{A}\mathbf{h})(0,y) &= Q\mathbf{h}(0,y) + (\mathcal{A}_1\mathbf{h})(0,y), & 0 < y < K, && (5.93) \\ (\mathcal{A}\mathbf{h})(x,0) &= Q\mathbf{h}(x,0) + (\mathcal{A}_2\mathbf{h})(x,0), & x > 0, && (5.94) \\ (\mathcal{A}\mathbf{h})(0,K) &= Q\mathbf{h}(0,K) + (\mathcal{A}_3\mathbf{h})(0,K), & && (5.95) \end{aligned}$$

where Q is the generator of the process (M_t),

$$\mathcal{A}_0 = \begin{pmatrix} -d_-\frac{\partial}{\partial x} - c_-\frac{\partial}{\partial y} & 0 \\ 0 & d_+\frac{\partial}{\partial x} - c_-\frac{\partial}{\partial y} \end{pmatrix}, \tag{5.96}$$

$$\mathcal{A}_1 = \begin{pmatrix} c_+\frac{\partial}{\partial y} & 0 \\ 0 & d_+\frac{\partial}{\partial x} - c_-\frac{\partial}{\partial y} \end{pmatrix}, \tag{5.97}$$

$$\mathcal{A}_2 = \begin{pmatrix} -d_-^0\frac{\partial}{\partial x} & 0 \\ 0 & d_+^0\frac{\partial}{\partial x} \end{pmatrix}, \tag{5.98}$$

and

$$\mathcal{A}_3 = \begin{pmatrix} c_+\frac{\partial}{\partial y} & 0 \\ 0 & d_+\frac{\partial}{\partial x} - c_-\frac{\partial}{\partial y} \end{pmatrix}. \tag{5.99}$$

The operator $\mathcal{A}$ can be viewed as a generalisation of the Q-matrix corresponding to a continuous-time Markov process with a finite state space. In the latter context a probability measure π is stationary if and only if it satisfies $\pi Q = \mathbf{0}$, i.e. if $\pi Q \mathbf{v} = 0$ for all vectors $\mathbf{v}$. Likewise, here a measure $\mathbf{F}$ is stationary if and only if it satisfies $\mathbf{F}\mathcal{A}\mathbf{h} = 0$ for all (vector-valued) functions $\mathbf{h}$, i.e.,

$$\int_0^\infty \int_0^\infty \mathbf{F}^T(dx,dy)(\mathcal{A}\mathbf{h})(x,y) \;= 0, \tag{5.100}$$

(see, e.g., [38, page 239]). According to Corollary 5.2 a unique limiting distribution exists for any $a \in (0,a_1)$, regardless of the value of β. Moreover, we know that for $a \in (0,a_0)$ this distribution is given by the specific distribution we found in Steps 1 and 2. We will designate this distribution here by $\mathbf{F}_a$ to emphasize its dependence on the parameter a. Because the limiting distribution is stationary, we can conclude that for any suitable function $\mathbf{h}$ and any $a \in (0,a_0)$, equation (5.100) holds for $\mathbf{F} = \mathbf{F}_a$, that is,

$$\begin{aligned}
0 \;=\; & P_{CK}\, a\,(h_1-h_0)(0,K) \;+\; \int_0^K \sigma_0(y)\left(a(h_1-h_0)(0,y) + c_+\frac{\partial h_0}{\partial y}(0,y)\right)dy \\
&+ \int_0^{Kd_+/c_-} \sigma_1(x)\left(-b(h_1-h_0)(x,K-c_-x/d_+) + d_+\frac{\partial h_1}{\partial x}(x,K-c_-x/d_+)\right. \\
&\qquad \left. -\, c_-\frac{\partial h_1}{\partial y}(x,K-c_-x/d_+)\right)dx \\
&+ \int_0^K \int_0^{(K-y)d_+/c_-} \left[f_0(x,y)\left(a(h_1-h_0)(x,y) - d_-\frac{\partial h_0}{\partial x}(x,y) - c_-\frac{\partial h_0}{\partial y}(x,y)\right)\right. \\
&\qquad \left. +\, f_1(x,y)\left(-b(h_1-h_0)(x,y) + d_+\frac{\partial h_1}{\partial x}(x,y) - c_-\frac{\partial h_1}{\partial y}(x,y)\right)\right]dxdy \\
&+ \int_0^\infty \left[\mu_0(x)\left(a(h_1-h_0)(x,0) - d_-^0\frac{\partial h_0}{\partial x}(x,0)\right)\right. \\
&\qquad \left. +\, \mu_1(x)\left(-b(h_1-h_0)(x,0) + d_+^0\frac{\partial h_1}{\partial x}(x,0)\right)\right]dx\;.
\end{aligned} \tag{5.101}$$

To show that the above is also true for $a \in [a_0,a_1)$, we prove the following lemma, in which we will show that for certain $a \in \mathbb{C}$ the right hand side of (5.101) is a complex analytic function of a. Because it is hard to check whether the normalization constant P_{CK} is an analytic function of a, we set $P_{CK} = 1$ for a moment, thereby ignoring the probabilistic interpretation of $\mathbf{F}_a$ (and of P_{CK} itself).

Lemma 5.14 *For any entire function* $\mathbf{h} : \mathbb{C}^2 \to \mathbb{C}^2$*, the function*

$$a \mapsto \int_0^\infty \int_0^\infty \mathbf{F}_a^T(dx, dy)(\mathcal{A}\mathbf{h})(x, y)$$

with $P_{CK} = 1$ *is complex analytic for* $a \in \{z \in \mathbb{C} | \mathrm{Re}(z) < a_1\}$

Proof. First we note that the singularities of the functions H_0 and H_1 in (5.64) and (5.65) can be removed by writing

$$H_0(x, y) = \frac{1}{2}\sum_{k=0}^{\infty} \frac{(z/4)^k}{k!(k+1)!}, \tag{5.102}$$

$$H_1(x, y) = H_0(x, y) + \frac{\omega x y \gamma}{4} \sum_{k=0}^{\infty} \frac{(z/4)^k}{k!(k+2)!}, \tag{5.103}$$

with

$$z = \omega(y^2 + 2xy\gamma) = \frac{4bd_-d_+}{c_-^2(d_- + d_+)^2}\,(y^2 + 2xy\gamma)\,a\,.$$

Since the power series in (5.102) and (5.103) are uniformly converging for all $z \in \mathbb{C}$, they are entire functions of z. Furthermore, since

$$(a, u) \mapsto \frac{4bd_-d_+}{c_-^2(d_- + d_+)^2}\; a\; u^2\,,$$

is an entire function of a for fixed u, but also of u for fixed a ($a, u \in \mathbb{C}$), and since sums, products and concatenations of entire funtions are again entire functions, we conclude that the integrand in (5.58) is also an entire function of a (for fixed u) and of u (for fixed a). But then the integral in (5.58), and hence $(a, y) \mapsto \sigma_0(y)$ is an entire function of a for fixed y and of y for fixed a, since the same holds in general for

$$(a, y) \mapsto \int_0^y g(a, u)du$$

when g is an entire function of a for fixed u and of u for fixed a. Similar statements can be shown to hold for σ_1, f_0, f_1, J_1, J_2, μ_0 and μ_1.

The lemma now follows readily because the partial derivatives of $\mathbf{h}$ are entire functions of x for fixed y and of y for fixed x. The restriction to $\mathrm{Re}(a) < a_1$ is due to the divergence of the last integral in (5.101) for other values of a. □

By analytic continuation, we can now conclude that equation (5.101) holds, for any $a \in \mathbb{C}$ with $\mathrm{Re}(a) < a_1$, even for general P_{CK}. In particular, for a real, $a \in [a_0, a_1)$, we find $\mathbf{F}_a$ to be a stationary distribution, when we choose P_{CK} such that the total probability is 1, as before. The fact that $\mathbf{F}_a$ is the *only* stationary distribution is immediate, since we know that the process has a unique limiting distribution, regardless of the initial distribution.

This concludes the proof of Theorem 5.9.

5.6 Special case: infinite credit buffer

We will now consider the case in which the credit buffer is infinitely large. We will establish an explicit result for the stationary distribution, which turns out to be relatively simple. Before we do so, however, we determine the conditions under which this distribution exists.

Theorem 5.15 If the size K of the credit buffer is infinite, the process (M_t, D_t, C_t) is regenerative with regeneration cycles that have a non-lattice distribution and finite expectation if and only if

$$\alpha = \frac{b}{d_+^0} - \frac{a}{d_-^0} > 0, \tag{5.104}$$

and

$$\beta = \frac{bd_-}{c_-d_- + c_-d_+ + c_+d_+} - \frac{a}{c_+} < 0. \tag{5.105}$$

Proof. As before, we take $(0,0,0)$ as regeneration point and assume that $(M_0, D_0, C_0) = (0,0,0)$. Furthermore, we let the stochastic variables Y, $I_0, I_1, \ldots$, and $B_0, B_1, \ldots$ be as in Section 5.4.2. We first consider the process at embedded points in time. Specifically, Z_k will denote the content of the credit buffer at the *beginning* of the kth idle period, $k = 0, 1, \ldots$ (notice that in Section 5.4.2 we observed the system at the endings of the idle periods). Thus, we let $Z_0 = 0$ and

$$Z_{k+1} = [Z_k + c_+I_k - c_-B_k]^+, \ k = 0, 1, \ldots,$$

We form a Lindley process (Z'_k) with the property that the distribution of Z'_k is the same as the distribution of Z_k, by defining $Z'_0 = 0$ and

$$Z'_{k+1} = [Z'_k + c_+I_k - c_-Y_k]^+, \ k = 0, 1, \ldots,$$

where $\{Y_k\}$ is an i.i.d. sequence of random variables distributed as Y. Letting $\tau = \min\{k > 0 \,|\, Z_k = 0\}$, we find that $E\tau < \infty$ if and only if

$$c_- \, EY > c_+ \, EI, \tag{5.106}$$

since the regeneration periods of the processes (Z_i) and (Z'_i) have equal distributions. We now use Corollary 4.2 to assert that when $bd_- > ad_+$, then (5.106) is equivalent to

$$c_- \frac{d_- + d_+}{bd_- - ad_+} > \frac{c_+}{a},$$

while on the other hand, if $bd_- \leq ad_+$, then $EY = \infty$ and (5.106) is obviously satisfied. In fact, since $bd_- \leq ad_+$ implies $\beta < 0$, we can conclude that $E\tau < \infty$ if and only if $\beta < 0$.

We now consider the joint process (M_t, D_t, C_t) in continuous time, and let T and T_1 be as defined in (5.6) and (5.7). We want to show that T has a finite expectation if and only if (5.104) and (5.105) hold.

First, we note that the following inequalities hold,

$$I_0 + B_0 + \cdots + I_{\tau-1} < T_1 \leq T = I_0 + B_0 + \cdots + I_{\tau-1} + B_{\tau-1}.$$

Because the total inflow into the credit buffer during the time interval $[0, T_1]$ must be equal to the total outflow, we have

$$c_+(I_0 + I_1 + \cdots + I_{\tau-1}) = c_-(T_1 - I_0 - I_1 \cdots - I_{\tau-1}), \tag{5.107}$$

which immediately leads to

$$T_1 = \frac{c_- + c_+}{c_-}\,(I_0 + \cdots + I_{\tau-1}).$$

By applying Wald's Lemma we subsequently find

$$ET_1 = \frac{c_- + c_+}{c_-}\,\frac{E\tau}{a},$$

which is finite if and only if $\beta > 0$. It remains to be shown that $E[T - T_1]$ is finite. During the interval $[T_1, T]$, the data buffer has up- and down-rates d_+^0 and d_-^0, respectively, while a (rough) upperbound for the amount of data at time T_1 is d_+T_1. Therefore, by conditioning on T_1 and using Corollary 4.2, we have

$$E[T - T_1] \leq E\frac{d_-^0 + d_+^0 + (a+b)d_+T_1}{bd_-^0 - ad_+^0} < \infty,$$

provided that (5.104) holds. Reversely, if (5.104) does not hold, $E[T - T_1 \,|\, T_1]$ is not finite according to Corollary 4.2.

The proof that T has a non-lattice distribution is the same as for the case in which the credit buffer is finite. □

We will assume in the remainder of this section that $\alpha > 0$ and $\beta < 0$, and conclude that the stationary distribution $\mathbf{F}$ of the process (M_t, D_t, C_t) exists. In the proof of the following theorem we show how this distribution follows from the one in Theorem 5.9 by letting $K \to \infty$.

Theorem 5.16 *If the size K of the credit buffer is infinite, the stationary joint distribution $\mathbf{F}$ of the process (M_t, D_t, C_t) is given by*

$$\begin{aligned}
F_0(\{0\}, dy) &= \sigma_0(y)\,dy, && y > 0, && (5.108)\\
F_i(dx, dy) &= f_i(x,y)\,dx\,dy, && x, y > 0, \quad i = 0,1, && (5.109)\\
F_i(dx, \{0\}) &= \mu_i(x)\,dx, && x > 0, \quad i = 0,1, && (5.110)
\end{aligned}$$

where the densities σ_0, f_i and μ_i, $i = 0, 1$, are given by

$$\sigma_0(y) = \kappa\, e^{\beta y}, \tag{5.111}$$

$$f_0(x,y) = \kappa\, \frac{bc_+}{c_-d_- + c_-d_+ + c_+d_+}\, e^{-\zeta x + \beta y}, \tag{5.112}$$

$$f_1(x,y) = \kappa\, \frac{a}{d_+}\, e^{-\zeta x + \beta y}, \tag{5.113}$$

$$\mu_0(x) = \frac{\kappa}{d_-^0(\zeta - \alpha)} \left\{ \frac{a(c_-d_-^0 + c_+d_+)}{d_+d_-^0}\, e^{-\alpha x} \right.$$
$$\left. - \frac{bc_+(c_-d_- + c_-d_+ + c_+d_+ - c_-d_+^0)}{d_+^0(c_-d_- + c_-d_+ + c_+d_+)}\, e^{-\zeta x} \right\}, \tag{5.114}$$

$$\mu_1(x) = \kappa\, \frac{a(c_-d_-^0 + c_+d_+)}{d_-^0 d_+ d_+^0(\zeta - \alpha)}\, (e^{-\alpha x} - e^{-\zeta x}). \tag{5.115}$$

The constants α and β are as in (5.70) and (5.71), and

$$\zeta = \frac{ac_-}{c_+d_+} + \frac{bc_-}{c_-d_- + c_-d_+ + c_+d_+}, \tag{5.116}$$

$$\kappa = \frac{\alpha\beta\zeta d_+ d_-^0 d_+^0}{(a+b)(\beta(c_-d_-^0 + c_+d_+) - \alpha d_-^0 d_+^0)}. \tag{5.117}$$

Proof. We assume that $\alpha > 0$ and $\beta < 0$. It is clear that for $K \to \infty$ the state space of the process (M_t, D_t, C_t) becomes $\{(x,y) \mid x, y \geq 0\}$, and, hence, that the form of the distribution in (5.53) – (5.57) must simplify to (5.108) – (5.110). Notice that the probability mass in (5.53) and the density σ_1 in (5.56) indeed vanish, since $\lim_{K\to\infty} P_{CK} = 0$, which is intuitively clear and can be verified using (5.52).

We will now show how the expressions for the densities σ_0, f_i and μ_i, $i = 0, 1$, in Theorem 5.9, simplify to the ones above by taking the limit for $K \to \infty$. First we let $K \to \infty$ in the expression for σ_0 in (5.58), and use that

$$\omega \int_0^\infty e^{-(\theta-\beta)u} H_0(0,u)\, du = (\theta - \beta) - \sqrt{(\theta-\beta)^2 - \omega},$$

see [37, page 235]. This leads to (5.111) with

$$\begin{aligned} \kappa &= \left(\frac{a}{c_+} - \frac{c_-\nu}{2} \left(\theta - \beta - \sqrt{(\theta-\beta)^2 - \omega} \right) \right) \lim_{K\to\infty} P_{CK}\, e^{-\beta K} \\ &= \left(\frac{a}{c_+} - \frac{bc_+d_-d_+}{(c_-d_- + c_-d_+ + c_+d_+)^2} \right) \lim_{K\to\infty} P_{CK}\, e^{-\beta K}, \end{aligned} \tag{5.118}$$

supposing that the limit exists. Next, looking at the expression for f_0 in (5.60), we find that

$$\lim_{K\to\infty} e^{-\theta K} H_1(x,K) = 0,$$

where we used

$$I_i(z) \;=\; \left(\frac{z}{2}\right)^i \sum_{k=0}^{\infty} \frac{\left(\frac{z}{2}\right)^{2k}}{k!(k+i)!} \;<\; \left(\frac{z}{2}\right)^i e^z,$$

and $\theta > \sqrt{\omega}$. For the other two terms we find

$$1 + x\omega\gamma \int_0^\infty e^{-(\theta-\beta)u} H_0(x,u)du \;=\; e^{x\gamma\left(\theta-\beta-\sqrt{(\theta-\beta)^2-\omega}\right)}, \tag{5.119}$$

and

$$\omega \int_0^\infty e^{-(\theta-\beta)u} H_1(x,u)du \;=\; \left(\theta-\beta-\sqrt{(\theta-\beta)^2-\omega}\right) e^{x\gamma\left(\theta-\beta-\sqrt{(\theta-\beta)^2-\omega}\right)}, \tag{5.120}$$

see the proof of Theorem 4.11. This leads us immediately to (5.112), with κ as in (5.118). Equation (5.113) is checked in a similar way. Taking the limit in (5.62) and (5.63) yields

$$\begin{aligned}
\mu_0(x) \;&=\; \frac{c_- e^{-\alpha x}}{d_-^0}\Big\{ \int_{u=0}^{x} \{\eta_0(u) f_0(u,0) + \eta_1(u) f_1(u,0)\}\, du \\
&\qquad\qquad + \eta_1(x) \int_{u=x}^{\infty} \{f_0(u,0) + f_1(u,0)\}\, du \Big\} \\
\mu_1(x) \;&=\; \frac{d_-^0}{d_+^0}\mu_0(x) - \frac{c_-}{d_+^0} \int_{u=x}^{\infty} \{f_0(u,0) + f_1(u,0)\}\, du.
\end{aligned}$$

After substituting (5.66), (5.67), (5.112) and (5.113) and tedious rewriting we obtain (5.114) and (5.115). Finally, we find κ by normalization or by substituting (5.52) into (5.118). □

At the end of this section we like to mention that a more straightforward derivation of Theorem 5.16, using the Laplace approach, is described in [58]. Although some calculations are more laborious, the analysis is conceptually the same as that in Section 4.8. This could be expected since the dual model from Chapter 4 can be regarded as the model at hand, only without feedback.

5.7 Generalisation: finite data buffer

We will shortly discuss the model in which both the credit buffer *and* the data buffer have finite sizes, K and L respectively, while the rest of the system remains unchanged, as in Section 5.2. We will only consider the case for which $L > Kd_+/c_-$, since then the analysis in Section 5.5 carries through almost identically. In that case, the form of the stationary distribution will be as follows (see Figure 5.10).

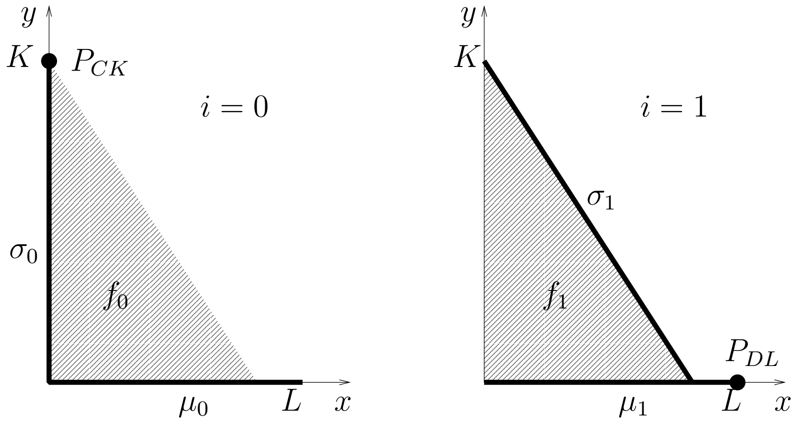


Figure 5.10: The stationary distribution for finite data buffer

$$
\begin{aligned}
F_0^{(L)}(\{0\},\{K\}) &= P_{CK}^{(L)}, && && (5.121)\\
F_i^{(L)}(dx,dy) &= f_i^{(L)}(x,y)\ dx\ dy, && (x,y)\in \mathring{S}^{(L)},\ i=0,1, && (5.122)\\
F_0^{(L)}(\{0\},dy) &= \sigma_0^{(L)}(y)\ dy, && y\in[0,K], && (5.123)\\
F_1^{(L)}(dx,K-c_-/d_+dx) &= \sigma_1^{(L)}(x)\ dx, && x\in[0,Kd_+/c_-], && (5.124)\\
F_i^{(L)}(dx,\{0\}) &= \mu_i^{(L)}(x)\ dx, && x\in[0,L),\ i=0,1. && (5.125)\\
F_1^{(L)}(\{L\},\{0\}) &= P_{DL}^{(L)}. && && (5.126)
\end{aligned}
$$

Here, the superscript is used to emphasize the dependence on L. The corresponding quantities from Section 5.5 will be written with superscript (∞) in the remainder of this section. The main result is stated in the following theorem.

Theorem 5.17 *If the size of the data buffer is $L > Kd_+/c_-$ and $\alpha > 0$, the stationary joint distribution $\mathbf{F}^{(L)}$ of the process $(M_t^{(L)}, D_t^{(L)}, C_t^{(L)})$ is of the form (5.121)–(5.126). The various probability masses and densities are given as follows.*

$$
\begin{aligned}
P_{CK}^{(L)} &= \psi\ P_{CK}^{(\infty)} && (5.127)\\
f_i^{(L)}(x,y) &= \psi\ f_i^{(\infty)}(x,y) && (5.128)\\
\sigma_0^{(L)}(y) &= \psi\ \sigma_0^{(\infty)}(y) && (5.129)\\
\sigma_1^{(L)}(x) &= \psi\ \sigma_1^{(\infty)}(x) && (5.130)\\
\mu_i^{(L)}(x) &= \psi\ \mu_i^{(\infty)}(x) && (5.131)\\
P_{DL}^{(L)} &= \psi\ \frac{d_-^0}{b}\mu_0^{(\infty)}(L), && (5.132)
\end{aligned}
$$

with

$$\psi = \left(1 - \frac{a+b}{\alpha b}\,\mu_0^{(\infty)}(L)\right)^{-1}.$$

Proof. With some modifications, the analysis in Section 5.5 can be carried out for the case of finite L. It follows that $\mathbf{F}^{(L)}$ is given in Theorem 5.9 if we superscribe P_{CK} and the densities with $^{(L)}$, and restrict x to the interval $[0, L)$ in (5.57). Therefore, if we interpret (5.127) as the definition of ψ, (5.128) – (5.131) follow immediately. It remains to find $P_{DL}^{(L)}$, which can be done by evaluating

$$P_{DL}^{(L)} = \frac{1}{ET^{(L)}} \sum_{j=0}^{1} \int_{u=0}^{Kd_+/c_-} E_{j,u} W_1^{(L)}(L)\ G_j^{(L)}(du),$$

where

$$W_1^{(L)}(L) = \int_{t=T_1^{(L)}}^{T^{(L)}} \mathbf{1}_{\{M_t^{(L)}=1, D_t^{(L)}=L\}} dt,$$

and the rest of the notation is similar to that of Section 5.5. Since

$$E_{j,u} W_1^{(L)}(L) = \frac{E_{j,u} N_1^{(L)}(x)}{b},$$

we can use Lemmas 5.11 and 5.12 to find that

$$\begin{aligned} P_{DL}^{(L)} \;&=\; \frac{e^{-\alpha L}}{b} \Big\{ c_- \int_{u=0}^{Kd_+/c_-} \Big\{ \eta_0(u) f_0^{(L)}(u,0) + \eta_1(u) f_1^{(L)}(u,0) \Big\} du \;+ \\ &\qquad\qquad \eta_1(Kd_+/c_-) \sigma_1^{(L)}(Kd_+/c_-) \Big\}. \end{aligned}$$

Using (5.62), (5.128), (5.130) and (5.131), we now find (5.132). Finally, ψ follows from the normalization condition

$$\psi \left(1 - \int_L^\infty (\mu_0^{(\infty)}(x) + \mu_1^{(\infty)}(x))\, dx \right) \;+\; P_{DL}^{(L)} \;=\; 1.$$

□

Obviously, the stationary distribution can also be shown to exist when $\alpha \leq 0$. If we set $P_{CK} = 1$, the expressions for the various densities remain valid for some normalization constant ψ (if we replace $\eta_0(u)$ in (5.66) by au/d_-^0 for $\alpha = 0$). Since Theorem 5.9 does not hold for this case, it is more difficult to find an explicit expression for this normalization constant.

Remark 5.2 In Section 5.2.3 it was shown how a tandem model with finite second reservoir follows easily from the model in Section 5.2 by choosing the parameters as in (5.5). In particular it is clear that the stationary distribution for such a model is given by $\mathbf{F}(dx, K - dy)$ with $\mathbf{F}$ as in Theorem 5.9, keeping the aforementioned parameter choice in mind.

In the same way Theorem 5.17 gives us the stationary distribution for a tandem fluid queue in which both buffers are finite, provided that the fluid rates are such that during long on-periods of the fluid source, the second buffer will be completely filled *before* the first buffer.

Bibliography

[1] S. Aalto. Characterization of the output rate process for a Markovian storage model. *J. Appl. Probab.*, 35(1):184–199, 1998.

[2] S. Aalto. Output of a multiplexer loaded by heterogeneous on-off sources. *Stochastic Models*, 14(4):993–1005, 1998.

[3] I.J.B.F. Adan, E.A. van Doorn, J.A.C. Resing, and W.R.W. Scheinhardt. Analysis of a single-server queue interacting with a fluid reservoir. *Queueing Systems*, to appear.

[4] I.J.B.F. Adan and J.A.C. Resing. Simple analysis of a fluid queue driven by an M/M/1 queue. *Queueing Systems*, 22:171–174, 1996.

[5] I.J.B.F. Adan and J.A.C. Resing. A two-level traffic shaper for an on-off source. COSOR memorandum, Department of Mathematics and Computing Science, Eindhoven University of Technology, The Netherlands, 1998. In preparation.

[6] D. Anick, D. Mitra, and M.M. Sondhi. Stochastic theory of a data-handling system with multiple sources. *Bell Syst. Tech. J.*, 61(8):1871–1894, 1982.

[7] S. Asmussen. *Applied Probability and Queues.* Wiley, New York, 1987.

[8] S. Asmussen. Busy period analysis, rare events and transient behavior in fluid flow models. *J. Appl. Math. Stoch. Anal.*, 7(3):269–299, 1994.

[9] S. Asmussen. Stationary distributions for fluid flow models with or without Brownian noise. *Stochastic Models*, 11(1):21–49, 1995.

[10] S. Asmussen and M. Bladt. A sample path approach to mean busy periods for Markov-modulated queues and fluids. *Adv. Appl. Probab.*, 26:1117–1121, 1994.

[11] S. Asmussen and R.Y. Rubinstein. Steady state rare events simulation in queueing models and its complexity properties. In J.H. Dshalalow, editor, *Advances in Queueing. Theory, Methods and Open Problems*, pages 429–461, Boca Raton, Florida, 1995. CRC Press.

[12] C. Berg. Markov's theorem revisited. *J. Approx. Theory*, 78:260–275, 1994.

[13] A.W. Berger. Performance analysis of a rate-control throttle where tokens and jobs queue. *IEEE J. Sel. Areas Commun.*, 9(2):165–170, 1991.

[14] A.W. Berger and W. Whitt. The impact of a job buffer in a token-bank rate-control throttle. *Stochastic Models*, 8:685–717, 1992.

[15] A.W. Berger and W. Whitt. The pros and cons of a job buffer in a token-bank rate-control throttle. *IEEE Trans. Commun.*, 42:857–861, 1994.

[16] S. Blaabjerg, H. Andersson, and H. Andersson. Approximating the heterogeneous fluid queue with a birth-death fluid queue. *IEEE Trans. Commun.*, 43(5):1884–1887, 1995.

[17] C.-S. Chang, P. Heidelberger, S. Juneja, and P. Shahabuddin. Effective bandwidth and fast simulation of ATM intree networks. *Performance Eval.*, 20:45–65, 1994.

[18] H. Chen and D.D. Yao. A fluid model for systems with random disruptions. *Oper. Res.*, 40(S2):S239–S247, 1992.

[19] T.S. Chihara. *An Introduction to Orthogonal Polynomials.* Gordon and Breach, New York, 1978.

[20] B.D. Choi and K.B Choi. A markov modulated fluid queueing system with strict priority. *Telecommunication Systems*, 9:79–95, 1998.

[21] G.L. Choudhury, A. Mandelbaum, M.I. Reiman, and W. Whitt. Fluid and diffusion limits for queues in slowly changing environments. *Stochastic Models*, 13(1):121–146, 1997.

[22] E.G. Coffman, Jr., B.M. Igelnik, and Y.A. Kogan. Controlled stochastic model of a communication system with multiple sources. *IEEE Trans. Inf. Theory*, 37:1379–1387, 1991.

[23] J.W. Cohen. Single server queue with uniformly bounded virtual waiting time. *J. Appl. Probab.*, 5:93–122, 1968.

[24] J.W. Cohen. Superimposed renewal processes and storage with gradual input. *Stochastic Processes Appl.*, 2:31–57, 1974.

[25] J.W. Cohen. *The Single Server Queue.* North-Holland publ. comp., Amsterdam, The Netherlands, 1982.

[26] J.N. Daigle and J.D. Langford. Models for analysis of packet voice communications systems. *IEEE J. Sel. Areas Commun.*, SAC-4(6):847–855, 1986.

[27] M.H.A. Davis. *Markov Models and Optimization.* Chapman and Hall, 1993.

[28] K. Debicki and T. Rolski. A Gaussian fluid model. *Queueing Systems*, 20:433–452, 1995.

[29] L.E.N. Delbrouck. Closed-form approximations for the means of queue size buildups and depletions in a simple ATM system. *IEEE Trans. Commun.*, 42(8):2518–2520, 1994.

[30] E.A. van Doorn. Some new results for chain-sequence polynomials. *J. Comput. Appl. Math.*, 57:309–317, 1995.

[31] E.A. van Doorn, A.A. Jagers, and J.S.J de Wit. A fluid reservoir regulated by a birth-death process. *Stochastic Models*, 4(3):457–472, 1988.

[32] E.A. van Doorn and W.R.W. Scheinhardt. Analysis of birth-death fluid queues. In B.D. Choi, editor, *Proc. KAIST Applied Mathematics Workshop*, pages 13–29, Taejon, Korea, 1996.

[33] E.A. van Doorn and W.R.W. Scheinhardt. A fluid queue driven by an infinite-state birth-death process. In V. Ramaswami and P.E. Wirth, editors, *Teletraffic Contributions for the Information Age, Proc. ITC 15*, pages 465–475, Amsterdam, 1997. Elsevier.

[34] A.I. Elwalid and D. Mitra. Analysis and design of rate-based congestion control of high-speed networks. Part I: Stochastic fluid models, access regulation. *Queueing Systems*, 9:29–64, 1991.

[35] A.I. Elwalid and D. Mitra. Fluid models for the analysis and design of statistical multiplexing with loss priorities on mutiple classes of bursty traffic. In *Proc. IEEE INFOCOM '92*, pages 415–425, 1992.

[36] A.I. Elwalid and D. Mitra. Analysis, approximations and admission control of a multi-service multiplexing system with priorities. In *Proc. IEEE INFOCOM '95*, pages 463–472, 1995.

[37] A. Erdelyi, editor. *Bateman Manuscript Project, Tables of Integral Transform.* McGraw Hill, New York, 1954.

[38] S.N. Ethier and T.G. Kurtz. *Markov Processes: Characterisation and Convergence.* Wiley, New York, 1986.

[39] D.P. Gaver and J.P. Lehoczky. Channels that cooperatively service a data stream and voice messages. *IEEE Trans. Commun.*, 30(5), 1982.

[40] G. Hjálmtýsson and A.G. Konheim. Policing and traffic shaping at the user-network-interface (UNI). *Telecommunication Systems*, 6:261–288, 1996.

[41] B. Igelnik, Y. Kogan, V. Kriman, and D. Mitra. A new computational approach for stochastic fluid models of multiplexers with heterogeneous sources. *Queueing Systems*, 20(1–2):85–116, 1995.

[42] R. Izmailov. Deterministic service of identical and independent on-off fluid sources. *Stochastic Models*, 12(2):329–342, 1996.

[43] S. Karlin and J.L. McGregor. The differential equations of birth-and-death processes, and the Stieltjes moment problem. *Trans. Amer. Math. Soc.*, 85:489–546, 1957.

[44] O. Kella. Parallel and tandem fluid networks with dependent Lévy inputs. *Ann. Appl. Probab.*, 3(3):682–695, 1993.

[45] O. Kella. Stochastic storage networks: stationarity and the feedforward case. *J. Appl. Probab.*, 34(2):498–507, 1997.

[46] O. Kella and W. Whitt. A storage model with a two-state random environment. *Oper. Res.*, 40(S2):S257–S262, 1992.

[47] O. Kella and W. Whitt. A tandem fluid network with Lévy input. In I. Basawa and U. Bhat, editors, *Queues and Related Models*, pages 112–128, Oxford, England, 1992. Oxford University Press.

[48] I. Kino and M. Miyazawa. The stationary work in system of a G/G/1 gradual input queue. *J. Appl. Probab.*, 30:207–222, 1993.

[49] H. Kobayashi and Q. Ren. A mathematical theory for transient analysis of communication networks. *IEICE Trans. Commun.*, E75-B(12):1266–1276, 1992.

[50] K. Kobayashi and Y. Takahashi. Gaussian-type variable input rate processes for ATM multiplexer. In *Performance Models for Information Communication Networks*, pages 432–443, Japan, 1995. Makuhari.

[51] K.P. Kontovasilis and N.M. Mitrou. Markov-modulated traffic with nearly complete decomposability characteristics and associated fluid queueing models. *Adv. Appl. Probab.*, 27:1144–1185, 1995.

[52] L. Kosten. Stochastic theory of a multi-entry buffer, part 1. *Delft Progress Report, Series F*, 1:10–18, 1974.

[53] L. Kosten. Stochastic theory of a multi-entry buffer, part 2. *Delft Progress Report, Series F*, 1:44–50, 1974.

[54] L. Kosten. Stochastic theory of data handling systems, with groups of multiple sources. In H. Rudin and W. Bux, editors, *Performance of Computer-communication Systems*, pages 321–331, North-Holland, 1984. Elsevier Science Publishers B.V.

[55] L. Kosten. Liquid models for a type of information buffer problems. *Delft Progress Report*, 11:71–86, 1986.

[56] L. Kosten and O.J. Vrieze. Stochastic theory of a multi-entry buffer, part 3. *Delft Progress Report, Series F*, 1:103–115, 1975.

[57] A.S. Krishnakumar and M. Morf. Eigenvalues of a symmetric tridiagonal matrix: a divide-and-concquer approach. *Numer. Math.*, 48:349–368, 1996.

[58] D.P. Kroese and W.R.W. Scheinhardt. A Markov-modulated fluid system with two interacting reservoirs. Memorandum 1365, Faculty of Mathematical Sciences, University of Twente, Enschede, The Netherlands, 1997.

[59] D.P. Kroese and W.R.W. Scheinhardt. A fluid queue driven by a fluid queue. In B. Goldstein, A. Koucheryavy, and M. Shneps-Shneppe, editors, *Teletraffic theory as a base for QOS: monitoring, evaluation, decisions*, pages 389–400, St. Petersburg, 1998.

[60] D.P. Kroese and W.R.W. Scheinhardt. A Markov-modulated fluid tandem queue and a dual system. Manuscript, 1998.

[61] V. Kulkarni. Fluid models for single buffer systems. In J.H. Dshalalow, editor, *Frontiers in queueing. Models and Applications in Science and Engineering*, pages 321–338, Boca Raton, Florida, 1997. CRC Press.

[62] V. Kulkarni and T. Rolski. Fluid model driven by an Ornstein–Uhlenbeck process. *Probab. Eng. Inf. Sci.*, 8:403–417, 1994.

[63] K.K. Leung, B. Sengupta, and R.W. Yeung. A credit manager for traffic regulation in high-speed networks: a queueing analysis. *IEEE/ACM Trans. Netw.*, 1(2):236–245, 1993.

[64] B. Li, A.G. Law, H. Raafat, P.H. Nguyen, and Y.F. Yan. Eigenvalues of tridiagonal matrices: an alternative to Givens' method. *Computers Math. Appl.*, 19:89–94, 1990.

[65] R.S. Liptser and A.N. Shiryayev. *Statistics of Random Processes II: Applications.* Springer, New York, 1978.

[66] Z. Liu and D. Towsley. Burst reduction properties of rate-control throttles: downstream queue behaviour. *IEEE/ACM Trans. Netw.*, 3:82–90, 1995.

[67] M.R.H. Mandjes. Asymptotically optimal importance sampling for tandem queues with Markov fluid input. *AEÜ Int. J. Electr. Commun.*, 52(3):152–161, 1998.

[68] B.L. Mark and G. Ramamurthy. UPC based traffic descriptors for ATM: how to determine, interpret and use them. *Telecommunication Systems*, 5:109–122, 1996.

[69] D. Mitra. Stochastic theory of a fluid model of producers and consumers coupled by a buffer. *Adv. Appl. Probab.*, 20:646–676, 1988.

[70] M. Miyazawa. Rate conservation laws: a survey. *Queueing Systems*, 15:1–58, 1994.

[71] I. Norros, J.W. Roberts, A. Simonian, and J.T. Virtamo. The superposition of variable bit rate sources in an ATM multiplexer. *IEEE J. Sel. Areas Commun.*, 9(3):378–387, 1991.

[72] F.W.J. Olver. *Asymptotics and Special Functions.* Academic Press, New York, 1974.

[73] T.J. Ott and D.J. Daley. On a fluid flow model for a packet switch. Unpublished manuscript, Bellcore Communications Research, Morristown, NJ, 1990.

[74] A. Pacheco and N.U. Prabhu. A Markovian storage model. Technical report 1079, School of operations research and industrial engineering, Cornell University, Ithaca, NY, 1993.

[75] A. Pacheco and N.U. Prabhu. Markov-additive processes of arrivals. In J.H. Dshalalow, editor, *Advances in Queueing. Theory, Methods and Open Problems*, pages 167–194, Boca Raton, Florida, 1995. CRC Press.

[76] H. Pan, H. Okazaki, and I. Kino. Analysis of a gradual input model for bursty traffic in ATM. In *Proc. ITC 13*, pages 795–800, 1991.

[77] B.V. Patel and C.C. Bisdikian. On the performance behavior of ATM end-stations. In *Proc. IEEE INFOCOM '95*, pages 188–196, 1995.

[78] R.J. Pilc. A derivation of buffer occupancy statistics in an asynchronous time-division multiplexer used with bursty sources. Internal report, Bell Labs., 1968.

[79] M.A. Pinsky. *Lectures on Random Evolution.* World Scientific, Singapore, 1991.

[80] N.U. Prabhu. *Stochastic Storage Processes.* Springer Verlag, New York, 1980.

[81] G.J.K. Regterschot. *Wiener-Hopf Factorization Techniques in Queueing Models.* PhD thesis, University of Twente, Enschede, The Netherlands, 1987.

[82] M. Ritter. Performance analysis of the dual cell spacer in ATM systems. In *IFIP 6th International Conference on High Performance Networking*, page 510 ff., 1995.

[83] J.W. Roberts, U. Mocci, and J.T. Virtamo, editors. *Broadband Network Teletraffic – Final Report of Action COST 242.* Springer, Berlin, 1996.

[84] L.C.G. Rogers. Fluid models in queueing theory and Wiener-Hopf factorization of Markov chains. *Ann. Appl. Probab.*, 4(2):390–413, 1994.

[85] L.C.G. Rogers and Z. Shi. Computing the invariant law of a fluid model. *J. Appl. Probab.*, 31:885–896, 1994.

[86] G. Sansigre and G. Valent. A large family of semi-classical polynomials: the perturbed Tchebichev. *J. Comput. Appl. Math.*, 57:271–281, 1995.

[87] W.R.W. Scheinhardt and D.P. Kroese. A system of fluid queues with feedback. Memorandum 1422, Faculty of Mathematical Sciences, University of Twente, Enschede, The Netherlands, 1997.

[88] M. Sidi, W.-Z. Liu, I. Cidon, and I. Gopal. Congestion control through input rate regulation. *IEEE Trans. Commun.*, 41:471–477, 1993.

[89] K. Sigman and G. Yamazaki. Fluid models with burst arrivals: a sample path analysis. *Probab. Eng. Inf. Sci.*, 6:17–28, 1992.

[90] A. Simonian. Stationary analysis of a fluid queue with input rate varying as an Ornstein-Uhlenbeck process. *SIAM J. Appl. Math.*, 51(3):828–842, 1991.

[91] A. Simonian and J.T. Virtamo. Transient and stationary distributions for fluid queues and input processes with a density. *SIAM J. Appl. Math.*, 51(6):1732–1739, 1991.

[92] P. Sonneveld. Some properties of the generalized eigenvalue problem $Mx = \lambda(\Gamma - cI)x$, where M is the infinitesimal generator of a Markov process, and Γ is a real diagonal matrix. Preliminary report, Delft University of Technology, Delft, 1988.

[93] T.E. Stern and A. I. Elwalid. Analysis of separable Markov-modulated rate models for information-handling systems. *Adv. Appl. Probab.*, 23:105–139, 1991.

[94] T. Takine, B. Sengupta, and T. Hasegawa. A conformance measure for queues. *Stochastic Models*, 11(4):645–670, 1995.

[95] T. Tanaka, O. Hashida, and Y. Takahashi. Transient analysis of fluid model for ATM statistical multiplexer. *Performance Eval.*, 23:145–162, 1995.

[96] R.C.F. Tucker. Accurate method for analysis of a packet-speech multiplexer with limited delay. *IEEE Trans. Commun.*, 36(4):479–483, 1988.

[97] J.T. Virtamo and I. Norros. Fluid queue driven by an M/M/1 queue. *Queueing Systems*, 16:373–386, 1994.

[98] J. Wijngaard. The effect of interstage buffer storage on the output of two unreliable production units in series, with different production rates. *AIIE Trans.*, 11(1):42–47, 1979.

[99] J. Zhang. Transient solution of continuous flow Markov modulated rate models. In *Proc. 26th Annual Conference on Information Sciences and Systems*, Princeton, N.J., 1992.

[100] J. Zhang. Performance study of Markov modulated fluid flow models with priority traffic. In *Proc. IEEE INFOCOM '93*, pages 10–17, 1993.

[101] A.P. Zwart. Asymptotics for the probability distribution of the buffer content of a fluid queue driven by an M/M/1 queue. Personal communication, CWI, Amsterdam, 1998.

Index

Samenvatting

De afgelopen twintig jaar hebben *Markov-modulated fluid queues* (Markov-gemoduleerde vloeistof-wachtsystemen) veel aandacht gehad. In deze modellen stroomt vloeistof in en/of uit een reservoir met een snelheid die wordt bepaald door de huidige toestand van een achterliggend Markov-proces. In het eerste hoofdstuk van dit proefschrift geven we een korte inleiding op hoe de stationaire verdeling voor een dergelijk model gewoonlijk wordt gevonden, evenals een literatuuroverzicht over Markov-gemoduleerde en aanverwante vloeistofmodellen. De rest van het proefschrift behandelt de vraag hoe de stationaire verdeling kan worden gevonden voor sommige vloeistofmodellen die tot nu toe weinig of geen aandacht hebben gehad. De twee belangrijkste bijdragen zijn de volgende.

1. We richten ons met name op modellen waarin de toestandsruimte van het regulerende Markov-proces oneindig groot is, al dan niet aftelbaar. In het aftelbare geval kijken we voornamelijk naar regulerende processen die een geboorte-sterfte structuur hebben. We geven procedures om de stationaire verdeling te vinden met behulp van orthogonale polynomen. In het overaftelbare geval kijken we naar eenvoudige systemen van fluid queues, waarin één fluid queue het gedrag van een tweede reguleert. Een voorbeeld van zo'n systeem is een fluid tandem queue.

2. We beschouwen modellen waarin de toestand van het vloeistofreservoir het gedrag van het regulerende proces beïnvloedt, zodat dit laatste geen Markov-proces is. We noemen dergelijk systemen *feedback fluid queues* (vloeistof-wachtsystemen met terugkoppeling), om de wederzijdse afhankelijkheid te benadrukken tussen het vloeistofreservoir en het regulerende proces.

Over de toegepaste technieken waarmee de stationaire verdeling wordt bepaald in de diverse modellen merken we op dat naast de spectraal-expansie methode, die algemeen wordt gebruikt in dit verband, ook Laplace-transformatie technieken zijn gebruikt. Tevens is gebruik gemaakt van de relatie met traditionele wachtsystemen van $M/G/1$ of $G/M/1$ type. In twee gevallen is de inhoud van het vloeistofreservoir (resp. één van de reservoirs) gediscretiseerd, wat leidt tot benaderingen en tussenresultaten.

De reden dat Markov-gemoduleerde vloeistof-wachtsystemen zo populair zijn geworden is dat ze de voornaamste eigenschappen van veel situaties in telecommunicatiesystemen goed kunnen beschrijven. Hoewel dit proefschrift vooral theoretisch van aard is, is er ook enige aandacht voor het praktisch nut van de erin behandelde modellen. We laten met name zien dat feedback-vloeistofmodellen relevant zijn voor het modelleren van “two-level traffic shapers” die in het kader van zogenaamde ATM-telecommunicatienetwerken zijn voorgesteld als verkeers-regulator.

Summary

In the last twenty years the field of *Markov-modulated fluid queues* has received considerable attention. In these models a fluid reservoir receives and/or releases fluid at rates which depend on the actual state of a background Markov chain. In the first chapter of this thesis we give a short introduction on how the stationary distribution for such a model is usually found, as well as a literature overview on Markov-modulated and related fluid queues. The rest of the thesis is concerned with finding stationary distributions for some types of fluid models that have received little or no attention until now. The two main contributions are the following.

1. We focus on models in which the state space of the regulating Markov process is infinitely large, either denumerable or not. Regarding the first type, we mainly look into regulating processes that are of birth-death type. We present procedures to find the stationary distribution, using the theory of orthogonal polynomials. In the nondenumerable case, we look into simple systems of fluid queues, in which one fluid queue regulates the behaviour of another (one example being a fluid tandem queue).

2. We look into models in which the state of the fluid reservoir influences the behaviour of the regulating process, so that the latter does not constitute a Markov process. We call suchlike systems *feedback fluid queues*, to emphasize the two-way dependence between fluid reservoir and regulating process.

With respect to the techniques employed to obtain the stationary distribution in the various models, we mention that apart from the spectral expansion method, which has been widely used in this context, we also apply Laplace-transform techniques, and exploit the connection with traditional queueing systems of $M/G/1$ or $G/M/1$ type. In two cases we discretize the content of (one of) the fluid reservoir(s), leading to approximations and intermediate results.

The reason that Markov-modulated fluid queues have become so popular is that they can capture the basic characteristics of many situations in telecommunication systems. Although this thesis concentrates mainly on theoretical aspects, some attention has gone to the practical use of the models discussed in it. In particular, feedback fluid models are shown to be relevant in modeling two-level traffic shapers which have been proposed in the context of so-called ATM telecommunication networks.

Over de auteur

Werner Scheinhardt werd geboren op 24 februari 1969 in Santiago, Chili. Na op éénjarige leeftijd zijn eerste voet(je) op vaderlandse bodem te hebben gezet, beleefde hij zijn jeugd in Eindhoven, Son, Scheveningen, Moordrecht en Wassenaar. Hij bezocht de middelbare school op het St. Adelbert College in Wassenaar en, de laatste twee jaar, op de Winschoter Scholengemeenschap, waar hij zijn V.W.O. diploma ontving in 1987. Hij vervolgde zijn opleiding aan de Universiteit Twente als student Toegepaste Wiskunde en behaalde zijn ingenieursdiploma in 1994 met als afstudeeronderwerp de "Convergentiesnelheid van eindige Markov-ketens in discrete tijd". Hij besloot als AIO in Twente te blijven om het onderzoek uit te voeren dat in dit proefschrift is beschreven. De komende tijd zal hij als postdoc deel uitmaken van de onderzoeksgroep van Professor Boxma aan de Technische Universiteit Eindhoven.

About the author

Werner Scheinhardt was born on the 24th of Februari, 1969 in Santiago, Chili. After the arrival in his mother country (the Netherlands) at the age of one, he spent his childhood in the south and west parts of it. He attended secondary school at the St. Adelbert College in Wassenaar, and, for the last two years, at the Winschoter Scholengemeenschap, where he received his diploma in 1987. He continued his education as a student at the faculty of Applied Mathematics (as it was called then) of the University of Twente and received his masters degree in 1994, the title of his thesis being "Rate of convergence of finite Markov chains in discrete time". He decided to stay in Twente as a PhD student to carry out the research that is described in this dissertation. During the upcoming one and a half years he will be a postdoctoral fellow in the research group of Professor Boxma at the Eindhoven University of Technology.