Walter Peter Sendfeld

Two-dimensional Overflow Queueing Systems

July 2009

Doctoral Thesis
Institute for Mathematics
Department of Mathematics/Computer Science
University of Osnabrück

Angewandte Mathematik

Two-dimensional Overflow Queueing Systems

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften im Fachbereich
Mathematik/Informatik
der Universität Osnabrück

vorgelegt von
Walter Peter Sendfeld
aus Gronau
im Juli 2009

Dekan:	Prof. Dr. Heinz Spindler
Betreuer:	Prof. Dr. Wolfgang Stadje

Contents

Introduction

> *Well informed people know that it is impossible*
> *to transmit the human voice over wires.*
> – Boston Newspaper, 1867

Following Leonard Kleinrock [37], *queueing theory* is the study of the phenomena of standing, waiting, and serving. As simple as this definition might be, as manifold are the applications and real world problems that can be described by queueing models. The queueing theory started at the early beginning of the 20th century with the pioneer work of Frederik Ferdinand Wilhelm Johannsen [33], Tore Olaus Engset [21,22] and Agner Krarup Erlang [23] as the study of telephone networks with limited capacity. Numerous refinements and new applications of queueing theory have arisen since then. Nowadays, the rich and fertile theory is applied to the analysis of communication networks and computer systems for internet and data traffic or bandwidth management, to health care systems, traffic control, insurance mathematics, machine plants and almost every area of everyday life in which "standing, waiting, and serving" takes precedence. Although the work of Johannsen, Engset and Erlang might seem old-fashioned from todays point of view – and impossible if one believes a Boston newspaper from 1867 –, their results are 100 years after the publication of Erlang's first paper still perfectly applicable and applied in modern teletraffic engineering (see Stordahl [61] and ITU [32]).

Whenever a queueing system offers only a finite number of service and/or waiting positions, some service demands might be declined and some not. From an economic point of view, it might be useful to know the system's long-run behavior in order to balance for example the cost of lost demands and the benefits from service. Quantities like the overall blocking or overflow probability, the average departure rate from the waiting room and the servers and the average occupation proportion of the waiting and service positions are amongst others of special interest. A system designer for example has to know these characteristics in order to control and optimize a queueing system and to reach a certain cost or utilization level. However, these characteristics can only be calculated for a limited class of queueing systems and the more involved the system dynamics get, the more involved the analysis of the long run behavior usually becomes.

In this thesis, we present two fairly general classes of so called *overflow queueing networks*. These networks consist of two queues, where the capacity of the first queue is always finite. Customers arriving at the first queue have an overflow capability from the first to the second queue if the first queue operates at a certain fixed capacity, i.e., under certain conditions, demands arriving at the first queue are allowed to join the second queue. In every model, the dynamic of the first queue is or is at least similar to the famous Erlang and Engset loss systems. The overflow stream will additionally be weighted with a parameter $p \in [0, 1]$. The parameter p can be used as a control parameter or to model the customers' impatience.

The first chapter gives a brief overview of the general stochastic structure underlying these networks. In principle, each of the queues is fed by a Markovian Poisson process and the service times are exponential.

In the second chapter, we consider a generalization of a queueing model presented by Perel and Yechiali [56] and additionally append an overflow capability. In this generalized two-queue network, the arrival and service rates for the finite first queue are state-dependent and the customers in the first queue act as servers for the second queue, i.e., the service rate in the second queue depends on the state of the first queue. This is for example the case for file sharing or torrent systems, where customers receive data from customers that are already in service. The state-dependent rates in the first queue cover many prominent queueing systems. The first queue is considered to have finite capacity, the second queue has an infinite capacity. We further consider two variants of this model by allowing customers to jockey from the second queue to the first queue. In queueing theory, *jockeying* is called the possibility for waiting customers to move from one queue to another queue. These jockeying customers can then act as servers for the second queue. We reduce the number of unknown steady-state probabilities of this system in a considerable amount by a generating functions approach due to Avi-Itzhak and Mitrani [6]. Some steady-state quantities of interest are also derived.

In the third chapter, we cover a variety of different models of two-queue networks in which we equip each of the two queues with a finite number of servers and waiting positions. We consider different routines for the handling of arriving, blocked, overflowing and jockeying customers. These models can be used to analyze for example call centers, telecommunication systems or traffic flows. The main difference between the queueing models presented in the second and third chapter is the finiteness of the capacity of the second queue. Moreover, the customers in the first queue no longer serve the customers in the second queue. The finiteness of the state space of these queueing systems gives – in contrast to the models presented in the second chapter – rise to additional boundary conditions. These boundary conditions and the special structure of the steady-state equations make it impossible to carry out the approach from the second chapter. Nevertheless, the number of steady-state equations that describe the system's behavior can be reduced substantially by exploiting a separation method due to Morrison [46, 48]. By using this separation technique, we will reduce the problem of solving the steady-state equations to the problem of solving a substantially smaller number of homogeneous linear equations in two sets of unknowns. With this approach, explicit formulas depending on these unknowns can be given for various steady-state quantities in an elegant form. The basic technique is to partition the state space into certain regions and boundaries and to separate the stationary probabilities within these regions. In every model, the separation leads to a set of eigenvalue problems for the separation constants. The eigenvalues are given by the roots of polynomial equations and are the pairwise distinct eigenvalues of real tridiagonal symmetric matrices as well. They possess an interlacing property, called the "Sturm sequence property", which reduces the computational complexity considerably. The desired probabilities are expressed as sums of eigenfunctions in terms of the eigenvalues. The number of eigenfunctions and therefore the number of coefficients to be determined in these representations is in general substantially smaller than the number of stationary probabilities. The coefficients are determined by the normalization condition and a set of linear equations that stems from the boundary conditions. The desired probabilities and steady-state quantities can be numerically determined once the coefficients and eigenvalues are numerically calculated. Some of the results were published in Sendfeld [59].

We give a detailed model and literature review and name main applications of our models at the beginning of every chapter. Due to the natural occurrence of overflow queueing problems, the related literature is vast, see for example Disney and König [17] for a broad overview. Two chapters devoted to the theory of queueing networks with restricted accessibility and overflow are found in Syski [63]. In Kosten [38], some aspects of networks with restricted accessibility are considered. Additionally, Koury et al. [39]

and Krieger et al. [40] give reviews of iterative numerical methods for overflow queueing models. A brief discussion of numerical methods for some two-queue overflow systems and further references are given in Ching and Ng [10].

Introduction (german)

Wohlunterrichete Menschen wissen, dass es unmöglich ist,
die menschliche Stimme mit Kabeln zu übertragen.
Bostoner Tageszeitung, 1867

Nach Leonard Kleinrock [37] ist die *Warteschlangentheorie* das Studium der Phänomene des Stehens, Wartens und Bedienens. So einfach diese Definition auch sein mag, so mannigfaltig sind die Anwendungen und realen Probleme, die als Warteschlangenmodelle beschrieben werden können. Die Grundsteine der Warteschlangentheorie wurden Anfang des frühen 20. Jahrhunderts gelegt. Die Pionierarbeiten von Frederik Ferdinand Wilhelm Johannsen [33], Tore Olaus Engset [21, 22] und Agner Krarup Erlang [23] befassten sich mit dem Studium der Telefonnetzwerke mit begrenzten Kapazitäten. Seither entstehen unzählige Verallgemeinerungen und neue Anwendungen der Warteschlangentheorie. Noch heute zählen die Kommunikations- und Netzwerktechnik bei der Analyse von Telefon-, Internet- und Datenverkehr zu den wichtigsten Anwendungsbereichen der Warteschlangentheorie. Auch im Bandbreitenmanagement, im Gesundheitssystem, der Verkehrsregulierung, in der Versicherungsmathematik und bei der Analyse von Kundenströmen - also in fast jedem Bereich des alltäglichen Lebens, in dem „Stehen, Warten und Bedienen“ eine Rolle spielen - ist die Warteschlangentheorie unerlässlich. Obwohl die Arbeiten von Johannsen, Engset und Erlang aus heutiger Sicht altmodisch erscheinen mögen - sogar nutzlos, sofern man einer Bostoner Zeitung von 1867 glaubt -, so sind deren Resultate auch 100 Jahre nach der Veröffentlichung von Erlangs erster Arbeit anwendbar und werden nach wie vor in der modernen Telekommunikation angewendet (siehe Stordahl [61] und ITU [32]).

Verfügt ein Warteschlangensystem nur über eine begrenzte Kapazität an Service- oder Warteplätzen, so kann dies dazu führen, dass nur ein Teil der Serviceanfragen erfüllt wird. Aus ökonomischer Sicht ist es dann sinnvoll, das Langzeitverhalten des Systems zu studieren, um zum Beispiel die Kosten abgelehnter Anfragen und die Einnahmen angenommener Anfragen auszubalancieren. In dieser Hinsicht sind unter anderem die erwartete Blockierwahrscheinlichkeit, die durchschnittlichen Abgangsraten von Kunden aus den Warteräumen oder Servern und die durchschnittliche Anzahl besetzter Warteplätze und Server von besonderem Interesse. Sind diese Größen unter gegebenen Voraussetzungen bekannt, so können Sie genutzt werden, um ein neues System optimal zu planen oder ein bestehendes zu optimieren. Hierbei können Zielvorgaben, wie die Sollauslastung oder die Einhaltung einer Kostenobergrenze, durch entsprechende Wahl der beeinflussbaren Systemparameter erfüllt werden. Die zur optimalen Kontrolle eines solchen Systems notwendigen Größen können jedoch im Allgemeinen nur für eine begrenzte Klasse von Warteschlangensystemen effizient berechnet werden. Je komplizierter die Abhängikeiten und Kundenströme in einem System sind, desto aufwändiger - wenn nicht unmöglich - ist dessen Analyse.

In dieser Arbeit präsentieren wir zwei allgemeine Klassen sogenannter *Warteschlangennetzwerke mit Overflow*. Die Netzwerke in diesen Klassen bestehen jeweils aus zwei Warteschlangen. Die Kapazität der ersten Warteschlange ist stets endlich. In den betrachteten Modellen entspricht die Dynamik der ersten Warteschlange den bekannten Erlang- oder Engset-Verlustsystemen oder ist diesen ähnlich. Wie in der Warteschlangentheorie üblich, werden wir im Folgenden den Begriff *Kunde* synonym für Anfrage und den Begriff *Server* für die Bedieneinheit verwenden. Ist die erste Warteschlange bis zu einer festgelegten Kapazitätsauslastung belegt, werden ankommenden Kunden bestimmte Wechselmöglichkeiten (*Overflow*) zur zweiten Warteschlange eingeräumt. Ein Kundenwechsel findet in den betrachteten Netzwerken zusätzlich mit einer Wahrscheinlichkeit $p \in [0, 1]$ tatsäschlich statt, das heißt, der Wechselstrom wird mit dem Parameter p gewichtet. Der Parameter p kann als Steuer- oder Kontrollparameter verwendet werden oder den Grad der Ungeduld der wechselnden Kunden beschreiben.

Das erste Kapitel gibt einen kurzen Überblick über die allgemeine stochastische Struktur, die den betrachteten Netzwerken zu Grunde liegt. Prinzipiell verfügt jede der Warteschlangen über einen markovschen Poisson-Ankunftsprozess und exponentialverteilte Servicezeiten.

Im zweiten Kapitel betrachten wir eine starke Verallgemeinerung eines Warteschlangenmodells von Perel und Yechiali [56], indem wir unter anderem Overflow zulassen. Bei dieser Verallgemeinerung eines Netzwerks aus zwei Warteschlangen sind die Ankunfts- und Serviceraten in der ersten Warteschlange variabel und abhängig vom Zustand dieser Warteschlange. Ferner werden die Kunden der zweiten Warteschlange von den Kunden der ersten bedient, so dass die Servicerate der zweiten Warteschlange ebenfalls vom Zustand der ersten abhängt. Dies ist zum Beispiel der Fall bei Filesharing- oder Torrent-Systemen, in denen Kunden Daten von anderen Kunden herunterladen, die bereits Service erhalten. Die Variabilität der Ankunfts- und Serviceraten in der ersten Warteschlange ermöglicht die Analyse vieler prominenter Warteschlangensysteme. Die erste Warteschlange besitzt stets endliche Kapazität, die zweite hingegen unendliche Kapazität. Wir betrachten ferner zwei Varianten dieses Modells, in denen wir Kunden, die sich im Warteraum der zweiten Warteschlange befinden, die Möglichkeit geben, zur ersten Warteschlange zu wechseln. Diese Kunden wiederum können dann in der ersten Warteschlange als Server für die zweite Warteschlange fungieren. Wir reduzieren die Anzahl der unbekannten stationären Wahrscheinlichkeiten drastisch, indem wir die Struktur der erzeugenden Funktionen dieser Wahrscheinlichkeiten mit Hilfe eines Ansatzes von Avi-Itzhak und Mitrani [6] analysieren. Ferner leiten wir Formeln für einige der wichtigsten stationären Größen her.

Im dritten Kapitel untersuchen wir zahlreiche unterschiedliche Warteschlangennetzwerke aus zwei Warteschlangen, in denen jede der beiden Warteschlangen mit einem Warteraum mit endlicher Kapazität und einer begrenzten Anzahl an Servern ausstattet ist. Wir betrachten unterschiedliche Routinen für die Behandlung ankommender, blockierter und wechselnder Kunden. Diese Modelle können verwendet werden, um zum Beispiel Call Center, Telekommunikationssysteme oder Verkehrsflüsse zu analysieren. Der Hauptunterschied zwischen den Modellen im zweiten und dritten Kapitel ist daher die nun endliche Kapazität der zweiten Warteschlange. Ferner werden die Kunden in der zweiten Warteschlange nicht mehr von den Kunden in der ersten Warteschlange bedient. Im Gegensatz zu den Modellen aus dem zweiten Kapitel gibt die endliche Kapazität der zweiten Warteschlange hier Anlass zu zusätzlichen Randbedingungen. Diese Randbedingungen und die spezielle Struktur der Gleichgewichtsgleichungen machen es unmöglich, diese Systeme mit den Methoden aus dem zweiten Kapitel zu analysieren. Jedoch reduzieren wir die Anzahl der unbekannten stationären Wahrscheinlichkeiten wiederum drastisch, indem wir eine Separations-Technik von Morrison [46, 48] verwenden und verallgemeinern. Mit Hilfe dieser Technik reduzieren wir das Problem der Lösung der Gleichgewichtsgleichung auf das Problem der Lösung einer erheblich kleineren Anzahl an homogenen Gleichungen und eines Eigenwertproblems. Mit dieser Methode können explizite Formeln für die verschiedensten stationären Größen in Abhängigkeit von diesen Eigenwerten in eleganter Form angeben werden. Die grundlegende

Technik basiert auf einer Partition des Zustandsraums in Mengen bestimmter innerer Punkte und Randpunkte. In den inneren Bereichen werden die stationären Wahrscheinlichkeiten dann in eine Summen-Produkt-Form zerlegt. Diese Zerlegung führt zu Eigenwertproblemen für tridiagonale symmetrische Matrizen, zu deren Lösung die stationären Gleichungen in den Randpunkten verwendet werden. Die Eigenwerte werden dabei als Nullstellen polynomialer Gleichungen bestimmt und sind paarweise verschieden und reell. Sie besitzen die sogenannte *Sturm-Folgen-* oder *Verzahnungseigenschaft*, die den numerischen Aufwand bei deren Berechnung erheblich reduziert. Die gesuchten Wahrscheinlichkeiten werden als gewichtete Summen der Eigenfunktionen in Abhängigkeit der Eigenwerte dargestellt. Die Koeffizienten in dieser Darstellung werden mit Hilfe der Normierungsbedingung und der linearen Randbedingungen bestimmt. Die gesuchten Wahrscheinlichkeiten können mit numerischen Verfahren berechnet werden, sobald die Koeffizienten und Eigenwerte berechnet worden sind. Einige der Resultate wurden in Sendfeld [59] veröffentlicht.

Wir geben zu Beginn jedes Kapitels eine detaillierte Literaturübersicht und Anwendungsbeispiele für die präsentierten Modelle. Aufgrund des natürlichen Auftretens von Warteschlangennetzwerken mit Wechselmöglichkeiten ist die verwandte Literatur reichhaltig. Eine breite Übersicht ist zum Beispiel in Disney und König [17] gegeben. Syski widmet in [63] zwei Kapitel der Analyse von Warteschlangennetzwerken mit begrenzter Kapazität und Wechselmöglichkeiten. Weitere Aspekte von Netzwerken mit begrenzter Kapazität werden in Kosten [38] betrachtet. Koury et al. [39] und Krieger et al. [40] geben Überblicke über iterative numerische Methoden für Warteschlangen mit Wechselmöglichkeiten. Eine kurze Diskussion numerischer Methoden für Zwei-Server-Warteschlangen mit Wechselmöglichkeiten und weitere Referenzen finden sich in Ching und Ng [10].

Acknowledgements

There are many people who influenced my personal and academic life in every positive way, especially during the past four years of my studies in Osnabrück. Without the belief, patience and support of these people, the writing of this dissertation would not have been possible.

I am indebted to the Department of Mathematics/Computer Science, in particular the Institute for Mathematics, for leaving nothing to be desired and for giving me the opportunity to realize my wishes.

I especially and deeply thank my advisor Prof. Dr. Wolfgang Stadje. He gave me the freedom to explore on my own and recovered me whenever my studies faltered.

It is an honor for me to thank Prof. Dr. Uri Yechiali for his immediate willingness to review this dissertation and to incur the related efforts. I would like to thank Prof. Dr. David Perry for the time I could spend in Haifa and for the support and warm words he gave me when times where uncertain. I would like to express my gratitude to my teachers and mentors, Dr. Wolfgang Schulte Ladbeck and Prof. Dr. Lothar Rogge, who encouraged me to be persistent and who supported me when I needed their personal and mathematical advice. Additionally, I would like to thank Dr. Christian Strotmann for the hearty welcome to Osnabrück and Dr. Achim Wübker for the valueable mathematical and non-mathematical discussions.

I deeply thank Dr. Christian Wahle for being a friend, his wonderful wife Stefanie for bearing our peculiarities and both for the many precious moments we have shared.

At last, and surely most of all, I owe a dept of gratitude to my family – Winfried, Marlis, Michael and Eva-Maria – for their deep encouragement and unconditional love. Finally, I would like to thank Maike for showing me the beauty of life besides mathematics.

This thesis is dedicated to the most influencing persons of my life: To Marlis, Winfried and Maike.

Chapter 1

Preliminaries

This chapter serves as a brief introduction to the theory of Markov jump processes in continuous time in order to classify the stochastic models presented in this thesis. The well-known classical results on general Markov chains in continuous time in this chapter can be found in various textbooks and are taken from Alsmeyer [3], Asmussen [5] and Bremaud [7], whereas the results on quasi birth and death chains are due to Latouche and Ramaswami [41] and Neuts [53].

1.1 Markov chains in continuous time

In this section, we recapitulate the definition and the basic structure of a continuous time Markov process or Markov jump process on a countable state space. Let $\mathfrak{S}$ be a countable nonempty set, called the *state space*, and let $X = (X_t)_{t\in[0,\infty)}$ be a stochastic process on the probability space $(\Omega, \mathfrak{A}, P)$ with values in $\mathfrak{S}$. Let $\mathscr{F} = (\mathscr{F}_t)_{t\in[0,\infty)}$ be the canonical filtration of X, that is, $\mathscr{F}_t$ is the σ-algebra generated by all X_s for $s \leq t$. The process X is called *Markov process* if the *Markov property* holds:

$$P(X_t \in A \mid \mathscr{F}_s) = P(X_t \in A \mid X_s) \quad P\text{-a.s.}$$

for all $s, t \in [0, \infty)$ with $s < t$ and every $A \subset \mathfrak{S}$. The process is called *homogeneous* if the transition kernels $\mathbb{P}_{s,t}(X_s, A) = P(X_t \in A \mid X_s)$ can be chosen so as to depend on s and t only through the difference $t - s$, i.e., if

$$P(X_t \in A \mid X_s = x) = \mathbb{P}_{0,t-s}(x, A) \quad P^{X_s}\text{-a.s.}$$

This will be assumed in the following. We will write $\mathbb{P}_{0,t} = \mathbb{P}_t$, where $\mathbb{P}_0(x, \cdot)$ is the Dirac measure δ_x in x, and call $\mathbb{P}_t$ the *t-step transition kernel*. The family $(\mathbb{P}_t)_{t\in[0,\infty)}$ of the t-step transition kernels satisfies the *Kolmogorov-Chapman equations*:

$$\mathbb{P}_{s+t} = \mathbb{P}_s \mathbb{P}_t \tag{1.1.1}$$

or equivalently

$$\mathbb{P}_{s+t}(x, A) = \int\limits_{\mathfrak{S}} \mathbb{P}_t(y, A)\mathbb{P}_s(x, dy)$$

for all $s, t \in [0, \infty)$ and every $A \subset \mathfrak{S}$.

In our setting with denumerable state space $\mathfrak{S}$, we may identify the transition kernel $\mathbb{P}_t$ with the matrix $\boldsymbol{P}(t) = (p_{i,j}(t))_{i,j\in\mathfrak{S}}$ and interpret the Kolmogorov-Chapman equations in the sense of matrix multiplication. We have

(i) $\boldsymbol{P}(t)$ is a stochastic matrix,

(ii) $\boldsymbol{P}(0)$ is the identity matrix and

(iii) $\boldsymbol{P}(t + s) = \boldsymbol{P}(t)\boldsymbol{P}(s)$ for all $s, t \in [0, \infty)$.

Moreover, we suppose that $\lim_{t\to 0} \boldsymbol{P}(t) = \boldsymbol{P}(0)$. With this assumption we can show that $p_{i,j}(t)$ is continuously differentiable for $t > 0$ and differentiable from the right at 0, i.e., the limit

$$q_{i,j} = \lim_{t\to\infty} \frac{p_{i,j}(t) - p_{i,j}(0)}{t}$$

exists and is finite for $i \neq j$ but maybe infinite for $i = j$. We also have $q_{i,j} \geq 0$ for all $i \neq j$. The matrix $Q = (q_{i,j})_{i,j\in\mathfrak{S}}$ is called the *infinitesimal generator*, *rate matrix* or *Q-matrix* of the process X. Q is called *conservative* if

$$\sum_{j\neq i} q_{i,j} = -q_{i,i} < \infty$$

or equivalently, in matrix form, if $Q\mathbf{1} = 0$ holds, where $\mathbf{1}$ is the vector with all entries equal to 1. It is useful to let $q_i = -q_{ii}$ for $i \in \mathfrak{S}$. We assume in the following that Q is conservative.

The process X has a fundamental jump structure. Let $S_0 = 0 < S_1 < S_2 < \ldots$ be the times of successive jumps of X, let $T_n = S_{n+1} - S_n$ for $n \geq 0$ be the associated sojourn times and let the sequence of states visited be given by $Y_n = X_{S_n}$ for $n \geq 0$. There are two phenomena which desire further attention: absorption and explosion. The process may be absorbed in the sense that there is a last finite S_n. In this case we may set $T_k = \infty$ and $Y_k = X_{S_n}$ for $k \geq n$. The case of an explosion of the process, i.e., an accumulation of infinitely many jumps in finite time is more involved. The following condition for a conservative Q-matrix prevents the process X from being explosive.

Proposition 1.1.1. *Let Q be conservative. Then X is nonexplosive if the condition $\sup_{i \in \mathfrak{S}} q_i < \infty$ holds. That is for example the case if $\mathfrak{S}$ is finite.*

The basic structure of X up to the time of explosion is very simple (see Theorem II 1.2 in [5]):

Proposition 1.1.2. *The joint distribution of the sequences $(Y_n)_{n \geq 0}$ and $(T_n)_{n \geq 0}$ before explosion is given by:*

(i) The sequence $(Y_n)_{n \geq 0}$ of states visited is a Markov chain.

(ii) There exist $\lambda(i) \geq 0$ such that $T_0, T_1, \ldots$ are independent and T_k is exponentially distributed with parameter $\lambda(Y_k)$ for $k \geq 0$ given $(Y_n)_{n \geq 0}$.

The transition matrix $\hat{P} = (\hat{p}_{i,j})_{i,j \in \mathfrak{S}}$ of the embedded Markov chain $(Y_n)_{n \geq 0}$ is linked to the generator Q by

$$\hat{p}_{i,i} = 0 \quad \textit{and} \quad \hat{p}_{i,j} = \frac{q_{i,j}}{q_i},$$

if $0 < q_i < \infty$ and $\hat{p}_{i,j} = \delta_{ij}$ for all $j \in \mathfrak{S}$ if $q_i = 0$, where δ_{ij} is the *Kronecker function*, i.e., $\delta_{ij} = 1$ for $i = j$ and 0 otherwise. Thus, absorption is excluded if $0 < q_i < \infty$ for all $i \in \mathfrak{S}$.

This basic structure suggests that one can construct a Markov jump process in continuous time by specifying a conservative Q-matrix: Start the chain at an arbitrary state and let it reside in state i_0 for an exponential holding time with parameter $\lambda(i) = \sum_{j \neq i_0} q_{i_0,j} = q_{i_0}$ and jump to the state i_1 with probability $\hat{p}_{i_0,i_1} = q_{i_0,i_1}/q_{i_0}$. Indeed, under the condition that the resulting process is nonexplosive, one gets a continuous time Markov chain with infinitesimal generator Q. The assumption that the process in nonexplosive can of course be avoided. Moreover, the resulting Markov process is called the *minimal construction* and is, in addition to the conditions given in Proposition 1.1.1, nonexplosive if the embedded Markov chain is recurrent.

The concepts of irreducibility, recurrence and transience can be defined via the embedded Markov chain. In this manner, a Markov jump process has the respective property if and only if the embedded Markov chain does. An ergodic process is an irreducible recurrent process with stationary measure having finite mass. We have the following stability theorem (see Theorem II 4.3 in [5])

Theorem 1.1.3. *An irreducible nonexplosive Markov jump process is ergodic if and only if one can find a probability vector p with $pQ = 0$. In that case p is the stationary distribution, i.e., $p\boldsymbol{P}(t) = \boldsymbol{P}(t)$ for every $t \in [0, \infty)$.*

All stochastic processes considered in this thesis arise from a minimal construction based on a conservative rate matrix Q. The resulting processes are nonexplosive since one of the properties from Proposition 1.1.1 holds in every case. In the second chapter, we consider queueing systems with infinite state space. Thus, we have to guarantee the ergodicity of the stochastic processes. The main ergodicity criterion that serves this purpose is stated in the next section. Once ergodicity is established, we can focus on finding the unique normalized solution of the equation $pQ = 0$. In the third chapter, we consider finite queueing systems. Ergodicity is in this case given by the irreducibility of Q and the finiteness of the state space. The main objective is again to find the unique normalized solution of the equation $pQ = 0$.

1.2 Quasi birth and death processes

We consider a two-dimensional Markov jump process $L = (L_{1,t}, L_{2,t})_{t\in[0,\infty)}$ in continuous time on the state space $\mathfrak{S} = \{0, \ldots, N\} \times \mathbb{N}_0$. The first component $(L_{1,t})_{t\in[0,\infty)}$ is called the *phase* or the *phase process*. The second component $(L_{2,t})_{t\in[0,\infty)}$ is called the *level* or the *level process*. Let

$$l(m) = \{0, \ldots, N\} \times \{m\}$$

be the collection of states in level m for $m \geq 0$. We will also call $l(m)$ the *level m*

Definition 1.2.1. A Markov chain on $\mathfrak{S}$ is called a *quasi birth and death process (QBP)* if transitions are restricted to one-step transitions from states in one level to states in the same level or to states in the neighboring levels.

By this definition, the transitions of L are restricted to one-step transitions from states in one level to states in the same level or to states in the neighboring levels, i.e., transitions from the states in $l(m)$ to the states in $l(m')$ can occur if and only if $m' = m - 1$, m, or $m + 1$. Furthermore, we assume that the transitions are level-independent, i.e., the transition rate from states in $l(m)$ to states in $l(m')$ depends on m and m' only through the difference $m - m'$. The QBP is called *homogeneous* in this case.

The states can be ordered lexicographically with respect to the first component in the order

$$(0,0),(1,0,),\dots,(N,0),(0,1),(1,1),\dots,(N,1),\dots.$$

In this order, the transition rate matrix or infinitesimal generator Q has a block-tridiagonal form of the following type:

$$Q = \begin{pmatrix} B & A_0 & 0 & 0 & \cdots \\ A_2 & A_1 & A_0 & 0 & \cdots \\ 0 & A_2 & A_1 & A_0 & \cdots \\ 0 & 0 & A_2 & A_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

where B, A_0, A_1 and A_2 are quadratic matrices. The matrices A_0 and A_2 are nonnegative; the matrices B and A_1 have nonnegative off-diagonal entries and negative entries on their diagonals. The sum of the elements in each row of Q equals 0. The matrix B contains the transition rates from states in level $l(0)$ to states the level $l(0)$. The matrices A_0, A_1 and A_2 are generated by the rates from states in level $l(m)$ to states in level $l(m+1)$, $l(m)$ and $l(m-1)$, respectively, for $m \geq 0$.

Quasi birth and death processes belong to the class of stochastic processes that can be analyzed by the *matrix-analytic method*. The origins of these technique go back to Marcel Neuts (see [51], [52] and [53]), whose research originated the *matrix-geometric distribution* and *phase-type processes*. The term "geometric" stems from the generalization of the structure of the stationary measure of the one-dimensional birth and death process on the nonnegative integers to the two-dimensional case. The next theorem displays this geometric structure of the stationary probability measure of a positive-recurrent QBP. The theorem and its proof can be found in Latouche and Ramaswami [41] (see Theorem 6.4.1).

Theorem 1.2.2. *Assume that the continuous time quasi birth and death process is positive-recurrent. Let the stationary distribution p of the process be partitioned by levels into subvectors p_m, $m \geq 0$. Then, the stationary probability distribution is such that*

$$p_m = p_0 R^m \quad \textit{for } m \geq 0,$$

where the matrix R records the rate of sojourn in the states of level $l(m+1)$ per unit of the local time of $l(m)$. Furthermore, we have that $R = A_0S$, where the matrix S records the expected sojourn time in the states of $l(m)$, starting from $l(m)$, before the first visit to $l(m-1)$.

The following stability theorem taken from [41] (see Theorem 7.2.4) gives a necessary and sufficient condition for stability and the existence of a unique stationary probability measure. The proof will be omitted and can be found in [41]; an earlier reference is [53] (see Theorem 3.1.1).

Theorem 1.2.3. *Consider an irreducible, continuous time QBP with a finite number of phases and assume that the matrix $A = A_0 + A_1 + A_2$ is irreducible. The process is positive recurrent if and only if*

$$\pi A_0\mathbf{1} < \pi A_2\mathbf{1}, \tag{1.2.1}$$

where $\mathbf{1} = (1, \dots, 1)^\top \in \mathbb{R}^{N+1}$ and $\pi = (\pi_0, \dots, \pi_N)$ is the existing and unique solution of the equations $\pi A = 0$ and $\pi\mathbf{1} = 1$. The process is recurrent if $\pi A_0\mathbf{1} = \pi A_2\mathbf{1}$ and transient if $\pi A_0\mathbf{1} > \pi A_2\mathbf{1}$.

The stability condition $\pi A_0\mathbf{1} < \pi A_2\mathbf{1}$ states that the phase-averaged rate for level transitions from $l(m)$ to $l(m+1)$ must be smaller than the phase-averaged rate for level transitions from $l(m)$ to $l(m-1)$ for every $m \geq 1$.

Chapter 2

Overflow to an infinite queue and customers as servers

2.1 Model overview

In this chapter, we consider a fairly general open queueing network consisting of two queues having an overflow capability from the first to the second queue. The arrival and service rates are state-dependent and the customers in the first queue act as servers for the second queue, i.e., the service rate in the second queue depends on the state of the first queue.

Consider two queues Q_1 and Q_2. Let the total number of customers in queue Q_1 be given by the state of a finite birth and death chain in continuous time with state space $\{0, \ldots, N\}$, reflecting barriers and state-dependent birth and death rates. Let the birth rate be $\lambda_{1,n} > 0$ if the chain is in state $n = 0, \ldots, N-1$ and the death rate be $\mu_{1,n} > 0$ if the chain is in state $n = 1, \ldots, N$. The state of the birth and death chain corresponds to the total number of customers in the queue Q_1, i.e., the number of customers in the waiting room plus the number of customers in service. If this number is n, then the arrival rate of the queue is $\lambda_{1,n} > 0$, $n = 0, \ldots, N-1$, and service rate is $\mu_{1,n} > 0$, $n = 1, \ldots, N$. Let $\lambda_{1,N} > 0$ be the potential birth rate in state N, i.e., the potential arrival rate if Q_1 is in state N. Let $\mu_{1,0} = 0$ and let L_1 be the (stationary) total number of customers in Q_1, i.e., the number of occupied waiting positions plus the number of customers being served (under stationary conditions). If Q_1 is fully occupied, i.e., if $L_1 = N$, then the arrival stream of Q_1 is weighted with $p \in [0, 1]$ and directed to the second queue. This procedure is called overflow. Furthermore, the fraction $1 - p$ of arriving customers is lost in this case. The parameter p can be used as a control parameter or to model the customers' impatience. Assume that the second queue has no other external arrivals. In other words, the arrival rate at Q_2 is $p\lambda_{1,N}$ in the case $L_1 = N$ and 0 otherwise. Q_2 has one server with exponentially distributed service times and a waiting room with infinite capacity. The service rate is variable and depends on the number of customers L_1 in Q_1 and is $L_1\mu_2$, where $\mu_2 > 0$. The customers in each queue are served in their order of arrival. This basic model is discussed in the next section.

We further consider two variants of the basic model by allowing customers to jockey from the second queue to the first queue. In queueing theory, *jockeying* is called the possibility for waiting customers to move from one queue to another queue. In the first variant, the first customer or more generally the first k customers, $1 \leq k \leq N-1$, from Q_2, if one is present, are forced to move to Q_1 as soon as Q_1 empties. These customers can then act as servers for the second queue. We call this model the model with *limited jockeying* because the number of jockeying customers limited to a fixed value smaller than the capacity of the first queue. Therefore, some customers might have to stay in Q_2 although Q_1 has not reached its capacity bound. In the second variant, as soon as Q_1 empties, it is filled with the customers of Q_2 until it reaches its capacity bound or Q_2 empties. This jockeying procedure is called *unlimited jockeying*. The transferral of customers does not result in service interruptions, because the service rate in Q_2 is 0 at the time points of customer transferrals. See Figure 2.1 for a schematic overview of the three models.

As indicated in Figure 2.1 it is possible to let Q_2 have a Poisson arrival stream being independent of Q_1 with intensity $\lambda_2 > 0$, so that the arrival rate to Q_2 is λ_2 for $L_1 = 0, \ldots, N-1$ and $p\lambda_1 + \lambda_2$ for $L_1 = N$. In order to simplify the presentation of the results, we omit the derivations in this case and discuss the solution in Section 2.2.5.

Our basic model covers for example the case of Q_1 being an $M_{(n)}/M_{(n)}/1/N-1$-queue with one server and $N-1$ waiting positions or any $M_{(n)}/M_{(n)}/K/N-K$ queueing system with K servers and $N-K$ waiting positions, where $N \geq K > 0$. The index (n) indicates state-dependent rates. Following the notation of van Doorn [19], the queueing systems with one server and $N-1$ waiting positions in Q_1 for example might be labeled with $(M_{(n)}/M_{(n)}/1/N-1)_{p\text{−overflow}}/M_{(n)}/1$.

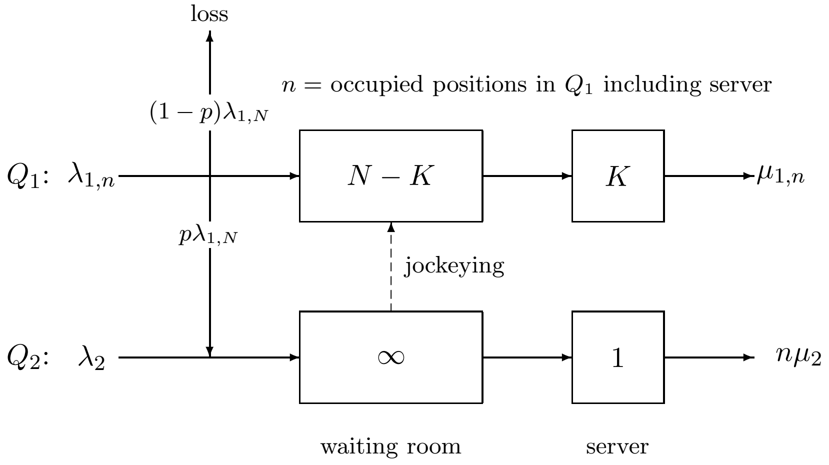


Figure 2.1: Model overview: Customers acting as servers, state-dependent service and arrival rates.

Solution approach

We are interested in the two-dimensional server and waiting room demand process of this model, embedded at the time instants of arrivals to Q_1 and Q_2 and departures from Q_1 and Q_2. This process is a Markov chain with state space $\mathfrak{S} = \{(n, m) \,|\, n = 0, \dots, N,\ m \geq 0\}$, where the first and second component of $(n, m) \in \mathfrak{S}$ correspond to the number of occupied servers and/or waiting positions in Q_1 and Q_2, respectively.

We solve the system of steady-state equations that describe the systems' dynamics by exploiting the probability generating functions of the number of customers in Q_2, see for example Avi-Itzhak and Mitrani [6] and Perel and Yechiali [56]. The solution of these equations is given in terms of only N unknowns $p_{1,0}, \dots, p_{N,0}$, where $p_{n,m}$ is the steady-state probability of having n customers in Q_1 and m customers in Q_2. These unknowns can then be determined by N linear equations in the unknowns given later. The model presented in [56] is an important special case of our birth and death queueing system for the case $p = 0$, $\lambda_{1,n} = \lambda_1 > 0$, $\mu_{1,n+1} = \mu_1 > 0$, $n = 0, \dots, N-1$ and $\lambda_2 > 0$.

Applications

Our models can be used to analyze for example file sharing or torrent systems. In these systems, data files are shared, exchanged and spread over the internet or a private intranet by letting downloading customers act as servers for further customers. Due to cost and bandwidth limitations it is a common case to have a limited number of primary download channels (i.e., the servers in Q_1), a limited number of primary queueing positions for these channels (i.e., the waiting positions in Q_1) and an unlimited number of secondary download channels (i.e., the positions in Q_2). A server or host provides a data file via the servers in the primary channels. This file is typically divided into several data portions which are downloaded by the customers in service in the primary channels. When the primary download and queueing positions are occupied, an incoming customer can decide whether to join one of the secondary download channels in Q_2 or not – the probability of joining may equal p for all potential secondary customers. Additionally, the parameter p can be regarded as a control parameter. The arrival stream of the secondary channels may consist of the overflow from the primary channels and/or an independent arrival stream. The secondary download channels are served by the customers that are present in the primary channels, i.e., the secondary customers can download the requested file or the data portions of this file from the customers in the primary channels. These torrent systems make hosting of a file with a potentially unlimited number of downloaders affordable because the costs of the secondary channels can be assigned to the primary and secondary customers. The download speed in the primary and secondary download channels can increase with the number of customers who are receiving service (in the primary channel). This is due to the fact that when customers download the same file at the same time, they can upload portions of the file to each other. In addition, the data portions of the file can be downloaded in an arbitrary order. Thus, a secondary customer can finish the service earlier than a primary customer who arrived earlier and who participated in serving this secondary customer. The download speed in the primary channel may also decrease with the number of customers because of a limitation of the bandwidth.

Further applications of our model are to the SETI@home and the GIMPS project. The SETI@home project (Search for ExtraTerrestrial Intelligence), initiated by the Space Sciences Laboratory of the University of California, Berkeley, is described in Perel and Yechiali [56]. This project is searching for extraterrestrial intelligence by radio telescopes. The GIMPS project (Great Internet Mersenne Prime Search) launched by the Mersenne Research Incorporation is a prominent and popular project searching for Mersenne prime numbers. In both projects, huge amounts of data have to be processed. This is done by institutions or private persons who install special computer programs on their computers. Whenever a computer equipped with this software is idle, these programs are activated and data can be processed.

Having these examples in mind, it is very natural to assume variable service and arrival rates due to bandwidth or CPU limitations. The models presented in this chapter achieve the purpose of having variable arrival and service rates.

Further related literature

The models presented in this chapter can be represented as quasi birth and death processes (QBP) presented in Section 1.2, see for example Latouche and Ramaswami [41]. Consequently, their analysis can be carried out using a matrix-geometric approach, see Neuts [53]. Furthermore, all models can be formulated as Markov-modulated queues. For related literature on both the matrix-geometric approach and Markov-modulated queues see for example Asmussen [4]. For the latter see Mahabhashyam et al. [43], Muscariello et al. [50] and Takine [64] and the references therein.

Overflow queueing models are widespread in literature. Van Doorn [19] and Parthasarathy and Sudhesh [55] study the interoverflow time distribution of a finite birth and death queue model as presented for Q_1. Koury et al. [39] and Krieger et al. [40] give reviews of iterative numerical methods for overflow queueing models. A brief discussion of numerical methods for some two-queue overflow systems and further references are given in Ching and Ng [10]. While most of these formulations are of primary interest when the focus is on numerical results, the method used in the following gives deep insight into the structure of the stability conditions and solutions. Nevertheless, we will use the stability theorem for QBP presented in Section 1.2 to derive necessary and sufficient conditions for stability in every model. Related overflow models are studied in Chapter 3 of this thesis, in van Doorn [19] and Guérin, Lien [27] and the referenced literature therein using a variety of different techniques. Further related literature is mentioned in Chapter 3, Section 3.1 of this thesis.

2.2 Customers as servers: Basic model

2.2.1 Model description and steady-state equations

Let the number of customers in Q_1 be given by the state of a finite birth and death chain with state space $\{0, \dots, N\}$, birth rate $\lambda_{1,n} > 0$ if the chain is in state $n = 0, \ldots, N-1$ and death rate $\mu_{1,n} > 0$ if the chain is in state $n = 1, \ldots, N$. Let $\lambda_{1,N} > 0$ be the potential birth rate if Q_1 is in state N. Let $\mu_{1,0} = 0$ and let L_1 be the (stationary) total number of customers in Q_1. Let Q_2 have one server with exponentially distributed service times and a waiting room with infinite capacity. The service rate in Q_2 is $L_1\mu_2$ if L_1 customers are present in Q_1, where $\mu_2 > 0$. Q_2 is fed by a portion of the rejected customers of Q_1. If queue one is fully occupied, i.e., if $L_1 = N$, then the arrival stream of Q_1 is weighted with $p \in (0, 1]$ and directed to Q_2 while the fraction $1 - p$ of arriving customers is lost in this case. The customers in each queue are served in their order of arrival. The underlying Markov process is irreducible since there exists a path with positive probability from the state $(0, 0)$ to every other state. Furthermore, the Markov process in non-explosive by Proposition 1.1.1. The transition rate diagram is given in Figure 2.2, where $\lambda_2 = 0$ in this section.

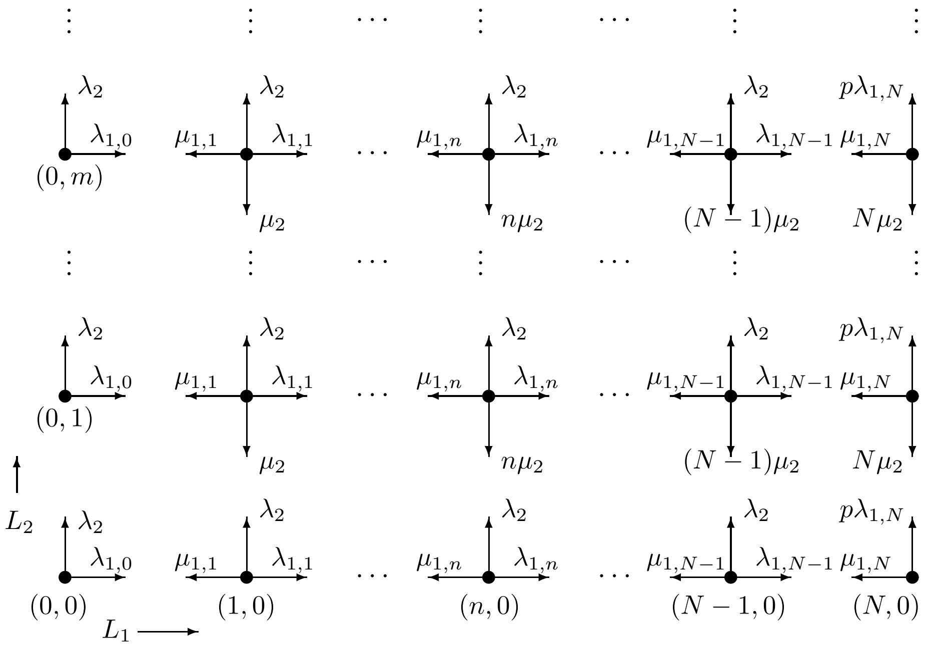


Figure 2.2: Basic model: Transition rate diagram.

The first queue is a finite one-dimensional birth and death chain and the distribution of L_1 is therefore well known, see for example or Cohen [12].

Setting $\rho_n = \lambda_{1,n}/\mu_{1,n+1}$, $n = 0, \dots, N-1$, we get

$$P(L_1 = n) = \frac{\prod_{i=0}^{n-1} \rho_i}{1 + \sum_{j=1}^{N} \prod_{i=0}^{j-1} \rho_i} \text{ and } EL_1 = \frac{\sum_{n=1}^{N} n \prod_{i=0}^{n-1} \rho_i}{1 + \sum_{n=1}^{N} \prod_{i=0}^{n-1} \rho_i}, \tag{2.2.1}$$

where the empty product is set to 1. For example, with constant service rates $\lambda_{1,n} = \lambda_1$, arrival rates $\mu_{1,n} = \mu_1$ and $\lambda_1 \neq \mu_1$ we arrive at the classical $M/M/1/N-1$ queue:

$$P(L_1 = n) = \frac{1 - \frac{\lambda_1}{\mu_1}}{1 - \left(\frac{\lambda_1}{\mu_1}\right)^{N+1}} \left(\frac{\lambda_1}{\mu_1}\right)^n \text{ and } EL_1 = \frac{\lambda_1}{\mu_1 - \lambda_1} - \frac{(N+1)\lambda_1^{N+1}}{\mu_1^{N+1} - \lambda_1^{N+1}}.$$

In the case $\lambda_1 = \mu_1$, we have $P(L_1 = n) = 1/(N+1)$ and $EL_1 = N/2$. For service rates $\lambda_{1,n} = \lambda_1$ and arrival rates $\mu_{1,n} = n\mu_1$, the first queue is an Erlang loss system, i.e., the $M/M/N/0$ queue:

$$P(L_1 = n) = \frac{\frac{\rho^n}{n!}}{1 + \sum_{n=1}^{N} \frac{\rho^n}{n!}} \text{ and } EL_1 = \frac{\sum_{n=1}^{N} \frac{\rho^n}{(n-1)!}}{1 + \sum_{n=1}^{N} \frac{\rho^n}{n!}},$$

where $\rho = \lambda/\mu$ (see Erlang [24]).

It is shown in van Doorn [19] that the overflow process from the first queue is a renewal process of hyperexponential type. An expression for the Laplace transform of the interoverflow time, i.e., the time between successive moments of overflow, can be given. The intensity Λ of the interoverflow process is

$$\Lambda = \frac{p\lambda_{1,N} \prod_{n=0}^{N-1} \rho_n}{1 + \sum_{n=1}^{N} \prod_{i=0}^{n-1} \rho_i}. \tag{2.2.2}$$

Parthasarathy and Sudhesh [55] express the interoverflow time distribution as a power series expansion and as a hyperexponential distribution in closed form. Additionally, a closed-form expression in terms of the system parameters for the r-th moment of the overflow process is given.

Let L_1 and L_2 be the (stationary) total number of customers in Q_1 and

Q_2, respectively, and

$$p_{n,m} = P(L_1 = n, L_2 = m)$$

for $n = 0, \ldots, N$ and $m \geq 0$. The unknown quantities $P(L_2 = m)$, $m \geq 0$, EL_2 and $\mathrm{Cov}(L_1, L_2)$ are of special interest and will be derived in the next section.

The bivariate server and waiting room demand distribution $(p_{n,m})_{n=0,\ldots,N,\, m\geq 1}$ in the case $\lambda_2 = 0$ is the unique nonnegative and normalized solution (i.e., $\sum_{n=0}^{N} \sum_{m\geq 0} p_{n,m} = 1$) of the following steady-state equations:

$$\begin{aligned}(\lambda_{1,n}(1-\delta_{nN}) &+ p\lambda_{1,N}\delta_{nN} + (1-\delta_{n0})\mu_{1,n} + (1-\delta_{m0})n\mu_2)p_{n,m} \\ &= (1-\delta_{n0})\lambda_{1,n-1}p_{n-1,m} + (1-\delta_{nN})\mu_{1,n+1}p_{n+1,m} \\ &\quad + (1-\delta_{m0})\delta_{nN}p\lambda_{1,N}p_{n,m-1} + (1-\delta_{n0}(1-\delta_{m0}))n\mu_2 p_{n,m+1}\end{aligned} \tag{2.2.3}$$

for $n = 0, \ldots, N$ and $m \geq 0$, where δ_{ij} is the *Kronecker function*, i.e., $\delta_{ij} = 1$ for $i = j$ and 0 otherwise.

2.2.2 Necessary and sufficient stability condition

In this section, we derive the necessary and sufficient condition for the existence of a normalized solution of the steady-state equations (2.2.3). First, we give an intuitive argument for the stability condition and the possibility of reducing the infinite number of unknowns in these equations to only N unknowns, namely $p_{1,0}, \ldots, p_{N,0}$.

The equations (2.2.3) for $m = 0$ and $n = 0, \ldots, N$, i.e.,

$$\lambda_{1,0}p_{0,0} = \mu_{1,1}p_{1,0}, \tag{2.2.4}$$

$$(\lambda_{1,n} + \mu_{1,n})p_{n,0} = \lambda_{1,n-1}p_{n-1,0} + \mu_{1,n+1}p_{n+1,0} + n\mu_2 p_{n,1}, \quad n = 1, \ldots, N-1, \tag{2.2.5}$$

$$(p\lambda_{1,N} + \mu_{1,N})p_{N,0} = \lambda_{1,N-1}p_{N-1,0} + N\mu_2 p_{N,1}, \tag{2.2.6}$$

give after summation over $n = 0, \ldots, N$

$$p\lambda_{1,N}p_{N,0} = \mu_2 \sum_{n=1}^{N} n p_{n,1}. \tag{2.2.7}$$

The equations (2.2.3) for $m \geq 1$ and $n = 0, \ldots, N$, i.e.,

$$\lambda_{1,0} p_{0,m} = \mu_{1,1} p_{1,m}, \tag{2.2.8}$$

$$(\lambda_{1,n} + \mu_{1,n} + n\mu_2) p_{n,m} = \lambda_{1,n-1} p_{n-1,m} + \mu_{1,n+1} p_{n+1,m} + n\mu_2 p_{n,m+1}, \quad n = 1, \ldots, N-1, \tag{2.2.9}$$

$$(p\lambda_{1,N} + \mu_{1,N} + N\mu_2) p_{N,m} = \lambda_{1,N-1} p_{N-1,m} + p\lambda_{1,N} p_{N,m-1} + N\mu_2 p_{N,m+1}, \tag{2.2.10}$$

yield by summing over $n = 0, \ldots, N$

$$p\lambda_{1,N} p_{N,m} + \mu_2 \sum_{n=1}^{N} n p_{n,m} = p\lambda_{1,N} p_{N,m-1} + \mu_2 \sum_{n=1}^{N} n p_{n,m+1}. \tag{2.2.11}$$

From (2.2.7) and (2.2.11) we obtain

$$p\lambda_{1,N} p_{N,m} = \mu_2 \sum_{n=1}^{N} n p_{n,m+1} \tag{2.2.12}$$

for all $m \geq 0$ by induction. Summation over $m \geq 0$ gives

$$p\lambda_{1,N} \sum_{m \geq 0} p_{N,m} = \mu_2 \left(\sum_{n=1}^{N} n \sum_{m \geq 0} p_{n,m} - \sum_{n=1}^{N} n p_{n,0} \right). \tag{2.2.13}$$

The equations (2.2.12) and (2.2.13) can (under stationary conditions) be written as

$$p\lambda_{1,N} p_{N,m} = \mu_2 P(L_2 = m+1) E(L_1 | L_2 = m+1)$$

and

$$p\lambda_{1,N} P(L_1 = N) = \mu_2 \big(EL_1 - P(L_2 = 0) E(L_1 | L_2 = 0) \big), \tag{2.2.14}$$

where $P(L_2 = m) E(L_1 | L_2 = m) = \sum_{n=1}^{N} n p_{n,m}$ for $m \geq 0$. The only term in (2.2.14) involving unknowns is the term

$$P(L_2 = 0) E(L_1 | L_2 = 0) = \sum_{n=1}^{N} n p_{n,0}.$$

This suggests that the set of unknown probabilities in the system of steady-state equations (2.2.3) can be reduced to $p_{1,0}, \ldots, p_{N,0}$. From equation (2.2.14) one might also predict from the following intuitive argument that the system is stable if

$$p\lambda_{1,N} < \mu_2 EL_1 \tag{2.2.15}$$

holds. Near saturation we have $P(L_1 = N) \sim 1$, $P(L_2 = 0) \sim 0$ and $E(L_1|L_2 = 0)$ is bounded since the system is stable. Hence by (2.2.14)

$$\frac{p\lambda_{1,N}}{\mu_2 EL_1} \sim 1 - \frac{P(L_2 = 0)E(L_1|L_2 = 0)}{\mu_2 EL_1 P(L_1 = N)} < 1.$$

Furthermore, the arrival and service rates in Q_1 have no influence on the stability of Q_1, since it is a loss system, but on the stability of Q_2. The arrival rate in Q_2 is $p\lambda_{1,N}$ while the average service rate is $\mu_2 EL_1$. The fraction of both should be smaller than 1 since Q_2 behaves like an $M/M/1$ queue in this case. Once again we arrive at (2.2.15). The stability condition should in general be weaker because the second queue has a non-zero arrival rate if and only if the first queue is fully occupied. By regarding the joint queue length process of Q_1 and Q_2 as a quasi birth and death process and using the stability Theorem 1.2.3 of Section 1.2, we will show that these presumptions are correct and that the necessary and sufficient stability condition is indeed the one given in the next proposition, i.e., $p\lambda_{1,N}P(L_1 = N) < \mu_2 EL_1$.

Proposition 2.2.1. *The system* (2.2.3) *has a unique nonnegative and normalized solution if and only if* $p\lambda_{1,N}P(L_1 = N) < \mu_2 EL_1$ *or equivalently*

$$\frac{p\lambda_{1,N}}{\mu_2} < \frac{\sum\limits_{n=1}^{N} n \prod\limits_{i=0}^{n-1} \rho_i}{\prod\limits_{n=0}^{N-1} \rho_n} \tag{2.2.16}$$

holds, where $\rho_n = \lambda_{1,n}/\mu_{1,n+1}$ *for* $n = 0, \dots, N-1$.

Proof. By regarding Q_1 as the phase and Q_2 as the level, the joint queue length process of Q_1 and Q_2 is a quasi birth and death process. The stability condition (2.2.16) is then derived from Theorem 1.2.3 in the following way. We assume a more general model for the moment by endowing Q_2 with an external arrival stream and service rates controlled by the state of Q_1. In this generalization of our basic model, the exponential arrival rate to Q_2 is $\lambda_{2,n}$ and the exponential service rate is $\mu_{2,n}$ given $L_1 = n$, $n = 0, \dots, N$. The rates for arrivals and service in Q_1 remain unchanged. In order to ensure irreducibility, we assume that there exists an $i \in \{0, \dots, N\}$ with $\lambda_{2,i} > 0$ and a $j \in \{0, \dots, N-1\}$ with $\mu_{2,j} > 0$. With this setting, the matrices A_0,

A_1 and A_2 are given by

$$A_0 = \operatorname{diag}(\lambda_{2,0}, \dots, \lambda_{2,N}), \quad A_2 = \operatorname{diag}(\mu_{2,0}, \dots, \mu_{2,N})$$

and $A_1 = A - A_0 - A_2$, where A is the rate matrix of the phase process governing Q_1, i.e., the standard birth and death process on $\{0, \dots, N\}$ with reflecting barriers and birth rates $\lambda_{1,n}$ in the states $n = 0, \dots, N-1$ and death rates $\mu_{1,n}$ in the states $n = 1, \dots, N$. The vector π is the stationary probability measure of the phase process in Q_1 and therefore $\pi_n = P(L_1 = n)$, $n = 0, \dots, N$, is given by (2.2.1). The stability condition (1.2.1) is equivalent to

$$\sum_{n=0}^{N} (\mu_{2,n} - \lambda_{2,n})\pi_n < 0. \tag{2.2.17}$$

Setting $\lambda_{2,n} = 0$ for $n = 0, \dots, N-1$, $\lambda_{2,N} = p\lambda_{1,N}$ and $\mu_{2,n} = n\mu_2 > 0$ we get stability if and only if (2.2.16) holds. □

We continue this section with a discussion of the stability condition. The stability condition can be written as

$$p\lambda_{1,N} P(L_1 = N) < \mu_2 EL_1 \tag{2.2.18}$$

as one might expect, since (2.2.1) states

$$\sum_{n=1}^{N} n \prod_{i=0}^{n-1} \rho_i = \frac{EL_1}{P(L_1 = 0)} \quad \text{and} \quad \prod_{n=0}^{N-1} \rho_n = \frac{P(L_1 = N)}{P(L_1 = 0)}.$$

The stability condition (2.2.18) can in this form be well interpreted: The system is stable if and only if the expected service rate $\mu_2 EL_1$ in Q_2 suffices to handle the average arrival or overflow rate $p\lambda_{1,N} P(L_1 = N)$ from Q_1 to Q_2.

Remark 2.2.2. The stability condition can be formulated independently of the arrival rate $\lambda_{1,0}$ in Q_1. This is explained by the fact that the arrival rate in Q_1 is $\lambda_{1,0}$ only in the case that no customers are present in the first queue. Therefore, the arrival rate $\lambda_{1,0}$ influences the length of the idle periods of the system (i.e., the periods in which no customers are served), but not the busy periods. The idle periods of this queueing system can be classified by the number of customers in Q_2 which are not served due to the lack of customers in Q_1.

When the first queue is an $M/M/1/N-1$ queue with one server, $N-1$

waiting positions and constant arrival and service rates, the stability condition simplifies:

Remark 2.2.3. In the case of constant service and arrival rates $\lambda_{1,n} = \lambda_1$ for $n = 0, \dots, N$ and $\mu_{1,n} = \mu_1$ for $n = 1, \dots, N$, the first queue is an $M/M/1/N-1$ queue and we get the stability condition

$$p\lambda_1 < \mu_2 \rho^{-N} \sum_{n=1}^{N} n\rho^n, \tag{2.2.19}$$

where $\rho = \lambda_1/\mu_1$ is the traffic intensity of Q_1. Observe that

$$EL_1 = \frac{\sum\limits_{n=1}^{N} n\rho^n}{(1 + \sum\limits_{n=1}^{N} \rho^n)} < \rho^{-N} \sum_{n=1}^{N} n\rho^n$$

holds in this case.

It follows from the condition (2.2.18) and (2.2.1) that a more restrictive and sufficient but not necessary condition for stability is

$$p\lambda_{1,N} < \mu_2 EL_1.$$

Observe that the state-dependent rates are involved in EL_1. (2.2.18) yields that the condition

$$\Lambda < \mu_2,$$

where Λ is given by (2.2.2), is necessary but not sufficient. Both results are not surprising: On the one hand, the arrival rate in Q_2 is only positive if $L_1 = N$. Therefore $p\lambda_{1,N}$ is in general greater than or equal to the actual mean service rate. On the other hand, the service rate in Q_2 is in general greater than or equal to μ_2 given that customers arrive to Q_2, i.e., given $L_1 = N$. Observe that the condition (2.2.18) complies with $\Lambda < \mu_2 EL_2$.

2.2.3 Generating functions and steady-state distribution

In this section, we exploit the steady-state equations (2.2.3) and derive recurrence equations for the probability generating functions

$$G_n(z) = \sum_{m=0}^{\infty} p_{n,m} z^m, \quad |z| \leq 1.$$

These relations are then used to reduce the set of unknown steady-state probabilities to the unknowns $p_{1,0}, \dots, p_{N,0}$ and to derive equivalent formulations for the stability condition (2.2.16). The stability condition will be related to the existence of a certain number of zeros of a function originating from the recurrence equations for the probability generating functions. This relation will give a deeper insight into the existence and uniqueness of a solution of the steady-state equations (2.2.3) and into the nature of the unknowns $p_{1,0}, \dots, p_{N,0}$.

We can assume for the moment that the system is stable and that (2.2.16) holds. In this case, the generating functions are well defined for $|z| \leq 1$. Exploiting the recurrence relations and using an inductive argument and Cramer's rule we derive N independent equations for the N unknowns and the stability condition.

By multiplying the equations (2.2.4), (2.2.5), (2.2.6), (2.2.8), (2.2.9) and (2.2.10) by z^m and summing over $m \geq 0$ we get, after simplifying,

$$\lambda_{1,0} G_0(z) = \mu_{1,1} G_1(z), \tag{2.2.20}$$

$$\begin{aligned}\big((\lambda_{1,n} + \mu_{1,n})z - n\mu_2(1-z)\big)G_n(z) &= \lambda_{1,n-1} z G_{n-1}(z) + \mu_{1,n+1} z G_{n+1}(z) \\ &\quad - n\mu_2(1-z)p_{n,0} \end{aligned} \tag{2.2.21}$$

for $n = 1, \dots, N-1$ and

$$\begin{aligned}&\big(\mu_{1,N} z + (p\lambda_{1,N} z - N\mu_2)(1-z)\big)G_N(z) \\ &\qquad = \lambda_{1,N-1} z G_{N-1}(z) - N\mu_2(1-z)p_{N,0}. \end{aligned} \tag{2.2.22}$$

In order to write these equations in matrix form, we define the vectors

$$G(z) = (G_0(z), \dots, G_N(z))^\top,$$
$$p = (0, p_{1,0}, 2p_{2,0}, \dots, Np_{N,0})^\top$$

and the matrix $A(z) \in \mathrm{Mat}(N+1, N+1, \mathbb{R})$ by

$$A(z) = \begin{pmatrix} \alpha_0(z) & -\mu_{1,1} & 0 & \dots & \dots & 0 \\ -\lambda_{1,0} z & \alpha_1(z) & -\mu_{1,2} z & \ddots & & \vdots \\ 0 & \ddots & \alpha_2(z) & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & -\mu_{1,N} z \\ 0 & \dots & \dots & 0 & -\lambda_{1,N-1} z & \alpha_N(z) \end{pmatrix}, \tag{2.2.23}$$

where

$$\alpha_0(z) = \lambda_{1,0}, \tag{2.2.24}$$

$$\alpha_n(z) = (\lambda_{1,n} + \mu_{1,n})z - n\mu_2(1-z) \text{ for } n = 1, \ldots, N-1 \text{ and} \tag{2.2.25}$$

$$\alpha_N(z) = \mu_{1,N}z + (p\lambda_{1,N}z - N\mu_2)(1-z). \tag{2.2.26}$$

In this notation, the equations (2.2.20), (2.2.21) and (2.2.22) are equivalent to

$$A(z)G(z) = -\mu_2(1-z)p. \tag{2.2.27}$$

Let $A_n(z)$ be the matrix obtained from $A(z)$ by replacing the $(n+1)$-th column by the vector $-\mu_2(1-z)p$ for $n = 0, \ldots, N$. By Cramer's rule we may write

$$\det(A(z))G_n(z) = \det(A_n(z)) \tag{2.2.28}$$

for every $n = 0, \ldots, N$ and all values z such that $A(z)$ is invertible. Since the functions A, A_n and G_n are continuous and bounded with at most finitely many zeros in the interval $[0,1]$, equation (2.2.28) must hold for all $z \in [0,1]$ and every $n = 0, \ldots, N$. Hence, the generating functions $G_0, \ldots, G_N$ are uniquely determined by the equations (2.2.28) and $p_{1,0}, \ldots, p_{N,0}$, since these are the only unknowns occurring in these equations (see also (2.2.20)-(2.2.22)).

$\det(A(z))$ is a polynomial in z of degree $N+1$. We will show in the following that $\det(A(z))$ has $N-1$ zeros in the open interval $(0,1)$ and one zero at $z=1$. Additionally, we will show that $\det(A(z))$ has another zero in the open interval $(1,\infty)$ if and only if the stability condition (2.2.16) holds. The $N-1$ zeros of $\det(A(z))$ in $(0,1)$ will provide us with $N-1$ linear homogeneous equations in the unknowns $p_{1,0}, \ldots, p_{N,0}$. Another linear equation yielding a system of N equations is (2.2.14). We are able to formulate this equation in terms of only $p_{1,0}, \ldots, p_{N,0}$ and the system parameters. The result is stated in Proposition 2.2.8.

Let $q_n(z)$ be the n-th minor of $A(z)$ for $n = 1, \ldots, N+1$ (the n-th minor of a matrix is the determinant of the n-th square sub-matrix). We have

$$q_1(z) = \alpha_0(z),\ q_2(z) = \det\begin{pmatrix} \alpha_0(z) & -\mu_{1,1} \\ -\lambda_{1,0}z & \alpha_1(z) \end{pmatrix}, \ldots, q_{N+1}(z) = \det(A(z)). \tag{2.2.29}$$

By Laplace expansion of the determinants we get

$$\begin{aligned} q_1(z) &= \alpha_0(z)q_0(z), \\ q_2(z) &= \alpha_1(z)q_1(z) - \lambda_{1,0}\mu_{1,1}zq_0(z) \quad \text{and} \\ q_n(z) &= \alpha_{n-1}(z)q_{n-1}(z) - \lambda_{1,n-2}\mu_{1,n-1}z^2q_{n-2}(z) \end{aligned} \tag{2.2.30}$$

for $n = 3, \dots, N+1$, where $q_0(z) = 1$.

In order to proceed with investigating the number and the location of the zeros of $\det(A) = q_{N+1}$, we have to study the system (2.2.30) in order to find the algebraic properties and to determine the shape of the functions $q_1, \dots, q_{N+1}$. The relevant properties are stated in the next proposition.

Proposition 2.2.4. *The function q_n is a polynomial in z of degree $n-1$ for $n = 1, \dots, N$ and of degree $N+1$ for $n = N+1$. The functions $q_0, \dots, q_{N+1}$ and $\alpha_0, \dots, \alpha_N$ have the following properties:*

(i) *q_n and q_{n+1} have no common root in $(0,1)$ for $n = 0, \dots, N$.*

(ii) $\operatorname{sgn}(\alpha_0(0)) = 1$ *and* $\operatorname{sgn}(\alpha_n(0)) = -1$ *for $n = 1, \dots, N$.*

(iii) $\operatorname{sgn}(q_n(0)) = (-1)^{n+1}$ *for $n = 1, \dots, N+1$.*

(iv) $q_n(1) = \prod_{i=0}^{n-1} \lambda_{1,i}$ *for $n = 0, \dots, N$ and $q_{N+1}(1) = 0$.*

(v) *For $n = 1, \dots, N$ the following implication holds: If $\tilde{z} > 0$ with $q_n(\tilde{z}) = 0$, then*

$$\operatorname{sgn}(q_{n-1}(\tilde{z})q_{n+1}(\tilde{z})) = -1.$$

(vi) *q_n has $n-1$ pairwise distinct zeros in $(0,1)$ for $n = 1, \dots, N$.*

(vii) $\lim_{z\to\infty} q_n(z) = \infty$ *for $n = 2, \dots, N$ and* $\lim_{z\to\infty} q_{N+1}(z) = -\infty$.

Proof. By the recursive definition (2.2.29), q_n is a polynomial in z of degree $n-1$ for $n = 1, \dots, N$ and of degree $N+1$ for $n = N+1$. We will prove the remaining properties.

(i) q_0 and q_1 have no zero since $\lambda_{1,0} > 0$. Suppose $q_n(z) = q_{n+1}(z) = 0$ for some $z > 0$ and some $n = 2, \dots, N$. Then $q_{n-1}(z) = 0$ follows from (2.2.30). We get $q_1(z) = 0$ by induction which is a contradiction.

(ii) $\operatorname{sgn}(\alpha_0(0)) = \operatorname{sgn}(\lambda_{1,0}) = 1$ since $\lambda_{1,0} > 0$, $\operatorname{sgn}(\alpha_n(0)) = \operatorname{sgn}(-n\mu_2) = -1$ for $n = 1, \dots, N$ since $\mu_2 > 0$.

(iii) By (ii) we have $\operatorname{sgn}(q_1(0)) = \operatorname{sgn}(\alpha_0(0)) = 1$ and

$$\operatorname{sgn}(q_2(0)) = \operatorname{sgn}(\alpha_1(0)q_1(0)) = -1.$$

By induction we get from (ii) that

$$\operatorname{sgn}(q_n(0)) = \operatorname{sgn}(\alpha_{n-1}(0))\operatorname{sgn}(q_{n-1}(1)) = (-1)^{n+1}$$

for $n = 0, \dots, N$. The functions α_n and therefore the functions q_n are continuous in 0.

(iv) Clearly, $q_0(1) = 1$ and $q_1(1) = \lambda_{1,0}$. For $n = 2, \dots, N$ we get from (2.2.30) by induction that

$$\begin{aligned} q_n(1) &= \alpha_{n-1}(1)q_{n-1}(1) - \lambda_{1,n-2}\mu_{1,n-1}q_{n-2}(1) \\ &= (\lambda_{1,n-1} + \mu_{1,n-1}) \prod_{i=0}^{n-2} \lambda_{1,i} - \lambda_{1,n-2}\mu_{1,n-1} \prod_{i=0}^{n-3} \lambda_{1,i} = \prod_{i=0}^{n-1} \lambda_{1,i}. \end{aligned}$$

$q_{N+1}(1) = 0$ follows from inserting $q_{N-1}(1) = \prod_{i=0}^{N-2} \lambda_{1,i}$ and $q_N(1) = \prod_{i=0}^{N-1} \lambda_{1,i}$ into (2.2.30) and simplifying.

(v) Let $n = 1, \dots, N$ and $\tilde{z} > 0$ with $q_n(\tilde{z}) = 0$. By (2.2.30) we get

$$\begin{aligned} \operatorname{sgn}(q_{n+1}(\tilde{z})q_{n-1}(\tilde{z})) &= \operatorname{sgn}\big((\alpha_n(\tilde{z})q_n(\tilde{z}) \\ &\qquad - \lambda_{1,n-1}\mu_{1,n-1}\tilde{z}^2 q_{n-1}(\tilde{z}))q_{n-1}(\tilde{z})\big) \\ &= \operatorname{sgn}(-\lambda_{1,n-2}\mu_{1,n-1}\tilde{z}^2 q_{n-1}^2(\tilde{z})) = -1. \end{aligned}$$

In the last line we used $\lambda_{1,n-2}$, $\mu_{1,n-1} > 0$ and the fact that q_n and q_{n-1} have no common zero which was shown in (i).

(vi) Once again we use an inductive argument to show this property. q_1 has no root in $(0, 1)$. q_2 has exactly one root in $(0, 1)$ since $\lambda_{1,1}$, $\mu_2 > 0$, namely $z = \frac{\mu_2}{\lambda_{1,1}+\mu_2}$. Let $n = 3, \dots, N$ and

$$0 < z_1 < \ldots < z_{n-2} < 1$$

be the pairwise distinct zeros of q_{n-1}. We show that (iii), (iv) and (v) imply that q_n has $n-1$ pairwise distinct zeros $z_{n,1}, \dots, z_{n,n-1}$ in $(0, 1)$ and that these zeros satisfy the interlacing property

$$z_0 = 0 < z_{n,1} < z_1 < z_{n,2} < \ldots < z_{n-2} < z_{n,n-1} < 1 = z_{n-1}. \quad (2.2.31)$$

Let $k = 1, \dots, n-2$. We obtain from (2.2.30) and (v) inductively that

$$\operatorname{sgn}(q_n(z_k)) = -\operatorname{sgn}(q_{n-2}(z_k)) = (-1)^{k+n+1}.$$

By (iii) and (iv) this is also true for $z_0 = 0$ and $z_{n-1} = 1$. Thus, q_n changes its sign $n-1$ times in the interval $(0,1)$. Therefore, the $n-1$ zeros of q_n are real, pairwise distinct and must be located in the interval $(0,1)$. Furthermore, q_n has alternating signs on the sequence of zeros of q_{n-1} since

$$\operatorname{sgn}(q_n(z_k)) = (-1)^{k+n+1}$$

for $k = 1, \dots, n-2$. Together with $\operatorname{sgn}(q_n(0)) = -\operatorname{sgn}(q_{n-1}(0))$, which follows from (iii), this finally yields (2.2.31).

(vii) Let $n = 2, \dots, N$. From (iv), (vi) and since q_n is a polynomial in z of degree $n-1$, we obtain immediately that $\lim_{z\to\infty} q_n(z) = \infty$. For $n = N+1$ we get $\lim_{z\to\infty} q_{N+1}(z) = -\infty$ from equation (2.2.30) since $\lim_{z\to\infty} \alpha_N(z) = -\infty$, $\lim_{z\to\infty} q_N(z) = \lim_{z\to\infty} q_{N-1}(z) = \infty$ and since the system parameters are positive. □

The proof of Proposition 2.2.4 (vi) shows that the $n-2$ pairwise distinct zeros $z_{n-1,1}, \dots, z_{n-1,n-2}$ of q_{n-1} and the $n-1$ pairwise distinct zeros $z_{n,1}, \dots, z_{n,n-1}$ of q_n satisfy the interlacing property

$$0 < z_{n,1} < z_{n-1,1} < z_{n,2} < \ldots < z_{n-1,n-2} < z_{n,n-1} < 1 \tag{2.2.32}$$

for every $n = 1, \dots, N$. We can expect a similar behavior for the zeros of q_N and q_{N+1}. In fact, it will turn out that an interlacing property holds for these zeros and that the sign of the slope of q_{N+1} at the point $z = 1$ plays an important role for the existence of roots of q_{N+1} in combination with the stability condition (2.2.16). Note that q_{N+1} has a zero at the point 1 according to Proposition 2.2.4 (iv). In this context, it seems natural to determine the function h_{N+1} with $q_{N+1}(z) = (1-z)h_{N+1}(z)$ for all $z > 0$.

Choose continuous functions $h_1, \dots, h_{N+1}$ on $[0,\infty)$ such that

$$\begin{aligned} q_n(z) &= z^{n-1} \prod_{i=0}^{n-1} \lambda_{1,i} + (1-z)h_n(z),\ n = 1, \dots, N, \\ q_{N+1}(z) &= (1-z)h_{N+1}(z). \end{aligned} \tag{2.2.33}$$

A system of recursive equations for $h_1(z), \dots, h_{N+1}(z)$ is given in the next proposition.

Proposition 2.2.5. *The functions $h_1, \ldots, h_{N+1}$ defined by* (2.2.33) *are given by $h_1(z) = 0$, $h_2(z) = -\lambda_{1,0}\mu_2$,*

$$\begin{aligned} h_n(z) = -(n-1)\mu_2 z^{n-2} \prod_{i=0}^{n-2} \lambda_{1,i} + \alpha_{n-1}(z) h_{n-1}(z) \\ - \lambda_{1,n-2}\mu_{1,n-1} z^2 h_{n-2}(z) \end{aligned} \tag{2.2.34}$$

for $n = 3, \ldots, N$ and

$$\begin{aligned} h_{N+1}(z) = (p\lambda_{1,N} z - N\mu_2) z^{N-1} \prod_{i=0}^{N-1} \lambda_{1,i} + \alpha_N(z) h_N(z) \\ - \lambda_{1,N-1}\mu_{1,N} z^2 h_{N-1}(z) \end{aligned} \tag{2.2.35}$$

for all $z \in [0, \infty)$.

Proof. Let $z > 0$ in the following and let the functions q_n, $n = 0, \ldots, N+1$, and h_n, $n = 1, \ldots, N+1$, be defined by (2.2.30) and (2.2.33), respectively, i.e.,

$$\begin{aligned} q_0(z) &= 1, \\ q_1(z) &= \alpha_0(z) q_0(z), \\ q_2(z) &= \alpha_1(z) q_1(z) - \lambda_{1,0}\mu_{1,1} z q_0(z) \quad \text{and} \\ q_n(z) &= \alpha_{n-1}(z) q_{n-1}(z) - \lambda_{1,n-2}\mu_{1,n-1} z^2 q_{n-2}(z) \end{aligned}$$

for $n = 3, \ldots, N+1$ and

$$\begin{aligned} q_n(z) &= z^{n-1} \prod_{i=0}^{n-1} \lambda_{1,i} + (1-z) h_n(z), \; n = 1, \ldots, N, \\ q_{N+1}(z) &= (1-z) h_{N+1}(z) \end{aligned}$$

for all $z \in [0, \infty)$. We will prove (2.2.34) by induction over n. We obtain immediately from the definition of q_1 and q_2 that $h_1(z) = 0$ and $h_2(z) = -\lambda_{1,0}\mu_2$. Now let $n = 3, \ldots, N$ and h_{n-1} and h_{n-2} be given by (2.2.34). By replacing $\alpha_{n-1}(z)$ given by (2.2.25) we get

$$\begin{aligned} q_n(z) &= \alpha_{n-1}(z) q_{n-1}(z) - \lambda_{1,n-2}\mu_{1,n-1} z^2 q_{n-2}(z) \\ &= \big((\lambda_{1,n-1} + \mu_{1,n-1}) z - (n-1)\mu_2(1-z)\big)\Big(z^{n-2} \prod_{i=0}^{n-2} \lambda_{1,i}\Big) \end{aligned}$$

$$
\begin{aligned}
&+ \alpha_{n-1}(z)(1-z)h_{n-1}(z) \\
&- \lambda_{1,n-2}\mu_{1,n-1}\Big(z^{n-1}\prod_{i=0}^{n-3}\lambda_{1,i} + (1-z)h_{n-3}(z)\Big) \\
= {}& z^{n-1}\prod_{i=0}^{n-1}\lambda_{1,i} + (1-z)\Big(\alpha_{n-1}(z)h_{n-1}(z) \\
&\qquad - (n-1)\mu_2 z^{n-2}\prod_{i=0}^{n-2} -\lambda_{1,n-2}\mu_{1,n-1}z^2 h_{n-2}(z)\Big) \\
= {}& z^{n-1}\prod_{i=0}^{n-1}\lambda_{1,i} + (1-z)h_n(z).
\end{aligned}
$$

The identity for $h_{N+1}(z)$ can be proved by inserting $q_{N-1}(z)$ and $q_N(z)$ into

$$q_{N+1}(z) = \alpha_N(z)q_N(z) - \lambda_{1,N-1}\mu_{1,N}z^2 q_{N-1}(z)$$

and simplifying. □

From Proposition 2.2.4 we obtain the main condition that relates the number of zeros of the function $\det(A(z))$ for $z \in (0,1)$ to $h_{N+1}(1)$. This condition will be shown to be the link between the number of zeros of $\det(A)$ and the stability condition (2.2.16).

Theorem 2.2.6. $\det(A(z))$ *is a polynomial of degree* $N+1$ *and has* $N-1$ *distinct zeros in the interval* $(0,1)$ *and one zero at* $z=1$*. Additionally,* $\det(A(z))$ *has another zero in the interval* $(1,\infty)$ *if and only if* $h_{N+1}(1) < 0$*, where* h_{N+1} *is defined by* $\det(A(z)) = (1-z)h_{N+1}(z)$ *for all* $z > 0$*.*

Proof. By (2.2.29) we have to prove that $q_{N+1}(z) = \det(A(z))$ has the claimed properties. Clearly, q_{N+1} has degree $N+1$. Without loss of generality, let N be odd in the following. The case of an even N can be handled analogously. By Proposition 2.2.4 (vi) the $N-1$ zeros $z_1,\dots,z_{N-1}$ of q_N can be labeled such that

$$0 < z_1 < \ldots < z_{N-1} < 1.$$

It follows from (2.2.30) and Proposition 2.2.4 (v) by induction that

$$\operatorname{sgn}(q_{N+1}(z_k)) = -\operatorname{sgn}(q_{N-1}(z_k)) = (-1)^{k+N} = (-1)^{k+1} \tag{2.2.36}$$

for $k = 1, \ldots, N-1$ since N is odd. Hence, by continuity and

$$\operatorname{sgn}(q_{N+1}(0)) = -\operatorname{sgn}(q_N(0))$$

(see Proposition 2.2.4 (iii)), there exist $N-1$ zeros $z_{N+1,1}, \ldots, z_{N+1,N-1}$ of q_{N+1} in $(0,1)$ with the interlacing property

$$0 < z_{N+1,1} < z_1 < z_{N+1,2} < \ldots < z_{N+1,N-1} < z_{N-1} < 1$$

regardless of the value of $h_{N+1}(1)$. $q_{N+1}(1) = 0$ was shown in Proposition 2.2.4 (iv). Now we show that q_{N+1} has one zero in $(1,\infty)$ if and only if $h_{N+1}(1) < 0$. By (2.2.36) and since q_{N+1} has at most $N+1$ zeros, the $(N+1)$-th zero has to be located in the interval (z_{N-1},∞). By Proposition 2.2.4 (vii) every other case would give a set of more than $N+1$ zeros which is impossible since q_{N+1} has degree $N+1$. Let q'_{N+1} be the derivative of q_{N+1}. Since $q_{N+1}(1) = 0$ by Proposition 2.2.4 (iv)) and

$$\operatorname{sgn}(q_{N+1}(z_{N-1})) = -\operatorname{sgn}(q_{N-1}(z_{N-1})) = (-1)^N = -1,$$

we must have $q'_{N+1}(1) > 0$. Each of the cases $q'_{N+1}(1) < 0$ and $q'_{N+1}(1) = 0$ would give a set of $N+1$ pairwise distinct zeros for q'_{N+1} which is impossible since q'_{N+1} is a polynomial of degree N. By $h_{N+1}(1) = -q'_{N+1}(1)$ we get that the $(N+1)$-th zero lies in $(1,\infty)$ if and only if $h_{N+1}(1) < 0$. We derive at the same condition in the case of an even N. □

By Theorem 2.2.6 we have to find $h_{N+1}(1)$. Evaluation of (2.2.34) and (2.2.35) at $z = 1$ gives a recursive system for $h_1(1), \ldots, h_{N+1}(1)$. We get $h_1(1) = 0$, $h_2(1) = -\lambda_{1,0}\mu_2$,

$$\begin{aligned} h_n(1) = -(n-1)\mu_2 \prod_{i=0}^{n-2} \lambda_{1,i} + (\lambda_{1,n-1} + \mu_{1,n-1}) h_{n-1}(1) \\ - \lambda_{1,n-2}\mu_{1,n-1} h_{n-2}(1) \end{aligned} \tag{2.2.37}$$

for $n = 3, \ldots, N$ and

$$\begin{aligned} h_{N+1}(1) = (p\lambda_{1,N} - N\mu_2) \prod_{i=0}^{N-1} \lambda_{1,i} + \mu_{1,N} h_N(1) \\ - \lambda_{1,N-1}\mu_{1,N} h_{N-1}(1). \end{aligned} \tag{2.2.38}$$

This system can be solved explicitly in terms of the system parameters. The solution is stated in the next proposition.

Proposition 2.2.7. *The functions $h_1, \dots, h_{N+1}$ defined by* (2.2.33) *satisfy*

$$h_n(1) = -\mu_2 \left(\prod_{i=0}^{n-1} \lambda_{1,i}\right) \left(\sum_{k=1}^{n-1} \frac{1}{\mu_{1,k+1}} \left(\prod_{i=0}^{k} \frac{\mu_{1,i+1}}{\lambda_{1,i}}\right) \left(\sum_{j=1}^{k} j \prod_{i=0}^{j-1} \frac{\lambda_{1,i}}{\mu_{1,i+1}}\right)\right) \tag{2.2.39}$$

for $n = 1, \dots, N$ and

$$h_{N+1}(1) = p \prod_{i=0}^{N} \lambda_{1,i} - \mu_2 \left(\prod_{i=0}^{N-1} \mu_{1,i+1}\right) \left(\sum_{n=1}^{N} n \prod_{i=0}^{n-1} \frac{\lambda_{1,i}}{\mu_{1,i+1}}\right), \tag{2.2.40}$$

where the empty sum and the empty product are defined to be 0 *and* 1*, respectively.*

Proof. (2.2.40) follows from (2.2.39) by inserting $h_N(1)$ and $h_{N-1}(1)$ into (2.2.38) and collecting terms. We will show (2.2.39) by induction over n using the recursion (2.2.37). Evaluating (2.2.39) for $n = 1$ and $n = 2$ we get $h_1(1) = 0$ and $h_2(1) = -\mu_2\lambda_{1,0}$. Let $n = 3, \dots, N$. We set

$$g_n = \sum_{k=1}^{n-1} \frac{1}{\mu_{1,k+1}} \left(\prod_{i=0}^{k} \frac{\mu_{1,i+1}}{\lambda_{1,i}}\right) \left(\sum_{j=1}^{k} j \prod_{i=0}^{j-1} \frac{\lambda_{1,i}}{\mu_{1,i+1}}\right)$$

so that $h_n(1) = -\mu_2 g_n \prod_{i=0}^{n-1} \lambda_{1,i}$. Let $h_{n-1}(1)$ and $h_{n-2}(1)$ be given by (2.2.39). Substituting these terms in (2.2.37) yields

$$\begin{aligned} h_n(1) &= -(n-1)\mu_2 \prod_{i=0}^{n-2} \lambda_{1,i} - \mu_2(\lambda_{1,n-1} + \mu_{1,n-1}) g_{n-1} \prod_{i=0}^{n-2} \lambda_{1,i} \\ &\quad + \mu_2 \lambda_{1,n-2} \mu_{1,n-1} g_{n-2} \prod_{i=0}^{n-3} \lambda_{1,i} \\ &= -\mu_2 \left(g_{n-1} + (n-1)\frac{1}{\lambda_{1,n-1}} + \frac{\mu_{1,n-1}}{\lambda_{1,n-1}}(g_{n-1} - g_{n-2})\right) \prod_{i=0}^{n-1} \lambda_{1,i}. \end{aligned} \tag{2.2.41}$$

Due to the additive structure of g_n we have

$$\frac{\mu_{1,n-1}}{\lambda_{1,n-1}}(g_{n-1} - g_{n-2}) = \frac{1}{\lambda_{1,n-1}} \left(\prod_{i=0}^{n-2} \frac{\mu_{1,i+1}}{\lambda_{1,i}}\right) \left(\sum_{j=1}^{n-2} j \prod_{i=0}^{j-1} \frac{\lambda_{1,i}}{\mu_{1,i+1}}\right)$$

$$= \frac{1}{\mu_{1,n}} \left(\prod_{i=0}^{n-1} \frac{\mu_{1,i+1}}{\lambda_{1,i}} \right) \left(\sum_{j=1}^{n-2} j \prod_{i=0}^{j-1} \frac{\lambda_{1,i}}{\mu_{1,i+1}} \right). \qquad (2.2.42)$$

Writing

$$(n-1)\frac{1}{\lambda_{1,n-1}} = \frac{1}{\mu_{1,n}} \left(\prod_{i=0}^{n-1} \frac{\mu_{1,i+1}}{\lambda_{1,i}} \right) \left((n-1) \prod_{i=0}^{n-2} \frac{\lambda_{1,i}}{\mu_{1,i+1}} \right),$$

the equation (2.2.42) gives

$$(n-1)\frac{1}{\lambda_{1,n-1}} + \frac{\mu_{1,n-1}}{\lambda_{n-1}}(g_{n-1} - g_{n-2})$$
$$= \frac{1}{\mu_{1,n}} \left(\prod_{i=0}^{n-1} \frac{\mu_{1,i+1}}{\lambda_{1,i}} \right) \left(\sum_{j=1}^{n-1} j \prod_{i=0}^{j-1} \frac{\lambda_{1,i}}{\mu_{1,i+1}} \right).$$

Substituting this in (2.2.41) and once again exploiting the additive structure of g_n we arrive at (2.2.39). □

We can derive $h_{N+1}(1)$ also in closed form in terms of the system parameters and $p_{1,0}, \dots, p_{N,0}$. First we determine $G_n(1)$ for $n = 0, \dots, N$ in terms of $G_0(1)$. Setting $z = 1$ in (2.2.20)-(2.2.22) gives

$$\lambda_{1,0}G_0(1) = \mu_{1,1}G_1(1), \qquad (2.2.43)$$
$$(\lambda_{1,n} + \mu_{1,n})G_n(1) = \lambda_{1,n-1}G_{n-1}(1) + \mu_{1,n+1}G_{n+1}(1) \quad \text{and} \qquad (2.2.44)$$
$$\mu_{1,N}G_N(z) = \lambda_{1,N-1}G_{N-1}(1), \qquad (2.2.45)$$

where $n = 1, \dots, N-1$ in the second equation. The solution of these equations in terms of $G_0(1)$ is clearly

$$G_n(1) = G_0(1) \prod_{i=0}^{n-1} \rho_i \text{ for } n = 0, \dots, N, \qquad (2.2.46)$$

where we have written $\rho_i = \lambda_{1,i}/\mu_{1,i+1}$ for $i = 0, \dots, N-1$. By

$$\sum_{n=0}^{N} \sum_{m \geq 0} p_{n,m} = \sum_{n=0}^{N} G_n(1) = 1,$$

summation over $n = 0, \ldots, N$ yields

$$G_0(1) = \left(\sum_{n=0}^{N} \prod_{i=0}^{n-1} \rho_i \right)^{-1} . \tag{2.2.47}$$

This can also be deduced directly from (2.2.1). By (2.2.14), (2.2.46) and since $P(L_1 = N) = G_N(1)$, $EL_1 = \sum_{n=1}^{N} nG_n(1)$ and

$$P(L_2 = 0)E(L_1|L_2 = 0) = \sum_{n=1}^{N} np_{n,0}$$

hold, we get

$$p\lambda_{1,N} G_N(1) = \mu_2 \left(G_0(1) \sum_{n=1}^{N} n \prod_{i=0}^{n-1} \rho_i - \sum_{n=1}^{N} np_{n,0} \right) .$$

By (2.2.46) the latter equation gives

$$\mu_2 \sum_{n=1}^{N} np_{n,0} = G_0(1) \left(\mu_2 \sum_{n=1}^{N} n \prod_{i=0}^{n-1} \rho_i - p\lambda_{1,N} \prod_{i=0}^{N-1} \rho_i \right) . \tag{2.2.48}$$

After rearranging terms and using (2.2.47) we get

$$h_{N+1}(1) = -\mu_2 \left(\prod_{n=0}^{N-1} \mu_{i+1} \right) \left(\sum_{n=0}^{N} \prod_{i=0}^{n-1} \rho_i \right) \sum_{n=1}^{N} np_{n,0} \tag{2.2.49}$$

from (2.2.40). By setting the right side of equation (2.2.40) and equation (2.2.49) equal, we get again the equation (2.2.14) in terms of $\sum_{n=1}^{N} np_{n,0}$ and the system parameters. This equation can also be derived from (2.2.40) and (2.2.48). The following proposition states the result.

Proposition 2.2.8. *Let a solution $p_{n,m}$ for $n = 0, \ldots, N$ and $m \geq 0$ of the equations* (2.2.3) *be given. Then the condition $\sum_{n=1}^{N} \sum_{m \geq 0} p_{n,m} = 1$ implies the equivalence of* (2.2.14) *and*

$$\sum_{n=1}^{N} np_{n,0} = \frac{\mu_2 \sum\limits_{n=1}^{N} n \prod\limits_{i=0}^{n-1} \rho_i - p\lambda_{1,N} \prod\limits_{i=0}^{N-1} \rho_i}{\mu_2 \sum\limits_{n=0}^{N} \prod\limits_{i=0}^{n-1} \rho_i} > 0. \tag{2.2.50}$$

Proposition 2.2.8 provides a useful numerical check for the stationary probabilities. Now we give equivalent formulations for the stability criterion

(2.2.16).

Proposition 2.2.9. *Let $\rho_n = \lambda_{1,n}/\mu_{1,n+1}$ for $n = 0, \dots, N-1$. The following conditions are equivalent:*

(i) The system of equations (2.2.3) *has a unique nonnegative and normalized solution.*

(ii) $h_{N+1}(1) < 0$ holds, where $h_{N+1}(1)$ is given by (2.2.40).

(iii) $p\lambda_1 P(L_1 = N) < \mu_2 EL_1$ holds, where $P(L_1 = N)$ and EL_1 are given by (2.2.1).

(iv) The system of equations (2.2.3) *has a solution with $\sum_{n=1}^{N} p_{n,0} > 0$.*

Proof. The equivalence of (i) and (ii) was shown in Proposition 2.2.1. (iii) is a reformulation of (ii). Also by Proposition 2.2.1, the system is irreducible and positive recurrent under condition (iii). We obtain immediately that $p_{n,m} > 0$ for $n = 0, \dots, N$ and $m \geq 0$, and hence (iv) follows.

We show now that (iv) implies (ii). Let (iv) hold. Note that it cannot be assumed that the Markov chain defined by the equations (2.2.3) is positive recurrent and hence the stationary measure (the solution of (2.2.3)) can have an infinite mass. Therefore, we cannot use the equations (2.2.13), (2.2.14) or (2.2.50) under condition (iv) in order to ensure (iii). Let the functions $\alpha_n(z)$ and $G_n(z)$, $n = 0, \dots, N$, be defined by (2.2.24)-(2.2.26) and (2.2.28), respectively, for $|z| \leq 1$. This ensures that (2.2.43)-(2.2.45) hold and that the functions $q_n(z)$ and $h_n(z)$, $n = 1, \dots, N+1$, are well defined for $|z| \leq 1$. By (2.2.28), (2.2.29) and (2.2.33) we get

$$\begin{aligned} \det(A_0(z)) &= \det(A(z))G_0(z) = q_{N+1}(z)G_0(z) \\ &= (1-z)h_{N+1}(1)G_0(z) \end{aligned} \tag{2.2.51}$$

for all $|z| \leq 1$. By writing down a recursive formula for $\det(A_0(z))$ by means of Cramer's rule and performing an inductive argument, it can be shown that

$$\left.\frac{\det(A_0(z))}{1-z}\right|_{z=1} = -\mu_2 \left(\prod_{n=0}^{N-1} \mu_{i+1}\right) \sum_{n=1}^{N} n p_{n,0}.$$

The calculations are straightforward but tedious and hence omitted. Using the above and (2.2.51) yields

$$G_0(1)h_{N+1}(1) = -\mu_2 \left(\prod_{n=0}^{N-1} \mu_{i+1}\right) \sum_{n=1}^{N} n p_{n,0}. \tag{2.2.52}$$

Together with $G_0(1) > 0$ and $\sum_{n=1}^{N} p_{n,0} > 0$ this shows that $h_{N+1}(1) < 0$ and therefore (ii) holds. □

The following consequence of Proposition 2.2.9 and the irreducibility of the system might be well known but is interesting to notice.

Corollary 2.2.10. *The system of equations* (2.2.3) *has a unique normalized solution if and only if it is stable. In this case $p_{n,m} > 0$ holds for all $n = 0, \dots, N$ and all $m \geq 0$. The system is unstable if and only if $p_{n,m} = 0$, $n = 0, \dots, N$, $m \geq 0$, is the only solution.*

Setting $h_{N+1}(1) < 0$ in (2.2.40) and simplifying gives the stability condition

$$p\lambda_{1,N} \prod_{n=0}^{N-1} \rho_n < \mu_2 \sum_{n=1}^{N} n \prod_{i=0}^{n-1} \rho_i. \tag{2.2.53}$$

The following statement links the number of zeros of $\det(A)$ to the stability condition and is an immediate consequence of Theorem 2.2.6 and Proposition 2.2.9.

Theorem 2.2.11. $\det(A(z))$ *is a polynomial of degree $N+1$ and has $N-1$ zeros in the interval* $(0,1)$ *and one zero at $z = 1$. Additionally,* $\det(A(z))$ *has another zero in the interval* $(1, \infty)$ *if and only if the system* (2.2.3) *is stable, i.e., if and only if*

$$\frac{p\lambda_{1,N}}{\mu_2} < \frac{\sum\limits_{n=1}^{N} n \prod\limits_{i=0}^{n-1} \rho_i}{\prod\limits_{n=0}^{N-1} \rho_n} \tag{2.2.54}$$

holds, where $\rho_n = \lambda_{1,n}/\mu_{1,n+1}$, $n = 0, \dots, N-1$.

Remark 2.2.12. By the condition given in Theorem 2.2.11, when the system is stable, meaning (2.2.54) holds, we can use the $N-1$ zeros $z_1, \dots, z_{N-1}$ of $\det(A(z)) = q_{N+1}(z)$ in $(0,1)$ to find the N unknown probabilities $p_{1,0}, \dots, p_{N,0}$. This can be done by inserting these values into (2.2.28) for $n = 0$. The $N-1$ pairwise distinct zeros then deliver $N-1$ equations for these unknowns, namely

$$\det(A_0(z_1)) = 0, \dots, \det(A_0(z_{N-1})) = 0.$$

One more equation relating the unknowns is (2.2.14) which is equivalent to (2.2.50). This provides us with N (independent) equations in the N

unknowns $p_{1,0},\ldots,p_{N,0}$. Observe that $\det(A_0(z_i))$ and $\det(A_n(z_i))$ for $i = 0,\ldots,N$ differ from each other by (2.2.28) only by a multiplicative constant, i.e.,

$$\det(A_0(z_i)) = \frac{G_n(z_i)}{G_0(z_i)} \det(A_n(z_i))$$

for $i = 0,\ldots,N$

A crucial point is to show that these equations are indeed linearly independent. For $N = 2,3$ it is possible to show this analytically. Numerical calculations support this conjecture for larger values of N. It is conjectured that these equations are indeed independent (see also Avi-Itzhak and Mitrani [6], Levy and Yechiali [42], Perel and Yechiali [56], Yechiali [66]).

2.2.4 Stationary quantities and numerical aspects

Once $p_{1,0},\ldots,p_{N,0}$ are determined, $G_0,\ldots,G_N$ are given and we can calculate EL_2, the mean number of customers in Q_2, by (2.2.14), i.e.,

$$EL_2 = \sum_{n=0}^{N} G_n'(1). \tag{2.2.55}$$

Furthermore, $\mathrm{Cov}(L_1, L_2)$ is of special interest because of the dependence of Q_2 on Q_1. By summing (2.2.20)-(2.2.22) over $n = 0,\ldots,N$ we get

$$p\lambda_{1,N} z G_N(1) - \mu_2 \sum_{n=1}^{N} G_n(z) = -\mu_2 \sum_{n=1}^{N} n p_{n,0}. \tag{2.2.56}$$

Differentiating this equation with respect to z at $z = 1$ leads to

$$p\lambda_{1,N}(G_N(1) + G_N'(1)) = \mu_2 \sum_{n=1}^{N} n G_n'(1)$$

and thus

$$\mu_2 E(L_1 L_2) = p\lambda_{1,N}(G_N(1) + G_N'(1))$$

since $\sum_{n=1}^{N} n G_n'(1) = E(L_1 L_2)$. The latter equation is equivalent to

$$\mu_2 E(L_1 L_2) = p\lambda_{1,N}\left(P(L_1 = N) + E(L_2|L_1 = N)P(L_1 = N)\right) \tag{2.2.57}$$

since $G_N(1) = P(L_1 = N)$ and $G_N'(1) = E(L_2|L_1 = N)P(L_1 = N)$. Observe that $P(L_1 = N)$ is known.

Remark 2.2.13. Numerical analysis shows that in general the covariance of L_1 and L_2 changes signs when the system parameters are varied. The numerical analysis rather suggests that $\mathrm{Cov}(L_1, L_2)$ as a function of p on the interval $[0, 1]$ is convex or concave and shows that it can be monotone decreasing, increasing or both with one or two zeros (where one zero is at $p = 0$). Furthermore, there exists a threshold parameter p^* for the overflow weight p that separates the case of positive and nonnegative correlation from each other as one might expect. See Figure 2.3 for the shape of the function $p \mapsto \mathrm{Cov}(L_1, L_2)$ for $N = 2$ with $\lambda_{1,0} = \lambda_{1,1} = \lambda_1$, $\mu_{1,1} = \mu_{1,2} = \mu_1$ and various parameter selections. An interesting case is $N = 1$, where we can show that $\mathrm{Cov}(L_1, L_2) = 0$ for all choices of these system parameters. Q_1 and Q_2 are in general not independent for $N = 1$. The analysis of this case is carried out in the next section.

Further numerical analysis suggests that $\mathrm{Cov}(L_1, L_2) \geq 0$ for every choice of the system parameters for which the stability condition (2.2.54) holds and for which Q_1 is stable, meaning $\prod_{n=0}^{N-1} \rho_n < 1$ (see Figure 2.3 and Table 2.1). This can be explained intuitively. Due to the overflow mechanism, the queue length in Q_2 can on the one hand only increase when the queue length in Q_1 increases and Q_1 reaches its capacity maximum. On the other hand, the service rate in Q_2 increases and therefore the queue length in Q_2 decreases simultaneously. Suppose that these two effects would overall lead to a decreasing queue length in Q_2. In this case it would be possible to raise $p\lambda_{1,N}$ slightly without leaving the stable regime in Q_2 and without affecting the queue length in Q_1. This would lead to a rising queue length in Q_2 and to $\mathrm{Cov}(L_1, L_2) > 0$, which would contradict the assumption. Consequently, $\mathrm{Cov}(L_1, L_2)$ must be nonnegative if the system and Q_1 are stable. Due to the mentioned “stability reserve”, this result will also hold if Q_1 is slightly unstable. This intuitive argument can be reversed in the case of a highly unstable Q_1, i.e., $\prod_{n=0}^{N-1} \rho_n >> 1$. In this case, the first queue is on average fully occupied and the probability of overflow is close to 1. Therefore, the queue length in Q_2 has a tendency to shorten if the queue length in Q_1 rises because otherwise Q_2 would become unstable.

In the case that (2.2.54) and $\prod_{n=0}^{N-1} \rho_n < 1$ hold, equation (2.2.57) provides an upper bound for EL_2 in terms of the unknowns $p_{1,0}, \ldots, p_{N,0}$. (2.2.54) and $\prod_{n=0}^{N-1} \rho_n < 1$ imply that $\mathrm{Cov}(L_1, L_2) \geq 0$ and therefore $E(L_1 L_2) \geq EL_1 EL_2$. Then (2.2.57) yields

$$EL_2 \leq \frac{p\lambda_{1,N} P(L_1 = N)}{\mu_2 EL_1}(1 + E(L_2 | L_1 = N)), \tag{2.2.58}$$

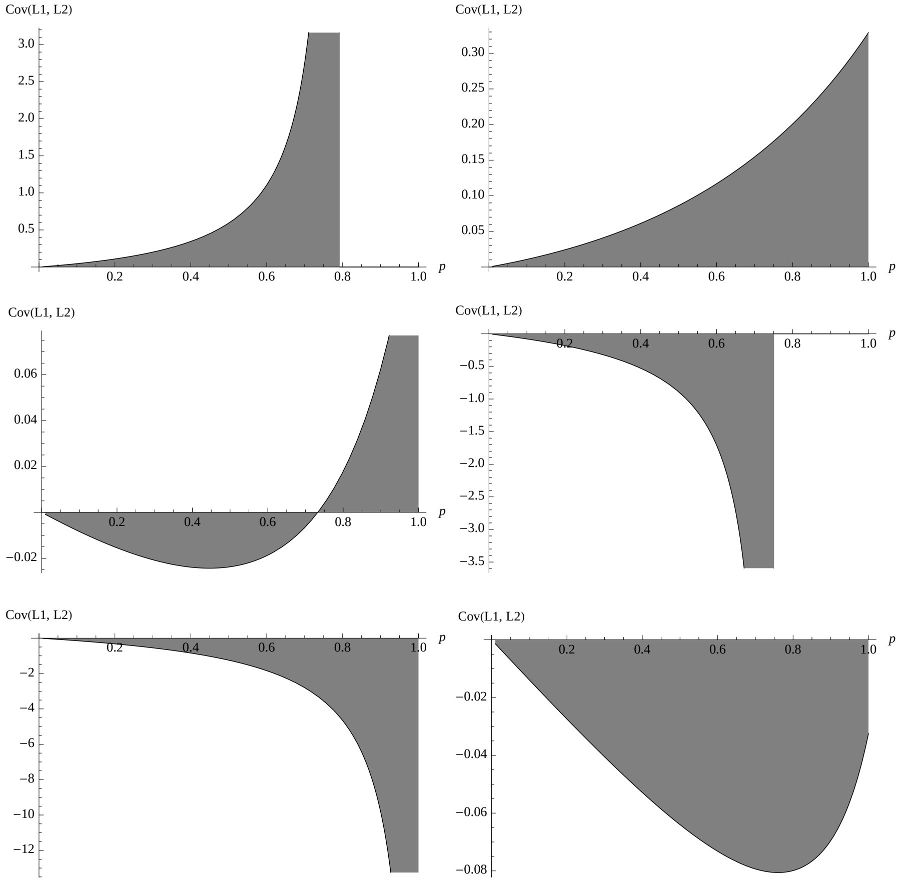


Figure 2.3: $p \mapsto \mathrm{Cov}(L_1, L_2)$ for $N = 2$, $\lambda_{1,0} = \lambda_{1,1} = \lambda_1$, $\mu_{1,1} = \mu_{1,2} = \mu_1$ and $(\lambda_1, \mu_1, \mu_2) = (2.5, 4, 0.5)$, $(3.8, 4, 2.5)$, $(5.3, 4, 2.5)$, $(4, 2, 1.2)$, $(5.4, 1, 2.5)$, $(5.6, 4, 2.2)$ (line by line from left to right, the gray shaded region is the stability region, i.e., all p that satisfy (2.2.19)).

where $P(L_1 = N)$ and EL_1 are known and

$$E(L_2|L_1 = N)P(L_1 = N) = G'_N(1)$$

has to be determined. The additional benefit of this bound compared to formula (2.2.55) is that there is only one unknown quantity, namely $G'_N(1)$. In case of heavy traffic for the first queue, i.e., $\prod_{n=0}^{N-1} \rho_n \to 1$, the first queue is always occupied implying $P(L_1 = N) = 1$ and $EL_2 = E(L_2|L_1 = N)$. Then (2.2.58) gives the bound

$$EL_2 \leq \frac{p\lambda_{1,N}}{N\mu_2 - p\lambda_{1,N}}, \tag{2.2.59}$$

λ_1	$\mu_1 = 0.1$	$\mu_1 = 0.2$	$\mu_1 = 0.3$	$\mu_1 = 0.4$	$\mu_1 = 0.5$	$\mu_1 = 0.6$	$\mu_1 = 0.7$
0.1	**0.02**	0.056	0.051	0.041	0.033	0.027	0.022
0.2	−0.061	**0.038**	0.069	0.074	0.07	0.063	0.056
0.3	−0.152	−0.013	**0.055**	0.083	0.092	0.091	0.087
0.4	−0.259	−0.085	0.018	**0.072**	0.098	0.109	0.111
0.5	−0.385	−0.176	−0.04	0.041	**0.088**	0.113	0.125
0.6	−0.532	−0.287	−0.118	−0.008	0.062	**0.105**	0.129
0.7	−0.705	−0.421	−0.217	−0.075	0.019	0.082	**0.122**
0.8	−0.908	−0.58	−0.338	−0.164	−0.041	0.043	0.101
0.9	−1.148	−0.77	−0.485	−0.274	−0.122	−0.012	0.066
1	−1.436	−0.998	−0.663	−0.412	−0.224	−0.086	0.014
1.1	−1.787	−1.275	−0.882	−0.581	−0.354	−0.183	−0.055
1.2	−2.221	−1.617	−1.151	−0.792	−0.517	−0.307	−0.148
1.3	−2.774	−2.048	−1.488	−1.056	−0.722	−0.465	−0.267
1.4	−3.498	−2.606	−1.921	−1.392	−0.984	−0.668	−0.422
1.5	−4.487	−3.353	−2.493	−1.833	−1.325	−0.931	−0.625
1.6	−5.918	−4.405	−3.28	−2.431	−1.782	−1.281	−0.894
1.7	−8.171	−5.991	−4.43	−3.281	−2.42	−1.764	−1.261
1.8	−12.237	−8.655	−6.261	−4.582	−3.366	−2.464	−1.783
1.9	−21.772	−14.05	−9.622	−6.809	−4.903	−3.556	−2.574
2	−70.383	−30.757	−17.789	−11.477	−7.82	−5.485	−3.899
2.1			−66.638	−27.383	−15.431	−9.77	−6.54
2.2					−84.64	−27.299	−14.299
2.3							−426.843
2.4							

Table 2.1: $\mathrm{Cov}(L_1, L_2)$ for $N = 2$, $\lambda_{1,0} = \lambda_{1,1} = \lambda_1$, $\mu_{1,1} = \mu_{1,2} = \mu_1$, $p = \mu_2 = 1$, $\lambda_1 = 0.1, 0.2 \ldots, 2.4$ and $\mu_1 = 0.1, 0.2 \ldots, 0.7$ (bold numbers mark the diagonal $\lambda_1 = \mu_1$, the cells are left blank in the cases where the system is unstable).

which is tight since there are always N customers present in Q_1 serving Q_2 with rate $N\mu_2$ and the arrival rate of Q_2 is $p\lambda_{1,N}$. In the case $N = 1$, equality also holds in (2.2.59) for arbitrary choice of the system parameters. This is shown in Section 2.2.6.

2.2.5 Model extension: External arrivals to second queue

The model can be generalized by equipping Q_2 with an arrival stream having exponentially distributed interarrival times with intensity $\lambda_2 > 0$. The arrival stream is independent of Q_1 and the arrival stream of Q_2 (see Figure 2.2 for the corresponding transition rate diagram). In this case, the generating functions (2.2.20)-(2.2.22) and the auxiliary α-functions (2.2.24)-(2.2.26) are given by

$$(\lambda_{1,0} + \lambda_2(1 - z))G_0(z) = \mu_{1,1}G_1(z),$$

$$\begin{aligned}\big((\lambda_{1,n}+\mu_{1,n})z+(\lambda_2 z-n\mu_2)(1-z)\big)G_n(z) &= \lambda_{1,n-1}zG_{n-1}(z)\\ &\quad+\mu_{1,n+1}zG_{n+1}(z)-n\mu_2(1-z)p_{n,0},\\ \big(\mu_{1,N}z+((p\lambda_{1,N}+\lambda_2)z-N\mu_2)(1-z)\big)G_N(z) &= \lambda_{1,N-1}zG_{N-1}(z)\\ &\quad-N\mu_2(1-z)p_{N,0},\end{aligned}$$

where the second equation holds for $n = 1,\dots,N-1$ and

$$\begin{aligned}\alpha_0(z) &= \lambda_{1,0}+\lambda_2(1-z),\\ \alpha_n(z) &= (\lambda_{1,n}+\mu_{1,n})z+(\lambda_2 z-n\mu_2)(1-z) \text{ for } n=1,\dots,N-1,\\ \alpha_N(z) &= \mu_{1,N}z+((p\lambda_{1,N}+\lambda_2)z-N\mu_2)(1-z).\end{aligned}$$

A procedure similar to the one for the basic model can be carried out, showing that $\det(A(z))$ is a polynomial of degree $2N+1$ with $N-1$ zeros in $(0,1)$, one zero at $z = 1$ and all other zeros in $(1,\infty)$ if and only if $h_{N+1}(1) < 0$, where

$$h_{N+1}(1) = p\prod_{i=0}^{N}\lambda_{1,i}-\left(\prod_{i=0}^{N-1}\mu_{1,i+1}\right)\left(\mu_2\sum_{n=1}^{N}n\prod_{i=0}^{n-1}\rho_i-\lambda_2\left(1+\sum_{n=1}^{N}\prod_{i=0}^{n-1}\rho_i\right)\right)$$

with $\rho_i = \lambda_{1,i}/\mu_{1,i+1}$ for $i = 0,\dots,N-1$. The stability condition $h_{N+1}(1) < 0$ (see also (1.2.1)) is equivalent to

$$\frac{p\lambda_{1,N}\prod\limits_{n=0}^{N-1}\rho_n}{\mu_2\sum\limits_{n=1}^{N}n\prod\limits_{i=0}^{n-1}\rho_i}+\frac{\lambda_2}{\mu_2 EL_1}<1$$

or $p\lambda_{1,N}P(L_1 = N) + \lambda_2 < \mu_2 EL_1$ where $P(L_1 = N)$ and EL_1 are given by (2.2.1). Setting $p = 0$, the stability condition becomes $\lambda_2 < \mu_2 EL_1$ which corresponds to the result in [56]. By letting $p = 0$, $\lambda_{1,n} = \lambda_1$ and $\mu_{1,n+1} = \mu_2$ for $n = 0,\dots,N-1$ in our model, both models coincide.

Remark 2.2.14. It is possible to further generalize the model by assuming that the service rate in Q_2 is $\mu_{2,n}$ (instead of $n\mu_2$) when Q_1 is in state n as done in Section 2.2.2 and suggested in Perel and Yechiali [56]. It has to be assumed that $\mu_{2,n} \geq 0$ for all $n \geq 0$ and $\mu_{2,n} > 0$ for at least one value $n = 1,\dots,N$. The stability condition in this case is given by (2.2.17). This will lead to a modified construction of the generating functions, the auxiliary functions and the stability condition in this sequel (we omit the details). In order to proceed and investigate the existence and number of zeros of $\det(A(z))$, it is necessary to specify $\mu_{2,n}$.

2.2.6 Closed-form solution in the case of capacity one

For the case $N = 1$, an analytic solution is available. While the case $N = 2$ is also analytically solvable, the solutions are lengthy and we focus on the case $N = 1$. Let $\lambda_{0,0}, \lambda_{1,0}, \mu_2 > 0$, $\mu_{1,1} = \mu_1 > 0$ and $p \in [0,1]$. For simplicity, we choose $\lambda_{0,0} = \lambda_{1,0} = \lambda_1$; by letting $p \in [0,\infty)$ the adequate choice of the parameter p can be used to treat the case $\lambda_{1,0} \neq \lambda_{1,1}$ because $p \in [0,1]$ is no necessary condition. The balance equations are

$$\begin{aligned} \lambda_1 p_{0,m} &= \mu_1 p_{1,m} \text{ for } n=0, m \geq 0, \qquad (2.2.60)\\ (p\lambda_1 + \mu_1)p_{1,0} &= \lambda_1 p_{0,0} + \mu_2 p_{1,1} \text{ for } n=1, m=0 \text{ and}\\ (p\lambda_1 + \mu_1 + \mu_2)p_{1,m} &= \lambda_1 p_{0,m} + \mu_2 p_{1,m+1} + p\lambda_1 p_1 \text{ for } n=1 \text{ and } m \geq 1. \end{aligned}$$

The equation (2.2.14) yields $p\lambda_1 P(L_1 = 1) = \mu_2(EL_1 - p_{1,0})$ and from

$$P(L_1 = 1) = EL_1 = \frac{\lambda_1}{\lambda_1 + \mu_1}$$

and (2.2.60) we get

$$\begin{aligned} p_{1,0} &= \frac{\mu_2 - p\lambda_1}{\mu_2} \cdot EL_1 = \frac{\lambda_1}{\lambda_1+\mu_1} \cdot \frac{\mu_2 - p\lambda_1}{\mu_2} \quad \text{and} \qquad (2.2.61)\\ p_{0,0} &= \frac{\mu_1}{\lambda_1} p_{1,0} = \frac{\mu_1}{\lambda_1+\mu_1} \cdot \frac{\mu_2 - p\lambda_1}{\mu_2}. \qquad (2.2.62) \end{aligned}$$

The generating functions G_0 and G_1 satisfy

$$\begin{aligned} \lambda_{1,0} G_0(z) &= \mu_1 G_1(z) \quad \text{and}\\ (z\mu_1 + (zp\lambda_1 - \mu_2)(1-z))G_1(z) &= z\lambda_1 G_0(z) - \mu_2 p_{1,0}(1-z) \end{aligned}$$

for $|z| \leq 1$. These equations give

$$G_1(z) = \frac{z\lambda_1 G_0(z) - \mu_2 p_{1,0}(1-z)}{z\mu_1 + (zp\lambda_1 - \mu_2)(1-z)} = \frac{z\mu_1 G_1(z) - \mu_2 p_{1,0}(1-z)}{z\mu_1 + (zp\lambda_1 - \mu_2)(1-z)}$$

which yields

$$\begin{aligned} G_1(z) &= \frac{\mu_2 p_{1,0}(1-z)}{(\mu_2 - zp\lambda_1)(1-z)} = \frac{\mu_2 - p\lambda_1}{\mu_2 - zp\lambda_1} \cdot \frac{\lambda_1}{\lambda_1+\mu_1} \quad \text{and}\\ G_0(z) &= \frac{\mu_1}{\lambda_1} G_1(z) = \frac{\mu_2 - p\lambda_1}{\mu_2 - zp\lambda_1} \cdot \frac{\mu_1}{\lambda_1+\mu_1}. \end{aligned}$$

Observe that the stability condition is $p\lambda_1 < \mu_2$ by Proposition 2.2.1, where $\lambda_1 = \lambda_{1,1}$ if the condition $\lambda_{1,0} = \lambda_{1,1} = \lambda_1$ is dropped. Then, $\mu_2 - zp\lambda_1 \neq 0$ for all $|z| \leq 1$. Differentiating and setting $z = 1$ leads to

$$\begin{aligned} G_0'(1) &= \frac{p\lambda_1\mu_1}{(\lambda_1+\mu_1)(\mu_2 - p\lambda_1)} \qquad \text{and} \\ G_1'(1) &= \frac{p\lambda_1^2}{(\lambda_1+\mu_1)(\mu_2 - p\lambda_1)}. \end{aligned}$$

Finally, we can calculate the expected queue length of the second queue:

$$EL_2 = G_0'(1) + G_1'(1) = \frac{p\lambda_1}{\mu_2 - p\lambda_1}.$$

Since $E(L_1L_2) = G_1'(1)$, we get

$$E(L_1L_2) = \frac{\lambda_1}{\lambda_1+\mu_1} \cdot \frac{p\lambda_1}{\mu_2 - p\lambda_1} = EL_1EL_2.$$

Surprisingly on first sight, we get $\mathrm{Cov}(L_1, L_2) = 0$ although L_1 and L_2 are in general not independent. The service rate in Q_2 for example equals 0 for $L_1 = 0$ and μ_2 for $L_1 = 1$. An explanation for this phenomenon is the following. Under stationary conditions, Q_1 is in state 1 for a fraction $\lambda_1/(\lambda_1 + \mu_1)$ of the time. In this case, Q_2 is busy for a fraction $p\lambda_1/\mu_2$ of the time. Since Q_2 is stable and only fed with customers or served if $L_1 = 1$ and remains unchanged if $L_1 = 0$, the queue length has to stay essentially the same and is not influenced by the queue length of Q_1. As mentioned in Remark 2.2.13 $\mathrm{Cov}(L_1, L_2) = 0$ does not hold in general for $N > 1$ because in this case the service rate in Q_2 is positive for $L_1 \geq 1$, whereas the arrival rate in Q_2 is positive only if $L_1 = N$. The service rate in Q_2 increases linearly with the queue length in Q_1 and the arrival rate of Q_2 is zero except in the case $L_1 = N$. Therefore, the queue length in Q_2 tends to shorten with increasing L_1 up to the point $L_1 = N$ where customers arrive to Q_2. In this case, the queue length in Q_2 can increase or decrease or stay unchanged in equilibrium, depending on the system parameters and the combination of these factors. This leads to a generally almost unpredictable behavior of $\mathrm{Cov}(L_1, L_2)$ (see also Remark 2.2.13).

2.3 Customers as servers: Models with jockeying

In this section, we discuss two variants of the basic model from Section 2.2. We let the assumptions be as for the basic model from Section 2.2 with the following exception. In the first variant, the first customer of Q_2, if one is present, is forced to move to Q_1 as soon as Q_1 empties. This jockeying customer can then act as a server for the second queue. The stochastic processes that describes the queue lengths in this model is a quasi birth and death process (QBP). This model can be further generalized by letting a fixed number $1 \leq k \leq N-1$ jockey from Q_2 to Q_1 if Q_1 becomes empty (see Remark 2.3.9). In this case, it cannot be represented as a QBP because the transitions from states in the level process, i.e., the number of customers in Q_2, are not restricted to transitions to the states in the two adjacent levels. In the second variant of the basic model, as soon as Q_1 empties, Q_1 is filled with the customers from Q_2 until it reaches its capacity bound or Q_2 empties. We called this jockeying procedure *unlimited jockeying*. The stochastic processes that describes the queue lengths in this model is again a QBP.

The approach for reducing the number of unknowns in the first model can be carried over to these model variations. In turn, the steady-state probabilities of the vector of the queue lengths are functions of $p_{1,0}, \ldots, p_{N,0}$ and we can give a set of N (independent) equations for these unknowns. Steady-state quantities of interest can be computed by means of the generating functions. For the model with unlimited jockeying a complete analytic solution is available, since in this case, Q_1 and Q_2 can be regarded as a special case of a single queue with infinite capacity and state-dependent service and arrival rates. The analysis of this model is carried out in Section 2.3.4.

2.3.1 Model description and steady-state equations

Let Q_1 be as in the basic model and let $N \geq 2$. We consider the first variant of the basic model. In this variant, the first customer of Q_2, if one is present, is forced to move to Q_1 as soon as Q_1 empties. The case $N = 1$ is equivalent to the second model variant. Let Q_2 be fed by the p-weighted overflow stream from Q_1, where $p \in [0, 1]$, and let the service rate in Q_2 be $n\mu_2$ if $L_1 = n$ where $\mu_2 > 0$. Modify Q_1 in the following way: If Q_1 becomes empty, then the first customer of Q_2 is instantly transferred to Q_1. Let $p_{n,m}$ be the steady-state probability of having n customers present in Q_1 and m customers in Q_2, $n = 0, \ldots, N$, $m \geq 0$. We must have $p_{0,m} = 0$ for $m \geq 1$, because in this case, one customer from Q_2 will be transferred instantly to Q_1. Let δ_{ij} be the Kronecker function. The balance equations for this model for $n = 0, \dots, N$ and $m = 0$ are

$$\lambda_{1,0}p_{0,0} = \mu_{1,1}p_{1,0}, \tag{2.3.1}$$

$$(\lambda_{1,n} + \mu_{1,n})p_{n,0} = \lambda_{1,n-1}p_{n-1,0} + \mu_{1,n+1}p_{n+1,0} + (n\mu_2 + \delta_{1n}\mu_{1,1})p_{n,1} \quad \text{for } n = 1, \dots, N-1 \text{ and} \tag{2.3.2}$$

$$(p\lambda_{1,N} + \mu_{1,N})p_{N,0} = \lambda_{1,N-1}p_{N-1,0} + N\mu_2 p_{N,1}. \tag{2.3.3}$$

The balance equations for $n = 0, \dots, N$ and $m \geq 1$ are

$$p_{0,m} = 0, \tag{2.3.4}$$

$$(\lambda_{1,n} + \mu_{1,n} + n\mu_2)p_{n,m} = \lambda_{1,n-1}p_{n-1,m}(1 - \delta_{1,n}) + \mu_{1,n+1}p_{n+1,m} + (n\mu_2 + \delta_{1n}\mu_{1,1})p_{n,m+1} \tag{2.3.5}$$

for $n = 1, \dots, N-1$ and

$$(p\lambda_{1,N} + \mu_{1,N} + N\mu_2)p_{N,m} = \lambda_{1,N-1}p_{N-1,m} + p\lambda_{1,N}p_{N,m-1} + N\mu_2 p_{N,m+1}. \tag{2.3.6}$$

The Markov chain defined by these balance equations is irreducible if the state space is restricted to $\{(0,0)\} \cup \{(n,m) \,|\, n = 1, \dots, N,\ m \geq 1\}$; the states $(0,m)$, $m \geq 1$, will not be visited. The transition rate diagram for this model is depicted in Figure 2.4, where $\lambda_2 = 0$. Observe that the states $(0,m)$, $m \geq 1$, will not be reached in this diagram if $\lambda_2 = 0$.

The main ideas for deriving the stability condition and reducing the number of unknowns can be carried over from the basic model without jockeying. One might predict that the possibility of jockeying influences the queue length of the first queue and therefore could inhibit the determination of a closed-form expression for the queue length probabilities in the first queue without using further unknown quantities. Nevertheless, this complication can be avoided. The approach is carried out in the next sections.

2.3.2 Necessary and sufficient stability condition

The balance equations for $m = 0$ and $n = 0, \dots, N$ give

$$p\lambda_{1,N}p_{N,0} = \mu_2 \sum_{n=1}^{N} np_{n,1} + \mu_{1,1}p_{1,1} \tag{2.3.7}$$

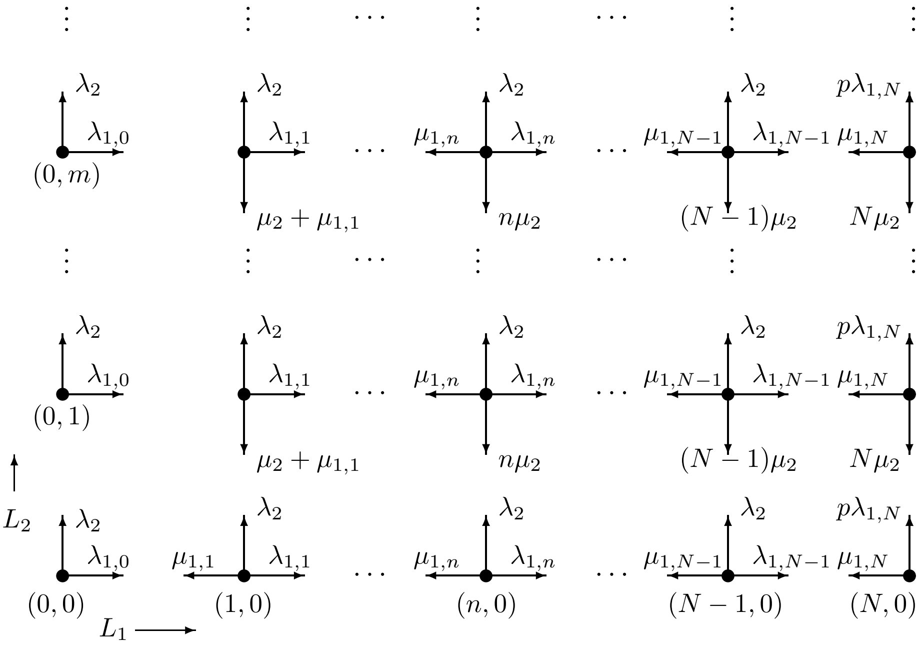


Figure 2.4: Limited jockeying: Transition rate diagram.

after summation over $n=0,\dots,N$. The balance equations for $m\geq 1$ and $n=0,\dots,N$ yield

$$\begin{aligned}p\lambda_{1,N}p_{N,m}+\mu_{1,1}p_{1,m}+\mu_2\sum_{n=1}^{N}np_{n,m}\\ =p\lambda_{1,N}p_{N,m-1}+\mu_{1,1}p_{1,m+1}+\mu_2\sum_{n=1}^{N}np_{n,m+1}\end{aligned} \tag{2.3.8}$$

by summing over $n=0,\dots,N$. From (2.3.7) and (2.3.8) we obtain analogously to the derivations for the basic model that

$$p\lambda_{1,N}P(L_1=N)=\mu_{1,1}\sum_{m\geq 1}p_{1,m}+\mu_2\big(EL_1-\sum_{n=1}^{N}np_{n,0}\big). \tag{2.3.9}$$

All terms in the above equation (except the system parameters) are unknown for this model. Nevertheless, the set of unknown probabilities in the system of steady-state equations can be reduced to $p_{1,0},\dots,p_{N,0}$. This is done in the next section.

From Theorem 1.2.3 we get the necessary and sufficient stability condition.

Proposition 2.3.1. *The system of equations* (2.3.1)-(2.3.6) *has a unique nonnegative and normalized solution if and only if*

$$p\lambda_{1,N} < \frac{\mu_2 \sum\limits_{n=1}^{N} n \prod\limits_{i=1}^{n-1} \rho_i + \mu_{1,1}}{\prod\limits_{n=1}^{N-1} \rho_n} \tag{2.3.10}$$

holds, where $\rho_n = \lambda_{1,n}/\mu_{1,n+1}$ *for* $n = 0, \dots, N-1$.

Proof. The proof is an almost verbatim repetition of the proof of Proposition 2.2.1. We regard Q_1 as the phase and Q_2 as the level of the underlying quasi birth and death process with phase and level given by L_1 and L_2, respectively. The stability condition (2.3.10) is then derived from Theorem 1.2.3 in the following way. The exponential arrival rate in Q_2 is $p\lambda_{1,N}$ if $L_1 = N$ and 0 if $L_1 < N$. The exponential service rate in Q_2 is $n\mu_2$ given $L_1 = n$, $n = 0, \dots, N$. The rates for arrivals in Q_1 are $\lambda_{1,n}$ given $L_1 = n$ for $n = 0$ and $n = 2, 3, \dots, N-1$ and $\lambda_{1,1} + \mu_{1,1}$ for $L_1 = 1$. Observe that the arrival stream in Q_1 consists of the original external arrival stream and the jockeying customers from Q_2. The service rate in Q_1 is $\mu_{1,n}$ given $L_1 = n$ for $n = 0, \dots, N$, where $\mu_{1,0} = 0$. Observe that the states $(0, m)$ for $m \geq 1$ are not visited by the quasi birth and death process. Therefore, we can restrict the state space of the process to $\{(n,m) \mid n = 1, \dots, N,\ m \geq 1\}$ and consequently, the phase process has the state space $\{1, \dots, N\}$. This ensures the irreducibility of the whole process. With this setting, the $(N \times N)$-matrices A_0, A_1 and A_2 from Theorem 1.2.3 are given by

$$A_0 = \text{diag}(0, \dots, 0, p\lambda_{1,N}), \quad A_2 = \text{diag}(\mu_2, 2\mu_2, \dots, N\mu_2)$$

and $A_1 = A - A_0 - A_2$, where A is the rate matrix of the phase process governing Q_1, i.e., the standard birth and death process on $\{1, \dots, N\}$ with birth rate $\lambda_{1,1} + \mu_2$ in state 1, birth rates $\lambda_{1,n}$ in the states $n = 2, \dots, N-1$ and death rates $\mu_{1,n}$ in the states $n = 2, \dots, N$. The vector π is the stationary probability measure of the phase process in Q_1, i.e., $\pi_n = P(L_1 = n)$, $n = 1, \dots, N$, is given by (2.2.1) with a suitable normalization factor. The stability condition (1.2.1) is then easily computed as (2.3.10). □

One might predict from equation 2.3.9 and

$$P(L_1 = 1) = \sum_{m \geq 1} p_{1,m} = \sum_{m \geq 0} p_{1,m} - p_{1,0}$$

that the system is stable if $p_{1,0} > 0$ or $\sum_{n=1}^{N} p_{n,0} > 0$. In this case a stability condition could be

$$p\lambda_{1,N} P(L_1 = N) < \mu_{1,1} P(L_1 = 1) + \mu_2 EL_1. \qquad (2.3.11)$$

In addition to this condition, the stability condition has an interesting interpretation in view of the results of Section 2.2 and the proof of Proposition 2.3.1. Let L_1^* be the stationary queue length of the first queue in the basic model, i.e., the queue without arrivals from jockeying customers from Q_2 and distribution given by (2.2.1). By multiplying the inequality (2.3.10) with $\prod_{n=1}^{N-1} \rho_n$ we can rephrase the stability condition as

$$p\lambda_{1,N} P(L_1^* = N) < \mu_{1,1} P(L_1^* = 1) + \mu_2 EL_1^*, \qquad (2.3.12)$$

where we have used $\sum_{n=1}^{N} n \prod_{i=0}^{n-1} \rho_i = EL_1^*/P(L_1^* = 0)$, $\prod_{n=0}^{N-1} \rho_n = P(L_1^* = N)/P(L_1^* = 0)$ and $\rho_0 = P(L_1^* = 1)/P(L_1^* = 0)$. This equation states that the average arrival rate in Q_2 should be smaller than the average departure rate in Q_2 in the case of no jockeying. The average departure rate is the sum of the average service rate delivered by the servers in Q_2 and the average departure rate of jockeying customers. One might expect that the conditions (2.3.11) and (2.3.12) are equivalent. We will show in the next section that they differ only by a positive constant and are therefore indeed equivalent. As for the basic model, one can show that a unique normalized solution of the steady-state equations (2.3.1)-(2.3.6) exists if and only if an $n \in \{0, \dots, N\}$ and an $m \geq 0$ exist with $p_{n,m} > 0$ (see also Proposition 2.2.9 and Corollary 2.2.10).

2.3.3 Generating functions and steady-state distribution

In this section, we derive the recurrence equations for the probability generating functions

$$G_n(z) = \sum_{m=0}^{\infty} p_{n,m} z^m, \quad |z| \leq 1.$$

As in the basic model, these relations are used to reduce the set of unknown steady-state probabilities to the unknowns $p_{1,0}, \dots, p_{N,0}$. The stability condition will then be related to the existence of $N - 1$ pairwise distinct zeros of a function originating from the recurrence equations for the probability generating functions.

We can assume for the moment that the system is stable and that (2.3.10) holds. In this case, the generating functions are well defined on $|z| \leq 1$. By multiplying the equations (2.3.1)-(2.3.6) by z^m and summing over $m \geq 0$ we get, after simplifying,

$$\lambda_{1,0} G_0(z) = \mu_{1,1} p_{1,0} = \lambda_{1,0} p_{0,0}, \tag{2.3.13}$$

$$\begin{aligned} \big(\lambda_{1,1} z - (\mu_{1,1} + \mu_2)(1-z)\big) G_1(z) &= \mu_{1,2} z G_2(z) \\ &\quad - (\mu_{1,1} + \mu_2)(1-z) p_{1,0}, \end{aligned} \tag{2.3.14}$$

$$\begin{aligned} \big((\lambda_{1,n} + \mu_{1,n}) z - n\mu_2(1-z)\big) G_n(z) &= \lambda_{1,n-1} z G_{n-1}(z) \\ &\quad + \mu_{1,n+1} z G_{n+1}(z) \\ &\quad - n\mu_2(1-z) p_{n,0}, \end{aligned} \tag{2.3.15}$$

$$\begin{aligned} \big(\mu_{1,N} z + (p\lambda_{1,N} z - N\mu_2)(1-z)\big) G_N(z) &= \lambda_{1,N-1} z G_{N-1}(z) \\ &\quad - N\mu_2(1-z) p_{N,0}, \end{aligned} \tag{2.3.16}$$

where the third equation holds for $n = 2, \ldots, N-1$ and vanishes in the case $N = 2$. In order to write these equations in matrix from, we define the vectors

$$G(z) = (G_1(z), \ldots, G_N(z))^\top,$$
$$p = \big((\mu_{1,1} + \mu_2) p_{1,0}, 2\mu_2 p_{2,0}, \ldots, N\mu_2 p_{N,0}\big)^\top$$

and the matrix $A(z) \in \mathrm{Mat}(N, N, \mathbb{R})$ by

$$A(z) = \begin{pmatrix} \alpha_1(z) & -\mu_{1,2} z & 0 & \ldots & \ldots & 0 \\ -\lambda_{1,0} z & \alpha_2(z) & -\mu_{1,3} z & \ddots & & \vdots \\ 0 & \ddots & \alpha_3(z) & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & -\mu_{1,N} z \\ 0 & \ldots & \ldots & 0 & -\lambda_{1,N-1} z & \alpha_N(z) \end{pmatrix}, \tag{2.3.17}$$

where

$$\alpha_1(z) = \lambda_{1,1} z - (\mu_{1,1} + \mu_2)(1-z), \tag{2.3.18}$$
$$\alpha_n(z) = (\lambda_{1,n} + \mu_{1,n}) z - n\mu_2(1-z) \text{ for } n = 2, \ldots, N-1 \text{ and} \tag{2.3.19}$$
$$\alpha_N(z) = \mu_{1,N} z + (p\lambda_{1,N} z - N\mu_2)(1-z). \tag{2.3.20}$$

For simplicity let $\alpha_0(z) = 0$. The dimension of the matrix is N whereas it was $N+1$ in (2.2.23). We dropped the first line of the matrix in (2.2.23) because the generating function G_0 does only depend on $p_{1,0}$ which itself is determined by the functions $G_1, \ldots, G_N$.

The equations (2.3.14), (2.3.15) and (2.3.16) are equivalent to

$$A(z)G(z) = -(1-z)p. \tag{2.3.21}$$

By Cramer's rule we have

$$\det(A(z))G_n(z) = \det(A_n(z)) \tag{2.3.22}$$

for all values z such that $A(z)$ is invertible, where $A_n(z)$ is the matrix obtained from $A(z)$ by replacing the n-th column with the vector $-(1-z)p$. The generating functions $G_0, \ldots, G_N$ are uniquely determined by the equations (2.3.22) and $p_{1,0}, \ldots, p_{N,0}$, since these are the only unknowns occurring in these equations.

$\det(A(z))$ is a polynomial in z of degree $N+1$. We will show in the following that $\det(A(z))$ has $N-1$ zeros in the open interval $(0,1)$ and one zero at $z = 1$. Additionally, we will show that $\det(A(z))$ has another zero in the open interval $(1, \infty)$ if and only the stability condition (2.3.10) holds. The $N-1$ zeros of $\det(A(z))$ in $(0,1)$ will provide us with $N-1$ linear homogeneous equations in the unknowns $p_{1,0}, \ldots, p_{N,0}$. Another linear equation yielding a system of N linear equations can be derived from (2.3.9) and (2.3.41) in a similar manner as for the basic model from Section 2.2.

Observe that (2.3.9) can be written as

$$p\lambda_{1,N}G_N(1) = \mu_{1,1}G_1(1) + \mu_2 \sum_{n=1}^{N} nG_n(1) - \mu_{1,1}p_{1,0} - \mu_2 \sum_{n=1}^{N} np_{n,0} \tag{2.3.23}$$

and does not involve $G_0(1)$. The normalization condition which is

$$\sum_{n=0}^{N} \sum_{m \geq 0} p_{n,m} = \sum_{n=0}^{N} G_n(1) = 1 \tag{2.3.24}$$

in terms of the generating functions is equivalent to

$$G_1(1) = \frac{\lambda_{1,0} - \mu_{1,1}p_{1,0}}{\lambda_{1,0} \sum\limits_{n=1}^{N} \prod\limits_{i=1}^{n-1} \rho_i}, \tag{2.3.25}$$

where $\mu_{1,1}p_{1,0} = \lambda_{1,0}p_{0,0} = \lambda_{1,0}G_0(1)$ and where we have written $\rho_i = \lambda_{1,i}/\mu_{1,i+1}$ for $i = 0, \ldots, N-1$. This can be shown by solving the equations (2.3.13)-(2.3.16) for $z = 1$ in terms of $G_0(1)$ and $G_1(1)$ and using the normalization condition. These equations are

$$\lambda_{1,0}G_0(1) = \mu_{1,1}p_{1,0}, \tag{2.3.26}$$
$$\lambda_{1,1}G_1(1) = \mu_{1,2}G_2(1), \tag{2.3.27}$$
$$(\lambda_{1,n} + \mu_{1,n})G_n(1) = \lambda_{1,n-1}G_{n-1}(1) + \mu_{1,n+1}G_{n+1}(1) \quad \text{and} \tag{2.3.28}$$
$$\mu_{1,N}G_N(z) = \lambda_{1,N-1}G_{N-1}(1), \tag{2.3.29}$$

where $n = 2, \ldots, N-1$ in the third equation. The solution of the equations (2.3.27)-(2.3.29) in terms of $G_1(1)$ is given by

$$G_n(1) = G_1(1)\prod_{i=1}^{n-1}\rho_i \text{ for } n = 1, \ldots, N. \tag{2.3.30}$$

By the normalization condition (2.3.24), summation of (2.3.26) and (2.3.30) over $n = 1, \ldots, N$ yields (2.3.25). Substituting (2.3.25) and (2.3.30) in (2.3.23) and collecting terms finally yields the additional equation

$$\frac{\mu_{1,1}p_{1,0} + \mu_2\sum_{n=1}^{N} np_{n,0}}{\lambda_{1,0} - \mu_{1,1}p_{1,0}} = \frac{\mu_{1,1} + \mu_2\sum_{n=1}^{N} n\prod_{i=1}^{n-1}\rho_i - p\lambda_{1,N}\prod_{i=1}^{N-1}\rho_i}{\lambda_{1,0}\sum_{n=1}^{N}\prod_{i=1}^{n-1}\rho_i} \tag{2.3.31}$$

that relates the unknowns $p_{1,0}, \ldots, p_{N,0}$ to the system parameters.

Let $G_n^*(1)$ be the generating function of the phase in the basic model at $z = 1$, i.e., let $G_0^*(1)$ be given by (2.2.47) and let $G_n^*(1)$ be given by (2.2.46) for $n = 1, \ldots, N$. Equation (2.3.30) shows that $G_n(1)$ and $G_n^*(1)$ differ only by a constant independent of n, namely

$$G_n(1) = G_n^*(1)\left(\frac{G_1(1)}{G_0^*(1)\rho_0}\right) \tag{2.3.32}$$

for $n = 1, \ldots, N$. We have shown the following result.

Remark 2.3.2. The equations (2.3.11) and (2.3.12) are equivalent given $p_{1,0} > 0$.

The equation (2.3.32) has another consequence that yields a lower bound for EL_1. A real-valued random variable X is called *stochastically larger* than the real-valued random variable Y, if $P(X > t) \geq P(Y > t)$ for all $t \in \mathbb{R}$.

Proposition 2.3.3. *Under stationary conditions, i.e., if* (2.3.10) *is fulfilled, the queue length* L_1 *of the first queue with jockeying is stochastically larger than the queue length* L_1^* *of the first queue without jockeying. Thus, a lower bound for* EL_1 *is given by*

$$EL_1 \geq EL_1^* = \frac{\sum\limits_{n=1}^{N} n \prod\limits_{i=0}^{n-1} \rho_i}{1 + \sum\limits_{n=1}^{N} \prod\limits_{i=0}^{n-1} \rho_i}.$$

We will not give a formal proof but point out the idea. Under stationary conditions, the equation (2.3.32) implies that $G_n(1) \geq G_n^*(1)$ for all $n = 1, \ldots, N$ if this is the case for at least one $n \in \{1, \ldots, N\}$ or if $G_0(1) \leq G_0^*(1)$ which implies the first condition. Therefore,

$$EL_1 = \sum_{n=1}^{N} nG_n(1) \geq \sum_{n=1}^{N} nG_n^*(1) = EL_1^*$$

holds.

Now we turn our attention to the derivation of the auxiliary functions $q_1, \ldots, q_N$ and their properties. This leads to the stability theorem concerning the zeros of $\det(A)$. For $n = 1, \ldots, N+1$ we let

$$q_1(z) = \alpha_1(z),\ q_2(z) = \det\begin{pmatrix} \alpha_1(z) & -\mu_{1,2}z \\ -\lambda_{1,1}z & \alpha_2(z) \end{pmatrix}, \ldots, q_N(z) = \det(A(z)). \tag{2.3.33}$$

By Laplace expansion we get

$$\begin{aligned} q_1(z) &= \alpha_1(z)q_0(z) \quad \text{and} \\ q_n(z) &= \alpha_n(z)q_{n-1}(z) - \lambda_{1,n-1}\mu_{1,n}z^2 q_{n-2}(z) \end{aligned} \tag{2.3.34}$$

for $n = 2, \ldots, N$, where $q_0(z) = 1$.

The crucial properties of the functions $q_1, \ldots, q_N$ and $\alpha_1, \ldots, \alpha_N$ are derived in the same manner as in Proposition 2.2.4.

Proposition 2.3.4. *The function* q_n *is a polynomial in* z *of degree* n *for* $n = 1, \ldots, N-1$ *and of degree* $N+1$ *for* $n = N$*. The functions* q_n *and* α_n*,* $n = 1, \ldots, N$*, satisfy the following properties:*

(i) q_n *and* q_{n+1} *have no common root in* $(0,1)$ *for* $n = 0, \ldots, N-1$*.*

(ii) $\operatorname{sgn}(\alpha_0(0)) = 1$ *and* $\operatorname{sgn}(\alpha_n(0)) = -1$ *for* $n = 1, \ldots, N$*.*

(iii) $\operatorname{sgn}(q_n(0)) = (-1)^{n+1}$ *for* $n = 1, \ldots, N$.

(iv) $q_n(1) = \prod_{i=0}^{n-1} \lambda_{1,i}$ *for* $n = 0, \ldots, N-1$ *and* $q_N(1) = 0$.

(v) For $n = 1, \ldots, N-1$ *the following implication holds: If* $\tilde{z} > 0$ *with* $q_n(\tilde{z}) = 0$*, then*

$$\operatorname{sgn}(q_{n-1}(\tilde{z})q_{n+1}(\tilde{z})) = -1.$$

(vi) q_n *has* $n-1$ *pairwise distinct zeros in* $(0,1)$ *for* $n = 1, \ldots, N-1$.

(vii) $\lim_{z\to\infty} q_n(z) = \infty$ *for* $n = 2, \ldots, N-1$ *and* $\lim_{z\to\infty} q_N(z) = -\infty$.

As for the basic model, we can determine functions $h_1, \ldots, h_N$ on $[0, \infty)$ satisfying

$$\begin{aligned} q_n(z) &= z^n \prod_{i=1}^{n} \lambda_{1,i} + (1-z)h_n(z) \text{ for } n = 1, \ldots, N-1 \text{ and} \\ q_N(z) &= (1-z)h_N(z) \end{aligned} \tag{2.3.35}$$

inductively from (2.3.34). These functions are given by

$$h_1(z) = -(\mu_{1,1} + \mu_2), \tag{2.3.36}$$

$$h_2(z) = \alpha_2(z)h_n(z)(1-z), \tag{2.3.37}$$

$$h_n(z) = -n\mu_2 z^{n-1} \prod_{i=1}^{n-1} \lambda_{1,i} + \alpha_n(z)h_{n-1}(z) - \lambda_{1,n-1}\mu_{1,n}z^2 h_{n-2}(z) \tag{2.3.38}$$

for $n = 3, \ldots, N-1$ and

$$\begin{aligned} h_N(z) = (p\lambda_{1,N}z - N\mu_2)z^{N-1} &\prod_{i=0}^{N-1} \lambda_{1,i} + \alpha_N(z)h_{N-1}(z) \\ &-\lambda_{1,N-1}\mu_{1,N}z^2 h_{N-2}(z) \end{aligned} \tag{2.3.39}$$

for all $z \in [0, \infty)$. See Proposition 2.2.5 in the previous section for a similar derivation for the basic model. From the previous proposition we can derive the stability condition in terms of $h_N(1)$ in the same manner as in Theorem 2.2.6.

Theorem 2.3.5. $\det(A(z))$ *is a polynomial of degree* $N+1$ *and has* $N-1$ *distinct zeros in the interval* $(0,1)$ *and one zero at* $z = 1$*. Additionally,* $\det(A(z))$ *has another zero in the interval* $(1, \infty)$ *if and only if* $h_N(1) < 0$*, where* h_N *is defined by* $\det(A(z)) = (1-z)h_N(z)$ *for all* $z > 0$.

It remains to determine $h_N(1)$. Evaluating (2.3.36)-(2.3.38) and (2.3.39) at $z = 1$ gives a recursive system of equations for $h_1(1), \dots, h_N(1)$. We omit the details and state the solution:

$$\begin{aligned} h_n(1) = &-\mu_2 \left(\prod_{i=1}^{n} \lambda_{1,i} \right) \left(\sum_{k=1}^{n} \frac{1}{\mu_{1,k+1}} \left(\prod_{i=1}^{k} \frac{\mu_{1,i+1}}{\lambda_{1,i}} \right) \left(\sum_{j=1}^{k} j \prod_{i=1}^{j-1} \frac{\lambda_{1,i}}{\mu_{1,i+1}} \right) \right) \\ &- \mu_{1,1} \left(\prod_{i=1}^{n} \lambda_{1,i} \right) \left(\sum_{k=1}^{n} \frac{1}{\mu_{1,k+1}} \prod_{i=1}^{k} \frac{\mu_{1,i+1}}{\lambda_{1,i}} \right) \end{aligned} \tag{2.3.40}$$

for $n = 1, \dots, N-1$, where the empty sum and empty product are defined to be 0 and 1, respectively. Inserting $h_{N-1}(1)$ and $h_{N-2}(1)$ from (2.3.40) into (2.3.39) and simplifying finally leads to

$$h_N(1) = p \prod_{i=1}^{N} \lambda_{1,i} - \mu_2 \left(\prod_{i=1}^{N-1} \mu_{1,i+1} \right) \left(\sum_{n=1}^{N} n \prod_{i=1}^{n-1} \frac{\lambda_{1,i}}{\mu_{1,i+1}} \right) - \prod_{i=0}^{N-1} \mu_{1,i+1}. \tag{2.3.41}$$

Setting $h_N(1) < 0$ and simplifying gives the stability condition from Proposition 2.3.1. The next theorem combines Proposition 2.3.1, Remark 2.3.2 and Theorem 2.3.5 and relates the number of zeros of $\det(A(z))$ to this stability condition.

Theorem 2.3.6. $\det(A(z))$ *is a polynomial of degree* $N+1$ *and has* $N-1$ *distinct zeros in the interval* $(0,1)$ *and one zero at* $z = 1$*. Additionally,* $\det(A(z))$ *has another zero in the interval* $(1,\infty)$ *if and only if the system of equations* (2.3.1)-(2.3.6) *has a unique nonnegative and normalized solution, i.e., if and only if* $p\lambda_{1,N} P(L_1 = N) < \mu_{1,1} P(L_1 = 1) + \mu_2 E L_2$ *or equivalently*

$$p\lambda_{1,N} < \frac{\mu_2 \sum\limits_{n=1}^{N} n \prod\limits_{i=1}^{n-1} \rho_i + \mu_{1,1}}{\prod\limits_{n=1}^{N-1} \rho_n} \tag{2.3.42}$$

holds, where $\rho_n = \lambda_{1,n}/\mu_{1,n+1}$, $n = 0, \dots, N-1$.

Remark 2.3.7. The stability condition is independent of the arrival rate $\lambda_{1,0}$ in Q_1. This is on the one hand due to the fact that the stability is derived from the generating functions $G_1, \dots, G_N$ which are independent of $\lambda_{1,0}$. On the other hand, the arrival rate in Q_1 is $\lambda_{1,0}$ only in the case that no customers are present in both queues. Therefore, the arrival rate $\lambda_{1,0}$ influences the length of the idle period of the system, but not the busy period. Due to the jockeying discipline, Q_1 is in state 0 if and only if Q_2 is empty.

Remark 2.3.8. By the condition given in Theorem 2.2.11, when the system is stable, meaning (2.2.54) holds, we can use the $N-1$ zeros of $\det(A(z)) = q_N(z)$ in $(0,1)$ to find the N unknown probabilities $p_{1,0}, \ldots, p_{N,0}$. The $N-1$ pairwise distinct zeros deliver $N-1$ equations for these unknowns. One more equation relating the unknowns is (2.3.31). This provides us with N (independent) equations in the N unknowns $p_{1,0}, \ldots, p_{N,0}$. For further comments on the procedure see Remark 2.2.12.

Once $p_{1,0}, p_{2,0}, \ldots, p_{N,0}$ are determined, $G_0, \ldots, G_N$ are given and we can calculate EL_1, EL_2 and $\mathrm{Cov}(L_1, L_2)$:

$$EL_1 = \sum_{n=1}^{N} nG_n(1),\ EL_2 = \sum_{n=0}^{N} G_n'(1) \text{ and } E(L_1L_2) = \sum_{n=1}^{N} nG_n'(1).$$

An equation similar to (2.2.57) for the basic model relating these terms is the equation

$$\begin{aligned}\mu_2 E(L_1L_2) + \mu_{1,1}E(L_2|L_1 = 1)P(L_1 = 1)&\\ = p\lambda_{1,N}\left(P(L_1 = N) + E(L_2|L_1 = N)(P(L_1 = N))\right),&\end{aligned}$$

where $E(L_1L_2) = \sum_{n=1}^{N} nG_n'(1)$, $E(L_2|L_1 = n)P(L_1 = n) = G_n'(1)$ for $n = 1, \ldots, N$ and $P(L_1 = N) = G_N(1)$. The equation is shown by summing the equations (2.3.13)-(2.3.16) and differentiating the result at $z = 1$.

Remark 2.3.9. It is possible to further generalize the model by:

1. Equipping Q_2 with an exponential arrival stream with rate $\lambda_2 > 0$ (see Figure 2.4).
2. Assuming that the service rate in Q_2 is $\mu_{2,n}$ (instead of $n\mu_2$) when Q_1 is in state n (see also Remark 2.2.14).
3. Letting a fixed number $1 \leq k \leq N-1$ jockey to Q_1 if Q_1 becomes empty.

In all cases, this will lead to a modified construction of the generating functions, the auxiliary functions and the stability condition (we omit the details). In the second case, the stability condition involves additional terms depending on the system parameters (compare (2.2.40) with (2.3.41) and (2.2.54) with (2.3.42)). The steady-state equations simplify substantially for $k = N$. The solution in this case is given in the next section.

2.3.4 Closed-from solution in the case of unlimited jockeying

For this model a complete analytic solution is available. In order to simplify the resulting formulas we assume constant arrival and service rates in Q_1. Let $N > 1$ and let Q_1 be an $M/M/1/N-1$ queue with arrival rate $\lambda_1 > 0$ and service rate $\mu_1 > 0$. Let Q_2 be fed by the p-weighted overflow stream from Q_1, $p \in [0,1]$, and let the service rate in Q_2 be $n\mu_2$ if $L_1 = n$ where $\mu_2 > 0$. Modify Q_1 in the following way: If Q_1 becomes empty, then the first customers waiting in Q_2 are instantly transferred to Q_1 until Q_1 is fully occupied or Q_2 empties, whatever happens first. Let $p_{n,m}$ be the steady-state probability of having n customers present in Q_1 and m customers in Q_2, $n = 0, \ldots, N$, $m \geq 0$. We must have $p_{n,m} = 0$ for $m \geq 1$ and $n < N$, because in this case, some customers will be transferred instantly from Q_2 to Q_1. The balance equations for this model are

$$\lambda_1 p_{0,0} = \mu_1 p_{1,0}, \tag{2.3.43}$$

$$(\lambda_1 + \mu_1) p_{n,0} = \lambda_1 p_{n-1,0} + \mu_1 p_{n+1,0}, \; n = 1, \ldots, N-1, \tag{2.3.44}$$

$$(p\lambda_1 + \mu_1) p_{N,0} = \lambda_1 p_{N-1,0} + (\mu_1 + N\mu_2) p_{N,1}, \; n = N, \, m = 0, \tag{2.3.45}$$

$$(p\lambda_1 + \mu_1 + N\mu_2) p_{N,m} = p\lambda_1 p_{N,m-1} + (\mu_1 + N\mu_2) p_{N,m+1}, \; n = N, \, m \geq 1, \tag{2.3.46}$$

$$p_{n,m} = 0, \; m \geq 1, \, n = 0, \ldots, N-1. \tag{2.3.47}$$

Solving (2.3.43) and (2.3.44) yields $p_{n,0} = \left(\frac{\lambda_1}{\mu_1}\right)^n p_{0,0}$ for $n = 0, \ldots, N$. Inserting

$$p_{N-1,0} = \left(\frac{\lambda_1}{\mu_1}\right)^{N-1} p_{0,0} \quad \text{and} \quad p_{N,0} = \left(\frac{\lambda_1}{\mu_1}\right)^{N} p_{0,0}$$

into (2.3.45) gives

$$p_{N,1} = \frac{(p\lambda_1 + \mu_1)\left(\frac{\lambda_1}{\mu_1}\right)^N - \lambda_1 \left(\frac{\lambda_1}{\mu_1}\right)^{N-1}}{\mu_1 + N\mu_2} p_{0,0} = \left(\frac{\lambda_1}{\mu_1}\right)^N \left(\frac{p\lambda_1}{\mu_1 + N\mu_2}\right) p_{0,0}.$$

This and (2.3.46) lead to

$$p_{N,m} = \left(\frac{\lambda_1}{\mu_1}\right)^N \left(\frac{p\lambda_1}{\mu_1 + N\mu_2}\right)^m p_{0,0}$$

for all $m \geq 0$. Thus all non-zero probabilities are expressed in terms of $p_{0,0}$ and we can determine $p_{0,0}$ from the normalization condition

$$\sum_{n=0}^{N} \sum_{m=0}^{\infty} p_{n,m} = 1.$$

Substituting the appropriate terms in the normalization condition and simplifying we get

$$p_{0,0} = \left(\left(\frac{\lambda_1}{\mu_1}\right)^N \frac{\mu_1 + N\mu_2}{\mu_1 + N\mu_2 - p\lambda_1} + \frac{\mu_1 - \lambda_1 \left(\frac{\lambda_1}{\mu_1}\right)^{N-1}}{\mu_1 - \lambda_1} \right)^{-1}.$$

The necessary and sufficient stability condition is clearly $p\lambda_1 < \mu_1 + N\mu_2$ (see also Theorem 1.2.3).

It is obvious that we can model the two queues as an ordinary single $M_{(n)}/M_{(n)}/1$ queue, where the subscript (n) stands for state-dependent service and arrival rates. The state space of this queue is $\mathbb{N}_0$, where the states $0, \ldots, N$ correspond to the states $(0,0), \ldots, (N,0)$ and the states $m \geq N+1$ correspond to the states (N,m), $m \geq 1$. The arrival rate is λ_1 in the states $0, \ldots, N-1$ and $p\lambda_1$ in the states $m \geq N$. The service rate is μ_1 in the states $1, \ldots, N$ and $\mu_1 + N\mu_2$ in the states $m \geq N+1$. By classical formulas for the steady-state distribution of infinite birth and death chains (see for example Cohen [12]) we get the same result.

Chapter 3

Overflow to a finite queue with waiting room

3.1 Model overview

As in the previous chapter, we consider an open queueing network consisting of two queues Q_1 and Q_2 with an overflow capability from Q_1 to Q_2. The main difference between the queueing models presented in the following and those presented in the first part of this thesis is the finiteness of the capacity of the second queue. Moreover, customers in the first queue no longer serve the customers in the second queue and the service and arrival rates are not variable. The finiteness gives rise to additional boundary conditions. These boundary conditions and the special structure of the steady-state equations make it impossible to carry out the approach from the second chapter. Nevertheless, the number of steady-state equations that describe the system's behavior can be reduced substantially. The steady-state probabilities and quantities of interest can be stated in an elegant way that reveals the underlying structure of the solutions. The separation approach used in this chapter is, due to the terms in the steady-state equations in the previous chapter that arise from the serving customers, not applicable to the models presented in the second chapter.

Now we give an overview over the models considered in this chapter and the methods used for deriving the quantities of interest. Each queue is equipped with a finite number of servers and a waiting room with finite capacity. We assume independent Poisson arrivals at each queue and exponential service times with server-dependent parameters. Q_i, $i = 1, 2$, possesses $n_i \geq 1$ servers and $q_i \geq 0$ waiting positions and is fed by a Poisson arrival stream with intensity $\lambda_i > 0$. In contrast to the previous chapter, we use lower case letters for the capacities of the queues in order to simplify the resulting formulas. The service requirement of a customer at Q_i is exponentially distributed with mean $1/\mu_i$, where $\mu_i > 0$ for $i = 1, 2$. A server serves exactly one customer at a time, in case one is present. Within the queues, customers are served in their order of arrival.

Since there is a limited number of servers and the waiting rooms are finite in each queue, arriving customers may be blocked at one of the queues depending on whether all servers and/or waiting positions are occupied. Moreover, blocked customers may overflow to the other queue if its capacity is exhausted. Since blocked customers may be allowed to overflow to the second queue, it seems appropriate to also allow waiting Q_1-customers to jockey to the second queue if there is capacity available. We will consider a variety of different models by combining different blocking, overflow and jockeying rules. These rules are explained in the following.

Blocking rules

We consider two blocking models for arriving Q_1-customers. In the first model, an arriving Q_1-customer is blocked if all n_1 servers are busy. In the second model, blocking takes place if all n_1 servers and all q_1 waiting positions are occupied.

A customer who arrives at Q_2 is served in Q_2 if less than n_2 servers are busy in Q_2 at the time of his arrival. The customer is queued in one of the waiting positions in Q_2 if all servers are busy and at least one waiting position is available in Q_2, otherwise the customer is blocked and cleared from the system.

Overflow rules

Blocked customers from Q_1 are treated with respect to two different routines, which will be called overflow rules. In the first routine, a blocked customer is served by one of the servers in Q_2 if at least one is available. If no server is available in Q_2, then the blocked customer is rejected and queued at Q_1 if a waiting position is available. The customer leaves the system otherwise. In the second routine, the blocked stream from Q_1 is directed to the waiting room in Q_2. If no waiting position is available in Q_2, then a blocked customer is rejected at Q_2 and queued in Q_1 if a waiting position is available. The blocked Q_1-customers are lost if the waiting rooms are fully occupied in both Q_1 and Q_2. Note that the blocked traffic and the overflow traffic from Q_1

are in general not identical, because blocked customers are allowed to join Q_1 if they are rejected at Q_2.

The *overflow stream* is the fraction of those arriving Q_1-customers who are blocked at Q_1 and who can potentially join Q_2. This overflow stream will additionally be weighted with a parameter $p \in [0, 1]$, i.e., an arriving customer who is blocked at Q_1 and overflows to Q_2 according to the blocking and overflow rules, joins Q_2 with probability p and leaves the system with probability $1 - p$. This overflow model with parameter p will be called *p-overflow model* or *weighted overflow model* in the following. The model with $p = 1$ will be called *deterministic overflow model.* Observe, that the queue sizes at Q_1 and Q_2 are in general not independent for every choice of the model configurations, not even for $p = 0$.

Jockeying rules

The possibility of waiting customers to move to another queue is called *jockeying*. Jockeying will be restricted to the customers of the first queue and therefore, Q_2-customers are not allowed to leave the second queue in order to receive service in the first queue. The treatment of waiting Q_1-customers can be classified according to three different cases. In the first case, a waiting Q_1-customer is placed in service in the second queue if at least one server and all waiting positions are available in Q_2. In the second case, waiting Q_1-customers must wait for a server in Q_1 to become free, i.e., jockeying is not allowed. In the third case, waiting Q_1 customers are placed in Q_2 if at least one waiting position or one server is available in Q_2.

Basic model notation

In order to distinguish between the different models and rules for blocking, overflow and jockeying, we denote an overflow system with this characteristics by a string

$$\alpha/\beta/\gamma$$

similar to Kendell's notation. In our notation, α refers to the blocking rule, β to the treatment of overflowing customers and γ is the indicator for the jockeying rule. The blocking rule indicator α is either S or W, where S indicates that arriving Q_1-customers are blocked if all servers in Q_1 are busy and W indicates that the customers are blocked if all servers and all waiting positions in Q_1 are occupied. The second indicator β is S or W, if the blocked customers overflow to the servers or the waiting rooms, respectively, in Q_2. Finally, γ, the jockeying rule indicator, is S or W if the waiting Q_1-customers are allowed to jockey to the servers or to the waiting room in Q_2, respectively. γ is N if no jockeying is allowed. The models S/S/γ and W/S/γ for γ = S, W, N are depicted in Figure 3.1. A schematic overview of the models S/W/γ and W/W/γ for γ = S, W, N is shown in Figure 3.2.

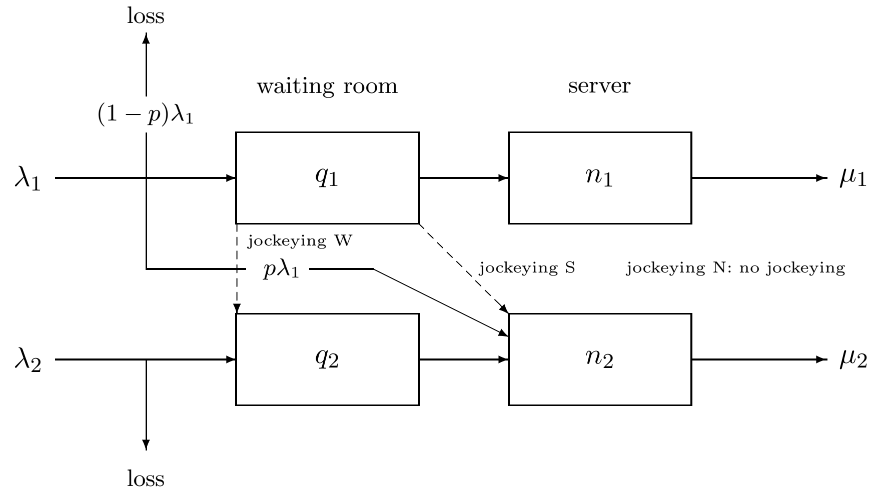


Figure 3.1: One-way p-overflow, $p \in [0, 1]$: Overflow to the servers in Q_2 (routine S)

Solution approach

We are interested in the two-dimensional server and waiting room demand process of the one-way p-overflow model, embedded at the time instants of arrivals to Q_1 and Q_2 and departures from Q_1 and Q_2. This process is a Markov chain with state space $\mathfrak{S} = \{0, \ldots, k_1\} \times \{0, \ldots, k_2\}$, where $k_i = n_i + q_i$ for $i = 1, 2$ and where we have labeled the servers and waiting rooms in Q_i with $1, \ldots, n_i$ and $n_i + 1, \ldots, k_i$, respectively, for $i = 1, 2$.

Remark 3.1.1. For every choice of $\lambda_1, \lambda_2, \mu_1, \mu_2 > 0$ in each model described above, the state space is finite and the fact that there exists a path from any state to $(0, 0)$ (and vice versa) having positive probability, yields irreducibility. Thus, the Markov chain is stationary with unique stationary distribution.

In some models it is possible to allow $q_1 = \infty$. In these cases we give a necessary condition for stability. The stationary distribution is uniquely

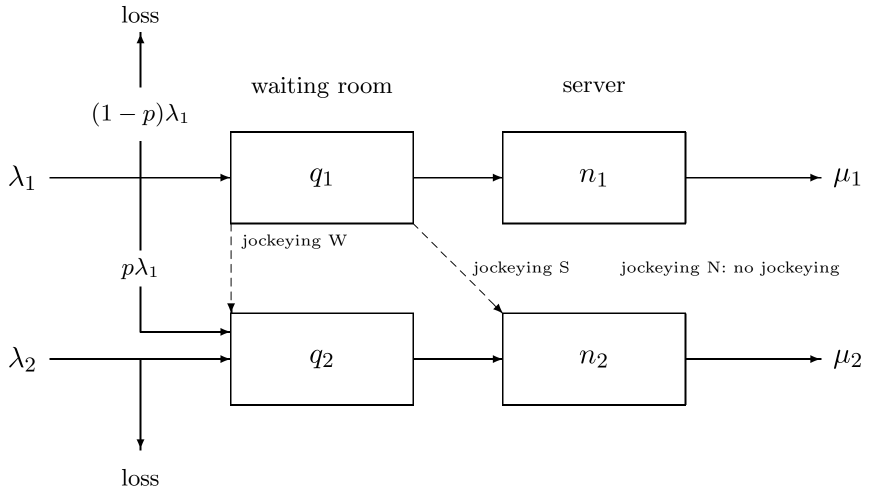


Figure 3.2: One-way p-overflow, $p \in [0,1]$: Overflow to the waiting room in Q_2 (routine W)

determined by the $(k_1+1)(k_2+1)$ balance equations and the normalization condition. To the best of our knowledge, no closed-form expression for the stationary distribution is known. By using a technique due to Morrison [46,48], we will reduce the problem of solving these $(k_1+1)(k_2+1)+1$ equations to the problem of solving a substantially smaller number of homogeneous linear equations in two sets of unknowns. With this approach, explicit formulas depending on these unknowns can be given for various steady-state quantities. Morrison considers the deterministic models S/S/S, S/S/N, W/S/N and W/W/N with $p=1$ and service rates $\mu_1=\mu_2$ and reduces the number of equations and unknowns by a separation method. The basic technique is to partition the state space into certain regions and boundaries and to separate the stationary probabilities within these regions. In every model, the separation leads to a set of eigenvalue problems for the separation constants. The eigenvalues are given by the roots of polynomial equations and are the pairwise distinct eigenvalues of real tridiagonal symmetric matrices as well. They possess an interlacing property, called the "Sturm sequence property", which reduces the computational complexity considerably. The desired probabilities are expressed as sums of eigenfunctions in terms of the eigenvalues. The number of eigenfunctions and therefore the number of coefficients to be determined in these representations is in general substantially smaller than the number of stationary probabilities.

The coefficients are determined by the normalization condition and a set of linear equations that stems from the boundary conditions. The desired probabilities and steady-state quantities can be numerically determined once the coefficients and eigenvalues are numerically calculated.

We will extend the results of Morrison in three directions. Firstly, we allow different service rates at the two queues, i.e., customers at Q_i have service rate μ_i, $i = 1, 2$, with arbitrary $\mu_1, \mu_2 > 0$. Secondly, the overflow process is weighted with a parameter $p \in [0, 1]$. In one model it is also possible to weight the stream of jockeying customers with the same parameter p. Thirdly, we consider several variations of the blocking and overflow routines as described above. It should be mentioned that the resulting formulas do not scale in the system parameters in general, i.e., for example the case of arbitrary service rates cannot be reduced to the case $\mu_1 = \mu_2 = 1$ by an adequate choice of the remaining parameters and the separation constants. Only the case $\mu_1 = \mu_2$ can be reduced to $\mu_1 = \mu_2 = 1$ by dividing the arrival rates and the separation constants by the service rate.

Stationary quantities

Once the stationary probabilities are specified in terms of the eigenvalues, it is possible to derive formulas for various steady-state quantities depending on the unknowns, like the loss probabilities, i.e., the probability that an arriving Q_i-customer is lost, $i = 1, 2$, the overflow and jockeying probability, the probability that an arriving customer is queued upon arrival, the mean departure rate from the waiting rooms to the servers, the average number of customers waiting and in service, the average waiting times and many more. These formulas have the advantage that the desired stationary quantities can be calculated directly in terms of the unknowns without computing the steady-state probabilities. We will display this possibility for the two basic models described in Section 3.2 and Section 3.3.1. It is immediately seen from Figure 3.1 that Q_1 in isolation is an Erlang loss system in the models without waiting room in Q_1, i.e., $q_1 = 0$. In these cases, the overflow stream from Q_1 to Q_2 consists of the p-fraction of blocked customers at Q_1. Let O_{12} be the expected (stationary) number of demands per unit time, which flow over from Q_1 to Q_2 (see Figure 3.1). With $\rho_1 = \lambda_1/\mu_1$, we must have

$$O_{12} = p\lambda_1 \cdot \frac{\rho_1^{n_1}}{n_1! \sum\limits_{i=0}^{n_1} \frac{\rho_1^i}{i!}}.$$

This is of course the loss probability for the famous Erlang loss system with capacity n_1 (see Erlang [24] and (2.2.1)). In general, the loss probabilities for Q_1 and Q_2, the blocking probabilities for customers at Q_2 and therefore the expected number of demands per unit time that flow over from Q_1 to Q_2 are unknown.

It is known from Cohen [11], Hordijk and Ridder [30] and van Marion [44] that the loss probabilities are in general sensitive with respect to the service-time distribution. In the literature, probabilistic and numerical bounds for the loss probabilities are known in some special cases. Van Dijk [15] derives upper and lower bounds for the loss probabilities by approximating the one-way overflow model by models with modified input and interconnection characteristics. Hordijk and Ridder [30, 31] and Ridder [57, 58] derive upper and lower bounds for weighted stationary probabilities by constructing approximating reversible Markov chains. These bounds are insensitive with respect to the service time distributions and are used to derive upper and lower bounds for the loss probabilities.

Applications

The most apparent application of this type of overflow queueing models is of course to telecommunication systems. As an example for a feasible application, the models presented in this chapter might be used to model call centers. In a specific call center telecommunication system, incoming telephone calls are answered by a limited number of operators (i.e., the servers). An incoming call is put on hold if all operators are busy. The number of calls on hold are in most practical situations finite and the on hold positions correspond to the waiting rooms in our queueing systems. Consider two incoming lines for calls. The first and second line are the queues Q_1 and Q_2, respectively. It may be permitted that customers from the first line have a certain type of priority, i.e., they may be allowed to overflow to the second line under one of the overflow rules specified earlier. The different rules for blocking, overflow and jockeying can be used to represent a certain constellation given in practice. Furthermore, the parameter p can be viewed as a control parameter or can be used to model the impatience rate of the overflowing calls. In another considerable application, one could assign costs or a premium to the parameter p that has to be paid by the priority customers from the first queue in order to receive service at the second queue. An increase of p results in a higher probability of acceptance at the second queue and a higher premium should be paid in order to receive a higher value for p. Thus, a customer or a system designer has to find a cost optimal value for p depending on his point of view.

Another application that can be slightly generalized to match our models is described in Altman et al. [2] and Hassin [28]. Consider two gas stations located at the same main road. Every gas station has a finite number of gas pumps and waiting positions. The two gas stations are located one after another, where the first and second station correspond to Q_1 and Q_2, respectively. Consider that some customers prefer the first and some customers prefer the second station. Additionally, if we consider one-way traffic, which is for example the case on an highway, then we get an arrival stream for each station that is independent from the other one. In the one-way setting, customers that find the first station occupied, can drive by in order to be served at the second station. Some of these overflowing customers might also reject service at the second station and drive by to another station they prefer. In this case, the overflow parameter p reflects this behavior. It should be noted, that some of the blocking, overflow and jockeying rules are unnatural in this context. Nevertheless, due to the diversity of the models presented in the following, many applications for example to communication, computer or traffic control systems are conceivable.

Further related literature

Some of the results developed in the following sections where published by the author in [59]. In the literature, numerical computations for some of the deterministic overflow models, i.e., with $p = 1$, are available. Kaufman [35] (model W/W/S and W/W/N) uses block iterative techniques and successive overrelaxation techniques for the numerical derivation of the steady-state probabilities. Chan [8, 9] (model W/W/N and W/W/S) uses block iterative techniques and preconditioned conjugate gradient methods for the numerical computation of the steady-state probabilities. The deterministic model with arbitrary service rates is treated in [9] in the context of overflow queueing networks with an arbitrary but finite number of queues. For some models considered in Chan [8], Kaufman [35] and Morrison [46], i.e., S/S/S, S/S/N and W/W/N, numerical results are presented in Kaufman et al. [34]. The overflow model with deterministic overflow and without waiting rooms is known as the *one-way overflow queueing model* (van Dijk [15,16], Doremalen [18], Hordijk and Ridder [30]). In telecommunication theory, it is a special type of an *asymmetric grading* (Kosten [36], van Marion [44] and Syski [63]).

Overflow queueing models are widespread in literature. We already mentioned Disney and König [17] for a broad overview and Koury et al. [39]

and Krieger et al. [40] for reviews of iterative numerical methods for overflow queueing models. A brief discussion of numerical methods for some two-queue overflow systems and further references are given in Ching and Ng [10]. Related overflow models are studied in van Doorn [19], El-Taha and Heath [20], Parthasarathy and Sudhesh [55] and the referenced literature therein using a variety of different techniques (for some comments on [19] and [55] see page 13). An overflow model with multiple primary queues, finite waiting rooms and a shared secondary overflow queue is numerically studied in Guérin and Lien [27] and related models are reviewed. Early references are given in van Marion [44], Morrison [46] and Morisson [47]. Further related literature is mentioned in Chapter 2, Section 2.1 of this thesis.

3.2 Overflow without waiting rooms

3.2.1 Steady-state equations and separation approach

In this section, we consider the basic p-overflow model without waiting rooms, that is, the capacity q_i of the waiting room in Q_i is 0 and the total capacity of Q_i is $k_i = n_i$ for $i = 1, 2$. In this case, all model variations coincide. This basic model and the derivation of the steady-state probabilities will be the starting point for the derivations in the more sophisticated models. Their analysis is a generalization of this basic approach.

Let Q_i, $i = 1, 2$, be fed by a Poisson arrival stream with intensity $\lambda_i > 0$ and let the service times at Q_i be independently identically and exponentially distributed with mean $1/\mu_i > 0$. The arrival stream from Q_1 is blocked and directed to Q_2 if all servers in the first queue are busy. Overflowing Q_1-customers and arriving Q_2-customers are lost if all servers in Q_2 are busy. The bivariate server and waiting room demand distribution is the unique nonnegative and normalized solution of the following equilibrium equations:

$$\begin{aligned}(\lambda_1(1-\delta_{in_1}) + p\lambda_1\delta_{in_1}(1-\delta_{jn_2}) + \lambda_2(1-\delta_{jn_2}) + i\mu_1 + j\mu_2)p_{i,j} \\ = \lambda_1(1-\delta_{i0})p_{i-1,j} + (1-\delta_{in_1})(i+1)\mu_1 p_{i+1,j} \qquad (3.2.1) \\ + (1-\delta_{j0})(p\lambda_1\delta_{in_1} + \lambda_2)p_{i,j-1} + (1-\delta_{jn_2})(j+1)\mu_2 p_{i,j+1}\end{aligned}$$

for $i = 0, \ldots, n_1$ and $j = 0, \ldots, n_2$, where δ_{ij} is the Kronecker symbol, i.e., $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise. Equation (3.2.1) for $i \neq n_1$ characterizes the flow into and out of states which is not due to an overflow of customers. In contrast to that, equation (3.2.1) for $i = n_1$ characterizes the flow which is caused by the overflow mechanism.

For $i \neq n_1$ the equations (3.2.1) are given by

$$\begin{aligned}(\lambda_1 + \lambda_2(1-\delta_{jn_2}) + i\mu_1 + j\mu_2)p_{i,j} = \lambda_1(1-\delta_{i0})p_{i-1,j} + (i+1)\mu_1 p_{i+1,j} \\ + (1-\delta_{j0})\lambda_2 p_{i,j-1} + (1-\delta_{jn_2})(j+1)\mu_2 p_{i,j+1}\end{aligned} \tag{3.2.2}$$

for $i = 0, \ldots, n_1 - 1$ and $j = 0, \ldots, n_2$. From these equation one might predict that the queue lengths of Q_1 and Q_2 behave independent from each other given that Q_1 is not fully occupied. In fact, it turns out that the variables in (3.2.2) can be separated, in the sense that there are solutions of the form $p_{i,j} = \alpha_i \beta_j$ for $i = 0, \ldots, n_1$ and $j = 0, \ldots, n_2$ as we will show in the following.

Setting $p_{i,j} = \alpha_i \beta_j$ for $i = 0, \ldots, n_1$ and $j = 0, \ldots, n_2$ in (3.2.2) and evaluating the result in $j = 0$ results in the equations

$$(\lambda_1 + i\mu_1 + c)\alpha_i = \lambda_1(1-\delta_{i0})\alpha_{i-1} + (i+1)\mu_1\alpha_{i+1} \tag{3.2.3}$$

for $i = 0, \ldots, n_1 - 1$ with the separation constant $c = \lambda_2 - \mu_2\frac{\beta_1}{\beta_0}$. On the other hand, the evaluation of the result in $i = 0$ leads to

$$(\lambda_2(1-\delta_{jn_2}) + j\mu_2 - c)\beta_j = \lambda_2(1-\delta_{j0})\beta_{j-1} + (1-\delta_{jn_2})(j+1)\mu_2\beta_{j+1} \tag{3.2.4}$$

for $j = 0, \ldots, n_2$ with the separation constant $c = -\lambda_1 + \mu_1\frac{\alpha_1}{\alpha_0}$. Note that

$$c = \lambda_2 - \mu_2\frac{\beta_1}{\beta_0} = -\lambda_1 + \mu_1\frac{\alpha_1}{\alpha_0}, \tag{3.2.5}$$

which follows from (3.2.2) with $i = j = 0$. By comparing (3.2.3) and (3.2.4) we get the following Lemma.

Lemma 3.2.1. *If $\alpha_i = s_i(c, \lambda_1, \mu_1)$, $i = 0, \ldots, n_1$, is a solution of* (3.2.3) *for $i = 0, \ldots, n_1 - 1$, then a solution of* (3.2.4) *for $j = 0, \ldots, n_2 - 1$ is given by $\beta_j = s_j(-c, \lambda_2, \mu_2)$, $j = 0, \ldots, n_2$.*

By this lemma it is only necessary to find a solution of the equations (3.2.3) for $\lambda_1, \mu_1 > 0$ and $c \in \mathbb{R}$. Their solution instantaneously yields a solution of the equations (3.2.4). In fact, the solution can be given in closed form in terms of the system parameters as a polynomial in the separation variable c. We will prove this fact and state the solution and some useful properties of these polynomials in the next section.

3.2.2 Solution for the separation variables

We derive the solution and some useful properties of the solutions of the basic recurrence equation (3.2.3) in this section. Let $c \in \mathbb{R}$ and $\lambda, \mu > 0$ be arbitrary and define $\alpha_i = s_i(c) = s_i(c, \lambda, \mu)$, $i \geq 0$, recursively by $s_0(c) = 1$ and (3.2.3), i.e.,

$$(\lambda + i\mu + c)s_i(c) = \lambda(1 - \delta_{i0})s_{i-1}(c) + (i+1)\mu s_{i+1}(c), \quad i \geq 0. \tag{3.2.6}$$

Multiplying this recurrence relation by z^i, $z \in \mathbb{R}$, and summing over $i \geq 0$ yields the differential equation

$$\mu(1-z)f_c'(z) = (\lambda(1-z) + c)f_c(z), \tag{3.2.7}$$

where f_c is the generating function of the sequence $(s_i(c, \lambda, \mu))_{i \geq 0}$, that is,

$$f_c(z) = \sum_{i=0}^{\infty} s_i(c, \lambda, \mu) z^i \tag{3.2.8}$$

for $z \in \mathbb{R}$ with $|z| < 1$. The condition $s_0(c, \lambda, \mu) = 1$ yields the boundary condition $f_c(0) = 1$ for the differential equation (3.2.7). The solution of (3.2.7) with $f_c(0) = 1$ is given by

$$f_c(z) = e^{\frac{\lambda}{\mu}z}(1-z)^{-\frac{c}{\mu}}. \tag{3.2.9}$$

Using the power series expansions

$$e^{az} = \sum_{n=0}^{\infty} \frac{a^n}{n!} z^n \quad \text{and}$$
$$(1-z)^b = \sum_{n=0}^{\infty} \binom{b}{n} (-1)^n z^n,$$

where $\binom{b}{n} = \frac{b(b-1)\cdot\ldots\cdot(b-n+1)}{n!}$ for $b \in \mathbb{R}$ is the generalized binomial coefficient, we get a closed-form expression for the coefficients $s_i(c, \lambda, \mu)$, $i \geq 0$, by the identity theorem for power series:

Proposition 3.2.2. *Let $c \in \mathbb{R}$, $\lambda > 0$ and $\mu > 0$ be arbitrary, then the solution of (3.2.6) with $s_0(c, \lambda, \mu) = 1$ is given by*

$$s_i(c, \lambda, \mu) = \frac{1}{i!}\left(\frac{\lambda}{\mu}\right)^i \sum_{k=0}^{i} \binom{i}{k} \lambda^{-k} \prod_{j=0}^{k-1}(j\mu + c) \tag{3.2.10}$$

$$= \left(\frac{\lambda}{\mu}\right)^i \sum_{k=0}^{i} \frac{\lambda^{-k}}{(i-k)!\,k!} \left(\frac{c}{\mu}\right)_k \tag{3.2.11}$$

for $i \geq 1$, where $(\alpha)_0 = 1$ and $(\alpha)_k = \prod_{j=0}^{k-1}(j+\alpha)$, $k \geq 1$, is the Pochhammer symbol.

Note that we have assumed that the sequence $(s_i(c,\lambda,\mu))_{i\geq 0}$ is indeed absolutely summable. This is immediately clear if we define f_c for $|z| < 1$ by (3.2.9) and develop f_c into the power series (3.2.8). The functions $s_i(c,\lambda,\mu)$, $i \geq 0$, are known as the *Poisson-Charlier polynomials* in a more general setting (see Abramowitz and Stegun [1]) and are connected to a family of special functions, the *confluent hypergeometric functions of the second kind*. For $a, b, z \in \mathbb{C}$ with $b \neq 0, -1, -2, \ldots$ and $|\arg(z)| < \pi$ the confluent hypergeometric function of the second kind $U(a,b,z)$ is a solution of the Kummer differential equation

$$zy'' + (b-z)y' - ay = 0$$

with boundary conditions $y(a,b,0) = 1$ and $\frac{\partial}{\partial z} y(a,b,z)|_{z=0} = a/b$. For $\mathrm{Re}(a) > 0$ and $\mathrm{Re}(z) > 0$, it has the integral representation

$$U(a,b,z) = \frac{1}{\Gamma(a)} \int_0^\infty e^{-zt} t^{a-1} (1+t)^{b-a-1} dt$$

(see Abramowitz and Stegun [1], Chapter 13 and Srivastava and Kashyap [60], Chapter II.3). The following proposition shows the connection between the functions $s_i(c,\lambda,\mu)$, $i \geq 0$, the confluent hypergeometric function and the *signless Stirling numbers of the first kind.*

Proposition 3.2.3. *Let $c \in \mathbb{R}$, $\lambda > 0$ and $\mu > 0$ be arbitrary and $s_i(c,\lambda,\mu)$, $i \geq 0$, be defined by* (3.2.10).

a) For every $c \notin \{\mu(n+1-i) \mid n \in \mathbb{N}_0\}$ it holds

$$s_i(c,\lambda,\mu) = \frac{1}{i!}\, U\left(-i, 1-i-\frac{c}{\mu}, \frac{\lambda}{\mu}\right).$$

b) For $0 \leq n \leq k$ let $\sigma(k,n)$ equal the number of permutations of k elements which contain exactly n permutation cycles. Then

$$s_i(c,\lambda,\mu) = \frac{1}{i!} \sum_{n=0}^{i} \left(\frac{c}{\mu}\right)^n \sum_{k=n}^{i} \binom{i}{k} \left(\frac{\lambda}{\mu}\right)^{i-k} \sigma(k,n) \tag{3.2.12}$$

holds. The numbers $\sigma(k,n)$*,* $0 \leq n \leq k$*, are called the* signless Stirling number of the first kind.

Proof. For (a) see Abramowitz and Stegun [1], Chapter 13. To show (b) we use (3.2.10) and $\prod_{j=0}^{k-1}(j-b) = \sum_{n=0}^{k} \sigma(k,n)b^n$, $b \in \mathbb{R}$, (see Abramowitz and Stegun [1], Chapter 24.1.3) □

The next Proposition states some properties of the functions $s_i(c, \lambda, \mu)$ for $i \geq 0$. These properties will be useful in the sequel.

Proposition 3.2.4. *Let* $\lambda, \mu > 0$ *and* $c \in \mathbb{R}$*. The solution* $s_i(c, \lambda, \mu)$*,* $i \geq 0$*, of the equations* (3.2.6) *satisfies the recurrence relations*

$$\mu(i+1)s_{i+1}(c) = \lambda s_i(c) + cs_i(c+\mu) \quad \textit{and} \tag{3.2.13}$$

$$s_i(c) = s_i(c+\mu) - s_{i-1}(c+\mu)(1-\delta_{i0}) \tag{3.2.14}$$

for $i \geq 0$*. Moreover, for every* $n \geq 0$*, the equations*

$$\sum_{i=0}^{n} s_i(c) = s_n(c+\mu), \tag{3.2.15}$$

$$\sum_{i=0}^{n} (n-i)s_i(c) = (1-\delta_{n0})s_{n-1}(c+2\mu) \quad \textit{and} \tag{3.2.16}$$

$$\sum_{i=1}^{n} is_i(c) = ns_n(c+\mu) - (1-\delta_{n0})s_{n-1}(c+2\mu) \tag{3.2.17}$$

hold.

Proof. Rewriting the differential equation (3.2.7) using (3.2.9) yields

$$f_c'(z) = \frac{\lambda}{\mu} f_c(z) + \frac{c}{\mu} f_{c+\mu}(z).$$

The power series expansions of f_c and $f_{c+\mu}$ imply

$$\sum_{i=1}^{\infty} is_i(c)z^{i-1} = \frac{\lambda}{\mu}\sum_{i=0}^{\infty} s_i(c)z^i + \frac{c}{\mu}\sum_{i=0}^{\infty} s_i(c+\mu)z^i$$

for all $|z| < 1$. Therefore, by multiplying this equation with μ and using the identity theorem for power series, we conclude that

$$\mu(i+1)s_{i+1}(c) = \lambda s_i(c) + cs_i(c+\mu), \quad i \geq 0,$$

i.e., that (3.2.13) holds. In the same manner, (3.2.14) follows from

$$f_c(z) = (1-z)f_{c+\mu}(z).$$

Summing (3.2.14) over $i = 0, \dots, n$ gives immediately (3.2.15). (3.2.16) follows easily from (3.2.13) by summing over $i = 0, \dots, n$ and using (3.2.15). Finally, (3.2.17) follows from (3.2.15) together with (3.2.16). □

3.2.3 Boundary condition and separation constants

We have shown in the previous section in Lemma 3.2.1 and Proposition 3.2.2 that a solution of (3.2.3) for $i = 0, \dots, n_1$ and (3.2.4) for $j = 0, \dots, n_2 - 1$ with $\alpha_0 = \beta_0 = 1$ is given by $\alpha_i = s_i(c, \lambda_1, \mu_1)$ for $i = 0, \dots, n_1$ and $\beta_j = s_j(-c, \lambda_2, \mu_2)$ for $j = 0, \dots, n_2$. Consequently, a natural candidate for the solution of the steady-state equations (3.2.1) is

$$p_{i,j} = \alpha_i \beta_j = s_i(c, \lambda_1, \mu_1) s_j(-c, \lambda_2, \mu_2)$$

for $i = 0, \dots, n_1$ and $j = 0, \dots, n_2$.

It still remains to determine the separation constant c, which can in view of (3.2.5) be reformulated as

$$c = \lambda_2 - \mu_2 s_1(-c, \lambda_2, \mu_2) = -\lambda_1 + \mu_1 s_1(c, \lambda_1, \mu_1).$$

This equation is a tautology as one could expect and therefore gives no additional information about c. In the previous sections, all steady-state equations despite the equation for $j = n_2$ have been utilized to find α_i and β_j. By taking the boundary equation for $j = n_2$ into account, we will derive the crucial condition in the following lemma that will be used to determine the separation constant c – or more precisely as we will see – the feasible separation constants $c_1, \dots, c_{n_2}$.

Lemma 3.2.5. *Let* $\beta_j = s_j(-c, \lambda_2, \mu_2)$ *for* $j \geq 0$ *and* $\lambda_2, \mu_2 > 0$ *and* $c \in \mathbb{R}$ *be given by* (3.2.10)*. Then the equation* (3.2.4) *for* $j = n_2$ *is equivalent to*

$$c s_{n_2}(\mu_2 - c, \lambda_2, \mu_2) = 0. \tag{3.2.18}$$

Proof. Setting $\beta_j = s_j(-c, \lambda_2, \mu_2)$ for $j = n_2 - 1$ and $j = n_2$ in (3.2.4), $\lambda = \lambda_2$, $\mu = \mu_2$ and substituting c with $-c$ immediately yields

$$(n_2\mu_2 - c) s_{n_2}(-c, \lambda_2, \mu_2) = \lambda_2 s_{n_2-1}(-c, \lambda_2, \mu_2). \tag{3.2.19}$$

Additionally, setting $i = n_2$ in (3.2.13) gives

$$n_2\mu_2 s_{n_2}(-c, \lambda_2, \mu_2) = \lambda_2 s_{n_2-1}(-c, \lambda_2, \mu_2) - c s_{n_2-1}(\mu_2 - c, \lambda_2, \mu_2).$$

Substituting $n_2\mu_2 s_{n_2}(-c, \lambda_2, \mu_2)$ from the latter equation on the right side of equation (3.2.19) yields

$$c\,(s_{n_2}(-c, \lambda_2, \mu_2) + s_{n_2-1}(\mu_2 - c, \lambda_2, \mu_2)) = 0.$$

Together with (3.2.14) this finally shows (3.2.18). The argument can simply be reversed to show the other direction of the equivalence. □

Equation (3.2.18) provides a condition from which the feasible candidates for the separation constant c can be derived. In fact, it is immediately seen from Proposition 3.2.2 that $s_{n_2}(\mu_2 - c, \lambda_2, \mu_2)$ is a polynomial of degree n_2 in c for every $n_2 \geq 1$. Furthermore, it follows from Proposition 3.2.2 that $s_{n_2}(\mu_2 - c, \lambda_2, \mu_2)$ has n_2 negative and pairwise distinct roots $c_1, \ldots, c_{n_2}$ for every $n_2 \geq 1$, as we will show in the following. For the moment we will suppress the indexes of the system parameters λ_i, μ_i and n_i for $i = 1, 2$ and substitute $\mu_2 - c$ with c.

As before let $s_i(c, \lambda, \mu)$ for $i \geq 0$ be given by $s_0(c, \lambda, \mu) = 1$ and the recurrence relations (3.2.6). Now we investigate the connection between the zeros of $s_i(c, \lambda, \mu)$ and those of $s_{i+1}(c, \lambda, \mu)$ as functions of c for $i \geq 1$. We will show that these two sets of zeros satisfy an interlacing property similar to the one observed in the previous chapter in (2.2.32) and the proof of Theorem 1.2.3. Although the two sets of zeros satisfy a similar interlacing property, the proof of this property has to be carried out in a different way. This can be done by using the so called *Sturm sequence property* or *roots separation theorem*. For the sake of completeness, we will state the theorem with a short proof (see Theorem 8.4.1 in Golub and Loan [26] and Chapter 5, §37 in Wilkinson [65]).

Theorem 3.2.6 (Sturm sequence property)**.** *For every $n \geq 1$ let*

$$A_n = (a_{i,j})_{i,j=1,\ldots,n}$$

be an $n \times n$ symmetric tridiagonal matrix with real entries. Further suppose that A_n is unreduced, i.e., A_n has nonnegative elements on the secondary diagonals. Then A_n has n pairwise distinct real eigenvalues. Let the eigenvalues

$$\lambda_1(A_n), \ldots, \lambda_n(A_n)$$

of A_n be ordered in increasing order, then the eigenvalues of A_n strictly separate the eigenvalues of A_{n+1}:

$$\lambda_1(A_{n+1}) < \lambda_1(A_n) < \lambda_2(A_{n+1}) < \ldots < \lambda_n(A_{n+1}) < \lambda_n(A_n) < \lambda_{n+1}(A_{n+1}). \tag{3.2.20}$$

This relation is called the strict interlacing property*.*

Proof. By Chapter 2, §47 in Wilkinson [65], the eigenvalues of A_n separate the eigenvalues of A_{n+1} at least in the weak sense, meaning

$$\lambda_1(A_{n+1}) \leq \lambda_1(A_n) \leq \lambda_2(A_{n+1}) \leq \ldots \leq \lambda_n(A_{n+1}) \leq \lambda_n(A_n) \leq \lambda_{n+1}(A_{n+1}). \tag{3.2.21}$$

Let χ_n be the characteristic polynomial of A_n, i.e.,

$$\chi_n(t) = \det(tE_n - A_n)$$

for $t \in \mathbb{R}$ and set $b_n = a_{n,n+1} = a_{n+1,n}$ for $n \geq 1$. By Laplace expansion of χ_n we get the recursive equations

$$\chi_n(t) = (t - a_{n,n})\chi_{n-1}(t) - b_{n-1}^2\chi_{n-2}(t) \tag{3.2.22}$$

for $n \geq 2$, where we have set $\chi_0(t) = 1$. Suppose that $\chi_n(t_0) = \chi_{n-1}(t_0) = 0$ for some $t_0 \in \mathbb{R}$ and $n \geq 2$. Then it follows from (3.2.21) and the assumption that A_k is unreduced, i.e., $b_k \neq 0$ for all $k \geq 1$, by induction that

$$\chi_0(t_0) = \chi_1(t_0) = \ldots = \chi_n(t_0) = 0.$$

This is a contradiction to $\chi_0(t_0) = 1$ and thus we must have strict inequalities in (3.2.21). □

By the roots separation theorem, the goal is to describe the recurrence relations (3.2.6) by a symmetric tridiagonal matrix. Therefore, we start with symmetrizing these relations.

Lemma 3.2.7. *Set*

$$u_i(c) = u_i(c, \lambda, \mu) = \left(\frac{i!\ \mu^{i+1}}{\lambda^i}\right)^{\frac{1}{2}} s_i(c, \lambda, \mu) \tag{3.2.23}$$

for $i \geq 0$. Then $s_0(c, \lambda, \mu) = 1$ and the equations (3.2.6) *are equivalent to*

$u_0(c) = \mu^{\frac{1}{2}}$ *and the symmetrized equations*

$$(1-\delta_{i0})i^{\frac{1}{2}}u_{i-1}(c) - \left(i + \frac{\lambda + c}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^{-\frac{1}{2}} u_i(c) + (i+1)^{\frac{1}{2}}u_{i+1}(c) = 0 \tag{3.2.24}$$

for $i \geq 0$.

In order to write the symmetrized recurrence relations (3.2.24) from the previous Lemma in matrix form, we introduce the matrices

$$M_n = M_n(\lambda, \mu) = (m_{i,j})_{i,j=1,\ldots,n}$$

for $n \geq 1$ defined by

$$\begin{aligned} m_{i,i} &= -(i-1)\left(\frac{\lambda}{\mu}\right)^{-\frac{1}{2}} && \text{for } i = 1, \ldots, n, \\ m_{i,i+1} = m_{i+1,i} &= i^{\frac{1}{2}} && \text{for } i = 1, \ldots, n-1 \text{ and} \\ m_{i,j} &= 0 && \text{for } |i-j| > 1. \end{aligned} \tag{3.2.25}$$

M_n is given in matrix form by

$$\begin{pmatrix} 0 & 1 & 0 & \cdots & \cdots & 0 \\ 1 & -\left(\frac{\lambda}{\mu}\right)^{-\frac{1}{2}} & \sqrt{2} & \ddots & & \vdots \\ 0 & \sqrt{2} & -2\left(\frac{\lambda}{\mu}\right)^{-\frac{1}{2}} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & -(n-2)\left(\frac{\lambda}{\mu}\right)^{-\frac{1}{2}} & \sqrt{n-1} \\ 0 & \cdots & \cdots & 0 & \sqrt{n-1} & -(n-1)\left(\frac{\lambda}{\mu}\right)^{-\frac{1}{2}} \end{pmatrix}$$

for $n \geq 1$. M_n is a real symmetric matrix. Thus M_n is diagonalizable and possesses only real eigenvalues. Furthermore, M_n is tridiagonal, that is $m_{i,j} = 0$ for $|i-j| > 1$, and has nonzero elements on the secondary diagonals, i.e., $m_{i,i+1}, m_{i+1,i} \neq 0$ for $i = 1, \ldots, n-1$. Then it follows from Theorem 3.2.6 that M_n has n pairwise distinct eigenvalues.

Let $u(c)^\top = (u_0(c), \ldots, u_{n-1}(c))^\top \in \mathbb{R}^n$ be given by (3.2.23) and consider the following system of linear equations:

$$\left(M_n - \frac{\lambda + c}{\mu}\left(\frac{\lambda}{\mu}\right)^{-\frac{1}{2}} E_n\right) u(c) = 0, \tag{3.2.26}$$

where E_n is the identity matrix in $\mathrm{Mat}(n, n, \mathbb{R})$. From (3.2.24) and (3.2.26) we immediately obtain the following result.

Lemma 3.2.8. *Let $u(c)^\top = (u_0(c), \ldots, u_{n-1}(c))^\top \in \mathbb{R}^n$ and $u_n(c) = 0$. Then the following statements are equivalent:*

(i) $u(c)$ is a nontrivial solution of (3.2.26)*.*

(ii) $u_0(c), \ldots, u_n(c)$ is a solution of (3.2.24) *for $i = 0, \ldots, n-1$ with $u_0 \neq 0$.*

From the roots separation theorem and Lemma 3.2.8 we get the desired statement about the zeros of $s_n(c, \lambda, \mu)$:

Theorem 3.2.9. *For $n \geq 1$, $c \in \mathbb{R}$, $\lambda > 0$ and $\mu > 0$ let $s_n(c, \lambda, \mu)$ be defined by* (3.2.10)*. Then the equation*

$$s_n(c, \lambda, \mu) = 0$$

has n negative and distinct solutions $c_1, \ldots, c_n$ in the variable c given by

$$c_i = (\lambda\mu)^{\frac{1}{2}} \varepsilon_i - \lambda \tag{3.2.27}$$

for $i = 1, \ldots, n$, where $\varepsilon_1, \ldots, \varepsilon_n$ are the pairwise distinct eigenvalues of the matrix $M_n(\lambda, \mu)$.

Proof. Let $n \geq 1$. The matrix M_n satisfies the assumptions of the roots reparation theorem and therefore has n real distinct eigenvalues $\varepsilon_1, \ldots, \varepsilon_n$. Thus, the equation

$$\left(M_n - \frac{\lambda + c}{\mu} \left(\frac{\lambda}{\mu} \right)^{-\frac{1}{2}} E_n \right) u = 0$$

with $c \in \mathbb{R}$ and $u^\top = (u_0, \ldots, u_{n-1})^\top \in \mathbb{R}^n$ has n real and linearly independent solutions. The solutions are of the form (c_j, u^j) for $j = 1, \ldots, n$, where

$$c_j = (\lambda\mu)^{\frac{1}{2}} \varepsilon_j - \lambda$$

and $u^j = (u_0^j, \ldots, u_{n-1}^j)$ is an eigenvector of M_n to the eigenvalue ε_j for $j = 1, \ldots, n$. We define $u_n^j = 0$. By Remark 3.2.8 (i) u_0^j is nonzero for all j. Thus, we can normalize the eigenvector u^j such that $u_0^j = \mu^{\frac{1}{2}}$. We further obtain from Lemma 3.2.8 that $c_j, u_0^j, \ldots, u_n^k$ is a solution of (3.2.24) for all $j = 1, \ldots, n$. If we define $s_i(c_j, \lambda, \mu)$ by (3.2.23) with $u_0, \ldots, u_n$ replaced by $u_0^j, \ldots, u_n^j$, then $s_0(c_j, \lambda, \mu) = 1$ and the equations (3.2.6) are fulfilled. Then Proposition 3.2.2 yields that $s_i(c_j, \lambda, \mu)$ is of the form (3.2.10) for $i = 1, \ldots, n$. In particular $s_n(c_j, \lambda, \mu)$ is of the form (3.2.10), a polynomial of degree n and $u_n^j = 0$ gives $s_n(c_j, \lambda, \mu) = 0$ for $j = 1, \ldots, n$. It follows from (3.2.10) that $s_n(c, \lambda, \mu) > 0$ for $c \geq 0$ so that the numbers $c_1, \ldots, c_n$ must be negative. □

Using the Sturm sequence property one can show another result that reduces the computational requirements for the derivation of the eigenvalues of a symmetric tridiagonal matrix with nonzero elements on the secondary diagonals substantially: Consider the conditions of Theorem 3.2.6 and let $s(t)$ be the number of sign changes of the sequence $\chi_0(t), \ldots, \chi_n(t)$, where $\chi_k(t)$ is defined by (3.2.22) with the additional convention that the sign of $\chi_k(t)$ is $-\operatorname{sgn}(\chi_{k-1}(t))$ if $\chi_k(t) = 0$. Then $s(t)$ equals the number of eigenvalues of A_n that are less than t. Based on this observation, fast and accurate algorithms - like bisection, multisection, polysection or Godunov-inverse iteration - for the numerical computation of the eigenvalues have been developed. A huge literature is available for real symmetric tridiagonal eigenvalue problems (see Cullum [14], Chapter 3, Section 3.5, Golub and Loan [26], §8.4, Matsekh [45] and Swarztrauber [62]).

We summarize the results for the number and location of the separation constants described by equation (3.2.18) in the following conclusion of Theorem 3.2.9.

Corollary 3.2.10. *For $n_2 \geq 0$, $\lambda_2 > 0$ and $\mu_2 > 0$ let $s_{n_2}(\mu_2 - c, \lambda_2, \mu_2)$ be defined by* (3.2.10). *Then the equation*

$$c s_{n_2}(\mu_2 - c, \lambda_2, \mu_2) = 0$$

has $n_2 + 1$ and distinct solutions $c_0, \ldots, c_{n_2}$ in the variable c given by $c_0 = 0$ and

$$c_i = \lambda_2 + \mu_2 - (\lambda_2 \mu_2)^{\frac{1}{2}} \varepsilon_i \tag{3.2.28}$$

for $i = 1, \ldots, n$, where $\varepsilon_1, \ldots, \varepsilon_{n_2}$ are the pairwise distinct eigenvalues of the matrix $M_n(\lambda_2, \mu_2)$ defined by (3.2.25). *Moreover, $c_1, \ldots, c_{n_2}$ are positive.*

3.2.4 Steady-state probabilities

We have shown in Corollary 3.2.10 that $s_{n_2}(\mu_2 - c, \lambda_2, \mu_2)$ is a polynomial in c of degree n_2 with n_2 positive and distinct zeros which can be found by solving the eigenvalue problem (3.2.26), i.e.,

$$\left(M_n(\lambda_2, \mu_2) - \frac{\lambda_2 + c}{\mu_2} \left(\frac{\lambda_2}{\mu_2} \right)^{-\frac{1}{2}} E_n \right) u = 0.$$

By denoting the zeros by $c_0, \ldots, c_{n_2}$ as in Corollary 3.2.10, we have

$$c_m s_{n_2}(\mu_2 - c_m, \lambda_2, \mu_2) = 0 \tag{3.2.29}$$

for $m = 0, \ldots, n_2$. For every $m = 0, \ldots, n_2$, we get a solution of the steady-state equations (3.2.1) by setting

$$p_{i,j} = s_i(c_m, \lambda_1, \mu_1) s_j(-c_m, \lambda_2, \mu_2) \tag{3.2.30}$$

for $i = 0, \ldots, n_1$ and $j = 0, \ldots, n_2$. Therefore, we can represent the existing and unique stationary probabilities $p_{i,j}$ as linear combinations of the solutions (3.2.30) of the separation approach in the form

$$p_{i,j} = \sum_{m=0}^{n_2} a_m s_i(c_m, \lambda_1, \mu_1) s_j(-c_m, \lambda_2, \mu_2) \tag{3.2.31}$$

for $i = 0, \ldots, n_1$ and $j = 0, \ldots, n_2$. The constants $a_0, \ldots, a_{n_2}$, have to be chosen such that the boundary conditions in (3.2.1) corresponding to $i = n_1$ and the normalization condition are satisfied.

The boundary conditions for $i = n_1$, $j = 0, \ldots, n_2 - 1$ are

$$\begin{aligned} &(p\lambda_1 + \lambda_2 + n_1\mu_1 + j\mu_2) p_{n_1,j} \\ &\quad = \lambda_1 p_{n_1-1,j} + (1 - \delta_{j0})(p\lambda_1 + \lambda_2) p_{n_1,j-1} + (j+1)\mu_2 p_{n_1,j+1}. \end{aligned} \tag{3.2.32}$$

The boundary condition for $i = n_1$ and $j = n_2$ is

$$(n_1\mu_1 + n_2\mu_2) p_{n_1,n_2} = \lambda_1 p_{n_1-1,n_2} + (p\lambda_1 + \lambda_2) p_{n_1,n_2-1}. \tag{3.2.33}$$

Inserting (3.2.31) into (3.2.32) and (3.2.33) and simplifying with the help of the recurrence relations (3.2.13) and (3.2.14) leads to

$$\sum_{m=0}^{n_2} a_m \left(c_m s_{n_1}(c_m + \mu_1) s_j(-c_m) + p\lambda_1 s_{n_1}(c_m) s_j(-c_m - \mu_2) \right) = 0 \tag{3.2.34}$$

for $i = n_1$, $j = 0, \dots, n_2 - 1$ and

$$\sum_{m=0}^{n_2} a_m \left(c_m s_{n_1}(c_m + \mu_1) s_{n_2-1}(-c_m + \mu_2) + p\lambda_1 s_{n_1}(c_m) s_{n_2-1}(-c_m)\right) = 0 \tag{3.2.35}$$

with the abbreviated notation $s_{n_1}(\cdot) = s_{n_1}(\cdot, \lambda_1, \mu_1)$ and $s_j(\cdot) = s_j(\cdot, \lambda_2, \mu_2)$ for $j = 0, \dots, n_2$. Together with the normalization condition there are n_2+2 linear equations for the unknowns $a_0, \dots, a_{n_2}$. Summing (3.2.34) over $j = 0, \dots, n_2 - 1$ and using (3.2.15) yields the redundancy of (3.2.35). Furthermore, inserting (3.2.31) into the normalization condition $\sum_{i=0}^{n_1} \sum_{j=0}^{n_2} p_{i,j} = 1$ and using (3.2.29) and (3.2.15) leads to an explicit expression for a_0:

$$a_0 = \left(s_{n_1}(\mu_1, \lambda_1, \mu_1) s_{n_2}(\mu_2, \lambda_2, \mu_2)\right)^{-1}. \tag{3.2.36}$$

Remark 3.2.11. It is seen from (3.2.31) and (3.2.34) that the probabilities $p_{i,j}$ depend on the parameter $p \in [0, 1]$ only through the coefficients $a_0, \dots, a_{n_2}$. This is due to the fact that the stationary probabilities have been calculated from the nonoverflow balance equations, i.e., (3.2.1) for $i \neq n_1$, which are independent of p. Whereas the equations for the coefficients $a_0, \dots, a_{n_2}$ arise from the overflow balance equations, i.e., (3.2.1) for $i = n_1$, which depend on p.

We summarize the results in the following theorem.

Theorem 3.2.12. *The unique nonnegative and normalized solution of the steady-state equations* (3.2.1) *is given by*

$$p_{i,j} = \sum_{m=0}^{n_2} a_m s_i(c_m, \lambda_1, \mu_1) s_j(-c_m, \lambda_2, \mu_2) \tag{3.2.37}$$

for $i = 0, \dots, n_1$ *and* $j = 0, \dots, n_2$ *if the coefficients* $a_0, \dots, a_{n_2}$ *are determined by* (3.2.34) *and* (3.2.36), *where* $c_0 = 0$ *and* $c_1, \dots, c_{n_2}$ *are the by Corollary 3.2.10 positive and pairwise distinct solutions of the equation* $s_{n_2}(\mu_2 - c_m, \lambda_2, \mu_2) = 0$, $m = 1, \dots, n_2$.

The problem of determining the $(n_1 + 1)(n_2 + 1)$ unknowns $p_{i,j}$, $i = 0, \dots, n_1$, $j = 0, \dots, n_2$, has now been reduced to the problem of determining n_2 eigenvalues from (3.2.29) and $n_2 + 1$ unknowns from the homogeneous linear equations (3.2.34) and (3.2.36).

The separation approach is displayed in Figure 3.3. The single line characterizes the boundary conditions, the unfilled endpoint characterizes the redundant boundary condition at $(i, j) = (n_1, n_2)$.

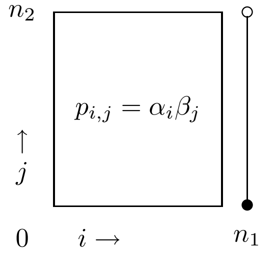


Figure 3.3: Separation scheme: Overflow without waiting rooms

Remark 3.2.13. By setting $n_2 = 0$, we can include the case that the second queue has no servers, i.e., the Erlang loss system. In this case, we obtain from (3.2.36) and (3.2.10) the standard formula for the stationary probabilities of the number of busy servers in the Erlang loss system with n_1 servers and traffic intensity $\rho = \lambda_1/\mu_1$:

$$P(L_1 = i) = \frac{\rho^i/i!}{1 + \rho + \ldots + \rho^{n_1}/n_1!}, \quad i = 0, \ldots, n_1.$$

Remark 3.2.14. For $n_2 > 0$ the steady-state probability that i customers are served at Q_1 is $\sum_{j=0}^{n_2} p_{i,j}$, $i = 0, \ldots, n_1$. Using (3.2.29), (3.2.36), (3.2.10) and (3.2.15) we obtain again the standard Erlang loss formula

$$P(L_1 = i) = \sum_{j=0}^{n_2} p_{i,j} = \frac{\rho^i/i!}{1 + \rho + \ldots + \rho^{n_1}/n_1!}, \quad i = 0, \ldots, n_1.$$

3.2.5 Stationary quantities and numerical results

At the end of this section, we show how the results from the separation approach can be used to derive steady-state quantities of interest. The main characteristics of a queueing network with loss and overflow are, amongst others, the average blocking and loss probabilities for arriving customers, the overflow probability and the average number of customers in service. These quantities can be given in closed form in terms of the system parameters and the unknowns described in Theorem 3.2.12.

Let B_i, $i = 1, 2$, be the probability that an arriving Q_i-customer is blocked at Q_i and let O_{12} be the expected stationary number of customers per unit time, which flow over from Q_1 to Q_1, then

$$B_1 = \sum_{j=0}^{n_2} p_{n_1,j} = \sum_{j=0}^{n_2}\sum_{m=0}^{n_2} a_m s_{n_1}(c_m, \lambda_1, \mu_1) s_j(-c_m, \lambda_2, \mu_2),$$

$$B_2 = \sum_{i=0}^{n_1} p_{i,n_2} = \sum_{i=0}^{n_1}\sum_{m=0}^{n_2} a_m s_i(c_m, \lambda_1, \mu_1) s_{n_2}(-c_m, \lambda_2, \mu_2),$$

$$O_{12} = \sum_{j=0}^{n_2-1} p_{n_1,j} = \sum_{j=0}^{n_2-1}\sum_{m=0}^{n_2} a_m s_{n_1}(c_m, \lambda_1, \mu_1) s_j(-c_m, \lambda_2, \mu_2).$$

By (3.2.15) we get

$$B_1 = \sum_{m=0}^{n_2} a_m s_{n_1}(c_m, \lambda_1, \mu_1) s_{n_2}(\mu_2 - c_m, \lambda_2, \mu_2),$$

$$B_2 = \sum_{m=0}^{n_2} a_m s_{n_1}(\mu_1 + c_m, \lambda_1, \mu_1) s_{n_2}(-c_m, \lambda_2, \mu_2),$$

$$O_{12} = \sum_{m=0}^{n_2} a_m s_{n_1}(c_m, \lambda_1, \mu_1) s_{n_2-1}(\mu_2 - c_m, \lambda_2, \mu_2).$$

The loss probability $P_{\text{Loss},1}$ for arriving Q_1-customers is then given by

$$P_{\text{Loss},1} = p_{n_1,n_2} + (1-p)O_{12} = B_1 - pO_{12},$$

the loss probability $P_{\text{Loss},2}$ for arriving Q_2-customers is equal to B_2.

EL_i is the expected average queue length of Q_i for $i = 1, 2$. A closed-form expression for EL_1 can be deduced directly from Remark 3.2.14. Moreover, EL_2 is given by

$$EL_2 = \sum_{j=1}^{n_2} j \sum_{i=0}^{n_1} p_{i,j} = \sum_{m=0}^{n_2} a_m \sum_{i=0}^{n_1} s_i(c_m, \lambda_1, \mu_1) \sum_{j=1}^{n_2} j s_j(-c_m, \lambda_2, \mu_2).$$

In order to derive EL_2, we can use (3.2.15) and (3.2.17) and get

$$\begin{aligned} EL_2 &= \sum_{m=0}^{n_2} a_m s_{n_1}(\mu_1 - c_m, \lambda_1, \mu_1)\big(n_2 s_{n_2}(\mu_2 - c_m, \lambda_2, \mu_2) \\ &\qquad - s_{n_2-1}(2\mu_2 - c_m, \lambda_2, \mu_2)\big) \\ &= n_2 - \sum_{m=0}^{n_2} a_m s_{n_1}(\mu_1 - c_m, \lambda_1, \mu_1) s_{n_2-1}(2\mu_2 - c_m, \lambda_2, \mu_2), \end{aligned}$$

where we have used Corollary 3.2.10 and equation (3.2.36) for the last equality. However, by Little's law (see for example Asmussen [5], Theorem 4.1), the expected number of customers in the second queue is equal to the expected arrival rate to that queue multiplied with the average time spent in the queue. Therefore, we have

$$EL_2 = \frac{p\lambda_1 O_{12} + \lambda_2(1 - B_2)}{\mu_2}. \tag{3.2.38}$$

Note that the numerical efforts for deriving one of the quantities B_1, B_2, O_{12}, EL_1 and EL_2 is the same as deriving one of the stationary probabilities (3.2.37). Furthermore, these quantities can be calculated directly without computing the stationary probabilities.

These results can for example be used to minimize the total average costs of lost customers subject to the overflow parameter p: The expected number of lost Q_1- and Q_2-customers is given by $\lambda_1 P_{\text{Loss},1} = \lambda_1(B_1 - pO_{12})$ and $P_{\text{Loss},1} = \lambda_2 B_2$, respectively. The total average costs of lost customers $C(p)$ is then given by

$$C(p) = C_1\lambda_1(B_1 - pO_{12}) + C_2\lambda_2 B_2,$$

where $C_i \in \mathbb{R}$, $i = 1, 2$, are cost parameters. It is evident that an increase of p results on the one hand in a decrease of the number of lost Q_1-customers because more of these customers will join Q_2. On the other hand, these overflowing Q_1-customers increase the occupation rate of Q_2 and thus, more Q_2-customers will be lost. In this view, the total costs $C(p)$ of lost customers should be a convex function of p and a cost-minimal overflow parameter p^* should exist. A numerical example is displayed in Figure 3.4. The cost minimizing values p^* in this example are $p^* = 0.9887$, 0.4838, 0.1899 and 0 (line by line from left to right) with associated minimal costs of $C(p^*) =$ 1.3714, 1.4288, 1.4775 and 1.52. Thus, by inspecting the maxima of the cost functions, it is seen that the optimal choice of the overflow parameter p can reduce the costs in a considerable amount.

The mean queue length EL_2 of the second queue can also be determined numerically from (3.2.38). Some numerical examples are given in Tables 3.1-3.6. For $p = 0$, both queues are independent and EL_2 is given by the Erlang loss formula:

$$EL_2 = \frac{\sum\limits_{i=1}^{n_1} \frac{\rho_2^i}{(i-1)!}}{1 + \sum\limits_{i=1}^{n_1} \frac{\rho_2^i}{i!}},$$

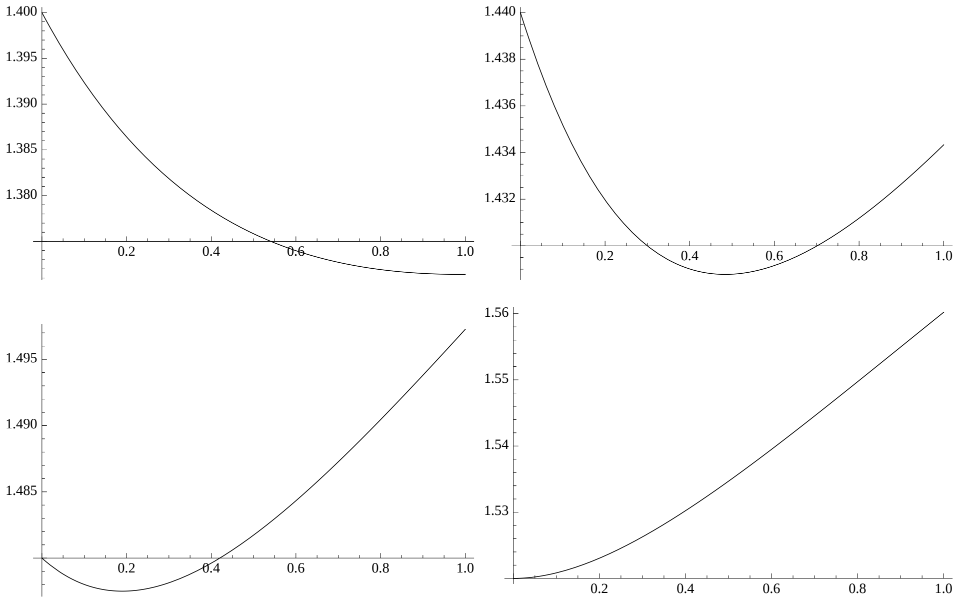


Figure 3.4: $p \mapsto C(p)$ for $(n_1, n_2, \lambda_1, \mu_1, \lambda_2, \mu_2, C_1) = (2, 2, 3, 3, 1, 1, 1.2)$ and $C_2 = 3.4$, 3.6, 3.8 and 4 (line by line from left to right)

where $\rho_2 = \lambda_2/\mu_2$. The first row of each of the Tables 3.1-3.5 reflects this observation. By inspecting the numerical results displayed in Tables 3.1-3.6, one finds that EL_2 is increasing in λ_2, n_2 and p and decreasing in μ_2. Furthermore, EL_2 is increasing for $p > 0$ in λ_1 and decreasing for $p > 0$ in n_1 and μ_1. This is on the one hand explained by the fact that letting all other values fixed, the mean queue length must increase if the capacity n_2 or the offered workload λ_2 increase and decreases if the service rate μ_2 is raised. Besides the usual influence of the arrival and service rates and the capacity on the queue length, on the other hand, an overflow is more likely if λ_1 or p increase or n_1 decreases. Thus, EL_2 increases in these cases.

p	$n_2 = 2$	$n_2 = 4$	$n_2 = 6$	$n_2 = 8$	$n_2 = 10$	$n_2 = 12$
0	1.71004	3.27076	4.59956	5.60233	6.22761	6.52748
0.2	1.73063	3.33039	4.72756	5.83362	6.58125	6.98453
0.4	1.74819	3.38066	4.83522	6.03102	6.89379	7.40765
0.6	1.76333	3.42349	4.92648	6.19970	7.16787	7.79408
0.8	1.77652	3.46035	5.00446	6.34428	7.40718	8.14308
1	1.78810	3.49235	5.07162	6.46874	7.61572	8.45559

Table 3.1: Table of EL_2 for $n_1 = 2$ and $(\lambda_1, \mu_1, \lambda_2, \mu_2) = (1, 0.1, 2, 0.3)$

p	$n_2 = 2$	$n_2 = 4$	$n_2 = 6$	$n_2 = 8$	$n_2 = 10$	$n_2 = 12$
0	1.71004	3.27076	4.59956	5.60233	6.22761	6.52748
0.2	1.72637	3.31814	4.70143	5.78646	6.50874	6.88981
0.4	1.74043	3.35861	4.78854	5.94651	6.76133	7.22921
0.6	1.75265	3.39350	4.86349	6.08565	6.98673	7.54366
0.8	1.76337	3.42382	4.92838	6.20682	7.18698	7.83228
1	1.77285	3.45038	4.98490	6.31267	7.36445	8.09517

Table 3.2: Table of EL_2 for $n_1 = 4$ and $(\lambda_1, \mu_1, \lambda_2, \mu_2) = (1, 0.1, 2, 0.3)$

p	$n_2 = 2$	$n_2 = 4$	$n_2 = 6$	$n_2 = 8$	$n_2 = 10$	$n_2 = 12$
0	1.71004	3.27076	4.59956	5.60233	6.22761	6.52748
0.2	1.72233	3.30649	4.67649	5.74141	6.43967	6.80011
0.4	1.73301	3.33738	4.74327	5.86430	6.63303	7.05812
0.6	1.74236	3.36429	4.80151	5.97283	6.80826	7.30014
0.8	1.75063	3.38790	4.85256	6.06875	6.96642	7.52542
1	1.75798	3.40876	4.89753	6.15369	7.10878	7.73374

Table 3.3: Table of EL_2 for $n_1 = 6$ and $(\lambda_1, \mu_1, \lambda_2, \mu_2) = (1, 0.1, 2, 0.3)$

p	$n_2 = 2$	$n_2 = 4$	$n_2 = 6$	$n_2 = 8$	$n_2 = 10$	$n_2 = 12$
0	1.71004	3.27076	4.59956	5.60233	6.22761	6.52748
0.2	1.71866	3.29586	4.65365	5.70014	6.37656	6.71854
0.4	1.72621	3.31779	4.70124	5.78782	6.51411	6.90094
0.6	1.73287	3.33708	4.74327	5.86635	6.64047	7.07384
0.8	1.73879	3.35416	4.78052	5.93670	6.75611	7.23671
1	1.74409	3.36936	4.81367	5.99978	6.86166	7.38927

Table 3.4: Table of EL_2 for $n_1 = 8$ and $(\lambda_1, \mu_1, \lambda_2, \mu_2) = (1, 0.1, 2, 0.3)$

p	$n_2 = 2$	$n_2 = 4$	$n_2 = 6$	$n_2 = 8$	$n_2 = 10$	$n_2 = 12$
0	1.71004	3.27076	4.59956	5.60233	6.22761	6.52748
0.2	1.71553	3.28676	4.63408	5.66475	6.32256	6.64904
0.4	1.72037	3.30088	4.66481	5.72141	6.41119	6.76592
0.6	1.72467	3.31342	4.69226	5.77279	6.49357	6.87769
0.8	1.72852	3.32460	4.71683	5.81937	6.56988	6.98402
1	1.73198	3.33463	4.73891	5.86162	6.64037	7.08470

Table 3.5: Table of EL_2 for $n_1 = 10$ and $(\lambda_1, \mu_1, \lambda_2, \mu_2) = (1, 0.1, 2, 0.3)$

λ_2	$\lambda_1 = 0.5$	$\lambda_1 = 1$	$\lambda_1 = 1.5$	$\lambda_1 = 2$	$\lambda_1 = 2.5$	$\lambda_1 = 3$
0.5	4.94770	5.74973	6.94810	7.83830	8.40927	8.77512
1	7.86791	8.13140	8.51837	8.82503	9.04536	9.20428
1.5	8.84975	8.94551	9.09722	9.23094	9.33734	9.42111
2	9.24300	9.28876	9.36491	9.43657	9.49736	9.54799
2.5	9.44250	9.46869	9.51365	9.55769	9.59659	9.63022
3	9.56073	9.57756	9.60704	9.63670	9.66362	9.68750

Table 3.6: Table of EL_2 for $n_1 = n_2 = 10$ and $(\mu_1, \mu_2, p) = (0.1, 0.1, 0.5)$

3.3 Overflow with waiting rooms: From servers to servers

3.3.1 Jockeying to servers

3.3.1.1 Steady-state equations and separation approach

Now we consider the p-overflow model S/S/S with waiting rooms. The waiting room at Q_i has capacity $q_i \geq 0$ for $i = 1, 2$. In this model, an arriving Q_1-customer is blocked and directed to Q_2 if all n_1 servers are busy in Q_1. Blocked customers are served by one of the servers in Q_2 if at least one is available, are queued in Q_1 if all servers in Q_2 are busy and a waiting position is available in Q_1 and are lost otherwise. The jockeying discipline ascertains that waiting Q_1-customers will be served at Q_1 if a Q_1-server becomes available or at Q_2 if a Q_2-server becomes available, all Q_1-servers are busy and no customers are waiting at Q_2, whatever happens first. With $k_i = q_i + n_i$ for $i = 1, 2$, the balance equations are

$$\begin{aligned}
&(\lambda_1(1-\delta_{ik_1})(1-\chi_{i-n_1}(1-\chi_{j-n_2})) + p\lambda_1\chi_{i-n_1}(1-\chi_{j-n_2})) \\
&\quad + \lambda_2(1-\delta_{jk_2}) + (i \wedge n_1)\mu_1 + (j \wedge n_2)\mu_2)p_{i,j} \\
&= (1-\chi_{i-n_1-1}\chi_{n_2-j-1}) \\
&\qquad \times \big(\lambda_1(1-\delta_{i0})p_{i-1,j} + (1-\delta_{jk_2})\left((j+1) \wedge n_2\right)\mu_2 p_{i,j+1}\big) \\
&\quad + (1-\delta_{j0})\big(p\lambda_1\delta_{in_1}\chi_{n_2-j} + \lambda_2(1-\chi_{i-n_1-1}\chi_{n_2-j})\big)p_{i,j-1} \\
&\quad + (1-\delta_{ik_1})\big((1-\chi_{i-n_1}\chi_{n_2-1-j})\left((i+1) \wedge n_1\right)\mu_1 + n_2\mu_2\delta_{jn_2}\chi_{i-n_1}\big)p_{i+1,j}
\end{aligned} \tag{3.3.1}$$

for $i = 0, \ldots, k_1$ and $j = 0, \ldots, k_2$, where $i \wedge j = \min\{i, j\}$ and $\chi_{i-n} = \mathbf{1}_{[n,\infty)}(i)$ for $i, j, n \in \mathbb{N}$.

For $i = n_1+1, \ldots, k_1$ and $j = 0, \ldots, n_2-1$ these equations imply $p_{i,j} = 0$, since it is impossible that customers are waiting at Q_1 while there is at least one server available at Q_2.

It will turn out that a similar but more complex separation approach than the one for the model without waiting rooms can be carried out for this model. This separation approach is depicted in Figure 3.5. The single lines correspond to the resulting boundary conditions and the unfilled circle highlights a redundant boundary condition at $(i, j) = (n_1, n_2)$.

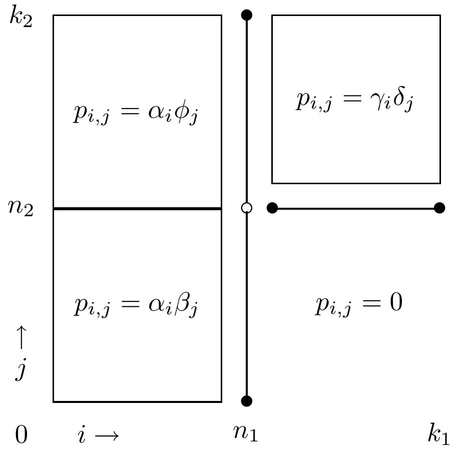

k_2
$p_{i,j} = \alpha_i \phi_j$
$p_{i,j} = \gamma_i \delta_j$
n_2
$p_{i,j} = \alpha_i \beta_j$
$p_{i,j} = 0$
↑
j
0
$i \rightarrow$
n_1
k_1

Figure 3.5: Separation scheme: Model S/S/S

For $i = 0, \ldots, n_1 - 1$ and $j = 0, \ldots, k_2$ the balance equations take the form

$$(\lambda_1 + \lambda_2(1 - \delta_{jk_2}) + i\mu_1 + (j \wedge n_2)\mu_2)p_{i,j} = \lambda_1(1 - \delta_{i0})p_{i-1,j} \\ + (i+1)\mu_1 p_{i+1,j} + (1 - \delta_{j0})\lambda_2 p_{i,j-1} + (1 - \delta_{jk_2})((j+1) \wedge n_2)\mu_2 p_{i,j+1} \tag{3.3.2}$$

for $i = 0, \ldots, n_1$ and $j = 0, \ldots, k_2$. By the results of Section 3.2, there are solutions of the form $p_{i,j} = \alpha_i \beta_j$ in the region $i = 0, \ldots, n_1$ and $j = 0, \ldots, k_2$, where α_i satisfies (3.2.3) for $i = 0, \ldots, n_1$ and is given by $\alpha_i = s_i(c, \lambda_1, \mu_1)$ for $i = 0, \ldots, n_1$. For $j = 0, \ldots, n_2$, β_j is given by or at least proportional to $\beta_j = s_j(-c, \lambda_2, \mu_2)$. c is the separation constant for this region and will be determined in the next section. Moreover, for $i = 0, \ldots, n_1$ and $j = n_2, \ldots, k_2$, the approach $p_{i,j} = \alpha_i \beta_j$ gives $\alpha_i = s_i(c, \lambda_1, \mu_1)$ and β_j satisfies

$$(\lambda_2(1 - \delta_{jk_2}) + n_2\mu_2 - c)\beta_j = \lambda_2 \beta_{j-1} + n_2\mu_2(1 - \delta_{jk_2})\beta_{j+1} \tag{3.3.3}$$

for $j = n_2, \dots, k_2$. We assume $q_1 \geq 1$ for the moment. In the region $i = n_1 + 1, \dots, k_1$ and $j = n_2 + 1, \dots, k_2$, the steady-state probabilities $p_{i,j}$ can be separated into the form $p_{i,j} = \gamma_i \delta_j$. This leads to the equations

$$\begin{aligned} &(\lambda_1(1 - \delta_{ik_1}) + n_1\mu_1 + d)\gamma_i = \lambda_1\gamma_{i-1} + (1 - \delta_{ik_1})n_1\mu_1\gamma_{i+1} \text{ and} \\ &(\lambda_2(1 - \delta_{jk_2}) + n_2\mu_2 - d)\delta_j = \lambda_2\delta_{j-1} + (1 - \delta_{jk_2})n_2\mu_2\delta_{j+1} \end{aligned} \tag{3.3.4}$$

for $i = n_1 + 1, \dots, k_1$ and $j = n_2 + 1, \dots, k_2$ with the separation constant

$$d = n_2\mu_2 - \lambda_2\frac{\delta_{k_2-1}}{\delta_{k_2}} = -n_1\mu_1 + \lambda_1\frac{\gamma_{k_1-1}}{\gamma_{k_1}}.$$

We will derive the solution of (3.3.3) and (3.3.4) and determine the separation constants in the next sections.

3.3.1.2 Solution of the separation approach

We have already shown that there are solutions of the form $p_{i,j} = \alpha_i\beta_j$ for $i = 0, \dots, n_1$ and $j = 0, \dots, n_2$. α_i equals $s_i(c, \lambda_1, \mu_1)$ for $i = 0, \dots, n_1$ and β_j is proportional to $s_i(-c, \lambda_2, \mu_2)$ for $j = 0, \dots, n_2$. c is a separation constant. Moreover, β_j has to satisfy (3.3.3) for $j = n_2, \dots, k_2$. In the region $i = n_1 + 1, \dots, k_1$ and $j = n_2 + 1, \dots, k_2$, the separation approach $p_{i,j} = \gamma_i\delta_j$ led to the equations (3.3.4). We now show that the solutions of equations (3.3.3) and (3.3.4) can be expressed in terms of the *Chebyshev polynomials of the second kind*. First, we give a formal definition of these polynomials (see for example Abramowitz and Stegun [1]).

Definition 3.3.1. The *Chebyshev polynomials of the second kind* are defined by the recurrence relation

$$2xU_l(x) = U_{l+1}(x) + U_{l-1}(x), \quad l \geq 0, \tag{3.3.5}$$

with the boundary conditions $U_{-1} \equiv 0$ and $U_0 \equiv 1$.

It is well known (see Abramowitz and Stegun [1]) that

$$U_l\left(\frac{1}{2}\left(z + \frac{1}{z}\right)\right) = \sum_{k=0}^{l} z^{2k-l} \tag{3.3.6}$$

for $z \neq 0$. The following proposition states the solution of (3.3.3).

Proposition 3.3.2. *A solution of* (3.3.3) *is given by* $\beta_j = \phi_j(c)$ *for* $j =$

$n_2-1,\dots,k_2$, where the function $\phi_i(c)$ is defined by

$$\phi_j(c) = \Psi_{k_2-j}(c) - \Psi_{k_2-j-1}(c), \quad j = n_2-1,\dots,k_2, \tag{3.3.7}$$

with

$$\Psi_l(c) = \Psi_l(c,\lambda_2,\mu_2) = \left(\frac{n_2\mu_2}{\lambda_2}\right)^{\frac{l}{2}} U_l\left(\frac{\lambda_2+n_2\mu_2-c}{2\sqrt{\lambda_2 n_2\mu_2}}\right) \tag{3.3.8}$$

for $l=-1,\dots,q_2$. Moreover, $\Psi_{-1}(c)=0$, $\Psi_0(c)=1$ and

$$(\lambda_2+n_2\mu_2-c)\Psi_l(c) = \lambda_2\Psi_{l+1}(c) + n_2\mu_2\Psi_{l-1}(c) \tag{3.3.9}$$

holds for every $l \geq 0$

Proof. $\Psi_{-1}(c)=0$, $\Psi_0(c)=1$ and (3.3.9) follow directly from (3.3.5). With $\phi_j(c)$ defined by (3.3.7) for $j=n_2-1,\dots,k_2$, we obtain (3.3.3), i.e.,

$$(\lambda_2(1-\delta_{jk_2})+n_2\mu_2-c)\phi_j(c) = \lambda_2\phi_{j-1}(c) + n_2\mu_2(1-\delta_{jk_2})\phi_{j+1}(c)$$

for $j=n_2,\dots,k_2$ from (3.3.9). □

It is immediately seen by comparing the equations (3.3.3) and (3.3.4) that a solution of (3.3.3) yields a solution of (3.3.4). For convenience, it is reasonable to set $\beta_j=\phi_j(c)$ for $j=n_2,\dots,k_2$. The solution of the equations (3.3.4) can be deduced directly from Proposition 3.3.2:

Corollary 3.3.3. *A solution of (3.3.4) is given by $\gamma_i=\theta_i(d)$ for $i=n_1,\dots,k_1$ and $\delta_j=\phi_j(d)$ for $j=n_2,\dots,k_2$ and every $d\in\mathbb{R}$, where the function $\phi_j(d)$ is defined by (3.3.7) and $\theta_i(d)$ is similarly to $\phi_j(c)$ defined by*

$$\theta_i(d) = \Omega_{k_1-i}(d) - \Omega_{k_1-i-1}(d), \quad i=0,\dots,k_1, \tag{3.3.10}$$

with

$$\Omega_l(d) = \Omega_l(d,\lambda_1,\mu_1) = \left(\frac{n_1\mu_1}{\lambda_1}\right)^{\frac{l}{2}} U_l\left(\frac{\lambda_1+n_1\mu_1+d}{2\sqrt{\lambda_1 n_1\mu_1}}\right) \tag{3.3.11}$$

for $l=-1,\dots,q_1$.

The solution of the steady-state equations can be deduced from Proposition 3.3.2 and Corollary 3.3.2 by taking the specific values of the solutions of the separation approach at the borders of the boundary region into account. This is done in the next section. We conclude this section with a note concerning $\phi_{k_2-1}(c)$ and a helpful relation between θ_{n_1} and θ_{n_1+1}.

Lemma 3.3.4. *(i) The function ϕ_{k_2-1} defined by* (3.3.7) *satisfies*

$$\phi_{k_2-1}(c) = \frac{n_2\mu_2 - c}{\lambda_2} \tag{3.3.12}$$

for every $c \in \mathbb{R}$.

(ii) The functions θ_{n_1} and θ_{n_1+1} defined by (3.3.10) *satisfy*

$$\lambda_1\theta_{n_1}(d) - n_1\mu_1\theta_{n_1+1}(d) = d\Omega_{q_1-1}(d) \tag{3.3.13}$$

for every $d \in \mathbb{R}$.

Proof. $U_1(x) = 2x$ follows from (3.3.5) and that shows

$$\begin{aligned}\phi_{k_2-1}(c) = \Psi_1(c) - \Psi_0(c) &= \left(\frac{n_2\mu_2}{\lambda_2}\right)^{\frac{1}{2}} U_1\left(\frac{\lambda_2 + n_2\mu_2 - c}{2\sqrt{\lambda_2 n_2\mu_2}}\right) - 1\\ &= \frac{n_2\mu_2 - c}{\lambda_2}.\end{aligned}$$

Now we show (iii). It follows analogously to the proof of Proposition (3.3.2) that $\theta_l(d)$, $l \geq 0$, defined by (3.3.10) and (3.3.11) satisfies

$$(\lambda_1(1-\delta_{ik_1}) + n_1\mu_1 + d)\theta_i(d) = \lambda_1\theta_{i-1}(d) + n_1\mu_1(1-\delta_{ik_1})\theta_{i+1}(d) \tag{3.3.14}$$

for $i = n_1, \ldots, k_1$. Then, (3.3.13) follows from (3.3.11) and (3.3.14). □

3.3.1.3 Boundary condition and separation constants

By Proposition 3.3.2, β_j must by proportional to $s_j(-c, \lambda_2, \mu_2)$ for $j = 0, \ldots, n_2$ and to $\phi_j(c)$ for $j = n_2 - 1, \ldots, k_2$ so that we can choose

$$\beta_j = \begin{cases} s_j(-c, \lambda_2, \mu_2)\phi_{n_2}(c), & j = 0, \ldots, n_2. \\ s_{n_2}(-c, \lambda_2, \mu_2)\phi_j(c), & j = n_2 - 1, \ldots, k_2. \end{cases} \tag{3.3.15}$$

The separation constant c has to be chosen such that the definitions of β_{n_2-1} and β_{n_2} match, i.e.,

$$s_{n_2-1}(-c, \lambda_2, \mu_2)\phi_{n_2}(c) = s_{n_2}(-c, \lambda_2, \mu_2)\phi_{n_2-1}(c) \tag{3.3.16}$$

has to be fulfilled. The following lemma gives an equivalent formulation of this condition.

Lemma 3.3.5. *Let $s_j(\mu_2 - c, \lambda_2, \mu_2)$ for $j = n_2 - 1, n_2$ be given by* (3.2.11) *and let $\Psi_j(c)$ for $j = q_2 - 1, q_2$ be given by* (3.3.8). *Then the equation* (3.3.16) *is equivalent to*

$$c\big[s_{n_2}(\mu_2 - c, \lambda_2, \mu_2)\Psi_{q_2}(c) - s_{n_2-1}(\mu_2 - c, \lambda_2, \mu_2)\Psi_{q_2-1}(c)\big] = 0. \quad (3.3.17)$$

Proof. We write $s_n(c)$ for $s_n(c, \lambda_2, \mu_2)$ in the following. From the definition (3.3.7) of $\phi_{n_2-1}(c)$ and (3.3.9) we get

$$\begin{aligned}\phi_{n_2-1}(c) &= \Psi_{q_2+1}(c) - \Psi_{q_2}(c)\\ &= \frac{\lambda_2 + n_2\mu_2 - c}{\lambda_2}\Psi_{q_2}(c) - \frac{n_2\mu_2}{\lambda_2}\Psi_{q_2-1}(c) - \Psi_{q_2}(c)\\ &= \frac{n_2\mu_2 - c}{\lambda_2}\Psi_{q_2}(c) - \frac{n_2\mu_2}{\lambda_2}\Psi_{q_2-1}(c).\end{aligned}$$

Inserting this and $\phi_{n_2}(c) = \Psi_{q_2}(c) - \Psi_{q_2-1}(c)$ in (3.3.16) gives after simplifying

$$\begin{aligned}&\Psi_{q_2}(c)\big(\lambda_2 s_{n_2-1}(-c) - (n_2\mu_2 - c)s_{n_2}(-c)\big)\\ &\qquad = \Psi_{q_2-1}(c)\big(\lambda_2 s_{n_2-1}(-c) - n_2\mu_2 s_{n_2}(-c)\big).\end{aligned} \quad (3.3.18)$$

With (3.2.13) and (3.2.14) we get expressions for the terms in the brackets in the equation above, namely

$$\begin{aligned}\lambda_2 s_{n_2-1}(-c) - (n_2\mu_2 - c)s_{n_2}(-c) &= c s_{n_2}(\mu_2 - c) \quad \text{and}\\ \lambda_2 s_{n_2-1}(-c) - n_2\mu_2 s_{n_2}(-c) &= c s_{n_2-1}(\mu_2 - c).\end{aligned}$$

Hence, (3.3.18) is equivalent to

$$c s_{n_2}(\mu_2 - c)\Psi_{q_2}(c) = c s_{n_2-1}(\mu_2 - c)\Psi_{q_2-1}(c)$$

and therefore equivalent to (3.3.17), too. □

The function in the square brackets in equation (3.3.17) is a polynomial of degree k_2 in c with k_2 positive and distinct zeros. These zeros are given by an eigenvalue problem for a tridiagonal real symmetric matrix (see Theorem 3.3.6 in the following). In order to prove this fact, it is necessary to symmetrize the recurrence equations (3.2.6) that define $s_i(c, \mu_2, \lambda_2)$ in a suitable way. We did this in a similar way in Section 3.2.3 for the basic model without waiting rooms. The corresponding result was stated in Lemma 3.2.7. We will start by symmetrizing the equations (3.2.6) and define the corresponding tridiagonal matrix.

We suppress the indexes of λ_2, μ_2, n_2 and q_2 for the moment. Let $\lambda, \mu > 0$, $c \in \mathbb{R}$ and $n \geq 1$, $q \geq 0$ and set

$$v_i(c) = \begin{cases} u_i(\mu - c), & i = 0, \dots, n, \\ v_{n+q-1}(c) U_{n+q-j-1}\left(\dfrac{\lambda + n\mu - c}{2\sqrt{\lambda n \mu}}\right), & j = n-1, \dots, n+q, \end{cases} \tag{3.3.19}$$

where $u_i(c)$, $i \geq 0$, is given by (3.2.23). It follows immediately from the definition of $u_i(c)$, $i \geq 0$, that $v_{n+q-1}(c) \neq 0$ and furthermore $v_{n+q}(c) = 0$ follows from $U_{-1} \equiv 0$. By considering the equations for $v_n(c)$ and $v_{n-1}(c)$ and the definition of Ψ_q and Ψ_{q-1} it is seen that the existence of $v_0(c), \dots, v_{n+q}(c)$ fulfilling (3.3.19) is equivalent to

$$s_n(\mu - c, \lambda, \mu)\Psi_q(c) - s_{n-1}(\mu - c, \lambda, \mu)\Psi_{q-1}(c) = 0.$$

We get from (3.2.24) and (3.3.5) the symmetrized set of equations

$$(1-\delta_{i0}) i^{\frac{1}{2}} v_{i-1}(c) - \left(i + \frac{\lambda - c}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^{-\frac{1}{2}} v_i(c) + (i+1)^{\frac{1}{2}} v_{i+1}(c) = 0 \tag{3.3.20}$$

for $i = 0, \dots, n-1$ and

$$n^{\frac{1}{2}} v_{i-1}(c) - \left(n + \frac{\lambda - c}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^{-\frac{1}{2}} v_i(c) + n^{\frac{1}{2}} v_{i+1}(c) = 0 \tag{3.3.21}$$

for $i = n, \dots, n+q-1$. Define the matrix $R_{n,q} = R_{n,q}(\lambda, \mu) = (r_{i,j})_{i,j=1,\dots,n+q}$ for $\lambda, \mu > 0$ by $r_{i,j} = m_{i,j}$ for $i, j = 1, \dots, n$ (see (3.2.25)) and

$$\begin{aligned} r_{i,i} &= -(n-1)\left(\frac{\lambda}{\mu}\right)^{-\frac{1}{2}} && \text{for } i = n, \dots, n+q, \\ r_{i,i+1} = r_{i+1,i} &= n^{\frac{1}{2}} && \text{for } i = n, \dots, n+q-1 \text{ and} \\ r_{i,j} &= 0 && \text{for } |i-j| > 1. \end{aligned} \tag{3.3.22}$$

$R_{n,q}$ is a real symmetric tridiagonal matrix and has therefore $n+q$ real and pairwise distinct eigenvalues. We get the following result concerning the solutions of (3.3.17) by generalizing Theorem 3.2.9, where the latter follows itself from Theorem 3.3.6 with $q = 0$ and c replaced by $\mu - c$.

Theorem 3.3.6. *For $n \geq 1$, $q \geq 0$, $c \in \mathbb{R}$, $\lambda > 0$ and $\mu > 0$. Then the* *equation*

$$s_n(\mu - c, \lambda, \mu)\Psi_q(c) - s_{n-1}(\mu - c, \lambda, \mu)\Psi_{q-1}(c) = 0$$

has $n + q$ positive and distinct solutions $c_1, \ldots, c_{n+q}$ in the variable c given by

$$c_i = \lambda + \mu - (\lambda\mu)^{\frac{1}{2}}\eta_i, \quad i = 1, \ldots, n + q, \tag{3.3.23}$$

where $\eta_1, \ldots, \eta_{n+q}$ are the pairwise distinct eigenvalues of the matrix $R_{n,q}(\lambda, \mu)$.

Proof. The first part of the proof is essentially the same as the proof of Theorem 3.2.9. Let

$$p_{n,q}(c) = p_{n,q}(c, \lambda, \mu) = s_n(\mu - c, \lambda, \mu)\Psi_q(c) - s_{n-1}(\mu - c, \lambda, \mu)\Psi_{q-1}(c).$$

It remains to show that the zeros of $p_{n,q}(c, \lambda, \mu) = 0$ are positive for $q \geq 0$. We show $p_{n,q}(c, \lambda, \mu) \neq 0$ for $c \leq 0$. Let $c \leq 0$. For $q = 0$, $\Psi_{-1}(c) = 0$ and $\Psi_0(c) = 1$ yield $p_{n,0}(c) = s_n(\mu - c, \lambda, \mu)$. From (3.2.10) it follows $s_n(\mu - c, \lambda, \mu) > 0$ and $s_n(-c, \lambda, \mu) > 0$. The latter yields together with (3.2.14) that

$$\frac{s_n(\mu - c, \lambda, \mu)}{s_{n-1}(\mu - c, \lambda, \mu)} > 1 \tag{3.3.24}$$

holds. Now let $q \geq 1$. With

$$z_0 = \frac{\lambda - c + n\mu}{2\sqrt{\lambda n\mu}} + \left(\left(\frac{\lambda - c + n\mu}{2\sqrt{\lambda n\mu}}\right)^2 - 1\right)^{\frac{1}{2}} > 0, \tag{3.3.25}$$

we get from (3.3.8) and (3.3.6) that

$$\Psi_l(c) = \left(\frac{n\mu}{\lambda}\right)^{\frac{l}{2}} U_l\left(\frac{1}{2}\left(z_0 + \frac{1}{z_0}\right)\right) = \left(\frac{n\mu}{\lambda}\right)^{\frac{l}{2}} \sum_{k=1}^{l} z_0^{2k-l} > 0$$

for $l = q - 1, q$. This yields

$$\frac{\Psi_q(c)}{\Psi_{q-1}(c)} = \left(\frac{n\mu}{\lambda}\right)^{\frac{1}{2}} \frac{U_q\left(\frac{1}{2}\left(z_0 + \frac{1}{z_0}\right)\right)}{U_{q-1}\left(\frac{1}{2}\left(z_0 + \frac{1}{z_0}\right)\right)} = z_0\left(\frac{n\mu}{\lambda}\right)^{\frac{1}{2}} \frac{\sum_{k=0}^{q} z_0^{2k-q}}{\sum_{k=1}^{q} z_0^{2k-q}}$$

$$> z_0\left(\frac{n\mu}{\lambda}\right)^{\frac{1}{2}} > 1. \tag{3.3.26}$$

by (3.3.25). Now (3.3.24) and (3.3.26) give $p_{n,q}(c, \lambda, \mu) \neq 0$ for $c \leq 0$ and

$n, q \geq 1$. □

3.3.1.4 Steady-state probabilities

By Theorem 3.3.6, the separation equation (3.3.17) has $k_2 + 1$ distinct solutions $c_0 = 0, c_1, \ldots, c_{k_2}$, where $c_1, \ldots, c_{k_2}$ are positive. We can represent the stationary probabilities $p_{i,j}$ for $i = 0, \ldots, n_1$ and $j = 0, \ldots, k_2$ as linear combinations of all feasible solutions of the separation approach $p_{i,j} = \alpha_i \beta_j$ where α_i is given by $s_i(c_m, \lambda_1, \mu_1)$ for $m = 0, \ldots, k_2$ and β_j is given by (3.3.15). We get

$$p_{i,j} = \begin{cases} \displaystyle\sum_{m=0}^{k_2} b'_m s_i(c_m, \lambda_1, \mu_1) s_j(-c_m, \lambda_2, \mu_2) \phi_{n_2}(c_m), & j = 0, \ldots, n_2. \\ \displaystyle\sum_{m=0}^{k_2} b'_m s_i(c_m, \lambda_1, \mu_1) s_{n_2}(-c_m, \lambda_2, \mu_2) \phi_j(c_m), & j = n_2, \ldots, k_2. \end{cases} \tag{3.3.27}$$

The constants b'_m, $m = 0, \ldots, k_2$, have to be chosen such that the boundary conditions in (3.3.1) corresponding to $i = n_1$, $j = 0, \ldots, k_2$ and the normalization condition are satisfied.

For $i = n_1, \ldots, k_1$ and $j = n_2, \ldots, k_2$, we examined the approach $p_{i,j} = \gamma_i \delta_j$ and showed in Corollary 3.3.3 that γ_i must be proportional to $\theta_i(d)$ defined by (3.3.10) for $i = n_1, \ldots, k_1$ and that δ_j must be proportional to $\phi_j(d)$ defined by (3.3.7) for $j = n_2, \ldots, k_2$. Furthermore, $p_{i,j} = \gamma_i \delta_j$ has to be chosen such that it matches (3.3.27) on the boundary $i = n_1$, $j = n_2, \ldots, k_2$. For this reason we must ensure $\theta_{n_1}(d) = 0$. The equation $\theta_{n_1}(d) = 0$ on its part has q_1 negative and distinct zeros as the following theorem shows. Once again, we drop the indexes of λ_1, μ_1, n_1 and q_1 for the moment.

Theorem 3.3.7. *For $n \geq 1$, $q \geq 0$, $d \in \mathbb{R}$, $\lambda > 0$ and $\mu > 0$ let $\theta_n(d, \lambda, \mu)$ be defined by* (3.3.29)*. Then the equation*

$$\theta_n(d, \lambda, \mu) = 0$$

has q negative and distinct solutions in the variable d given by the eigenvalues of the matrix $T_{n,q}(\lambda, \mu) = (t_{i,j})_{i,j=1,\ldots,q}$, where

$$\begin{aligned} t_{i,i} &= -\big((1 - \delta_{i1})\lambda + n\mu\big) && \textit{for } i = 1, \ldots, q, \\ t_{i,i+1} = r_{i+1,i} &= \sqrt{\lambda n \mu} && \textit{for } i = 1, \ldots, q-1 \textit{ and} \\ t_{i,j} &= 0 && \textit{for } |i - j| > 1. \end{aligned} \tag{3.3.28}$$

Proof. For $\lambda, \mu > 0$ and $n, q \geq 1$ recall the definitions (3.3.10) and (3.3.11):

$$\theta_i(d) = \theta_i(d, \lambda, \mu) = \Omega_{n+q-i}(d) - \Omega_{n+q-i-1}(d), \quad i = 0, \ldots, n+q, \tag{3.3.29}$$

and

$$\Omega_l(d) = \Omega_l(d, \lambda, \mu) = \left(\frac{n\mu}{\lambda}\right)^{\frac{l}{2}} U\left(\frac{\lambda + n\mu + d}{2\sqrt{\lambda n\mu}}\right) \tag{3.3.30}$$

for $l = -1, \ldots, q$. Define

$$w_i(d) = w_i(d, \lambda, \mu) = U_i\left(\frac{\lambda + n\mu + d}{2\sqrt{\lambda n\mu}}\right) - \left(\frac{\lambda}{n\mu}\right)^{\frac{1}{2}} U_{i-1}\left(\frac{\lambda + n\mu + d}{2\sqrt{\lambda n\mu}}\right) \tag{3.3.31}$$

for $i \geq 0$. Obviously, $\theta_n(d) = 0$ is equivalent to $w_q(d) = 0$ by (3.3.29). Additionally, (3.3.5) yields

$$(1 - \delta_{i0})\sqrt{\lambda n\mu}\, w_{i-1}(d) - \big((1 - \delta_{i0})\lambda + n\mu + d\big) w_i(d) + \sqrt{\lambda n\mu}\, w_{i+1}(d)$$

for $i \geq 0$. $T_{n,q} = T_{n,q}(\lambda, \mu)$ is defined such that the solutions of $\theta_n(d) = 0$ are exactly the eigenvalues of $T_{n,q}$. $T_{n,q}$ itself is a real symmetric tridiagonal matrix of rank q and has therefore q real and pairwise distinct eigenvalues by Theorem 3.2.6. In remains to show that the eigenvalues of $T_{n,q}$ are negative. Using the equivalence of $\theta_n(d) = 0$ and $w_n(d) = 0$, (3.3.31) and (3.3.26) it is seen analogously to the proof of the positivity argument in Theorem 3.3.6 that $\theta_n(d) \neq 0$ for $d \geq 0$. □

We denote the solutions of $\theta_{n_1}(d) = 0$ by $d_1, \ldots, d_{q_1}$. By Theorem 3.3.7 these solutions are negative and distinct. The roots of (3.3.17) are non-negative and distinct so that $c_0, \ldots, c_{k_2}, d_1, \ldots, d_{q_1}$ are distinct. Thus, the probabilities $p_{i,j}$ for $i = n_1, \ldots, k_1$ and $j = n_2, \ldots, k_2$ are given by

$$p_{i,j} = \sum_{m=0}^{k_2} b_m s_{n_1}(c_m, \lambda_1, \mu_1) s_{n_2}(-c_m, \lambda_2, \mu_2) \theta_i(c_m) \phi_j(c_m) + \sum_{l=1}^{q_1} e_l \theta_i(d_l) \phi_j(d_l) \tag{3.3.32}$$

for arbitrary $q_1 \geq 0$. In order to match (3.3.27) and (3.3.32) it is convenient to choose $b'_m = b_m \theta_{n_1}(c_m)$, $m = 0, \ldots, k_2$, so that (3.3.27) can be restated as

$$p_{i,j} = \begin{cases} \displaystyle\sum_{m=0}^{k_2} b_m s_i(c_m, \lambda_1, \mu_1) s_j(-c_m, \lambda_2, \mu_2) \theta_{n_1}(c_m) \phi_{n_2}(c_m), \ j = 0, \dots, n_2, \\ \displaystyle\sum_{m=0}^{k_2} b_m s_i(c_m, \lambda_1, \mu_1) s_{n_2}(-c_m, \lambda_2, \mu_2) \theta_{n_1}(c_m) \phi_j(c_m), \ j = n_2, \dots, k_2, \end{cases} \tag{3.3.33}$$

for $i = 0, \dots, n_1$.

The constants $b_0, \dots, b_{k_2}$ and $e_1, \dots, e_{q_1}$ in (3.3.32) and (3.3.33) have to be chosen such that the boundary conditions in (3.2.1) for 1) $i = n_1$ and $j = 0, \dots, k_2$ and 2) $i = n_1 + 1, \dots, k_1$ and $j = n_2$ and the normalization condition are satisfied. The boundary conditions are

$$\begin{aligned} (p\lambda_1 + \lambda_2 + n_1\mu_1 + j\mu_2)p_{n_1,j} &= \lambda_1 p_{n_1-1,j} \\ &\quad + (1 - \delta_{j0})(p\lambda_1 + \lambda_2)p_{n_1,j-1} \\ &\quad + (j+1)\mu_2 p_{n_1,j+1}, \\ &\qquad i = n_1, j = 0, \dots, n_2 - 1, \quad (3.3.34) \\ (\lambda_1 + \lambda_2 + n_1\mu_1 + n_2\mu_2)p_{n_1,n_2} &= \lambda_1 p_{n_1-1,n_2} + (p\lambda_1 + \lambda_2)p_{n_1,n_2-1} \\ &\quad + n_2\mu_2 p_{n_1,n_2+1} \\ &\quad + (n_1\mu_1 + n_2\mu_2)p_{n_1+1,n_2}, \\ &\qquad i = n_1, j = n_2, \quad (3.3.35) \\ (\lambda_1 + \lambda_2(1 - \delta_{jk_2}) + n_1\mu_1 + n_2\mu_2)p_{n_1,j} &= \lambda_1 p_{n_1-1,j} \\ &\quad + (1 - \delta_{jk_2})n_2\mu_2 p_{n_1,j+1} \\ &\quad + \lambda_2 p_{n_1,j-1} + n_1\mu_1 p_{n_1+1,j}, \\ &\qquad i = n_1, j = n_2 + 1, \dots, k_2, \\ &\qquad (3.3.36) \\ (\lambda_1(1 - \delta_{ik_1}) + \lambda_2 + n_1\mu_1 + n_2\mu_2)p_{i,n_2} &= \lambda_1 p_{i-1,n_2} + n_2\mu_2 p_{i,n_2+1} \\ &\quad + (1 - \delta_{ik_1})(n_1\mu_1 + n_2\mu_2)p_{i+1,n_2}, \\ &\qquad i = n_1 + 1, \dots, k_1, j = n_2. \\ &\qquad (3.3.37) \end{aligned}$$

The equations (3.3.34) are identical to (3.2.32) so that (3.3.33), (3.3.34) and

(3.2.34) yield

$$\sum_{m=0}^{k_2} b_m\big(c_m s_{n_1}(c_m+\mu_1)s_j(-c_m) \\ + p\lambda_1 s_{n_1}(c_m)s_j(-c_m-\mu_2)\big)\theta_{n_1}(c_m)\phi_{n_2}(c_m) = 0 \quad (3.3.38)$$

for $i = n_1$ and $j = 0, \dots, n_2 - 1$. By substituting (3.3.32) and (3.3.33) into (3.3.36) and (3.3.37) and using (3.3.13) and the recurrence relations (3.2.13) and (3.2.14) we get

$$\sum_{m=0}^{k_2} b_m c_m s_{n_2}(-c_m)\big(s_{n_1}(c_m+\mu_1)\theta_{n_1}(c_m) + s_{n_1}(c_m)\Omega_{q_1-1}(c_m)\big)\phi_j(c_m) \\ = n_1 \sum_{l=1}^{q_1} e_l\theta_{n_1+1}(d_l)\phi_j(d_l) \quad (3.3.39)$$

for $i = n_1$, $j = n_2 + 1, \dots, k_2$ and

$$\sum_{m=0}^{k_2} b_m s_{n_1}(c_m)s_{n_2}(-c_m)\big(\lambda_2\theta_i(c_m)\phi_{n_2-1}(c_m) - n_2(1-\delta_{ik_1})\theta_{i+1}(c_m)\phi_{n_2}(c_m)\big) \\ = \sum_{l=1}^{q_1} e_l\big(n_2(1-\delta_{ik_1})\theta_{i+1}(d_l)\phi_{n_2}(d_l) - \lambda_2\theta_i(d_l)\phi_{n_2-1}(d_l)\big) \quad (3.3.40)$$

for $i = n_1+1, \dots, k_1$ and $j = n_2$. By summing (3.3.38), (3.3.39) and (3.3.40) and once again using the recurrence relations (3.2.13) and (3.2.14) it is seen that equation (3.3.35) is redundant.

Together with the normalization condition, we have $k_2 + q_1 + 1$ linear equations for the unknowns $b_0, \dots, b_{k_2}$ and $e_1, \dots, e_{q_1}$. Inserting (3.3.32) and (3.3.33) into the normalization condition $\sum_{i=0}^{k_1}\sum_{j=0}^{k_2} p_{i,j} = 1$ and using (3.2.16) results in an equation for b_0:

$$b_0 s_{n_1}(\mu)\theta_{n_1}(0)\big(s_{n_2}(\mu_2)\Psi_{q_2}(0) - s_{n_2-1}(\mu_2)\Psi_{q_2-1}(0)\big) \\ + \sum_{m=0}^{k_2} b_m s_{n_1}(c_m)s_{n_2}(-c_m)\Omega_{q_1-1}(c_m)\Psi_{q_2}(c_m) \quad (3.3.41) \\ + \sum_{l=1}^{q_1} e_l\Omega_{q_1-1}(d_l)\Psi_{q_2}(d_l) = 1.$$

Remark 3.3.8. It is seen that the probabilities $p_{i,j}$ depend on the parameter $p \in [0,1]$ only through the coefficients $b_0, \dots, b_{k_2}$ and $e_1, \dots, e_{q_1}$. This is due to the fact that the stationary probabilities have been calculated from the non-overflow balance equations, i.e., (3.3.1) for $i = n_1$, $j = 0, \dots, k_2$ and $i = n_1+1, \dots, k_1$, $j = n_2$, which are independent of p, whereas the equations for the coefficients arise from the p-dependent balance equations in (3.3.1).

We summarize the result in the following theorem.

Theorem 3.3.9. *The unique nonnegative and normalized solution of the steady-state equations* (3.3.1) *is given by*

$$p_{i,j} = \begin{cases} \displaystyle\sum_{m=0}^{k_2} b_m s_i(c_m, \lambda_1, \mu_1) s_j(-c_m, \lambda_2, \mu_2) \theta_{n_1}(c_m) \phi_{n_2}(c_m), & j = 0, \dots, n_2, \\ \displaystyle\sum_{m=0}^{k_2} b_m s_i(c_m, \lambda_1, \mu_1) s_{n_2}(-c_m, \lambda_2, \mu_2) \theta_{n_1}(c_m) \phi_j(c_m), & j = n_2, \dots, k_2, \end{cases}$$

for $i = 0, \dots, n_1$ *and*

$$p_{i,j} = \sum_{m=0}^{k_2} b_m s_{n_1}(c_m, \lambda_1, \mu_1) s_{n_2}(-c_m, \lambda_2, \mu_2) \theta_i(c_m) \phi_j(c_m) + \sum_{l=1}^{q_1} e_l \theta_i(d_l) \phi_j(d_l)$$

for $i = n_1, \dots, k_1$, $j = n_2, \dots, k_2$ *and every* $q_1 \geq 0$ *if the coefficients* $b_0, \dots, b_{k_2}$ *and* $e_1, \dots, e_{q_1}$ *are determined by* (3.3.38)-(3.3.40) *and* (3.3.41). $c_1, \dots, c_{k_2}$ *are the by Theorem 3.3.6 positive and pairwise distinct solutions of* (3.3.17), $c_0 = 0$ *and* $d_1, \dots, d_{q_1}$ *are the by Theorem 3.3.7 negative and pairwise distinct solutions of* $\theta_{n_1}(d_l) = 0$, $l = 1, \dots, q_1$.

The problem of determining the $(k_1 + 1)(k_2 + 1)$ unknowns $p_{i,j}$, $i = 0, \dots, k_1$, $j = 0, \dots, k_2$, has now been reduced to the problem of determining $k_2 + q_1$ eigenvalues from (3.3.17) and the equation $\theta_{n_1}(d) = 0$ and $k_2 + q_1 + 1$ unknowns from the linear equations (3.3.38)-(3.3.40) and (3.3.41). Alternative separation approaches for the case $\mu_1 = \mu_2$ and $p = 1$ can be found in Morrison [47] and Morrison and Wright [49]. The approach in [47] is preferable, when q_2 is large compared to q_1, n_1 and n_2 or infinite. The approach in [49] is preferable, when q_1 is large compared to q_2 and n_2 or infinite. Both approaches can be generalized to the case of arbitrary service rates and weighted overflow traffic, too.

Remark 3.3.10. In the case $q_2 = 0$, (3.3.17) reduces to (3.2.18) because of $\Psi_{-1} \equiv 0$ and $\Psi_0 \equiv 1$. If in addition $q_1 = 0$, then (3.3.27) and (3.3.32) reduce to (3.2.31). Moreover, equation (3.3.38) reduces to (3.2.34) because of $\theta_{n_1} \equiv \phi_{n_2} \equiv 1$, while equations (3.3.39) and (3.3.40) vanish. Finally, (3.3.41) reduces to (3.2.36) because of $\Psi_0 \equiv \Omega_0 \equiv 1$ and $\Psi_{-1} \equiv \Omega_{-1} \equiv 0$.

3.3.1.5 Stationary quantities

The results of the separation approach can be used to derive steady-state quantities of interest. The characteristics of this queueing network can be calculated in a similar manner as for the model without waiting rooms. Most of these characteristics can be derived conformably to the results in Morrison [46]; therefore, we omit the details. The average blocking probabilities B_1 and B_2 for arriving customers, the loss probabilities $P_{\text{Loss},1}$ and $P_{\text{Loss},2}$ for arriving customers and the overflow probability O_{12} are given by

$$B_1 = \sum_{j=0}^{n_2} p_{n_1,j}, \qquad B_2 = \sum_{i=0}^{k_1} p_{i,k_2},$$

$$P_{\text{loss},1} = \sum_{j=n_2}^{k_2} p_{k_1,j} + (1-p)O_{12}, \qquad P_{\text{loss},2} = B_2,$$

$$O_{12} = \sum_{j=0}^{n_2-1} p_{n_1,j}.$$

Additionally, let $P_{\text{queue},1}$ be the probability that an arriving Q_i-customer is queued in the waiting room in Q_i, $i = 1, 2$, then

$$P_{\text{queue},1} = \sum_{i=n_1}^{k_1-1} \sum_{j=0}^{k_2} p_{i,j} \quad \text{and} \quad P_{\text{queue},2} = \sum_{i=0}^{k_1-1} \sum_{j=n_2}^{k_2-1} p_{i,j}.$$

The mean departure rates R_{11}, R_{12} and R_{22} from the waiting room in Q_1 to the servers in Q_1 and Q_2, respectively, and from the servers in Q_2 to Q_2 are given by

$$R_{11} = n_1\mu_1 \sum_{i=n_1+1}^{k_1} \sum_{j=n_2}^{k_2} p_{i,j}, \qquad R_{12} = n_2\mu_2 \sum_{i=n_1+1}^{k_1} p_{i,n_2},$$

$$R_{22} = n_2\mu_2 \sum_{i=0}^{k_1} \sum_{j=n_2+1}^{k_2} p_{n_1,j}.$$

The mean number EL_i of customers in Q_i, $i = 1, 2$, is given by

$$EL_1 = \sum_{i=0}^{n_1} \sum_{j=0}^{k_2} i p_{i,j} + n_1 \sum_{i=n_1+1}^{k_1} \sum_{j=n_2}^{k_2} p_{i,j} + \sum_{i=n_1+1}^{k_1} \sum_{j=n_2}^{k_2} (i-n_1) p_{i,j},$$

$$EL_2 = \sum_{i=0}^{n_1} \sum_{j=0}^{n_2-1} j p_{i,j} + n_2 \sum_{i=0}^{k_1} \sum_{j=n_2}^{k_2} p_{i,j} + \sum_{i=0}^{k_1} \sum_{j=n_2+1}^{k_2} (j-n_2) p_{i,j}.$$

By (3.2.15), (3.3.7) and (3.3.10) we get

$$\begin{aligned}
B_1 &= \sum_{m=0}^{k_2} b_m s_{n_1}(c_m, \lambda_1, \mu_1) s_{n_2}(\mu_2 - c_m, \lambda_2, \mu_2)\theta_{n_1}(c_m)\phi_{n_2}(c_m),\\
B_2 &= \sum_{m=0}^{k_2} b_m s_{n_2}(-c_m, \lambda_2, \mu_2)\big(s_{n_1}(c_m + \mu_1, \lambda_1, \mu_1)\theta_{n_1}(c_m)\\
&\qquad\qquad + s_{n_1}(c_m, \lambda_1, \mu_1)\Omega_{q_1-1}(c_m)\big) + \sum_{l=1}^{q_1} e_l \Omega_{q_1-1}(d_l),\\
P_{\text{loss},1} &= \sum_{m=0}^{k_2} b_m s_{n_1}(c_m, \lambda_1, \mu_1) s_{n_2}(c_m, \lambda_2, \mu_2)\Psi_{q_2}(c_m)\\
&\qquad\qquad + \sum_{l=1}^{q_1} e_l \Psi_{q_2}(d_l) + (1-p)O_{12},\\
O_{12} &= \sum_{m=0}^{k_2} b_m s_{n_1}(c_m, \lambda_1, \mu_1) s_{n_2-1}(\mu_2 - c_m, \lambda_2, \mu_2)\theta_{n_1}(c_m)\phi_{n_2}(c_m),\\
R_{11} &= n_1\mu_1\Big(\sum_{m=0}^{k_2} b_m s_{n_1}(c_m, \lambda_1, \mu_1) s_{n_2}(-c_m, \lambda_2, \mu_2)\Omega_{q_1-1}(c_m)\Psi_{q_2}(c_m)\\
&\qquad\qquad + \sum_{l=1}^{q_1} e_l \Omega_{q_1-1}(d_l)\Psi_{q_2}(d_l)\Big),\\
R_{12} &= n_2\mu_2\Big(\sum_{m=0}^{k_2} b_m s_{n_1}(c_m, \lambda_1, \mu_1) s_{n_2}(-c_m, \lambda_2, \mu_2)\Omega_{q_1-1}(c_m)\phi_{n_2}(c_m)\\
&\qquad\qquad + \sum_{l=1}^{q_1} e_l \Omega_{q_1-1}(d_l)\phi_{n_2}(d_l)\Big) \quad \text{and}\\
R_{22} &= n_2\mu_2 \sum_{j=n_{2+1}}^{k_2} \left(\frac{n_2\mu_2}{\lambda_2}\right)^{k_2-j} P_{\text{Loss},2}.
\end{aligned}$$

where we have used $\phi_{k_2} \equiv \theta_{k_1} \equiv 1$. For the last equality, we exploited the sum of (3.3.1) over $i = 0, \ldots, n_1$. It is also possible to give an expression for EL_1 and EL_2. We omit the details and refer to [46].

3.3.1.6 Alternative approach without waiting room in second queue

In the special case that the second queue has no waiting room, i.e., in the case $q_2 = 0$, another approach is feasible to derive the stationary probabilities as functions of only $2n_2 + 1$ unknowns (see [46] for the case $p = 1$ and $\mu_1 = \mu_2$). The approach from the previous section gives the probabilities as functions of $2(n_2 + q_1) + 1$ unknowns in this case. The approach is sketched in Figure 3.6. Compared with Figure 3.5, it is seen that the balance equations on the boundary line at $i = n_1 + 1, \dots, k_1$ and $j = n_2$, are no longer required to determine the separation constants in the region above the boundary line. Actually, these equations lead to a recursive formula for the stationary probabilities on the boundary line. The redundant boundary condition corresponds to the state (n_1, n_2), which is represented by an unfilled circle in Figure 3.6.

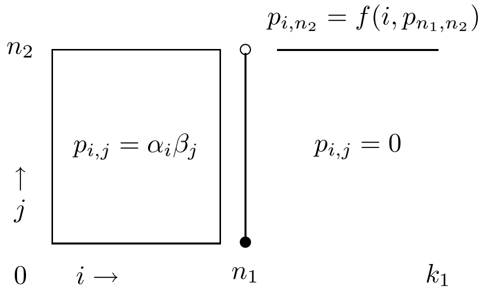


Figure 3.6: Separation scheme: Model S/S/S with $q_2 = 0$

Let $q_2 = 0$ in the following. The balance equations are given by (3.3.1) with k_2 replaced by n_2, i.e.,

$$\begin{aligned}
&(\lambda_1(1-\delta_{ik_1})(1-\chi_{i-n_1}(1-\delta_{jn_2})) + p\lambda_1\chi_{i-n_1}(1-\delta_{jn_2})) \\
&\quad + \lambda_2(1-\delta_{jn_2}) + (i \wedge n_1)\mu_1 + (j \wedge n_2)\mu_2)p_{i,j} \\
&\quad = (1-\chi_{i-n_1-1}(1-\delta_{jn_2})) \qquad\qquad\qquad\qquad\qquad\qquad (3.3.42) \\
&\qquad \times \big(\lambda_1(1-\delta_{i0})p_{i-1,j} + (1-\delta_{jn_2})\left((j+1) \wedge n_2\right)\mu_2 p_{i,j+1}\big) \\
&\quad + (1-\delta_{j0})\big(p\lambda_1\delta_{in_1}\chi_{n_2-j} + \lambda_2(1-\chi_{i-n_1-1}\chi_{n_2-j})\big)p_{i,j-1} \\
&\quad + (1-\delta_{ik_1})\big((1-\chi_{i-n_1}\chi_{n_2-1-j})\left((i+1) \wedge n_1\right)\mu_1 + n_2\mu_2\delta_{jn_2}\chi_{i-n_1}\big)p_{i+1,j}
\end{aligned}$$

for $i = 0, \dots, k_1$ and $j = 0, \dots, n_2$. Observe that $\delta_{jn_2} = \chi_{j-n_2}$ and $1-\delta_{jn_2} = \chi_{n_2-j-1}$ and note that $p_{i,j} = 0$ for $i > n_1$ and $j < n_2$ because it is impossible for customers to wait at Q_1 while there is at least one server available at Q_2.

For $i = 0, \dots, n_1-1$ and $j = 0, \dots, n_2$ the balance equations are identical to (3.3.2) with k_2 replaced by n_2 so that we establish from (3.2.31) that

$$p_{i,j} = \sum_{m=0}^{n_2} a_m s_i(c_m, \lambda_1, \mu_1) s_j(-c_m, \lambda_2, \mu_2) \tag{3.3.43}$$

for $i = 0, \ldots, n_1$ and $j = 0, \ldots, n_2$, where $c_0, \ldots, c_{n_2}$ are given by (3.2.29). It remains to determine p_{i,n_2} for $i = n_1, \ldots, k_1$. Let $q_1 \geq 1$ in the following. The balance equations

$$\begin{aligned}(\lambda_1(1-\delta_{ik_1}) + n_1\mu_1 + n_2\mu_2)p_{i,n_2} &= \lambda_1 p_{i-1,n_2} \\ &\quad + (1-\delta_{ik_1})(n_1\mu_1 + n_2\mu_2)p_{i+1,n_2}\end{aligned} \tag{3.3.44}$$

for $i = n_1+1, \ldots, k_1$ and $j = n_2$ yield

$$\lambda_1 p_{i-1,n_2} = (n_1\mu_1 + n_2\mu_2)p_{i,n_2}$$

for $i = n_1+1, \ldots, k_1$. It follows

$$p_{i,n_2} = \left(\frac{\lambda_1}{n_1\mu_1 + n_2\mu_2}\right)^{i-n_1} p_{n_1,n_2} \tag{3.3.45}$$

for $i = n_1, \ldots, k_1$, where p_{n_1,n_2} is given by (3.3.43).

It remains to satisfy the boundary conditions at $i = n_1$ and $j = 0, \ldots, n_2$ and the normalization condition. For $i = n_1$ and $j = 0, \ldots, n_2 - 1$ these boundary conditions are given by (3.2.32). We deduce from (3.3.43), (3.2.31) and (3.2.34) that

$$\sum_{m=0}^{n_2} a_m \left(c_m s_{n_1}(c_m + \mu_1)s_j(-c_m) + p\lambda_1 s_{n_1}(c_m)s_j(-c_m - \mu_2)\right) = 0 \tag{3.3.46}$$

must hold for $i = n_1$ and $j = 0, \ldots, n_2 - 1$. The boundary condition

$$\begin{aligned}(\lambda_1 + n_1\mu_1 + n_2\mu_2)p_{n_1,n_2} &= \lambda_1 p_{n_1-1,n_2} \\ &\quad + (n_1\mu_1 + n_2\mu_2)p_{n_1+1,n_2} + (p\lambda_1 + \lambda_2)p_{n_1,n_2-1}\end{aligned} \tag{3.3.47}$$

for $i = n_1$ and $j = n_2$ reduces with the help of (3.3.45) to

$$(n_1\mu_1 + n_2\mu_2)p_{n_1,n_2} = \lambda_1 p_{n_1-1,n_2} + (p\lambda_1 + \lambda_2)p_{n_1,n_2-1},$$

which was shown to be equivalent to (3.2.35). The latter equation was seen to be redundant. Substituting (3.3.43) and (3.3.45) into the normalization condition yields after simplification

$$\begin{aligned} &a_0 s_{n_1}(\mu_1, \lambda_1, \mu_1) s_{n_2}(\mu_2, \lambda_2, \mu_2) \\ &\quad + \left(\sum_{m=0}^{n_2} a_m s_{n_1}(c_m, \lambda_1, \mu_1) s_{n_2}(-c_m, \lambda_2, \mu_2) \right) \sum_{l=1}^{q_1} \left(\frac{\lambda_1}{n_1\mu_1 + n_2\mu_2} \right)^l = 1. \end{aligned} \tag{3.3.48}$$

In summary, we can record the following result.

Theorem 3.3.11. *The unique nonnegative and normalized solution of the steady-state equations* (3.3.42) *is given by*

$$p_{i,j} = \sum_{m=0}^{n_2} a_m s_i(c_m, \lambda_1, \mu_1) s_j(-c_m, \lambda_2, \mu_2)$$

for $i = 0, \ldots, n_1$ *and* $j = 0, \ldots, n_2$ *and*

$$p_{i,n_2} = \left(\frac{\lambda_1}{n_1\mu_1 + n_2\mu_2} \right)^{i-n_1} p_{n_1,n_2}$$

for $i = n_1, \ldots, k_1$ *if the coefficients* $a_0, \ldots, a_{n_2}$ *are determined by* (3.3.46) *and* (3.3.48)*, where* $c_0 = 0$ *and* $c_1, \ldots, c_{n_2}$ *are the by Corollary 3.2.10 positive and pairwise distinct solutions of the equation* $s_{n_2}(\mu_2 - c_m, \lambda_2, \mu_2) = 0$*,* $m = 1, \ldots, n_2$*.*

The number of unknowns has now been reduced from $(n_1+1)(n_2+1)+q_1$ to $2n_2 + 1$.

Remark 3.3.12. For $q_1 = 0$ the results reduce to those from Section 3.2.

Remark 3.3.13. By (3.3.45) and (3.3.48) the results even hold in the case $q_1 = \infty$ as long as the stability condition

$$\lambda_1 < n_1\mu_1 + n_2\mu_2 \tag{3.3.49}$$

is fulfilled.

The stability condition (3.3.49) states that the servers of Q_1 and Q_2 must together be able to handle the arrival stream of Q_1. Moreover, the stability condition is independent of p. This can be explained analytically by the fact that the stationary probabilities have been calculated from the p-independent balance equations, which led to the stationary probabilities (3.3.43) and (3.3.45). These probabilities and therefore equation (3.3.48) are independent of p. Heuristically, this is due to the fact that on the one hand in the case $q_1 = \infty$, the capacity of Q_2 is still limited by n_2 so that in the long-run average, the queue size of Q_2 is negligible and the arrival stream of Q_2 has no influence on the stability of the system. On the other hand, queued Q_1-customers overflow to the servers in Q_2 as soon as one is available. But once again in the long-run average, the queue size of Q_2 is negligible so that queued Q_1-customers must be handled by the servers of both Q_1 and Q_2 to guarantee stability. Furthermore, near saturation, there are always customers waiting in Q_1, who move to Q_2 if a server becomes available so that near saturation, the overflow stream is negligible, too.

3.3.2 No jockeying

We investigate the structure of the steady-state equations for the model S/S/N in this section. This model differs from the model S/S/S of Section 3.3.1 with jockeying only in the fact that waiting Q_1-customers are served exclusively by the Q_1-servers, i.e., no jockeying is allowed (see Figure 3.1). Recall that an arriving Q_1-customer is blocked and directed to Q_2 if all n_1 servers are busy in Q_1. Blocked customers are served by one of the servers in Q_2 if at least one is available, are queued in Q_1 if all servers in Q_2 are busy and a waiting position is available in Q_1 and are lost otherwise. We suppose that the number of waiting positions in Q_1 and Q_2 is positive, i.e. $q_1 \geq 1$ and $q_2 \geq 1$. This model is also treated in Morrison [46] for the case $p = 1$ and $\mu_1 = \mu_2$ and numerically in Kaufman et al. [34] for arbitrary service rates and $p = 1$. In our case of arbitrary $\mu_1, \mu_2 > 0$ and p-weighted overflow traffic the balance equations are given by

$$\begin{aligned}
&(\lambda_1(1-\delta_{ik_1}\chi_{j-n_2})(1-\chi_{i-n_1}(1-\chi_{j-n_2})) + p\lambda_1\chi_{i-n_1}(1-\chi_{j-n_2})) \\
&\quad + \lambda_2(1-\delta_{jk_2}) + (i \wedge n_1)\mu_1 + (j \wedge n_2)\mu_2)p_{i,j} \\
&= (1-\chi_{i-n_1-1}\chi_{n_2-j-1})(\lambda_1(1-\delta_{i0})p_{i-1,j}) \qquad (3.3.50)\\
&\quad + (1-\delta_{ik_1})((i+1) \wedge n_1)\mu_1 p_{i+1,j} \\
&\quad + (1-\delta_{j0})\big(p\lambda_1\chi_{i-n_1}\chi_{n_2-j} + \lambda_2\big)p_{i,j-1} + (1-\delta_{jk_2})((j+1) \wedge n_2)\mu_2 p_{i,j+1}
\end{aligned}$$

for $i = 0, \ldots, k_1$ and $j = 0, \ldots, k_2$. It is seen that the term $p\lambda_1\chi_{i-n_1}(1 - \chi_{j-n_2})$ on the left side of (3.3.50) and $p\lambda_1\chi_{i-n_1}\chi_{n_2-j}$ on the right side of (3.3.50) do not vanish in the regions $i = n_1, \ldots, k_1$, $j = 0, \ldots, n_2 - 1$ and $i = n_1, \ldots, k_1$, $j = 0, \ldots, n_2$, respectively. Consequently, in contrast to the previous models, the probabilities in these regions depend on the parameter p not only through the coefficients determined by the boundary conditions (see also (3.3.53) and (3.3.54)). The separation approach for model S/S/S (see Figure 3.5) can be extended for this model to the region $i = n_1 + 1, \ldots, k_1$, $j = 0, \ldots, n_2 - 1$. The extended approach is illustrated in Figure 3.7 and leads to a reduction of the number of probabilities to be calculated in this region.

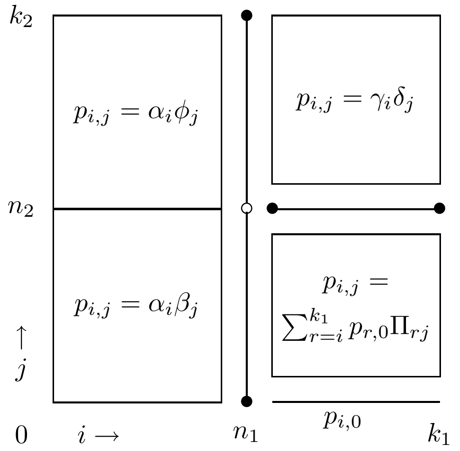

k_2
$p_{i,j} = \alpha_i \phi_j$
$p_{i,j} = \gamma_i \delta_j$
n_2
$p_{i,j} =$
$\sum_{r=i}^{k_1} p_{r,0} \Pi_{rj}$
$p_{i,j} = \alpha_i \beta_j$
↑
j
0
$i \rightarrow$
n_1
$p_{i,0}$
k_1

Figure 3.7: Separation scheme: Model S/S/N

Analogously to the model with jockeying of Section 3.3.1 the balance equations lead to

$$p_{i,j} = \begin{cases} \displaystyle\sum_{m=0}^{k_2} b_m s_i(c_m, \lambda_1, \mu_1) s_j(-c_m, \lambda_2, \mu_2) \theta_{n_1}(c_m) \phi_{n_2}(c_m), \ j = 0, \ldots, n_2, \\ \displaystyle\sum_{m=0}^{k_2} b_m s_i(c_m, \lambda_1, \mu_1) s_{n_2}(-c_m, \lambda_2, \mu_2) \theta_{n_1}(c_m) \phi_j(c_m), \ j = n_2, \ldots, k_2, \end{cases} \tag{3.3.51}$$

for $i = 0, \ldots, n_1$ and

$$p_{i,j} = \sum_{m=0}^{k_2} b_m s_{n_1}(c_m, \lambda_1, \mu_1) s_{n_2}(-c_m, \lambda_2, \mu_2) \theta_i(c_m) \phi_j(c_m) + \sum_{l=1}^{q_1} e_l \theta_i(d_l) \phi_j(d_l) \tag{3.3.52}$$

for $i = n_1, \ldots, k_1$ and $j = n_2, \ldots, k_2$. Again, $c_0, \ldots, c_{k_2}$ and $d_1, \ldots, d_{q_1}$ are the roots of (3.3.17) and $\theta_{n_1}(d_l) = 0$, $l = 1, \ldots, q_1$, respectively.

For $i = n_1 + 1, \dots, k_1$ and $j = 0, \dots, n_2$ (3.3.50) yields

$$p_{i,j} = \sum_{r=i}^{k_1} p_{r,0} \Pi_{r-i,j}, \tag{3.3.53}$$

where $\Pi_{r,j}$, $r = 0, \dots, k_1$, $j = 0, \dots, n_2$, are the solutions of the p-dependent equations

$$\begin{aligned}(p\lambda_1 + \lambda_2 + n_1\mu_1 + j\mu_2)\Pi_{r,j} \\ = (p\lambda_1 + \lambda_2)(1 - \delta_{j0})\Pi_{r,j-1} + n_1\mu_1(1 - \delta_{r0})\Pi_{r-1,j} + (j + 1)\Pi_{r,j+1}\end{aligned} \tag{3.3.54}$$

with boundary conditions $\Pi_{r0} = \delta_{r0}$ for $r = 0, \dots, k_1$, where δ_{r0} is the Kronecker function. It is shown in Morrison [46] that further investigations of these quantities lead successively to formulas for $p_{n_1+1,0}, \dots, p_{k_1,0}$ depending on $b_0, \dots, b_{k_2}$ and $e_1, \dots, e_{q_1}$, while these coefficients can be calculated from the boundary conditions in (3.3.50). In particular $b_0, \dots, b_{k_2}$, $e_1, \dots, e_{q_1}$ and the probabilities $p_{n_1+1,0}, \dots, p_{k_1,0}$ are uniquely determined up to a multiplicative constant by the equations

$$\sum_{m=0}^{k_2} b_m s_{n_1}(c_m) s_{n_2}(-c_m) \theta_i(c_m) \phi_{n_2}(c_m) + \sum_{l=1}^{q_1} e_l \theta_i(d_l) \phi_{n_2}(d_l) = \sum_{r=i}^{k_1} p_{r,0} \Pi_{r-i,n_2} \tag{3.3.55}$$

for $i = n_1 + 1, \dots, k_1$,

$$\begin{aligned}\sum_{m=0}^{k_2} b_m \big(c_m s_{n_1}(\mu_1 + c_m) s_j(-c_m) + p\lambda_1 s_{n_1}(c_m) s_j(-\mu_2 - c_m)\big) \theta_{n_1}(c_m) \phi_{n_2}(c_m) \\ = n_1 \sum_{r=n_1+1}^{k_1} p_{r,0} \Pi_{r-n_1-1,j}\end{aligned} \tag{3.3.56}$$

for $j = 0, \dots, n_2 - 1$,

$$\begin{aligned}\sum_{m=0}^{k_2} b_m c_m s_{n_2}(-c_m) \big(s_{n_1}(\mu_1 + c_m) \theta_{n_1}(c_m) + s_{n_1}(c_m) \Omega_{q_1-1}(c_m)\big) \phi_j(c_m) \\ = n_1 \sum_{l=1}^{q_1} e_l \theta_{n_1+1}(d_l) \phi_j(d_l)\end{aligned} \tag{3.3.57}$$

for $j = n_2 + 1, \dots, k_2$ and

$$\lambda_2 \sum_{m=0}^{k_2} b_m s_{n_1}(c_m) s_{n_2}(-c_m)\theta_i(c_m)\phi_{n_2-1}(c_m) \\ + \lambda_2 \sum_{l=1}^{q_1} e_l \theta_i(d_l)\phi_{n_2-1}(d_l) = (p\lambda_1 + \lambda_2)\sum_{r=i}^{k_1} p_{r,0}\Pi_{r-i,n_2-1} \quad (3.3.58)$$

for $i = n_1+1, \dots, k_1$. It can be shown that the boundary equation for $i = n_1$ and $j = n_2$ is redundant. The normalization condition yields

$$b_0 s_{n_1}(\mu)\theta_{n_1}(0)\big(s_{n_2}(\mu_2)\Psi_{q_2}(0) - s_{n_2-1}(\mu_2)\Psi_{q_2-1}(0)\big) \\ + \frac{\lambda_1}{n_1}\sum_{m=0}^{k_2} b_m s_{n_1}(c_m)s_{n_2}(-c_m)(\Omega_{q_1}(c_m) - 1)\Psi_{q_2}(c_m) \quad (3.3.59) \\ + \frac{\lambda_1}{n_1}\sum_{l=1}^{q_1} e_l(\Omega_{q_1}(d_l) - 1)\Psi_{q_2}(d_l) = 1.$$

The result is outlined in the following theorem.

Theorem 3.3.14. *The unique nonnegative and normalized solution of the steady-state equations* (3.3.50) *is given by*

$$p_{i,j} = \begin{cases} \displaystyle\sum_{m=0}^{k_2} b_m s_i(c_m, \lambda_1, \mu_1) s_j(-c_m, \lambda_2, \mu_2)\theta_{n_1}(c_m)\phi_{n_2}(c_m), & j = 0, \dots, n_2, \\ \displaystyle\sum_{m=0}^{k_2} b_m s_i(c_m, \lambda_1, \mu_1) s_{n_2}(-c_m, \lambda_2, \mu_2)\theta_{n_1}(c_m)\phi_j(c_m), & j = n_2, \dots, k_2, \end{cases}$$

for $i = 0, \dots, n_1$,

$$p_{i,j} = \sum_{m=0}^{k_2} b_m s_{n_1}(c_m, \lambda_1, \mu_1) s_{n_2}(-c_m, \lambda_2, \mu_2)\theta_i(c_m)\phi_j(c_m) + \sum_{l=1}^{q_1} e_l \theta_i(d_l)\phi_j(d_l)$$

for $i = n_1, \dots, k_1$, $j = n_2, \dots, k_2$ *and*

$$p_{i,j} = \sum_{r=i}^{k_1} p_{r,0}\Pi_{r-i,j}$$

for $i = n_1 + 1, \dots, k_1$ *and* $j = 0, \dots, n_2$ *if the coefficients* $b_0, \dots, b_{k_2}$ *and* $e_1, \dots, e_{q_1}$ *and the probabilities* $p_{n_1+1,0}, \dots, p_{k_1,0}$ *are determined by the* $k_2 + 2q_1 + 1$ *equations* (3.3.55)-(3.3.58) *and* (3.3.59). $\Pi_{r,j}$, $r = 0, \dots, k_1$, $j = 0, \dots, n_2$, *are the solutions of* (3.3.54). $c_1, \dots, c_{k_2}$ *are the by Theorem 3.3.6 positive and pairwise distinct solutions of* (3.3.17), $c_0 = 0$ *and* $d_1, \dots, d_{q_1}$ *are* *the by Theorem 3.3.7 negative and pairwise distinct solutions of* $\theta_{n_1}(d_l) = 0$, $l = 1, \dots, q_1$.

The problem of determining the $(k_1 + 1)(k_2 + 1)$ unknowns $p_{i,j}$, $i = 0, \dots, k_1$, $j = 0, \dots, k_2$, has now been reduced to the problem of determining $k_2 + q_1$ eigenvalues from (3.3.17) and the equation $\theta_{n_1}(d) = 0$ and $k_2 + 2q_1 + 1$ unknowns from the linear equations (3.3.55) to (3.3.59). An alternative separation approach for the case $\mu_1 = \mu_2$ and $p = 1$ that is preferable when q_2 is infinite or large compared to q_1, n_1 and n_2, can be found in Morrison [47]. This approach allows for a generalization to the case of arbitrary service rates and weighted overflow traffic.

3.3.3 Jockeying to the waiting room

Now we consider model S/S/W. In this model, waiting customers from Q_1 move to Q_2 as soon as there is a position in the waiting room or a server available. An arriving Q_1-customer is blocked and directed to Q_2 if all n_1 servers are busy in Q_1. Blocked customers are served by one of the servers in Q_2 if at least one is available. They are queued in Q_1 if all servers in Q_2 are busy and a waiting position is available in Q_1 and are lost otherwise. Thus, $p_{i,j} = 0$ for $i > n_1$ and $j = 0, \dots, k_2$, because it is impossible for customers to wait at Q_1 while there is at least one waiting position or server available at Q_2. The balance equations for this model are

$$\begin{aligned}
&(\lambda_1(1-\delta_{ik_1})(1-\chi_{i-n_1}(1-\chi_{j-n_2})) + p\lambda_1\chi_{i-n_1}(1-\chi_{j-n_2}))\\
&\qquad + \lambda_2(1-\delta_{jk_2}) + (i \wedge n_1)\mu_1 + (j \wedge n_2)\mu_2)p_{i,j}\\
&\quad = \big(1-\chi_{i-n_1-1}(1-\delta_{jk_2})\big)\\
&\qquad\qquad \times \big(\lambda_1(1-\delta_{i0})p_{i-1,j} + (1-\delta_{jk_2})((j+1) \wedge n_2)\mu_2 p_{i,j+1}\big)\\
&\qquad + (1-\delta_{j0})\big(p\lambda_1\delta_{in_1}\chi_{n_2-j} + \lambda_1\delta_{in_1}(1-\chi_{n_2-j}) + \lambda_2(1-\chi_{i-n_1-1})\big)p_{i,j-1}\\
&\qquad + (1-\delta_{ik_1})\big((1-\chi_{i-n_1}(1-\delta_{jk_2}))((i+1) \wedge n_1)\mu_1\\
&\qquad + n_2\mu_2\delta_{jn_2}\chi_{i-n_1}\big)p_{i+1,j} \qquad\qquad (3.3.60)
\end{aligned}$$

for $i = 0, \dots, k_1$ and $j = 0, \dots, k_2$. The balance equations are constructed so as to imply $p_{i,j} = 0$ for $i > n_1$ and $j = 0, \dots, k_2$. The term $\lambda_1\delta_{in_1}(1-\chi_{n_2-j})$ in the second last line indicates that arriving Q_1-customers, who find all waiting positions in Q_1 unoccupied and all n_1 servers busy in Q_1, move instantaneously to Q_2 if there is a waiting position or server available. Hence, λ_1 is not weighted by p since the customers are jockeying immediately to Q_2 after their arrival to Q_1 instead of overflowing. By definition, overflow takes places in model-type S/S/γ if and only if all n_1 servers in Q_1 are busy and there is at least one server available in Q_2.

Remark 3.3.15. It might be useful in applications to weight the traffic that is due to jockeying with p. This is done by replacing the terms in the first line of (3.3.60) by $\lambda_1(1-\delta_{ik_1})\big(1-\chi_{i-n_1}(1-\delta_{jk_2})\big)+p\lambda_1\chi_{i-n_1}(1-\delta_{jk_2})$ and the term $\lambda_1\delta_{in_1}(1-\chi_{n_2-j})$ on the right side by $p\lambda_1\delta_{in_1}(1-\chi_{n_2-j})$. However, in this case, the model S/S/W is equivalent to model S/W/W from Section 3.4.2

Although the balance equations (3.3.60) are intricate at first glance, they are easily solved with an analog of the procedure of the alternative approach in Section 3.3.1.6. The separation approach for this model is shown in Figure 3.8.

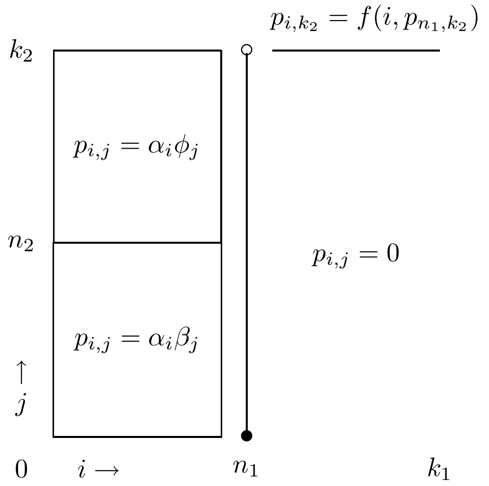

$p_{i,k_2} = f(i, p_{n_1,k_2})$
k_2
$p_{i,j} = \alpha_i\phi_j$
n_2
$p_{i,j} = 0$
$p_{i,j} = \alpha_i\beta_j$
$\uparrow$ j
0
$i \rightarrow$
n_1
k_1

Figure 3.8: Separation scheme: Model S/S/W

For $i = 0, \ldots, n_1 - 1$ and $j = 0, \ldots, k_2$ the balance equations are identical to (3.3.2). It then follows from (3.3.27) that

$$p_{i,j} = \begin{cases} \displaystyle\sum_{m=0}^{k_2} a_m s_i(c_m, \lambda_1, \mu_1) s_j(-c_m, \lambda_2, \mu_2)\phi_{n_2}(c_m), & j = 0, \ldots, n_2, \\ \displaystyle\sum_{m=0}^{k_2} a_m s_i(c_m, \lambda_1, \mu_1) s_{n_2}(-c_m, \lambda_2, \mu_2)\phi_j(c_m), & j = n_2, \ldots, k_2, \end{cases} \tag{3.3.61}$$

for $i = 0, \ldots, n_1$ with $c_0, \ldots, c_m$ such that (3.3.17) is fulfilled. It remains to determine $p_{i,j}$ for $j = k_2$ and $i = n_1 + 1, \ldots, k_1$ because the other probabilities vanish. The balance equations for $i = n_1 + 1, \dots, k_1$ and $j = k_2$ are identical to (3.3.44) with $p_{\cdot,n_2}$ replaced by $p_{\cdot,k_2}$. We obtain

$$p_{i,k_2} = \left(\frac{\lambda_1}{n_1\mu_1 + n_2\mu_2}\right)^{i-n_1} p_{n_1,k_2}, \quad i = n_1, \dots, k_1, \tag{3.3.62}$$

as in (3.3.45), where p_{n_1,k_2} is given by (3.3.61). As usual, the constants $a_0, \dots, a_{k_2}$ can be determined from the normalization condition and the boundary conditions at $i = n_1$, i.e.,

$$\begin{aligned}(p\lambda_1 + \lambda_2 + n_1\mu_1 + j\mu_2)p_{n_1,j} &= \lambda_1 p_{n_1-1,j} + (1 - \delta_{j0})(p\lambda_1 + \lambda_2)p_{n_1,j-1} \\ &\quad + (j+1)\mu_2 p_{n_1,j+1}, \quad j = 0, \dots, n_2 - 1,\end{aligned} \tag{3.3.63}$$

$$\begin{aligned}(\lambda_1 + \lambda_2 + n_1\mu_1 + n_2\mu_2)p_{n_1,n_2} &= \lambda_1 p_{n_1-1,n_2} + n_2\mu_2 p_{n_1,n_2+1} \\ &\quad + (p\lambda_1 + \lambda_2)p_{n_1,n_2-1}, \quad j = n_2,\end{aligned} \tag{3.3.64}$$

$$\begin{aligned}(\lambda_1 + \lambda_2 + n_1\mu_1 + n_2\mu_2)p_{n_1,j} &= \lambda_1 p_{n_1-1,j} + n_2\mu_2 p_{n_1,j+1} \\ &\quad + (\lambda_1 + \lambda_2)p_{n_1,j-1}, \quad j = n_2 + 1, \dots, k_2 - 1,\end{aligned} \tag{3.3.65}$$

$$\begin{aligned}(\lambda_1 + n_1\mu_1 + n_2\mu_2)p_{n_1,k_2} &= \lambda_1 p_{n_1-1,n_2} + (n_1\mu_1 + n_2\mu_2)p_{n_1+1,k_2} \\ &\quad + (\lambda_1 + \lambda_2)p_{n_1,k_2-1}, \quad j = k_2.\end{aligned} \tag{3.3.66}$$

Inserting (3.3.61) and (3.3.62) into (3.3.63) to (3.3.66) and using (3.3.16) and the recurrence relations (3.2.13), (3.2.14) and additionally $\phi_{k_2} \equiv 1$ and (3.3.12) for the last equation gives the conditions for $a_0, \dots, a_{k_2}$. These conditions are

$$\sum_{m=0}^{k_2} a_m\big(c_m s_{n_1}(c_m + \mu_1)s_j(-c_m) + p\lambda_1 s_{n_1}(c_m)s_j(-c_m - \mu_2)\big)\phi_{n_2}(c_m) = 0 \tag{3.3.67}$$

for $j = 0, \dots, n_2 - 1$ which follow from (3.3.63) and

$$\begin{aligned}\sum_{m=0}^{k_2} a_m\big(c_m s_{n_1}(c_m + \mu_1)s_{n_2}(-c_m) - p\lambda_1 s_{n_1}(c_m)s_{n_2-1}(-c_m) \\ + \lambda_1 s_{n_1}(c_m)s_{n_2}(-c_m)\big)\phi_{n_2}(c_m) = 0\end{aligned} \tag{3.3.68}$$

which follows from (3.3.64). The equations (3.3.65) yield

$$\sum_{m=0}^{k_2} a_m\big(c_m s_{n_1}(c_m+\mu_1)s_{n_2}(-c_m)\phi_j(c_m) \\ + \lambda_1 s_{n_1}(c_m)s_{n_2}(-c_m)(\phi_j(c_m)-\phi_{j-1}(c_m))\big) = 0 \tag{3.3.69}$$

for $j = n_2+1,\dots,k_2-1$ and finally (3.3.66) gives

$$\sum_{m=0}^{k_2} a_m\big(c_m s_{n_1}(c_m+\mu_1)s_{n_2}(-c_m) - \lambda_1 s_{n_1}(c_m)s_{n_2}(-c_m)\phi_{k_2-1}(c_m)\big) = 0. \tag{3.3.70}$$

Summing (3.3.67) over $j = 0,\dots,n_2-1$ and (3.3.69) over $j = n_2+1,\dots,k_2-1$ using (3.3.7) and (3.2.15) gives

$$\sum_{m=0}^{k_2} a_m\big(c_m s_{n_1}(c_m+\mu_1)s_{n_2-1}(-c_m+\mu_2) \\ + p\lambda_1 s_{n_1}(c_m)s_{n_2-1}(-c_m)\big)\phi_{n_2}(c_m) = 0 \tag{3.3.71}$$

and

$$\sum_{m=0}^{k_2} a_m\big(c_m s_{n_1}(c_m+\mu_1)s_{n_2}(-c_m)(\Psi_{q_2-1}(c_m)-\Psi_0(c_m)) \\ + \lambda_1 s_{n_1}(c_m)s_{n_2}(-c_m)(\phi_{k_2-1}(c_m)-\phi_{n_2}(c_m))\big) = 0. \tag{3.3.72}$$

Finally, adding (3.3.68), (3.3.71) and (3.3.72) and simplifying by once again using (3.3.17) and the recurrence relations (3.2.13) and (3.2.14) results in (3.3.70). Consequently, condition (3.3.70) is redundant and $a_0,\dots,a_{k_2}$ are determined by (3.3.67), (3.3.68), (3.3.69) and the normalization condition. The normalization condition can be expressed in terms of $a_0,\dots,a_{k_2}$ and gives after substituting (3.3.61) and (3.3.62) and reduction with the help of (3.3.17) and (3.2.15)

$$a_0 s_{n_1}(\mu_1,\lambda_1,\mu_1)\big(s_{n_2}(\mu_2,\lambda_2,\mu_2)\Psi_{q_2}(0) - s_{n_2-1}(\mu_2,\lambda_2,\mu_2)\Psi_{q_2-1}(0)\big) \\ + \sum_{m=0}^{k_2} a_m s_{n_1}(c_m,\lambda_1,\mu_1)s_{n_2}(-c_m,\lambda_2,\mu_2)\sum_{l=1}^{q_1}\left(\frac{\lambda_1}{n_1\mu_1+n_2\mu_2}\right)^l = 1. \tag{3.3.73}$$

The separation approach and the derivations of this section lead to the following theorem that summarizes the results of this section.

Theorem 3.3.16. *The unique nonnegative and normalized solution of the steady-state equations* (3.3.60) *is given by*

$$p_{i,j} = \begin{cases} \sum_{m=0}^{k_2} a_m s_i(c_m, \lambda_1, \mu_1) s_j(-c_m, \lambda_2, \mu_2) \phi_{n_2}(c_m), & j = 0, \dots, n_2, \\ \sum_{m=0}^{k_2} a_m s_i(c_m, \lambda_1, \mu_1) s_{n_2}(-c_m, \lambda_2, \mu_2) \phi_j(c_m), & j = n_2, \dots, k_2, \end{cases}$$

for $i = 0, \dots, n_1$ *and*

$$p_{i,k_2} = \left(\frac{\lambda_1}{n_1\mu_1 + n_2\mu_2} \right)^{i-n_1} p_{n_1,k_2}$$

for $i = n_1, \dots, k_1$ *if the coefficients* $a_0, \dots, a_{k_2}$ *are determined by the* $k_2 + 1$ *equations* (3.3.67), (3.3.68), (3.3.69) *and* (3.3.73). $c_0 = 0$ *and* $c_1, \dots, c_{k_2}$ *are the by Theorem 3.3.6 positive and pairwise distinct solutions of* (3.3.17).

Remark 3.3.17. As in Section 3.3.1.6 the results remain valid even if $q_1 = \infty$ as long as the stability condition

$$\lambda_1 < n_1\mu_1 + n_2\mu_2$$

is fulfilled. The stability condition is again independent of p.

The problem of determining the $(n_1 + 1)(k_2 + 1) + q_1$ nonzero unknowns $p_{i,j}$, $i = 0, \dots, k_1$, $j = 0, \dots, k_2$, has now been reduced to the problem of determining $k_2 + 1$ eigenvalues and $k_2 + 1$ unknowns.

3.4 Overflow with waiting rooms: From servers to waiting room

3.4.1 Jockeying to servers and no jockeying

In this section, we consider the deterministic overflow models S/W/S and S/W/N, i.e., we choose $p = 1$, and show exemplarily for model S/W/N that the technique used throughout the previous sections does not succeed for this models. However, it is possible to treat the case S/W/W because in this case certain stationary probabilities vanish. This is done in the next section.

The number of servers and the waiting room capacity of Q_i is n_i and $q_i \geq 1$, respectively, for $i = 1, 2$. In these models with blocking rule S, an arriving Q_1-customer is blocked if all n_1 servers are busy. Blocked Q_1-customers are treated with respect to overflow routine W. They are served by one of the servers in Q_2 if at least one is available and are queued in Q_2 if all Q_2-servers are busy and at least one waiting position is available. The blocked Q_1-customers who find Q_2 fully occupied are redirected to the waiting room in Q_1 and are lost if Q_1 is fully occupied. Hence, the overflow stream from Q_1 follows the same route through Q_2 as the Q_2-customers do. The jockeying discipline N ascertains that waiting Q_1-customers are served solely in Q_1. The balance equations for model S/W/N are given by

$$\begin{aligned}
&(\lambda_1(1-\delta_{ik_1}\delta_{jk_2}) + \lambda_2(1-\delta_{jk_2}) + (i \wedge n_1)\mu_1 + (j \wedge n_2)\mu_2)p_{i,j} \\
&\quad = (1-\delta_{i0})(1-\chi_{i-n_1-1}\delta_{jk_2})\lambda_1 p_{i-1,j} + (1-\delta_{ik_1})((i+1)\wedge n_1)\mu_1 p_{i+1,j} \\
&+ (1-\delta_{j0})\big(\lambda_1\chi_{i-n_1} + \lambda_2\big)p_{i,j-1} + (1-\delta_{jk_2})((j+1)\wedge n_2)\mu_2 p_{i,j+1} \qquad (3.4.1)
\end{aligned}$$

for $i = 0, \ldots, k_1$ and $j = 0, \ldots, k_2$. For $i = n_1 + 1, \ldots, k_1$ and $j = n_2 + 1, \ldots, k_2$ these equation reduce to

$$\begin{aligned}
&(\lambda_1(1-\delta_{ik_1}\delta_{jk_2}) + \lambda_2(1-\delta_{jk_2}) + n_1\mu_1 + n_2\mu_2)p_{i,j} \\
&\qquad = \delta_{jk_2}\lambda_1 p_{i-1,j} + (1-\delta_{ik_1})n_1\mu_1 p_{i+1,j} \\
&\qquad\quad + (\lambda_1+\lambda_2)p_{i,j-1} + (1-\delta_{jk_2})n_2\mu_2 p_{i,j+1}. \qquad (3.4.2)
\end{aligned}$$

Using the separation idea in this region, i.e., trying again $p_{i,j} = \alpha_i\beta_j$, gives the equations

$$(\lambda_1(1-\delta_{ik_1}) + n_1\mu_1 + n_2\mu_2 - \frac{\beta_{k_2-1}}{\beta_{k_2}}(\lambda_1+\lambda_2))\alpha_i = (1-\delta_{ik_1})n_1\mu_1\alpha_{i+1} \qquad (3.4.3)$$

for $i = n_1 + 1, \ldots, k_1$ and

$$\begin{aligned}
&\big((\lambda_1+\lambda_2)(1-\delta_{jk_2}) + n_1\mu_1 + n_2\mu_2 - (1-\delta_{jk_2})\frac{\alpha_{k_1-1}}{\alpha_{k_1}}\lambda_1\big)\beta_j \\
&\qquad = (\lambda_1+\lambda_2)\beta_{j-1} + (1-\delta_{jk_2})n_2\mu_2\beta_{j+1} \qquad (3.4.4)
\end{aligned}$$

for $j = n_2 + 1, \ldots, k_2$. It is seen that the separation approach leads to different separation constants in (3.4.3) and (3.4.4). Furthermore, a solution of (3.4.3) cannot be expressed in terms of a solution of (3.4.4). Instead of solving these equations simultaneously, equation (3.4.3) can be solved recursively, while (3.4.4) leads to another eigenproblem and consequently additional coefficients arise from the boundary conditions. Furthermore, it is necessary to piece together the solutions at the boundaries of the separation regions. This results in additional conditions for the sets of eigenvalues. Moreover, a sequential approach as in Section 3.3.2 is not possible in this region and in the region $i = n_1+1, \dots, k_1$, $j = 0, \dots, n_2$. Similar statements hold for the models W/W/S and W/W/W (see Section 3.6.1).

3.4.2 Jockeying to the waiting room

Now we consider the p-overflow model S/W/W. The number of servers and the waiting room capacity of Q_i is n_i and $q_i \geq 1$, respectively, for $i = 1, 2$. In this model with blocking rule S an arriving Q_1-customer is blocked and directed to Q_2 if all n_1 servers are busy in Q_1. Furthermore, blocked Q_1-customers are treated with respect to overflow routine W. Hence, the overflow stream from Q_1 follows the same route through the waiting room and servers in Q_2 as the Q_2-customers do. The jockeying discipline W ascertains that waiting Q_1-customers skip to Q_2 as soon as a waiting position or server is available in Q_2.

The balance equations are

$$\begin{aligned}
&(\lambda_1(1-\delta_{ik_1})(1-\chi_{i-n_1}(1-\delta_{jk_2})) + p\lambda_1\chi_{i-n_1}(1-\delta_{jk_2}) \\
&\qquad + \lambda_2(1-\delta_{jk_2}) + (i \wedge n_1)\mu_1 + (j \wedge n_2)\mu_2)p_{i,j} \\
&\quad = \big(1-\chi_{i-n_1-1}(1-\delta_{jk_2})\big) \\
&\qquad\qquad \times \big(\lambda_1(1-\delta_{i0})p_{i-1,j} + (1-\delta_{jk_2})((j+1) \wedge n_2)\mu_2 p_{i,j+1}\big) \qquad (3.4.5) \\
&\qquad + (1-\delta_{j0})\big(p\lambda_1\delta_{in_1} + \lambda_2(1-\chi_{i-n_1-1})\big)p_{i,j-1} \\
&\qquad + (1-\delta_{ik_1})\big((1-\chi_{i-n_1}(1-\delta_{jk_2}))((i+1) \wedge n_1)\mu_1 \\
&\qquad\qquad\qquad + n_2\mu_2\delta_{jk_2}\chi_{i-n_1}\big)p_{i+1,j}
\end{aligned}$$

for $i = 0, \dots, k_1$ and $j = 0, \dots, k_2$. Once again, these equations are constructed such that $p_{i,j} = 0$ for $i = n_1 + 1, \dots, k_1$ and $j \neq k_2$ since it is impossible for customers to wait at Q_1 while there is at least one waiting position available in Q_2.

The separation approach for this model is identical to the one for model S/S/W depicted in Figure 3.8. For $i = 0, \dots, n_1 - 1$ and $j = 0, \dots, k_2$ the balance equations are identical to those in (3.3.2) so that $p_{i,j}$ for $i = 0, \dots, n_1$

and $j = 0, \dots, k_2$ can be chosen as in (3.3.27):

$$p_{i,j} = \begin{cases} \displaystyle\sum_{m=0}^{k_2} b_m s_i(c_m, \lambda_1, \mu_1) s_j(-c_m, \lambda_2, \mu_2) \phi_{n_2}(c_m), & j = 0, \dots, n_2, \\ \displaystyle\sum_{m=0}^{k_2} b_m s_i(c_m, \lambda_1, \mu_1) s_{n_2}(-c_m, \lambda_2, \mu_2) \phi_j(c_m), & j = n_2, \dots, k_2, \end{cases} \tag{3.4.6}$$

for $i = 0, \dots, n_1$, where $b_0, \dots, b_{k_2}$ have to be determined from the boundary conditions in (3.4.5) and the normalization condition and $c_0, \dots, c_{k_2}$ are the solutions of (3.3.17). The balance equations for $i = n_1+1, \dots, k_1$ and $j = k_2$ are

$$(\lambda_1(1-\delta_{ik_1}) + n_1\mu_1 + n_2\mu_2) p_{i,k_2} = \lambda_1 p_{i-1,k_2} + (1-\delta_{ik_1})(n_1\mu_1 + n_2\mu_2) p_{i+1,k_2}. \tag{3.4.7}$$

These equations are equivalent to (3.3.44) with $p_{\cdot,n_2}$ replaced by $p_{\cdot,k_2}$. We get from (3.3.45) that

$$p_{i,k_2} = \left(\frac{\lambda_1}{n_1\mu_1 + n_2\mu_2} \right)^{i-n_1} p_{n_1,k_2}, \quad i = n_1, \dots, k_1, \tag{3.4.8}$$

where p_{n_1,k_2} is given by (3.4.6) (see also Section 3.3.3). Now we turn to the determination of the constants $b_0, \dots, b_{k_2}$ with the help of the boundary conditions for $i = n_1$, i.e.,

$$\begin{aligned} (p\lambda_1 + \lambda_2 + n_1\mu_1 + j\mu_2) p_{n_1,j} &= \lambda_1 p_{n_1-1,j} + (1-\delta_{j0})(p\lambda_1 + \lambda_2) p_{n_1,j-1} \\ &\quad + (j+1)\mu_2 p_{n_1,j+1}, \quad j = 0, \dots, n_2 - 1, \end{aligned} \tag{3.4.9}$$

$$\begin{aligned} (p\lambda_1 + \lambda_2 + n_1\mu_1 + n_2\mu_2) p_{n_1,n_2} &= \lambda_1 p_{n_1-1,n_2} + n_2\mu_2 p_{n_1,n_2+1} \\ &\quad + (p\lambda_1 + \lambda_2) p_{n_1,n_2-1}, \quad j = n_2, \end{aligned} \tag{3.4.10}$$

$$\begin{aligned} (p\lambda_1 + \lambda_2 + n_1\mu_1 + n_2\mu_2) p_{n_1,j} &= \lambda_1 p_{n_1-1,j} + n_2\mu_2 p_{n_1,j+1} \\ &\quad + (p\lambda_1 + \lambda_2) p_{n_1,j-1}, \; j = n_2 + 1, \dots, k_2 - 1, \end{aligned} \tag{3.4.11}$$

$$\begin{aligned} (\lambda_1 + n_1\mu_1 + n_2\mu_2) p_{n_1,k_2} &= \lambda_1 p_{n_1-1,n_2} + (n_1\mu_1 + n_2\mu_2) p_{n_1+1,k_2} \\ &\quad + (p\lambda_1 + \lambda_2) p_{n_1,k_2-1}, \quad j = k_2. \end{aligned} \tag{3.4.12}$$

Using (3.4.6) and (3.4.8) and paralleling the simplifications that lead from (3.3.63) - (3.3.66) to (3.3.67) - (3.3.70) it is seen that (3.4.9) - (3.4.12) are equivalent to the following set of equations:

$$\sum_{m=0}^{k_2} b_m\big(c_m s_{n_1}(c_m+\mu_1)s_j(-c_m)+p\lambda_1 s_{n_1}(c_m)s_j(-c_m-\mu_2)\big)\phi_{n_2}(c_m)=0 \tag{3.4.13}$$

for $j=0,\dots,n_2-1$,

$$\begin{aligned}\sum_{m=0}^{k_2} b_m\big(&c_m s_{n_1}(c_m+\mu_1)s_{n_2}(-c_m)\phi_j(c_m)\\ &+p\lambda_1 s_{n_1}(c_m)s_{n_2}(-c_m)(\phi_j(c_m)-\phi_{j-1}(c_m))\big)=0\end{aligned} \tag{3.4.14}$$

for $j=n_2,\dots,k_2-1$ and

$$\sum_{m=0}^{k_2} b_m\big(c_m s_{n_1}(c_m+\mu_1)s_{n_2}(-c_m)-p\lambda_1 s_{n_1}(c_m)s_{n_2}(-c_m)\phi_{k_2-1}(c_m)\big)=0. \tag{3.4.15}$$

Note that both (3.4.10) and (3.4.11) reduce to the form of (3.4.14). Summing (3.4.13) over $j=0,\dots,n_2-1$ and (3.4.14) over $j=n_2,\dots,k_2-1$ and adding the resulting equations yields the redundancy of (3.4.15). Finally, the normalization condition, (3.4.6) and (3.4.8) lead to

$$\begin{aligned}&b_0 s_{n_1}(\mu_1,\lambda_1,\mu_1)\big(s_{n_2}(\mu_2,\lambda_2,\mu_2)\Psi_{q_2}(0)-s_{n_2-1}(\mu_2,\lambda_2,\mu_2)\Psi_{q_2-1}(0)\big)\\ &+\sum_{m=0}^{k_2} b_m s_{n_1}(c_m,\lambda_1,\mu_1)s_{n_2}(-c_m,\lambda_2,\mu_2)\sum_{l=1}^{q_1}\left(\frac{\lambda_1}{n_1\mu_1+n_2\mu_2}\right)^l=1\end{aligned} \tag{3.4.16}$$

as in model S/S/W. The calculations of this section are summarized in the next theorem.

Theorem 3.4.1. *The unique nonnegative and normalized solution of the steady-state equations* (3.4.5) *is given by*

$$p_{i,j}=\begin{cases}\displaystyle\sum_{m=0}^{k_2} b_m s_i(c_m,\lambda_1,\mu_1)s_j(-c_m,\lambda_2,\mu_2)\phi_{n_2}(c_m), & j=0,\dots,n_2,\\ \displaystyle\sum_{m=0}^{k_2} b_m s_i(c_m,\lambda_1,\mu_1)s_{n_2}(-c_m,\lambda_2,\mu_2)\phi_j(c_m), & j=n_2,\dots,k_2,\end{cases}$$

for $i=0,\dots,n_1$ *and*

$$p_{i,k_2}=\left(\frac{\lambda_1}{n_1\mu_1+n_2\mu_2}\right)^{i-n_1}p_{n_1,k_2}$$

for $i = n_1, \dots, k_1$ *if the coefficients* $b_0, \dots, b_{k_2}$ *are determined by the* $k_2 + 1$ *equations* (3.4.13), (3.4.14) *and* (3.4.16). $c_0 = 0$ *and* $c_1, \dots, c_{k_2}$ *are the by Theorem 3.3.6 positive and pairwise distinct solutions of* (3.3.17).

The problem of determining the $(n_1 + 1)(k_2 + 1) + q_1$ nonzero unknowns has now been reduced to the problem of determining $k_2 + 1$ eigenvalues and $k_2 + 1$ unknowns.

Remark 3.4.2. The results remain valid even if $q_1 = \infty$ as long as the stability condition $\lambda_1 < n_1\mu_1 + n_2\mu_2$ is fulfilled.

3.5 Overflow with waiting rooms: From waiting room to server

In the following two sections, we consider p-overflow models with blocking rule W and overflow routine S. In these models, arriving Q_1-customers are blocked if all Q_1-servers are busy and the Q_1-waiting room is fully occupied, i.e., no overflow is allowed as long as the waiting room in Q_1 is not fully occupied. The blocked customers overflow to Q_2 and are served by a server in Q_2 if at least one is idle. In these models the (unweighted) overflow stream is identical to the stream of blocked customers.

3.5.1 Jockeying to servers and to waiting room

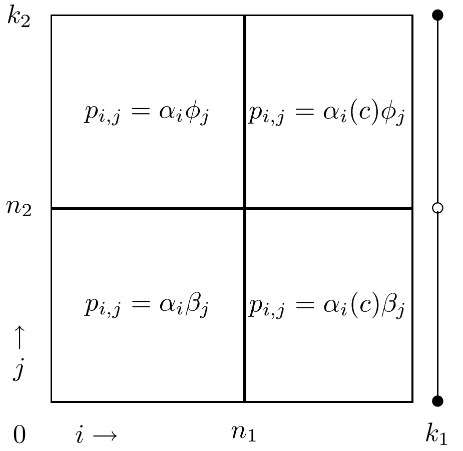
At first, it should be mentioned that the configurations W/S/S and W/S/W make sense only if the capacity of the waiting room in Q_1 is set to zero. This can be explained exemplarily for model W/S/S by the following observations. On the one hand, in model W/S/S with $q_1 \geq 1$, waiting Q_1-customers swap to Q_2 as soon as a Q_2-server becomes available so that $p_{i,j} = 0$ for $i = n_1 + 1, \dots, k_1$ and $j = 0, \dots, n_2 - 1$. On the other hand, arriving Q_1-customers overflow to Q_2 if all Q_1-servers are busy, the Q_1-waiting room is fully occupied and a server is available in Q_2, i.e., $i = k_1$ and $j = 0, \dots, n_2-1$, but in these cases $p_{i,j} = 0$. Consequently, no overflow occurs in W/S/S for $q_1 \geq 1$. The same observations can be made for jockeying discipline W, since it is a generalization of jockeying discipline S. Therefore, no overflow takes place in model W/S/W for the case $q_1 \geq 1$. However, with $q_1 = 0$, model W/S/S and W/S/W correspond to model S/S/S and S/S/W, respectively, which were treated in Section 3.3.1 and Section 3.3.3.

3.5.2 No jockeying

We examine model W/S/N by generalizing the results in Morrison [48] to the case of p-overflow, $p \in [0, 1]$, and arbitrary service rates $\mu_1, \mu_2 > 0$ (see also Kaufman et al. [34] for a numerical treatment for the case of arbitrary service rates and $p = 1$). In this model, overflow is allowed if and only if Q_1 is fully occupied. The blocked customers overflow to Q_2 and are served by a server in Q_2 if at least one is idle. Waiting Q_1-customers are served exclusively at Q_1. The balance equations are given by

$$\begin{aligned}(\lambda_1(1-\delta_{ik_1})\big(1-\delta_{ik_1}(1-\chi_{j-n_2})\big) + p\lambda_1\delta_{ik_1}(1-\chi_{j-n_2}) + \lambda_2(1-\delta_{jk_2})\\ + (i \wedge n_1)\mu_1 + (j \wedge n_2)\mu_2)p_{i,j}\\ = \lambda_1(1-\delta_{i0})p_{i-1,j} + (1-\delta_{ik_1})((i+1) \wedge n_1)\mu_1 p_{i+1,j}\\ + (1-\delta_{j0})(p\lambda_1\delta_{ik_1}\chi_{n_2-j} + \lambda_2)p_{i,j-1}\\ + (1-\delta_{jk_2})((j+1) \wedge n_2)\mu_2 p_{i,j+1} \qquad (3.5.1)\end{aligned}$$

for $i = 0, \ldots, k_1$ and $j = 0, \ldots, k_2$. Figure 3.9 displays the separation approach used for this model.



k_2
$p_{i,j} = \alpha_i\phi_j$
$p_{i,j} = \alpha_i(c)\phi_j$
n_2
$p_{i,j} = \alpha_i\beta_j$
$p_{i,j} = \alpha_i(c)\beta_j$
↑
j
0
$i \rightarrow$
n_1
k_1

Figure 3.9: Separation scheme: Model W/S/N

Separating the variables into $p_{i,j} = \alpha_i\beta_j$ for $i \neq k_1$ and $j = 0, \ldots, k_2$ leads analogously to Section 3.3.1 to

$$\beta_j = \begin{cases} s_j(-c, \lambda_2, \mu_2)\phi_{n_2}(c), & j = 0, \ldots, n_2, \\ s_{n_2}(-c, \lambda_2, \mu_2)\phi_j(c), & j = n_2 - 1, \ldots, k_2, \end{cases} \qquad (3.5.2)$$

where c is a solution of $s_{n_2-1}(-c,\lambda_2,\mu_2)\phi_{n_2}(c) = s_{n_2}(-c,\lambda_2,\mu_2)\phi_{n_2-1}(c)$ or equivalently (3.3.17). For $i = 0,\dots,n_1$ we get

$$\alpha_i = s_i(c,\lambda_1,\mu_1). \tag{3.5.3}$$

For $i = n_1,\dots,k_1$, the equations

$$(\lambda_1 + n_1\mu_1 + c)\alpha_i = \lambda_1\alpha_{i-1} + n_1\mu_1\alpha_{i+1} \tag{3.5.4}$$

must be satisfied with the boundary conditions $\alpha_{n_1} = s_{n_1}(c,\lambda_1,\mu_1)$ and $\alpha_{n_1-1} = s_{n_1-1}(c,\lambda_1,\mu_1)$. It follows from (3.3.11) and (3.3.5) that the functions $\Omega_i(c)$, $i \geq -1$, solve

$$(\lambda_1 + n_1\mu_1 + c)\Omega_i = \lambda_1\Omega_{i+1} + n_1\mu_1\Omega_{i-1}$$

for $i \geq 0$. Thus, a solution of (3.5.4) is given by

$$\alpha_i = \left(\frac{\lambda_1}{n_1\mu_1}\right)^{n_1-i} \big(s_{n_1}(c,\lambda_1,\mu_1)\Omega_{i-n_1}(c) - s_{n_1-1}(c,\lambda_1,\mu_1)\Omega_{i-n_1-1}(c)\big). \tag{3.5.5}$$

Plugging (3.5.3) and (3.5.4) together by an appropriate normalization yields the solutions for the first separation variable α_i:

$$\alpha_i(c) = \begin{cases} \left(\dfrac{\lambda_1}{n_1\mu_1}\right)^{q_1} s_i(c,\lambda_1,\mu_1), & i = 0,\dots,n_1. \\ \left(\dfrac{\lambda_1}{n_1\mu_1}\right)^{k_1-i} \big(s_{n_1}(c,\lambda_1,\mu_1)\Omega_{i-n_1}(c) & \\ \qquad -s_{n_1-1}(c,\lambda_1,\mu_1)\Omega_{i-n_1-1}(c)\big), & i = n_1,\dots,k_1. \end{cases} \tag{3.5.6}$$

Finally, together with (3.5.2), a solution of the balance equations (3.5.1) is given by

$$p_{i,j} = \begin{cases} \displaystyle\sum_{m=0}^{k_2} a_m\alpha_i(c_m)s_j(-c_m,\lambda_2,\mu_2)\phi_{n_2}(c_m), & j = 0,\dots,n_2, \\ \displaystyle\sum_{m=0}^{k_2} a_m\alpha_i(c_m)s_{n_2}(-c_m,\lambda_2,\mu_2)\phi_j(c_m), & j = n_2,\dots,k_2, \end{cases} \tag{3.5.7}$$

for $i = 0,\dots,k_1$. The boundary conditions at $i = k_1$ in (3.5.1) are

$$\begin{aligned}(p\lambda_1 + \lambda_2 + n_1\mu_1 + j\mu_2)p_{k_1,j} &= \lambda_1 p_{k_1-1,j} \\ &\quad + (1-\delta_{j0})(p\lambda_1 + \lambda_2)p_{k_1,j-1}\end{aligned}$$

$$
\begin{aligned}
&\qquad + (j+1)\mu_2 p_{k_1,j+1}, \quad j = 0, \ldots, n_2 - 1, \qquad (3.5.8)\\
(\lambda_2(1-\delta_{q_2 0}) + n_1\mu_1 + n_2\mu_2)p_{k_1,n_2} &= \lambda_1 p_{k_1-1,n_2} + (1-\delta_{q_2 0})n_2\mu_2 p_{k_1,n_2+1}\\
&\quad + (p\lambda_1 + \lambda_2)p_{k_1,n_2-1}, \quad j = n_2, \qquad (3.5.9)\\
(\lambda_2(1-\delta_{jk_2}) + n_1\mu_1 + n_2\mu_2)p_{k_1,j} &= \lambda_1 p_{k_1-1,j} + (1-\delta_{jk_1})n_2\mu_2 p_{k_1,j+1}\\
&\quad + \lambda_2 p_{k_1,j-1}, \quad j = n_2+1, \ldots, k_2. \qquad (3.5.10)
\end{aligned}
$$

Substituting (3.5.7) into (3.5.8) and (3.5.10) gives

$$
\begin{aligned}
&\sum_{m=0}^{k_2} a_m \Big[c_m s_j(-c_m)\big(s_{n_1}(c_m+\mu_1)\Omega_{q_1}(c_m) - s_{n_1-1}(c_m+\mu_1)\Omega_{q_1-1}(c_m)\big)\\
&+ p\lambda_1 s_j(-c_m-\mu_2)\big(s_{n_1}(c_m)\Omega_{q_1}(c_m) - s_{n_1-1}(c_m)\Omega_{q_1-1}(c_m)\big)\Big]\phi_{n_2}(c_m) = 0 \qquad (3.5.11)
\end{aligned}
$$

for $j = 0, \ldots, n_2 - 1$ and

$$
\sum_{m=0}^{k_2} a_m c_m s_{n_2}(-c_m)\big(s_{n_1}(c_m+\mu_1)\Omega_{q_1}(c_m) - s_{n_1-1}(c_m+\mu_1)\Omega_{q_1-1}(c_m)\big)\phi_j(c_m) \qquad (3.5.12)
$$

for $j = n_2 + 1, \ldots, k_2$. By summing (3.5.11) and (3.5.12) it is seen that (3.5.9) is redundant. The coefficients $a_0, \ldots, a_{k_2}$ are uniquely determined by (3.5.11), (3.5.12) and the normalization condition. This condition yields – after substituting (3.5.7) and simplifying – an explicit formula for a_0 in terms of the auxiliary functions. We arrive at the closed-form expression

$$
\begin{aligned}
a_0 = \Big(&\big(s_{n_1}(\mu_1,\lambda_1,\mu_1)\Omega_{q_1}(0) - s_{n_1-1}(\mu_1,\lambda_1,\mu_1)\Omega_{q_1-1}(0)\big)\\
&\times \big(s_{n_2}(\mu_2,\lambda_2,\mu_2)\Psi_{q_2}(0) - s_{n_2-1}(\mu_2,\lambda_2,\mu_2)\Psi_{q_2-1}(0)\big)\Big)^{-1}. \qquad (3.5.13)
\end{aligned}
$$

Theorem 3.5.1. *The unique nonnegative and normalized solution of the steady-state equations* (3.5.1) *is given by*

$$
p_{i,j} = \begin{cases} \displaystyle\sum_{m=0}^{k_2} a_m \alpha_i(c_m) s_j(-c_m,\lambda_2,\mu_2)\phi_{n_2}(c_m), & j = 0, \ldots, n_2,\\ \displaystyle\sum_{m=0}^{k_2} a_m \alpha_i(c_m) s_{n_2}(-c_m,\lambda_2,\mu_2)\phi_j(c_m), & j = n_2, \ldots, k_2, \end{cases}
$$

for $i = 0, \ldots, k_1$ *if the coefficients* $a_0, \ldots, a_{k_2}$ *are determined by* (3.5.11), (3.5.12) *and* (3.5.13). $c_0 = 0$ *and* $c_1, \ldots, c_{k_2}$ *are the by Theorem 3.3.6 positive and pairwise distinct solutions of* (3.3.17).

The number of unknowns has now been reduced from $(k_1+1)(k_2+1)$ to $2k_2+1$.

3.6 Overflow with waiting rooms: From and to waiting room

3.6.1 Jockeying to servers and to waiting room

In this section, we consider the deterministic overflow models W/W/S and W/W/W, i.e., we choose $p = 1$, and show exemplarily for model W/W/S that the technique used throughout the previous sections does not succeed for these models. Similar observations were made in Section 3.4.1 for models S/W/S and S/W/N. However, it is possible to treat the case W/W/N, because in this case certain stationary probabilities and inconsistencies at the boundaries of the balance equations vanish. This is done in the next section. Numerical methods can be found in Chan [9] and Kaufman [35] for model W/W/S and in Chan [8] and Kaufman [35] for model W/W/N

The number of servers and the waiting room capacity of Q_i is n_i and $q_i \geq 1$, respectively, for $i = 1, 2$. In these models with blocking rule W, an arriving Q_1-customer is blocked and directed to Q_2 if all n_1 servers are busy and all waiting positions in Q_1 are occupied. Blocked Q_1-customers are treated with respect to overflow routine W, i.e., they are served by one of the servers in Q_2 if at least one is available, are queued in Q_2 if all Q_2-servers are busy and at least one waiting position is available and are lost otherwise. According to jockeying discipline S, waiting Q_1-customers are forced to move to Q_2 as soon as a server becomes available in Q_2. The balance equations for model W/W/S are given by

$$\begin{aligned}
&(\lambda_1(1-\delta_{ik_1}\delta_{jk_2}) + \lambda_2(1-\delta_{jk_2}) + (i\wedge n_1)\mu_1 + (j\wedge n_2)\mu_2)p_{i,j} \\
&\quad = (1-\chi_{i-n_1-1}\chi_{n_2-1-j})(\lambda_1(1-\delta_{i0})p_{i-1,j}) \\
&\qquad + (1-\delta_{jk_2})((j+1)\wedge n_2)\mu_2 p_{i,j+1} \\
&\qquad + (1-\delta_{j0})\big(\lambda_1(\delta_{in_1}\chi_{n_2-j} + \delta_{ik_1}\chi_{j-n_2-1}) \\
&\qquad + \lambda_2(1-\chi_{i-n_1-1}\chi_{n_2-j})\big)p_{i,j-1} \\
&\qquad + (1-\delta_{ik_1})\big((1-\chi_{i-n_1}\chi_{n_2-j-1})((i+1)\wedge n_1)\mu_1 \\
&\qquad\qquad\qquad + n_2\mu_2\delta_{jn_2}\chi_{i-n_1}\big)p_{i+1,j}
\end{aligned} \tag{3.6.1}$$

for $i = 0, \ldots, k_1$ and $j = 0, \ldots, k_2$. Note that $p_{i,j} = 0$ for $i = n_1 + 1, \ldots, k_1$ and $j = 0, \ldots, n_2 - 1$. For $i = n_1 + 1, \ldots, k_1$ and $j = n_2 + 1, \ldots, k_2$ these equation reduce to

$$\begin{aligned}(\lambda_1(1 - \delta_{ik_1}\delta_{jk_2}) + \lambda_2(1 - \delta_{jk_2}) + n_1\mu_1 + n_2\mu_2)p_{i,j} \\ = \lambda_1 p_{i-1,j} + (1 - \delta_{jk_2})n_2\mu_2 p_{i,j+1} \\ + (\lambda_1\delta_{ik_1} + \lambda_2)p_{i,j-1} + (1 - \delta_{ik_1})n_1\mu_1 p_{i+1,j}. \quad (3.6.2)\end{aligned}$$

Separating the probabilities in this region into $p_{i,j} = \alpha_i\beta_j$ gives the equations

$$\begin{aligned}(\lambda_1(1 - \delta_{ik_1}) + n_1\mu_1 + n_2\mu_2 - \frac{\beta_{k_2-1}}{\beta_{k_2}}(\lambda_1\delta_{ik_1} + \lambda_2))\alpha_i \\ = \lambda_1\alpha_{i-1} + (1 - \delta_{ik_1})n_1\mu_1\alpha_{i+1} \quad (3.6.3)\end{aligned}$$

for $i = n_1 + 1, \ldots, k_1$ and

$$\begin{aligned}\big((\lambda_1 + \lambda_2)(1 - \delta_{jk_2}) + n_1\mu_1 + n_2\mu_2 - \frac{\alpha_{k_1-1}}{\alpha_{k_1}}\lambda_1\big)\beta_j \\ = (\lambda_1 + \lambda_2)\beta_{j-1} + (1 - \delta_{jk_2})n_2\mu_2\beta_{j+1} \quad (3.6.4)\end{aligned}$$

for $j = n_2 + 1, \ldots, k_2$. As in Section 3.4.1 it is seen that the separation approach leads to different separation constants in (3.6.3) and (3.6.4) so that a solution of (3.6.3) cannot be expressed in terms of a solution of (3.6.4). Instead of solving these equations simultaneously, different eigenproblems have to be solved and consequently additional coefficients arise from the boundary conditions. Furthermore, it is necessary to match the solutions at the boundaries of the separation regions. This results in additional conditions for the sets of eigenvalues.

3.6.2 No jockeying

We apply a separation approach to determine the equations for the reduced system of balance equations for model W/W/N with p-overflow, $p \in [0, 1]$, and arbitrary service rates $\mu_1, \mu_2 > 0$ in this section. In this model, overflow is allowed if and only if all servers and waiting positions in Q_1 are occupied. The blocked customers overflow to Q_2 and are queued in the waiting room or served by a server in Q_2 if at least one is available. Waiting Q_1-customers have to wait for service at Q_1. The deterministic model, i.e., the case $p = 1$, with arbitrary service rates is also treated in Chan [8] and Kaufman [35] with numerical methods. The balance equations for model W/W/N are

$$\begin{aligned}(\lambda_1(1-\delta_{ik_1}) &+ p\lambda_1\delta_{ik_1}(1-\delta_{jk_2}) + \lambda_2(1-\delta_{jk_2}) \\ &+ (i\wedge n_1)\mu_1 + (j\wedge n_2)\mu_2)p_{i,j} \\ &= \lambda_1(1-\delta_{i0})p_{i-1,j} + (1-\delta_{ik_1})((i+1)\wedge n_1)\mu_1 p_{i+1,j} \\ &\quad + (1-\delta_{j0})(p\lambda_1\delta_{ik_1}+\lambda_2)p_{i,j-1} + (1-\delta_{jk_2})((j+1)\wedge n_2)\mu_2 p_{i,j+1}\end{aligned} \tag{3.6.5}$$

or $i = 0,\dots,k_1$ and $j = 0,\dots,k_2$. The separation approach for model W/S/N can be used for this model. A schematic overview of this approach is given in Figure 3.9. For model W/W/N the redundant boundary equation is located at the state (k_1,k_2).

Separating the variables into $p_{i,j} = \alpha_i\beta_j$ for $i \neq k_1$ and $j = 0,\dots,k_2$ leads analogously to the derivations in Section 3.5.2 to

$$p_{i,j} = \begin{cases} \displaystyle\sum_{m=0}^{k_2} a_m\alpha_i(c_m)s_j(-c_m,\lambda_2,\mu_2)\phi_{n_2}(c_m), & j=0,\dots,n_2, \\ \displaystyle\sum_{m=0}^{k_2} a_m\alpha_i(c_m)s_{n_2}(-c_m,\lambda_2,\mu_2)\phi_j(c_m), & j=n_2,\dots,k_2, \end{cases} \tag{3.6.6}$$

for $i = 0,\dots,k_1$ with

$$\alpha_i(c) = \begin{cases} \left(\dfrac{\lambda_1}{n_1\mu_1}\right)^{q_1} s_i(c,\lambda_1,\mu_1), & i=0,\dots,n_1, \\ \left(\dfrac{\lambda_1}{n_1\mu_1}\right)^{k_1-i} \big(s_{n_1}(c,\lambda_1,\mu_1)\Omega_{i-n_1}(c) & \\ \qquad -s_{n_1-1}(c,\lambda_1,\mu_1)\Omega_{i-n_1-1}(c)\big), & i=n_1,\dots,k_1, \end{cases} \tag{3.6.7}$$

for $c \in \mathbb{R}$. The boundary conditions at $i = k_1$ are

$$\begin{aligned}(p\lambda_1+\lambda_2+n_1\mu_1+j\mu_2)p_{k_1,j} &= \lambda_1 p_{k_1-1,j} + (j+1)\mu_2 p_{k_1,j+1} \\ &\quad + (1-\delta_{j0})(p\lambda_1+\lambda_2)p_{k_1,j-1}, \\ &\quad j=0,\dots,n_2-1, \qquad (3.6.8) \\ \big((p\lambda_1+\lambda_2)(1-\delta_{q_20})+n_1\mu_1+n_2\mu_2\big)p_{k_1,n_2} &= \lambda_1 p_{k_1-1,n_2} \\ &\quad + (1-\delta_{q_20})n_2\mu_2 p_{k_1,n_2+1} \\ &\quad + (p\lambda_1+\lambda_2)p_{k_1,n_2-1}, \\ &\quad j=n_2, \qquad (3.6.9)\end{aligned}$$

$$\begin{aligned}(p\lambda_1+\lambda_2+n_1\mu_1+n_2\mu_2)p_{k_1,j} &= \lambda_1 p_{k_1-1,j}+n_2\mu_2 p_{k_1,j+1}\\ &\quad+(p\lambda_1+\lambda_2)p_{k_1,j-1}, \quad (3.6.10)\\ &\quad j=n_2+1,\dots,k_2-1,\\ (n_1\mu_1+n_2\mu_2)p_{k_1,k_2} &= \lambda_1 p_{k_1-1,k_2}\\ &\quad+(p\lambda_1+\lambda_2)p_{k_1,k_2-1},\\ &\quad j=k_2. \qquad (3.6.11)\end{aligned}$$

Substituting (3.6.6) into (3.6.8), (3.6.9) and (3.6.10) yield

$$\begin{aligned}&\sum_{m=0}^{k_2} a_m\Big[c_m s_j(-c_m)\big(s_{n_1}(c_m+\mu_1)\Omega_{q_1}(c_m)-s_{n_1-1}(c_m+\mu_1)\Omega_{q_1-1}(c_m)\big)\\ &+p\lambda_1 s_j(-c_m-\mu_2)\big(s_{n_1}(c_m)\Omega_{q_1}(c_m)-s_{n_1-1}(c_m)\Omega_{q_1-1}(c_m)\big)\Big]\phi_{n_2}(c_m)=0\end{aligned} \tag{3.6.12}$$

for $j=0,\dots,n_2-1$ (see also (3.5.11)) and

$$\begin{aligned}&\sum_{m=0}^{k_2} a_m\Big[c_m s_{n_2}(-c_m)\big(s_{n_1}(c_m+\mu_1)\Omega_{q_1}(c_m)\\ &\quad-s_{n_1-1}(c_m+\mu_1)\Omega_{q_1-1}(c_m)\big)\phi_j(c_m)+p\lambda_1 s_{n_2}(-c_m)\big(s_{n_1}(c_m)\Omega_{q_1}(c_m)\\ &\qquad-s_{n_1-1}(c_m)\Omega_{q_1-1}(c_m)\big)\big(\phi_j(c_m)-\phi_{j-1}(c_m)\big)\Big]=0\end{aligned} \tag{3.6.13}$$

for $j=n_2,\dots,k_2-1$. By summing (3.6.12) and (3.6.13) it is seen that (3.6.11) is redundant. (3.6.12), (3.6.13) and the normalization condition determine the coefficients $a_0,\dots,a_{k_2}$. Substituting (3.6.6) into the normalization condition and simplifying gives an explicit formula for a_0 in terms of the auxiliary functions. We arrive at the closed-form expression

$$\begin{aligned}a_0=\Big(&\big(s_{n_1}(\mu_1,\lambda_1,\mu_1)\Omega_{q_1}(0)-s_{n_1-1}(\mu_1,\lambda_1,\mu_1)\Omega_{q_1-1}(0)\big)\\ &\times\big(s_{n_2}(\mu_2,\lambda_2,\mu_2)\Psi_{q_2}(0)-s_{n_2-1}(\mu_2,\lambda_2,\mu_2)\Psi_{q_2-1}(0)\big)\Big)^{-1}\end{aligned} \tag{3.6.14}$$

as in (3.5.13).

By the approach above, the number of unknowns reduces from $(k_1+1)(k_2+1)$ to $2k_2+1$. The results in the case of blocking if all waiting positions are occupied, no jockeying and overflow to the waiting rooms of the second queue is summarized in the following theorem.

Theorem 3.6.1. *The unique nonnegative and normalized solution of the* *steady-state equations* (3.6.5) *is given by*

$$p_{i,j} = \begin{cases} \displaystyle\sum_{m=0}^{k_2} a_m \alpha_i(c_m) s_j(-c_m, \lambda_2, \mu_2) \phi_{n_2}(c_m), & j = 0, \dots, n_2, \\ \displaystyle\sum_{m=0}^{k_2} a_m \alpha_i(c_m) s_{n_2}(-c_m, \lambda_2, \mu_2) \phi_j(c_m), & j = n_2, \dots, k_2, \end{cases}$$

for $i = 0, \dots, k_1$ *if the coefficients* $a_0, \dots, a_{k_2}$ *are determined by* (3.6.12), (3.6.13) *and* (3.6.14). $c_0 = 0$ *and* $c_1, \dots, c_{k_2}$ *are the by Theorem 3.3.6 positive and pairwise distinct solutions of* (3.3.17).

List of Figures

List of Tables

Symbols

$A(z)$	matrix defined in (2.2.23) or (2.3.17)
A_0	rate matrix for transitions from level $l(m)$ to level $l(m+1)$
A_1	rate matrix for transitions from level $l(m)$ to level $l(m)$
A_2	rate matrix for transitions from level $l(m)$ to level $l(m-1)$
$A_n(z)$	matrix obtained from $A(z)$ in Section 2.2 by replacing the $(n+1)$-th column by the vector $-\mu_2(1-z)p$ or in Section 2.3 by replacing the n-th column with the vector $-(1-z)p$
B	rate matrix for transitions from level $l(0)$ to level $l(0)$
B_i	probability that an arriving Q_i-customer is blocked at Q_i, $i = 1, 2$
χ_{i-n}	defined as 1 for $i \geq n$ and 0 otherwise
$\chi_n(t)$	$= \det(tE_n - A_n)$, characteristic polynomial of A_n in $t \in \mathbb{R}$
δ_{ij}	Kronecker function, defined by 1 for $i = j$ and 0 otherwise
diag	diagonal matrix
E_n	identity $(n \times n)$-matrix
$i \wedge j$	minimum of i and j
k_i	$= n_i + q_i$, total number of positions in Q_i, $i = 1, 2$
L_1	phase of a quasi birth and death chain, (stationary) queue length of Q_1
L_2	level of a quasi birth and death chain, (stationary) queue length of Q_2
Λ	intensity of the interoverflow process
λ_1	arrival rate for Q_1

$\lambda_{1,n}$	service rate for Q_1 if Q_1 is in state n
λ_2	service rate for Q_2
$l(m)$	$= \{0, \dots, N\} \times \{m\}$, level of a quasi birth and death chain
$\mathrm{Mat}(n, n, C)$	$(n \times n)$-matrix with entries in the set C
M_n	$= M_n(\lambda, \mu)$, matrix defined in (3.2.25)
μ_1	service rate for Q_1
$\mu_{1,n}$	service rate for Q_1 if Q_1 is in state n
μ_2	service rate for Q_2
N	capacity of Q_1 in Chapter 2
$\mathbb{N}_0$	set of the nonnegative integers
n_i	number of servers in Q_i, $i = 1, 2$
N	in $\alpha/\beta/$N: indicator for "no jockeying", only in Chapter 3
O_{12}	expected (stationary) number of demands, which flow over from Q_1 to Q_2
Ω_l	defined in (3.3.11)
ϕ_j	defined in (3.3.7)
$\Pi_{r,j}$	defined in (3.3.54)
$P_{\mathrm{Loss},i}$	probability that an arriving Q_i-customer is lost, $i = 1, 2$
$p_{n,m}$	steady-state probability for $L_1 = n$ and $L_2 = m$
$P_{\mathrm{queue},i}$	probability that an arriving Q_i-customer is queued in the waiting room in Q_i, $i = 1, 2$
Ψ_l	defined in (3.3.8)
Q	$= (q_{i,j})_{i,j \in \mathfrak{S}}$, infinitesimal generator, rate matrix
Q_1	first queue
Q_2	second queue
q_i	number of waiting positions in Q_i, $i = 1, 2$
$\mathbb{R}$	set of the real numbers

R_{ij} mean departure rate from the waiting room in Q_i to the servers in Q_j

$s_i(c, \lambda, \mu)$ $= s_i(c)$, solution of (3.2.6) with $s_0(c, \lambda, \mu) = 1$

S in S/β/γ: indicator for "blocking if servers occupied", in α/S/γ: indicator for "overflow to servers", in α/β/S: indicator for "jockeying to servers"

$\mathfrak{S}$ state space of a Markov chain

θ_i defined in (3.3.10)

$u_i(c, \lambda, \mu)$ $= u_i(c)$, defined in (3.2.23)

U_l l-th Chebyshev polynomial of the second kind

$v_i(c)$ defined in (3.3.19)

W in W/β/γ: indicator for "blocking if waiting room occupied", in α/W/γ: indicator for "overflow to waiting room", in α/β/W: indicator for "jockeying to waiting room"

Index

Bibliography

[1] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, Wiley, New York, 1972

[2] E. Altman, T. Jiménez, R. Núñez-Queija and U. Yechiali, *Optimal routing among* $\cdot/M/1$ *queues with partial information*, Stoch. Models 20, 149-171, 2004

[3] G. Alsmeyer, *Erneuerungstheorie. Analyse stochastischer Regenerationsschemata*, Teubner Skripten zur Mathematischen Stochastik, Stuttgart, 1991

[4] S. Asmussen, *Matrix-analytic models and their analysis*, Scand. J. Stat. 27, 193-226, 2000

[5] S. Asmussen, *Applied probability and queues*, 2. ed., Springer, New York, 2003

[6] B. Avi-Itzhak and I. L. Mitrani, *A many-server queue with server interruptions*, Oper. Res. 16, 628-639, 1968

[7] P. Bremaud, *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*, Texts in applied mathematics 31, Springer, New York, 1998

[8] R. H. Chan, *Iterative methods for overflow queueing models I*, Numer. Math. 51, 143-180, 1987

[9] R. H. Chan, *Iterative methods for overflow queueing models II*, Numer. Math. 54, 57-78, 1988

[10] W. Ching and M. K. Ng, *Markov chains: Models, algorithms and applications*, Springer, New York, 2006

[11] J. W. Cohen, *Sensitivity and insensitivity*, Delft Prog. Rep. 5, 159-173, 1980

[12] J. W. Cohen, *The single server queue*, North-Holland Publ. Co., Amsterdam, 1982

[13] R. B. Cooper, *Introduction to queueing theory*, 2. ed., Edward Arnold, London, 1981

[14] J. K. Cullum and R. A. Willoughby, *Lanczos algorithms for large symmetric eigenvalue computations, Vol. 1: Theory, Progress in scientific computing 3*, Birkhäuser, Boston, 1985

[15] N. M. van Dijk, *Simple and insensitive bounds for a grading and an overflow model*, Oper. Res. Letters 6, 73-76, 1987

[16] N. M. van Dijk, *A proof of simple insensitive bounds for a pure overflow system*, J. Appl. Prob. 26, 113-120, 1989

[17] R. L. Disney and D. König, *Queueing networks: a survey of their random processes*, SIAM Review 27, 335-403, 1984

[18] J. van Doremalen, *Two parallel queues with one way overflow - a matrix structure approach*, in Operations research Proc. (Ohse et al, ed), Springer, Berlin, 563-570, 1984

[19] E. A. van Doorn, *On the overflow process from a finite markovian queue*, Perf. Evalf. 4, 233-240, 1984

[20] M. El-Taha and J. R. Heath, *Traffic overflow in loss systems with selective trunk reservation*, Perf. Eval. 41, 295-306, 2000

[21] T. O. Engset, *On the calculation of switches in an automatic telephone system*, unpublished report, translated to English by A. Myskja in Espvik and Myskja [25], 1915

[22] T. O. Engset, *Die Wahrscheinlichkeitrechnung zur Bestimmung der Wählerzahl in automatischen Fernsprechämtern*, Elektrotechnische Zeitschrift 31, 1918.

[23] A. K. Erlang, *The theory of probabilities and telephone conversations*, Nyt Tidsskrift for Matematik B, 20, 33-39, 1909

[24] A. K. Erlang, *Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges*, Elektroteknikeren 13, 5-22, 1917

[25] O. Espvik and A. Myskja, *Telektronikk and the History of Engset*, Kjeller, Telenor R&D, 78/2000, 2000

[26] G. H. Golub and C. F. van Loan, *Matrix computations*, 2. ed., Johns Hopkins series in the mathematical sciences 3, Hopkins Univ. Press, Baltimore Md., 1991

[27] R. Guérin and L. Y. C. Lien, *Overflow analysis for finite waiting room systems*, IEEE Transactions on Communications 38, 1569-1577, 1990

[28] R. Hassin, *On the advantage of being the first server*, Manage. Sci. 42, 618-623, 1996

[29] A. Hordijk and N. M. van Dijk, *Networks of queues. Part I: Job-local-balance and the adjoint process. Part II: General routing and service characteristics*, Modelling and performance evaluation methodology, Proc. Int. Semin., Paris, Lect. Notes Control Inf. Sci. 60, 151-205, 1983

[30] A. Hordijk and A. Ridder, *Stochastic inequalities for an overflow model*, J. Appl. Prob. 24, 696-708, 1987

[31] A. Hordijk and A. Ridder, *Insensitive bounds for the stationary distribution of non-reversible Markov chains*, J. Appl. Prob. 25, 9-20, 1988

[32] ITU, *Handbook Teletraffic Engineering*, ITUD Study group 2, Question 16/2, Geneve, 2005

[33] F. Johannsen, *Busy*, Ingeniørvitenskapelige Skrifter A 32, 1908

[34] L. Kaufman, J. A. Morrison and J. B. Seery, *Overflow models for Dimension PBX feature packages*, Bell Syst. Tech. J. 60, 661-676, 1981

[35] L. Kaufman, *Matrix methods for queueing problems*, SIAM J. Sci. Stat. Comput. 4, 525-552, 1983

[36] L. Kosten, *Behaviour of overflow traffic and the probabilities of blocking in simple gradings*, International Teletraffic Congress: ITC 8, Melbourne, Vol. 2, 425-425, 1976

[37] L. Kleinrock, *Queueing systems, Vol. 1: Theory*, Wiley-Interscience publications, Wiley, 1975

[38] L. Kosten, *Stochastic theory of service systems*, International series of monographs in pure and applied mathematics 103, Pergamon Press, Oxford, 1973

[39] J. R. Koury, D. F. McAllister and W. J. Stewart, *Iterative methods for computing stationary distributions of nearly decomposable Markov chains*, SIAM J. Alg. Disc. Meth. 5, 164-186, 1984

[40] U. R. Krieger, B. Müller-Clostermann and M. Sczittnick, *Modeling and analysis of communication systems based on computational methods for Markov chains*, IEEE Journal on selected areas in communications 8, 1630-1648, 1990

[41] G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modeling*, ASA, Alexandria, 1999

[42] Y. Levy and U. Yechiali, *An M/M/s Queue with servers vacations*, Can. J. Oper. Res. Inf. Process. 14, 153-163, 1976

[43] S. R. Mahabhashyam and N. Gautam, *On queues with Markov modulated service rates*, Queueing Syst. 51, 89-113, 2005

[44] E. W. van Marion, *Influence of holding time distributions on blocking probabilities of a grading*, TELE 20, 17-12, 1968

[45] A. M. Matsekh, *The Godunov-inverse iteration: A fast and accurate solution to the symmetric tridiagonal eigenvalue problem*, Appl. Numer. Math. 54, 208-221, 2005

[46] J. A. Morrison, *Analysis of some overflow problems with queueing*, Bell Syst. Tech. Journal 59, 1427-1462, 1980

[47] J. A. Morrison, *Some traffic overflow problems with a large secondary queue*, Bell Syst. Tech. Journal 59, 1463-1482, 1980

[48] J. A. Morrison, *An overflow system in which queueing takes precedence*, Bell Syst. Tech. Journal 60, 1-12, 1981

[49] J. A. Morrison and P. E. Wright, *A traffic overflow system with a large primary queue*, Bell Syst. Tech. Journal 61, 1487-1517, 1982

[50] L. Muscariello, M. Mellia, M. Meo, M. A. Marsan and R. Lo Cigno, *Markov models of internet traffic and a new hierarchical MMPP model*, Comp. Comm. 28, 1835-1851, 2005

[51] M. F. Neuts, *Computer power or the liberation of applied probability*, Technical Report 312, Department of Statistics, Purdue University, West Lafayette, IN, 1973

[52] M. F. Neuts, *Probability distributions of phase type*, In Liber Amicorum Prof. Emeritus H. Florin, University of Louvain, Belgium, 173-206, 1975

[53] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models - An Algorithmic Approach*, Johns Hopkins, Baltimore, 1981

[54] R. Núñez-Queija and O. J. Boxma, *Analysis of a multi-server queueing model of ABR*, J. Appl. Math. Stoch. Anal. 11, 339-354, 1998

[55] P. R. Parthasarathy and R. Sudhesh, *The overow process from a state-dependent queue*, Int. J. Comput. Math. 82, 1073-1093, 2005

[56] E. Perel and U. Yechiali, *Queues where customers of one queue act as servers of the other queue*, Queueing Syst. 60, 271-288, 2008

[57] A. Ridder, *A note on insensitive bounds for a grading*, Research report, University of Leiden, 1986

[58] A. Ridder, *Stochastic inequalities for queues*, PhD thesis, University of Leiden, 1987

[59] P. Sendfeld, *Two queues with weighted one-way overflow*, Method. Comp. Appl. Prob. 10, 531-555, 2008

[60] H. M. Srivastava and B. R. K. Kashyap, *Special functions in queueing theory and related stochastic processes*, Academic Press, New York, 1982

[61] K. Stordahl, *The history behind the probability theory and the queuing theory*, Teletronikk 2, 123-140, 2007

[62] P. N. Swarztrauber, *A parallel algorithm for computing the eigenvalues of a symmetric tridiagonal matrix*, Math. Comput. 60, No. 202, 651-668, 1993

[63] R. Syski, *Introduction to congestion theory in telephone systems*, 2. ed., Studies in telecommunication 4, North-Holland, Amsterdam, 1986

[64] T. Takine, *Single-server queues with Markov-modulated arrivals and service speed*, Queueing Syst. 49, 7-22, 2005

[65] J. H. Wilkinson, *The algebraic eigenvalue problem*, Monographs on numerical analysis, Clarendon Press, Oxford, 1988

[66] U. Yechiali, *A queueing-type birth-and-death process defined on a continuous Markov chain*, Oper. Res. 21, 604-609, 1973