Roland de Haan

Queueing Models for Mobile Ad Hoc Networks D120

Queueing Models for Mobile Ad Hoc Networks

University of Twente
Beta dissertation series D120
ISBN 978-90-365-2827-6

Roland de Haan

Queueing Models for Mobile Ad Hoc Networks

by

Roland de Haan

PhD dissertation committee:

prof.dr. J.L. van den Berg	(Universiteit Twente)
prof.dr. R.J. Boucherie	(Universiteit Twente)
prof.dr.ir. O.J. Boxma	(Technische Universiteit Eindhoven)
prof.dr. N.M. van Dijk	(Universiteit van Amsterdam)
prof.dr.ing. D. Fiems	(Universiteit Gent)
dr. J.C.W. van Ommeren	(Universiteit Twente)
prof.dr. A.A. Stoorvogel	(Universiteit Twente)

UT / EEMCS / AM / DMMP
P.O. Box 217, 7500 AE Enschede
The Netherlands

Centre for Telematics and Information Technology
CTIT PhD Thesis Series 09-143

BETA, Research School for Operations Management and Logistics.
Beta Dissertation Series D120

The research in this thesis is financially supported by Easy Wireless - Ministry of Economic Affairs, Department of Commerce, under Grant IS043014.

E-Quality, Expertise Centre on Performance and Quality of Service in ICT

This thesis was edited with WinEdt and typeset with LaTeX.
Printed by Wöhrmann Print Service, Zutphen, The Netherlands.

ISBN 978-90-365-2827-6 ISSN 1381-3617
http://dx.doi.org/10.3990/1.9789036528276

QUEUEING MODELS
FOR
MOBILE AD HOC NETWORKS

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof.dr. H. Brinksma,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op donderdag 4 juni 2009 om 13.15 uur

door

Roland de Haan

geboren op 25 oktober 1980
te Geldermalsen

Dit proefschrift is goedgekeurd door:

prof.dr. R.J. Boucherie	(promotor)
dr. J.C.W. van Ommeren	(assistent-promotor)

Acknowledgements

This monograph embodies about four years of research performed at the University of Twente. Commonly, monographs treat not only a single matter or subject, but also refer to something written by a single author. Although I am indeed the person that eventually put all these words down on paper, the selection of which words to put and at which position appears a much more crucial aspect in the entire process. The realization of this selection is definitely not an effort carried out all by myself. Therefore, I would like to thank here a number of people that were indispensable in the development of this booklet.

First of all, I am greatly indebted to my promoter Richard Boucherie for giving me the opportunity to pursue a PhD degree in the Stochastic Operations Research (SOR) group at all, the overwhelming number of ideas generated at each discussion, and his discerning attitude. Also, I owe many thanks to my assistant-promoter Jan-Kees van Ommeren for our constructive collaboration, his infinite patience and the many enjoyable chats. Besides, I would like to express my gratitude to the other members of the SOR group for the warm research environment. In particular, thanks to Yana for the pleasant company during all these years and the assistance in the final preparation of this monograph. I want also to thank Ahmad for the many interesting discussions and the fruitful cooperation, part of which can be found in this monograph. Then, thanks to Michela for her cheerful company and offering me the opportunity to fine-tune my Italian during her internship period: *Grazie mille!*

Also, I would like to mention a number of people outside the work environment who have been important for me during these past four years in Enschede. I have really appreciated the warm atmosphere of the triathlon club D.S.T.V. Aloha, so

thanks to all of you guys! In particular, I should mention Pieter Vernooij for making the “zaterdagochtendsjoktochten” bearable (and performing these at all!) and his everlasting competitive behavior. Also, thanks to Karin and Dannis for showing me that there is more in life than just swimming and running, namely cycling. Further, I want to thank the rest of my friends and family for their curiosity and interest in my work, their support and their willingness to travel regularly all the way to Enschede. Finally, I am mostly indebted to my parents for their unconditional love, support, and interest.

Roland de Haan
Enschede, April 2009

CONTENTS

CHAPTER

1

Introduction

1.1 Motivation

Data communication networks exist in a myriad of flavors. Well-known examples of such are cable television networks, satellite networks and office networks. Networks are typically formed by connecting a number of computer systems in some fashion. The number of connected devices can be quite small (e.g., a home network), but also extremely large (e.g., the Internet). Traditionally, computer networks are mainly used for communication (e-mail), file exchange, or sharing peripherals (e.g., a common printer in an office environment). For a long period of time these networks have primarily been wireline networks, but the last decade wireless networks have been introduced universally and have proven to be a prosperous communication medium. These networks have extended the applications for data communication enormously. Initially, all wireless communication took place between users (i.e., a notebook, PDA or cellular phone) and a base station (which grants the users access to other networks such as the Internet), even when two users wanted to communicate directly. However, the wireless communication medium offers other opportunities for communication between two users. Emphasizing the broader applicability of the wireless medium, the term mobile ad hoc networks (MANETs) was introduced and recently these networks have attracted an interest both from a practical and from a theoretical point of view. We will next give a brief description on the operational aspects of data communication networks below as to illustrate the aspects in which MANETs

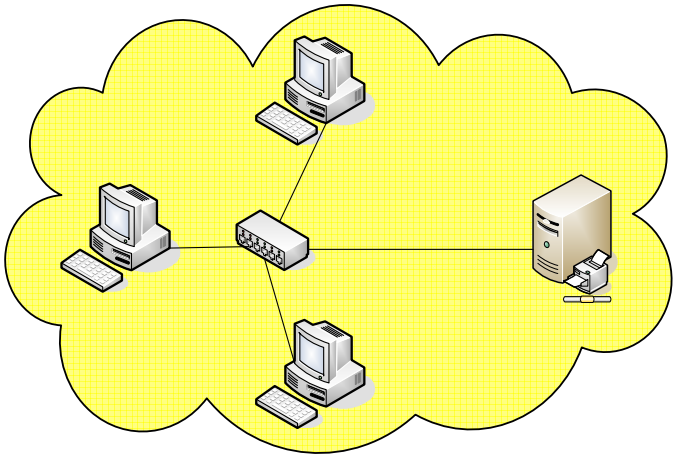

Figure 1.1: Communication in a classical wireline setting.

diverge from the other types of networks.

A wireline network consists of a set of computer systems that are connected via wired communication links. Such a network is shown in Fig. 1.1 in which the lines represent the wired connections between the different entities. Wireline networks allow for high-speed data communication in an error-free fashion. Stations located in the same network but not directly sharing a link may readily communicate via one or more routers. A disadvantage of these wireline networks is their inflexible and immobile character.

The introduction of wireless communication resolved some of these drawbacks. Classical wireless networks comprise several devices (e.g., PCs, notebooks) which are connected via the wireless medium to a base station (see Fig. 1.2). The base station provides connections to other networks (e.g., the Internet) often via a wireline network. Such a wireless connection allows a user to move within the communication range of a base station. Also, it allows for connecting to another base station nearby (and thus possibly to another network). However, the flexibility of wireless communication comes at the cost of lower data rates and an increase in transmission errors. The network is typically fully-connected meaning that all stations are aware of each other's presence and that only one transmission can take place successfully at a time. Although such networks are way more flexible than wireline networks, for many applications their flexibility is still too restricted.

Networks which go beyond this classical concept of wireless networks are the so-called mobile ad hoc networks. MANETs consist of mobile and fixed wireless stations and are characterized by the lack of infrastructure. In fact, wireless devices possess the power to create their own wireless network (also referred to as self-organizability property) in a distributional fashion. The term mobile in MANET

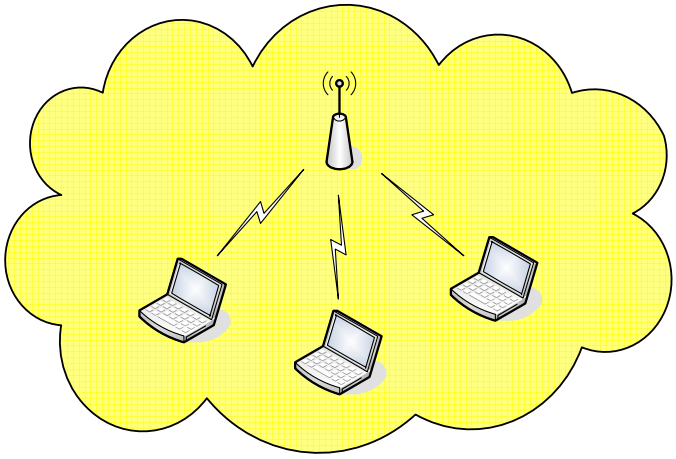

Figure 1.2: Communication in a classical wireless setting.

emphasizes the mobility of the users or devices in the network. Users may move and thereby change communication links in the network. A prerequisite of MANETs is that these networks should allow for multi-hop communication, while in the classical wireless concept mostly single-hop communication is used (from the base station to the user and vice versa). Also, the full-connectivity property (i.e., all stations hear any transmission in the network) is relaxed, so that also signal interference between different transmissions becomes an issue. Examples of such networks are animal-monitoring systems, collaborative conference computing, vehicular networks, peer-to-peer file-sharing and disaster-relief networks. Three of these examples will be highlighted in more detail next.

Example 1.1 (Animal-monitoring systems)**.** *Animal-monitoring systems (see, e.g., [53, 93]) monitor the nomadic behavior of groups of animals and individuals. Animals under research are equipped with small and simple communication means (e.g., a packet radio). Regularly or upon specific events, e.g., an encounter between two animals, (GPS) data is stored or exchanged. Researchers periodically collect the data and draw conclusions on the animal behavior.*

Clearly, the (inner) network is a mobile ad hoc network. A fixed infrastructure is lacking and communication between the stations (animals) occurs in a wireless fashion upon encounters. The frequency and duration of such encounters depends strongly on the mobility pattern of the animals present.

Example 1.2 (Vehicular networks)**.** *Vehicular networks (see, e.g., [10, 33]) are networks which are formed in a road-traffic situation. The mobile stations in such networks are the cars and trucks on the road. These vehicles can easily be equipped with communication equipment. Additionally, the network may comprise some static*

stations which can be road signs with a packet radio attached. Such a network is used to quickly distribute traffic information (on congestion, accidents, etc.), to improve the safety and comfort of drivers, but may also be used to offer in-car internet access.

Also, a vehicular network is a mobile ad hoc network. The network is infrastructureless and created on the fly by the vehicles on the road. Wireless communication links arise and disappear as vehicles move closer or farther apart. The link duration will depend on the mobility pattern of the objects. However, as vehicles can easily be equipped with larger, and thus also more powerful, radio equipment, the communication range is typically large and not so sensitive to small changes in the vehicles' positions.

Example 1.3 (Disaster-relief networks)**.** *Disaster-relief or emergency networks (see, e.g., [H4, H5]) come into play when a big disaster leads to the elimination of a complete communication infrastructure, such as the GSM network. Examples of such have been witnessed in the recent past during the bomb attack in the London metro and the hurricane Katrina (New Orleans, United States). To accommodate the rescue operation, it is of the utmost importance that rescue workers and operation leaders are still able to communicate. By equipping the rescue workers with appropriate light-weight communication radios and positioning static rescue equipment (e.g., fire engines) in strategic locations on the spot, a fully-operational network may quickly be deployed. Accordingly, relatively high bandwidth links can be created, so that besides voice traffic also data and video traffic can be sustained by the network.*

The mobile ad hoc network paradigm is the only feasible solution for communication in such a catastrophic situation. Although the static stations may be primarily used for coordination purposes, the mobile stations (i.e., rescue workers) should in fact supply the multi-hop connectivity between the stations and the quality of the links in the network. However, as the key focus of the rescue team is on casualty care and disaster relief in the first place, coordinating such an ad hoc network effectively at the same time becomes an extremely complex task (see Fig. 1.3 [H5]).

The organization of the remainder of this chapter is as follows. In Sect. 1.2, we discuss the main characteristics of MANETS, the related research challenges and we outline which challenges will be considered in this thesis. In Sect. 1.3, we explain the main concepts of polling systems, which emerge as a natural performance model for mobile ad hoc networking, and review the most relevant analytical results from the literature. We conclude this chapter in Sect. 1.4 with an outline of the thesis.

1.2 MANETs: characteristics and research issues

The specific MANET extensions with respect to the classical wireless concept broaden the set of applications over such networks considerably. From a practical point of view, it raises also many questions regarding the successful operability of such networks. For instance, regarding the mobility of the communication devices, it is clear

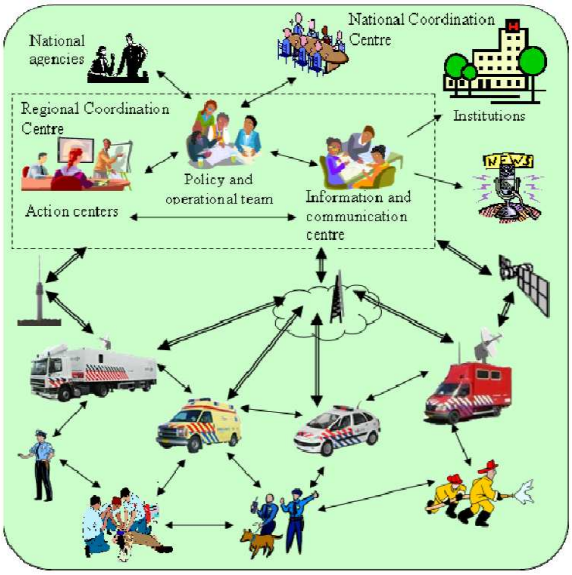

Figure 1.3: Communication in a mobile ad hoc network.

that on the one hand the wireless equipment should be small and light-weight, but on the other hand it should still be powerful enough for data transmission (see, e.g., Example 1.1). In vehicular networks (see Example 1.2), the dispersion rate of the information may be paramount to inform drivers about upcoming traffic jams and proposing alternative routes, while in a disaster-relief situation (see Example 1.3), the most relevant network properties will be robustness and stability. In that case, the strategically positioning of (static) rescue vehicles in the disaster area becomes essential, since it creates a kind of backbone for the mobile ad hoc network.

These practical implications trigger issues on an operational and technological level. Several issues for MANETs are inherited directly from the classical wireless setting, e.g., error-prone channels, sensitivity for security attacks, asymmetric channel conditions, and a low bandwidth. Moreover, the specific characteristics of MANETs induce a great number of novel problems on network performance. Before we come to those, we list these characteristics first (see, e.g., [21]).

- Energy constraints; wireless devices possess a battery with a limited lifetime. For instance, in animal-monitoring systems or sensor networks where batteries cannot simply be replaced this plays a crucial role.

- Lack of infrastructure; an ad hoc network is formed on the fly by stations in a local neighborhood, so it requires self-organizability of the wireless stations. Also, the network structure is decentralized in the sense that there is typically no central entity which controls the network and its traffic streams.

- Dynamically changing topology; the structure of the network varies over time. For instance, the breakdown of a centrally located station may lead to disconnectivity of the network. Similarly, a moving station may destroy existing communication links or create new ones. In any case, the network topology will change, possibly leading to packet drops and forcing stations to search for new routes.

- Multi-hop communication; end-to-end communication between two stations in the network may require traffic to cover multiple hops (a hop is here a single transmission over a wireless link). Contrary to the classical one-hop setting (see Fig. 1.2), wireless stations need also to operate as a relay or forwarding device.

- Ad hoc mode of operation; communication in MANETs is no longer necessarily between a base station and its users, but individual users can also communicate directly. Hence, within a local region multiple transmissions may take place simultaneously. This may cause interference problems leading to destruction of data packets or to a large decrease in transmission opportunities.

These characteristics lead to challenging issues both for network practitioners and for network researchers. The energy constraints create the need to manage the available energy (i.e., battery power) as efficiently as possible as to elongate the network life-time. Energy savings can be realized for instance by reducing transmission power or activating the sleep mode of a station more frequently. Excellent surveys on energy issues in MANETs can be found in [3, 52].

In the remainder of this thesis, we will focus on the issues departing from the last four listed characteristics and leave the energy issues untouched. These issues will be discussed on the basis of two sets of closely related performance measures, viz., on the one hand network capacity and stability and on the other hand transfer delays and buffer sizes.

Network capacity and stability The stability of a network is typically defined in terms of conditions for the amount of traffic offered to the system, while the network capacity in fact refers to this maximum amount of data traffic that can be sustained. Exceeding this amount of traffic leads to instability of (parts of) the network and is not desirable. Of course, network operators would like to push the usage of their network to its limit; however, a network operating continuously against its limits may yield poor performance for its users.

The capacity and stability of a wireless ad hoc network depend critically on the communication links that are available. In a single-channel environment, this unavailability may be caused by the interference of other, nearby transmissions. Such an environment restricts the number of transmissions that can take place simultaneously within a local region. However, transmissions that occur "sufficiently" distant from each other can be sustained together. These observations lead to interesting research

questions on the optimization of the number of simultaneous transmissions in a network. Through employing an adequate routing protocol (see [1] for a nice overview of routing protocols for MANETs) a station may be aware of nearby stations, though a station is typically not aware of the exact location of other stations and the intentions of these stations regarding data transmission. Thus, stations would autonomously and selfishly commence transmitting data which would readily lead to unsuccessful data reception at the receiver station due to interference of other transmissions. To alleviate this problem, distributed Medium Access Control protocols [60] are applied as to prevent unnecessary data-packet collisions to happen. Hence, in practice, the ad hoc mode of operation leads to a situation in which the available resources must be shared by several stations. Similarly, the dynamics of the network topology may yield capacity and stability problems. Stations that break down due to hardware failures may render the network temporarily disconnected and thus instable. Also, the mobility of the stations may lead to a time-varying availability of resources, so that capacity and stability issues are not readily resolved.

Regarding network capacity, it has been shown in the literature (see [45]) somewhat surprisingly that for dense wireless networks with mobile stations the capacity may in fact increase with respect to static wireless networks [46]. This was done for an asymptotically large number of stations with communication along a simple 2-hop relay scheme. For ad hoc networks with a finite population of stations, capacity questions have been studied by accounting specifically for multi-hop communication (see, e.g., [47, 51, 58]). On a more abstract level, stability questions have been addressed already a long time ago in the context of Jackson networks [50]. For such networks, the condition $\rho_i < 1, \forall_i$, where ρ_i is the offered traffic to station i, is a necessary and sufficient condition.

Part I of this thesis will be dedicated to the issues of capacity and stability.

Transfer delays and buffer sizes The time from generation of a data packet or file until it finally reaches its destination is referred to as the end-to-end transfer delay. The importance of the delay as a performance measure depends highly on the nature of the traffic, e.g., speech traffic is delay sensitive, while data traffic is not. The buffer size refers to the number of memory positions (typically in terms of packets) used by a station during network operation. This measure gives insight in the dimensioning of the buffers of the stations. Large buffers may relieve data-packet loss and lead to better delay figures for the users, but for the network operator these may also be quite costly.

The mobility of the stations puts restrictions on the size and the weight of communication equipment (see, e.g., [53]) and thus also puts bounds on the size of the buffer. Conversely, the multi-hop character of MANETs infers a relay function of the stations, such that larger buffers may be required. Regarding the transfer delays, mobility of the stations will increase the uncertainty in transmission times as individual links are not always available. Besides, as an end-to-end transmission consists in fact of multiple single-hop transmissions, the variability in the end-to-end delay

increases also significantly. Hence, the delay and buffer size measures may differ significantly under mobile ad hoc networking from the behavior under more established networking paradigms.

The original efforts on the network capacity in dense wireless networks [45, 46] optimized indeed the capacity but did so at the cost of an infinite end-to-end delay. Many authors considered afterwards this trade-off between capacity and delay in more detail (see, e.g., [5, 74, 92]). Also for finite-size networks, which are more realistic from an application viewpoint [24], delay performance has been studied. However, this has been done on quite strong assumptions, such as instantaneous transmissions [44, 94], only a single packet in the network [72] or stationary stations [8]. Abstracting from the world of MANETs, for a Jackson network it is well-known that the buffer-size distribution satisfies a simple product-form solution [50], i.e., the joint queue-length distribution is the product of the marginal distributions. However, if one wants to incorporate ad hoc network characteristics, such as the time-varying availability of servers at the stations into the concept of Jackson networks, then such simple solutions cannot be obtained. Thus, it might be wise to address first a simpler problem of a single station in isolation that wants to transmit over a wireless link to a neighboring station. Due to the variability in the network, the link will not be available continuously for transmissions. The presence or absence of a link in the wireless network model can then be mapped onto the availability of a server in a queueing model. More precisely, such queueing models are referred to as unreliable-server models. The unreliable-server model is a single-server single-queue model in which the server alternates availability periods with periods of unavailability (repair). The availability periods, i.e., the time until a next breakdown, have a random duration independent of the number of customers in the system. Moreover, the repair period has a random duration. An ad hoc network may be observed as a connection of several of these blocks comprising a single station and a link. From a queueing theoretical perspective, this leads quite naturally to the class of models known as polling systems. Polling systems are multi-queue models in which (from the perspective of the queues) server availability periods are alternated with random periods of server unavailability. The duration of the availability period, or also the visit period, is governed by the service discipline of the server. The discipline that fits seamlessly to the random topololgy changes (e.g., due to autonomous behavior of mobile stations) is the so-called *pure time-limited discipline*, which will be introduced below. As polling systems operating under this specific discipline will take a fundamental position in the remainder of this thesis, we will next introduce polling systems more formally, define our basic polling system and review related analysis of polling systems.

Parts II and III of this thesis will be dedicated to the issues of transfer delays and buffer sizes.

1.3 Polling systems

Polling systems are queueing systems consisting of multiple queues served by one or more servers. Systems with a single server have extensively been studied in the literature, whereas only little attention has been devoted to multi-server systems. For more details on a broad class of polling models and their analysis we refer to [97, 98, 102]. Here, we concentrate on the models and the results which are most relevant in light of this thesis. First, we will discuss the single-server case and afterwards review the analytical efforts on the multi-server case.

1.3.1 Single-server models

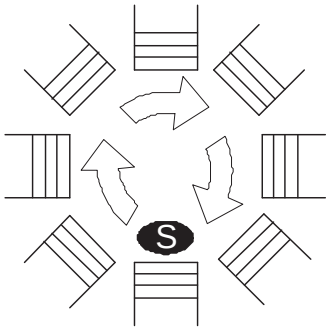

Figure 1.4: Single-server polling model.

Polling models are typically characterized by: (i) the arrival process of the customers to the system, (ii) the service requirements of the customers, (iii) the switch-over times of the server between visits to the queues, (iv) the visit strategy of the server, and (v) the servicing policy of the server (e.g., exhaustive or gated). Applications of polling systems are ubiquitous. For instance, traffic light systems, multiple-access protocols for communication networks (e.g., IEEE 802.11) and product-assembly systems can be modelled as a polling model.

Formally, the single-server polling model (see Fig. 1.4) can be described as follows. A polling model is a system consisting of M queues, which we denote by $Q_i,\ i = 1, \ldots, M$, each equipped with a buffer. The queues are served by a single server at unit rate. Throughout we will use the subscript i to refer to a queue and for convenience leave out its range ($i = 1, \ldots, M$) whenever this does not lead to ambiguity. The interarrival times of customers arriving to Q_i are distributed according to a generic random variable I_i with distribution function $I_i(t)$, Laplace-Stieltjes Transform (LST) $\tilde{I}_i(s)$ and mean $1/\lambda_i$. A customer arriving to Q_i requires an amount of service according to a generic random variable X_i with distribution function $X_i(t)$, LST $\tilde{X}_i(s)$, and mean $1/\mu_i$.

The server visits a queue, offers service to (a part of) the customers present at this queue, and then switches to a next queue. We denote the switch-over time $C_{i,j}$ as the time needed for the server to move from Q_i to Q_j. We assume that the switch-over times follow a general distribution $C_{i,j}(t)$, with LST $\tilde{C}_{i,j}(s)$, and mean $c_{i,j}$.

The server picks the next queue upon the end of a visit according to a specific visit (or polling) strategy. The most common strategy is the cyclic polling strategy. According to this strategy, the server visits the queues in the fixed order $Q_1, Q_2, \ldots, Q_M, Q_1$, etc. A generalization of the cyclic strategy is the periodic polling strategy. The server still visits the queues according to a fixed schedule, but not necessarily each queue equally often. Thus, schedules of the form $Q_1, Q_2, Q_1, Q_3, Q_1, Q_2$, etc., are also included. The Markovian polling strategy is a random visit strategy according to which the server switches queue in a probabilistic manner. More specifically, the probability of choosing the next queue upon a visit completion depends only on the queue left behind by the server.

The service discipline describes the behavior of the server at a queue. In fact, it determines the set of customers that will be attended during a visit of the server. Let us list the most common service disciplines below:

- Exhaustive discipline; the server serves all the customers at the queue (both the ones present upon arrival and the ones that arrive during a service of another customer) and departs from the queue only when it is empty.
- Gated discipline; upon arrival of the server to the queue a gate is placed behind the customers present at the queue. The customers in front of the gate will be served during the visit and customers which arrive during the course of the visit will be served only on a next occasion.
- k-limited discipline; the server serves k customers at a queue or leaves when then queue becomes empty.
- Exhaustive time-limited discipline; the server serves customers at the queue until a time limit expires or leaves when the queue becomes empty.

We note that the exhaustive time-limited discipline appears in the literature commonly as time-limited discipline. However, in this way it is easier to distinguish between this discipline and the pure time-limited discipline that will be introduced soon. Moreover, we should emphasize that there exist still many other service disciplines, such as the binomial-gated, the globally-gated or decrementing service disciplines (see, e.g., [102]).

1.3.2 The basic single-server polling model as a model for MANETs

A polling model emerges quite naturally as a packet-level performance model for MANETs (see also the end of Sect. 1.2). The dynamics of the stations drive in fact

a process of wireless communication links that originate and break down. Consequently, the lifetime of these links is random and in particular does not depend on the amount of traffic offered to or transmitted over such links. An active link in the ad hoc network can be mapped to a queue being served in the polling model and its lifetime to the visit time of the server to a queue. This visit time is controlled by the service discipline at the queue. Unfortunately, under the common service disciplines, the visit time depends on the evolution of the number of customers during this visit at the queue that is being served. Thus, such disciplines do not qualify to model the random link activation process in MANETs properly. Hence, we introduce here a novel service discipline. This novel discipline corresponds in a natural way to the random changes in resource availability in mobile ad hoc networks. In particular, this discipline neglects the state of the network in terms of queue lengths and it is defined as follows.

Definition 1.4 (Pure time-limited discipline)**.** *The server visits a queue for a random amount of time independent of anything else in the system, and then leaves for a next queue.*

This discipline enforces that the service at a queue will be preempted at the end of a visit of the server. At the beginning of the next visit, a service time will be redrawn from the original distribution; thus, we adopt the so-called *preemptive-repeat-random* strategy. We note that in a wireless environment the transmission rate (and thus the service time) is largely dominated by the highly dynamic channel conditions. This specific service strategy appears therefore the most appropriate one (rather than, e.g., a preemptive-resume strategy). Further, we emphasize that the server remains at a queue even if it becomes empty during a visit. Thus, the pure time-limited discipline is not a work-conserving discipline. The random time limit will be assumed throughout exponentially distributed unless explicitly mentioned otherwise.

Basic single-server polling model The basic single-server polling model that we will consider in this thesis is defined as follows. We consider a system of M queues each with infinite-sized buffer. The queues are served by a single server at unit rate. We assume that the interarrival time is exponentially distributed, i.e., the arrival process is Poisson with rate λ_i. A customer arriving to Q_i requires an amount of service with generic random variable X_i with distribution function $X_i(t)$, LST $\tilde{X}_i(s)$, and mean $1/\mu_i$. We assume that customers at a queue are served according to the First-Come-First-Served (FCFS) discipline. The server serves the queues according to the pure exponential time-limit discipline. The switch-over times of the server $C_{i,j}$ follow a general distribution $C_{i,j}(t)$, with LST $\tilde{C}_{i,j}(s)$, and mean $c_{i,j}$. Finally, we leave the server visit strategy unspecified as it does not play a critical role in the analysis.

1.3.3 Single-server analysis

The most celebrated approach to analyze polling systems is based on the construction of Markov chains at specific embedded epochs and subsequently relating the state space at these epochs. The approach was originally introduced by Eisenberg [29] to analyze a polling system with the exhaustive and the gated service discipline, but the main ideas apply to more general systems. These epochs refer to instants of visit beginnings, visit completions, service beginnings and service completions. The approach aims at finding expressions for the probability generating functions (p.g.f.'s) of the queue-length distribution at these epochs. Let us denote these p.g.f.'s of the queue length as follows for $i = 1, \ldots, M$:

$\alpha^i(\mathbf{z})$: p.g.f. of the queue-length distribution at visit beginnings,
$\beta^i(\mathbf{z})$: p.g.f. of the queue-length distribution at visit completions,
$\omega^i(\mathbf{z})$: p.g.f. of the queue-length distribution at service beginnings,
$\pi^i(\mathbf{z})$: p.g.f. of the queue-length distribution at service completions.

Subsequently, Eisenberg [29] established three relations (per queue) between these p.g.f.'s. The first relation is derived via some simple, but elegant, counting arguments:

$$\eta\alpha^i(\mathbf{z}) + \pi^i(\mathbf{z}) = \omega^i(\mathbf{z}) + \eta\beta^i(\mathbf{z}),$$

where η is a known positive constant. The second relation describes the queue-length evolution during the service of a customer:

$$\pi^i(\mathbf{z}) = \frac{\hat{X}_i(\mathbf{z})}{z_i} \cdot \omega^i(\mathbf{z}),$$

where $\hat{X}_i(\mathbf{z})$ denotes the p.g.f. of the number of arrivals to all queues during a service at Q_i. The final relation couples the queue length at the start of a visit to Q_{i+1} to the queue length at the end of a visit to Q_i, viz.,

$$\alpha^{i+1}(\mathbf{z}) = \hat{C}_{i,i+1}(\mathbf{z}) \cdot \beta^i(\mathbf{z}), \tag{1.1}$$

where $\hat{C}_{i,i+1}(\mathbf{z})$ denotes the p.g.f. of the number of arrivals to all queues during a switch-over time of the server from Q_i to Q_{i+1}. Equation (1.1) is independent of the service discipline, but depends on the server strategy. However, a similar relation can readily be established for other server visit strategies. Altogether, this yields in total $3 \cdot M$ equations between the $4 \cdot M$ p.g.f's of interest. Thus, it will require still an additional equation (for each queue) between these p.g.f.'s to fully determine all the p.g.f.'s above.

Eisenberg solved the complete system by deriving an explicit expression for $\beta^i(\mathbf{z})$. Unfortunately, this cannot be done for general service disciplines, so that we will pursue a different solution approach. We will concentrate on the key relation which

describes the queue-length evolution during a service visit and which can be written in the following general form:

$$\beta^i(\mathbf{z}) = \mathcal{F}(\alpha^i)(\mathbf{z}), \tag{1.2}$$

where $\mathcal{F}$ is some operator representing the evolution of the joint queue-length process during a visit and which depends on the assumed service discipline. The relations of Eqs. (1.1) and (1.2) for all queues in the system together give rise to a system of equations which may be solved in an iterative fashion. For service disciplines satisfying the so-called branching property (e.g., the exhaustive and gated disciplines), this leads to a closed-form solution for the joint queue-length distribution at the embedded epochs, while for other disciplines one must generally resort to numerical solution techniques.

We will continue by reviewing first the general solution concepts for branching and non-branching type disciplines, respectively. Finally, we zoom in on the analytical efforts for a specific class of non-branching type disciplines, viz., the pure and exhaustive time-limited disciplines.

1.3.3.1 Branching-type disciplines

In the analysis of polling systems a fundamental part is played by the branching property. A branching-type service discipline satisfies the following property [40]:

Property 1.5. *(Branching-type service disciplines) If the server arrives to Q_i to find k_i customers there, then during the course of the server's visit, each of these k_i customers will effectively be replaced in an i.i.d. manner by a random population having (say) p.g.f. $h_i(z_1, \ldots, z_M)$, which can be any M-dimensional p.g.f.*

Polling systems which operate under such service disciplines (e.g., the exhaustive and gated disciplines) are amenable to a tractable analysis, while the analysis of other disciplines (e.g., the k-limited and time-limited disciplines) is usually restricted to special cases or numerical approaches. This dichotomy is reflected in the operator $\mathcal{F}$ which for service disciplines satisfying the branching property is of a simple form, so that one obtains the following *direct* relation:

$$\beta^i(\mathbf{z}) = \alpha^i(z_1, \ldots, z_{i-1}, h_i(\mathbf{z}), z_{i+1}, \ldots, z_M), \tag{1.3}$$

where $h_i(\mathbf{z})$ is the p.g.f. of the random population which replaces a customer served at Q_i and depends on the specific service discipline. For the gated discipline, we have:

$$h_i(\mathbf{z}) = \tilde{X}_i \left(\sum_j \lambda_j (1 - z_j) \right),$$

while for the exhaustive discipline, we have

$$h_i(\mathbf{z}) = \tilde{U}_i \left(\sum_{j \neq i} \lambda_j (1 - z_j) \right),$$

where $\tilde{U}_i$ refers to the LST of the busy period in an M/G/1 queue with service requirement X_i and arrival rate λ_i. For disciplines satisfying the branching property, Eq. (1.3) together with Eq. (1.1) leads to a closed-form solution for the joint queue-length distribution at the embedded epochs. For instance, for a polling model with a cyclic server the solution reads (see, e.g., [29]):

$$\beta^i(\mathbf{z}) = \prod_{l=1}^{\infty} \tilde{C}_{i,i+1}(h_i^{(l)}(\mathbf{z})), \tag{1.4}$$

where $h_i^{(l)}(\mathbf{z})$ is an l-fold nested function defined as follows:

$$h_i^{(l)}(\mathbf{z}) \quad := \quad h_{i-l+1}(\cdots(h_{i-2}(h_{i-1}(h_i(\mathbf{z}))))\cdots), \; l = 1, 2, \ldots.$$

The expressions of Eq. (1.4) for the p.g.f. can be used to calculate moments of the queue length or waiting-time distribution (see, e.g., [25, 29]). It is good to notice at this point that exact closed-form expressions even for the mean queue-length or the mean waiting-time are only known for particular polling systems, such as fully-symmetric systems. As a result, a large number of methods (not directly based on Eq. (1.4)) have appeared in the literature for efficient moment computation for polling systems operating under branching-type service disciplines (see, e.g., [36, 59, 105]).

1.3.3.2 Non-branching type disciplines

For service disciplines that do not satisfy the branching property, such as the k-limited and time-limited disciplines, closed-form solutions of the form of Eq. (1.4) are not likely to exist. In particular, the key relation of Eq. (1.2) cannot be written in the direct form of Eq. (1.3) and for this reason a different solution approach is required.

In the literature, apart from many approximation and simulation efforts, several exact (numerical) methods have been used to study polling systems operating under these service disciplines. To compute the steady-state queue-length probabilities, Blanc [9] developed the power series algorithm which can be applied to a large variety of service disciplines (both branching and non-branching type). This technique essentially boils down to numerically solving a large multi-dimensional Markov chain in a computationally efficient way. Another, recursive, approach has been introduced by Leung [67]. Opposed to the direct relation from the start to the end of a visit (cf. Eq. (1.3)), he established an *indirect* relation by segmenting a visit according to service completions. We illustrate the main steps of this approach here, as it will return at several places in the remainder of this thesis.

To this end, let us denote the number of customers at all queues at the jth service completion at a visit to Q_i by $\psi_j^i = (\psi_j^i(1), \ldots, \psi_j^i(M))$. Accordingly, we denote the joint queue-length p.g.f. at these embedded instants by $\Psi_j^i(\mathbf{z}) := \mathbb{E}[\mathbf{z}^{\psi_j^i}]$. The p.g.f. $\Psi_j^i(\mathbf{z})$ satisfies the following recursive relation:

$$\Psi_j^i(\mathbf{z}) = \Psi_{j-1}^i(\mathbf{z})|_{z_i=0} + \frac{\hat{X}_i(\mathbf{z})}{z_i} \cdot \left(\Psi_{j-1}^i(\mathbf{z}) - \Psi_{j-1}^i(\mathbf{z})|_{z_i=0}\right), \; j = 1, 2, \ldots, \quad (1.5)$$

with initial value $\Psi_0^i(\mathbf{z}) = \alpha^i(\mathbf{z})$. It is shown in [67] that $\beta^i(\mathbf{z})$ can be expressed as:

$$\beta^i(\mathbf{z}) = \sum_{j=0}^{\infty} a_j^i \Psi_j^i(\mathbf{z}), \quad (1.6)$$

where a_j^i is a model parameter which refers to the probability of having service limit j at Q_i. For instance, the 1-limited discipline is fully characterized by:

$$a_j^i = \begin{cases} 1, & j = 1, \\ 0, & \text{otherwise.} \end{cases}$$

Hence, for this discipline, Eq. (1.6) can readily be rewritten in the general form of Eq. (1.2) as:

$$\beta^i(\mathbf{z}) = \frac{\hat{X}_i(\mathbf{z})}{z_i} \cdot \left(\alpha^i(\mathbf{z}) - \alpha^i(\mathbf{z})|_{z_i=0}\right) + \alpha^i(\mathbf{z})|_{z_i=0}, \quad (1.7)$$

where $\hat{X}_i(\mathbf{z})$ denotes the p.g.f. of the arrivals to the system during a service at Q_i. It is readily observed that in general the p.g.f. $\beta^i(\mathbf{z})$ cannot be obtained in closed form from Eqs. (1.1) and (1.7).

To resolve this difficulty, Leung [67] proposes to determine $\beta^i(\mathbf{z})$ numerically along an iterative algorithm which can be applied to any service discipline. This algorithmic scheme is constructed in terms of Discrete Fourier Transforms (DFTs) as these appear more convenient for computational purposes. To this end, replace z_i, $i = 1, \ldots, M$, in the expressions above by $\omega_i^{k_i}$, where $\omega_i = exp(-2\pi I/H_i)$, so that all expressions become functions of $\mathbf{k} = (k_1, \ldots, k_M)$. Here, I is the imaginary unit and H_i refers to the number of discrete points used for Q_i to determine the joint probabilities. In particular, we approximate the DFT of $\alpha^i(\mathbf{z})$ and $\beta^i(\mathbf{z})$ as:

$$\begin{aligned} \breve{\alpha}^i(\mathbf{k}) &\approx \sum_{n_1=0}^{H_1-1} \sum_{n_2=0}^{H_2-1} \cdots \sum_{n_M=0}^{H_M-1} \omega_1^{k_1 \cdot n_1} \omega_2^{k_2 \cdot n_2} \cdots \omega_M^{k_M \cdot n_M} P_{\alpha^i}(n_1, n_2, \ldots, n_M), \\ \breve{\beta}^i(\mathbf{k}) &\approx \sum_{n_1=0}^{H_1-1} \sum_{n_2=0}^{H_2-1} \cdots \sum_{n_M=0}^{H_M-1} \omega_1^{k_1 \cdot n_1} \omega_2^{k_2 \cdot n_2} \cdots \omega_M^{k_M \cdot n_M} P_{\beta^i}(n_1, n_2, \ldots, n_M), \end{aligned}$$

where $P_{\alpha^i}(n_1, n_2, \ldots, n_M)$ and $P_{\beta^i}(n_1, n_2, \ldots, n_M)$ refer to joint queue-length probabilities at a visit beginning and completion instant at Q_i, respectively. For convenience, let us assume the cyclic polling strategy and denote $\breve{C}_{i,i+1}(\mathbf{k})$ as the DFT

of $\hat{C}_{i,i+1}(\mathbf{z})$. The algorithm departs from an empty system with the server at Q_{i_1}. Thus, $\check{\alpha}^{i_1}(\mathbf{k}) = 1$, and $\check{\beta}^{i_1}(\mathbf{k})$ is computed according to Eq. (1.6). The value of $\check{\beta}^{i_1}(\mathbf{k})$ is stored and used to compute $\check{\alpha}^{i_1+1}(\mathbf{k})$ according to Eq. (1.1). Next, $\check{\beta}^{i_1+1}(\mathbf{k})$ is computed, and so on. Notice that due to the cyclic polling strategy, the algorithm returns in fact to Q_{i_1} after M steps. The pseudo-code of the iterative algorithm is presented in Algorithm 1.6. The standard values for the convergence parameters are $\epsilon = 10^{-6}$ and $\delta = 10^{-9}$. We note that the algorithm will always converge as long as the embedded queue-length process forms an ergodic Markov chain. Finally, via the Inverse Fourier Transform, the steady-state probabilities are obtained, i.e.,

$$P_{\beta^i}(n_1, n_2, \ldots, n_M) \approx \frac{1}{H_1 H_2 \cdots H_M} \sum_{k_1=0}^{H_1-1} \sum_{k_2=0}^{H_2-1} \cdots \sum_{k_M=0}^{H_M-1} \nu_1^{k_1 \cdot n_1} \nu_2^{k_2 \cdot n_2} \cdots \nu_M^{k_M \cdot n_M} \check{\beta}^i(\mathbf{k}),$$

where $\nu_j = exp(2\pi I/H_i)$, for $j = 1, \ldots, M$. It is good to observe that the probabilities $P_{\beta^i}(n_1, n_2, \ldots, n_M)$ are only exact for $H_i \to \infty,\ i = 1, \ldots, M$. However, the strength of the approach is that in general the probabilities are already close to the exact values for small values of H_i. It should also be noted that when the system load increases, these values H_i must be increased to guarantee the accurate computation of the probabilities. Thus, this iterative approach appears mainly applicable to systems with a light to moderate load.

Algorithm 1.6. *Pseudo-code of the iterative scheme for determining* $\check{\beta}^i(\boldsymbol{k}), \forall_i, \forall_{\boldsymbol{k}}$.

$\check{\beta}^{i_0}(\boldsymbol{k}) = 1,\ \forall_{i_0},\ \forall_{\boldsymbol{k}}$; *(start with an empty system)*
FOR $i_1 = 1, \ldots, M$
set $i_2 := i_1$;
REPEAT
$\bar{\beta}^{i_2}(\boldsymbol{k}) = \check{\beta}^{i_2}(\boldsymbol{k}),\ \forall_{\boldsymbol{k}}$;
set $j := 0$;
set $\Psi_0^{i_2}(\boldsymbol{k}) = \check{\beta}^{i_2-1}(\boldsymbol{k}) \cdot \check{C}_{i_2-1,i_2}(\boldsymbol{k})$;
REPEAT
set $j := j + 1$;
compute $\Psi_j^{i_2}(\boldsymbol{k}),\ \forall_{\boldsymbol{k}}$, *using Eq. (1.5);*
compute $\check{\beta}^{i_2}(\boldsymbol{k}) = \sum_{l=1}^{j} a_l^{i_2} \Psi_l^{i_2}(\boldsymbol{k}),\ \forall_{\boldsymbol{k}}$;
UNTIL $1 - Re(\check{\beta}^{i_2}(\boldsymbol{0})) < \delta$
set $i_2 := MOD(i_2, M) + 1$;
UNTIL $\lvert Re(\check{\beta}^{i_1}(\boldsymbol{k})) - Re(\bar{\beta}^{i_1}(\boldsymbol{k}))\rvert < \epsilon,\ \forall_{\boldsymbol{k}}$
END

1.3.3.3 Pure and exhaustive time-limited disciplines

There exists hardly any literature on single-server polling systems operating under the pure time-limited disciplines. The only work that includes both a given visit time and a patient server, i.e., a server which does not leave before the end of the visit time, is [108]. This work considers the workload process for a pure time-limited polling model with deterministic visit times and a cyclic visit schedule. Due to the deterministic nature of the model, the queue lengths at the different queues can be decoupled and each queue is modelled as an M/G/1 queue with server vacations. Using an approximate analysis, the mean workload and mean message delay are studied.

On the contrary, for the exhaustive time-limited discipline a large number of both approximative and exact analysis exists (see, e.g., [95, 31, 32, 39, 68]). Leung [68] analyzes the queue-length distribution at embedded epochs for a time-limited model in which the server remains an exponential time at a queue but service is non-preemptive. A deterministic time-limited polling model with preemptive service is studied by De Souza e Silva et al. [95] for exponential service times. Uniformization methods are employed to eventually obtain the queue-length distribution at specific embedded epochs. Frigui and Alfa [39] consider Markovian Arrival Processes for a polling system with a deterministic time-limit. The authors present an approximative analysis for the queue-length distribution and mean waiting time. Eliazar and Yechiali [31, 32] studied the exhaustive time-limited discipline with an exponential time limit and preemptive service. Observing that upon successful service completion at a queue the busy period in fact regenerates, the authors could obtain a closed-form relation between the joint queue length at the end and the start of a server visit of the following form:

$$\beta^i(\mathbf{z}) = c(\mathbf{z}) \cdot (\alpha_i(\mathbf{z}) - \alpha^i(\mathbf{z}_i^*)) + \alpha^i(\mathbf{z}_i^*), \tag{1.8}$$

where $\alpha^i(\mathbf{z}_i^*) := \alpha_i(z_1, \ldots, z_{i-1}, k_i(\mathbf{z}), z_{i+1}, \ldots, z_M)$, and $c(\mathbf{z})$ and $k_i(\mathbf{z})$ are functions of $\mathbf{z}$ with $k_i(\mathbf{z})$ being related to the LST of the busy period of a customer at Q_i.

1.3.4 Multi-server models

Polling systems may also comprise multiple, say $K \geq 2$, servers that serve the queues. These multi-server polling models expand the visit strategy of the single-server model. For this reason, we will describe the dynamics of the servers (hereby neglecting switch-over times) by a K-dimensional discrete-time Markov chain $X_n = (l_1^n, \ldots, l_S^n) \in \mathcal{L}_1 \times \cdots \times \mathcal{L}_K$, where $\mathcal{L}_1, \ldots, \mathcal{L}_K \subseteq \{1, \ldots, M\}$, $n \geq 0$, driven by the transition probability matrix $S = \{s_{l,j}\}_{l,j \in \mathcal{L}_1 \times \cdots \times \mathcal{L}_K}$. We assume that the Markov jump chain has a stationary distribution which we denote by τ_l, $l \in \mathcal{L}_1 \times \cdots \times \mathcal{L}_K$. In the sequel, we indicate $l = (l_1, \ldots, l_K)$ as *server-location* state, where l_j, $j = 1, \ldots, K$, is the location of server S_j in state l, and leave out the superscript n.

According to this description, the servers may visit the queues in many different ways. The most common server strategies for multi-server polling models are:

- coupled servers; i.e., the servers are coupled and move as a group along the queues (thus, $l_1^n = \ldots = l_K^n,\ n = 0, 1, \ldots$);

- individual servers; i.e, the servers move individually through the system.

In the first case, each server will visit all the queues, whereas in the second case each server might only serve a subset of the queues in the system. The coupled-server case resembles the single-server case. The main difference is that in the multi-server case multiple customers can be served simultaneously. In the individual-server case, each server will basically follow its own visit schedule. This schedule may either be fixed or random. Anyhow, it is essential for the stability of the system that each queue in the system will be visited with strictly positive probability (by at least one server) from time to time. Besides, it must be established how the system proceeds when a server polls a queue where a number of servers is already present. A common strategy is that if the number of servers present exceeds a specific limit then a server will jump over this queue and move immediately to the next queue in its schedule. However, note that under this strategy the *movements* of the servers are in fact not independent, unless this limit equals S. In this latter special case, a server will indeed move independently of the position of the other servers in the system and we refer to this visit strategy as *independent-server* strategy. Finally, it is good to notice that by appropriately setting the state space and transition probabilities any of these server strategies, viz., coupled servers and individual servers, can indeed be modelled.

1.3.5 The basic multi-server polling model as a model for MANETs

We have justified in Sect. 1.3.2 that the single-server basic polling model is an appropriate performance model for MANETs. In particular, it can be applied to study wireless networks with a single active link, such as is typical for small networks or (large) fully-connected networks. A logical next step is to consider a performance model which allows for studying scenarios with multiple links that can be active simultaneously in such a dynamic ad hoc network topology. Clearly, a polling model with multiple servers operating under the pure time-limited discipline satisfies these requirements in a natural way. Regarding the individual-server strategy, this means that each server will visit a queue for an amount of time and leaves the queue if and only if this time period has expired. For the coupled-server strategy, this means that the group of servers will visit a queue for an amount of time and this group leaves the queue together if and only if this time period has expired. The random time limit for the multi-server model will be assumed exponentially distributed. Similarly as for the single-server model, this discipline will lead to the *preemptive-repeat-random* service strategy.

Basic multi-server polling model The basic multi-server polling model that we will consider in this thesis is defined as follows. We consider a system of M queues each with infinite-sized buffer. The queues are served by $K \geq 2$ servers at unit rate. We assume that the interarrival time is exponentially distributed, i.e., the arrival process is Poisson with rate λ_i. A customer arriving to Q_i requires an exponential amount of service with mean $1/\mu_i$. We assume that customers at a queue are served according to the FCFS discipline. The server serves the queues according to the pure exponential time-limit discipline. We assume that the switch-over times for a server (in the individual-servers case) or a group of servers (in the coupled-servers case) to switch from Q_i to Q_j follow a general distribution $C_{i,j}(t)$, with LST $\tilde{C}_{i,j}(s)$, and mean $c_{i,j}$. Finally, we leave the server visit strategy unspecified, since we will consider both strategies described above.

1.3.6 Multi-server analysis

Multi-server polling models have been awarded little attention in the literature, especially when compared to their single-server counterparts. The principal reason being that such models do not seem to allow for nice exact solutions such as obtained for single-server polling models operating under a branching-type service discipline. As a consequence, the analytical attempts towards a better understanding of multi-server polling models are quite diverse and a general analytical framework is absent. Hence, we confine ourselves here to an overview of the literature on the performance analysis of multi-server polling models without displaying any explicit analysis.

1.3.6.1 Coupled servers

Browne and Weiss [20] extend the analysis of a multi-server single-queue model to a polling model with c coupled servers. In each cycle, only the queues with more than or equal to $c-1$ customers are served. For the gated and exhaustive service discipline, closed-form expressions for the mean cycle time are derived in terms of (the unknown) $Q_i(0)$, the number of customers present at Q_i at the start of a cycle. Borst [14] discusses multi-server polling models which allow for an exact analysis of distributional measures. The work builds on the analysis of an M/M/c queue with service interruptions by extending the decomposition ideas of Fuhrmann and Cooper [41]. As for the single-server model, the approach of relating the number of customers at the beginning and the end of a visit is followed leading to a system of equations. A number of special cases was discussed for which these equations could explicitly be solved. These cases include several one and two-queue systems with a finite number of servers, and larger systems with an infinite number of servers and deterministic service times.

For a two-queue system with an infinite number of servers and deterministic service at one queue the LST of the busy period is obtained in [19]. This is done as a special case of a study on M/G/∞ vacation models (according to the N-policy and to the multi-vacation policy). Also, Lee [65, 66] discusses a multi-queue system

served by an infinite group of servers. The service times are assumed deterministic and service is performed according to the globally-gated discipline. Transient and steady-state analysis are given for the mean waiting time of a customer. The results are expressed in terms of the probability of the system (and queues) being empty which follows from a system of equations. Another infinite-server polling model is studied by Vlasiou and Yechiali [103]. Their model assumes Poisson arrivals, general service times and the pure time-limited service discipline. The p.g.f. of the joint queue length at polling instants is obtained and also the LST of the sojourn time.

1.3.6.2 Individual servers

Morris and Wang [75] analyze a polling model with independent servers under the gated and a kind of limited discipline. Each server follows its own trajectory, but skips a queue that is being served already. The analysis regards a quite complex approximation for the mean sojourn time of a job. Experimental results show that servers coalesce when the same cyclic order is used. Bhuyan et al. [6] present a unified approximative analysis of various single-ring and multi-ring networks. Under quite strong assumptions, a closed-form approximation is derived for the mean waiting time and mean queue length for the multiple token ring. The model resembles a polling model with multiple independent servers. Another approximate analysis for a polling model with multiple independent servers is given in [2]. Closed-form expressions (with unknown parameter p) are derived for the mean (partial) cycle time, mean visit time and mean intervisit time. This parameter p refers to probability of an arriving server to a queue being allowed to serve this queue. Based on these exact closed-form expressions, an approximation for the mean waiting time is proposed. Exact results for multi-server polling systems served according to the Bernoulli discipline are presented by Van der Mei and Borst [73]. This service discipline includes the 1-limited and exhaustive discipline, while the gated discipline cannot be considered. Using the power series algorithm, the authors compute the joint distribution of the queue length and the position of the servers. Faced by the intrinsic difficulties to analyze multi-server polling systems in an exact fashion, the same authors present approximations for the mean waiting time [15] hereby focussing on the case of independent servers. Eventually, expressions for the mean waiting time are derived under the key assumption that all servers carry the same load.

1.4 Outline of the thesis

This thesis is organized in three parts.

1.4.1 Part I: Network capacity and stability

In the first part of the thesis, we consider the capacity and stability of performance models for mobile ad hoc networking. We will study the impact of signal interfer-

ence on network performance measures in Chapter 2. In particular, we focus on the capacity under interference hereby emphasizing the performance trade-off between single-path and multi-path routing. It may seem attractive to employ multi-path routing, but as all stations share a single channel, efficiency may drop due to increased interference levels thus yielding single-path performance for some network topologies. To this end, we develop a queueing model which characterizes explicitly the interference in ad hoc networks. We address the question of optimization of the network performance and formulate this as a nonlinear programming problem. It will be shown that for the network capacity the optimum could in principle be found by solving a number of linear programmes. However, this number increases exponentially in the number of stations in the network. Therefore, we propose a computationally attractive, greedy algorithm that efficiently searches these programmes to approximate this optimal solution. Numerical results for small topologies provide structural insight in optimal path selection and demonstrate the excellent performance of the proposed algorithm. In addition, larger networks and more advanced scenarios with multiple source-destination pairs and different radio ranges are analyzed. The insights gained from the numerical experiments may be applied in the development of routing protocols.

Next, in Chapter 3, we turn to the stability question for performance models for MANETs. More specifically, we will state and prove the stability conditions of single-server polling systems operating under the pure and exhaustive exponential time-limited service discipline. These conditions will be proven for the polling system operating under the periodic polling strategy and preemptive service. The stability proof of the pure time-limited discipline is straightforward as stability may be considered for each queue in isolation. The proof for the exhaustive time-limited discipline is more laborious. We follow the line of proof as introduced by Fricker and Jaïbi [37] for a large class of service disciplines. Unfortunately, the preemptive nature of the exhaustive time-limited discipline excludes it from this class and as a result substantial efforts are required to modify the proof as to allow for preemptive disciplines. Finally, the extension of the proofs to the Markovian polling strategy is discussed.

1.4.2 Part II: Single-server polling models

The second part regards exact and approximative analysis of single-server polling systems operating under a time-limited discipline. First, we present in Chapter 4 an exact analysis for the joint queue-length distribution of our basic polling system (see Sect. 1.3.2) which operates under the novel pure time-limited discipline. The analysis builds on the work of Eisenberg [29] which identified relations between the queue length at embedded epochs as discussed in Sect. 1.3.3. We extend this work to account for the preemptions which depart from the pure time-limited discipline, such that in total we consider eight p.g.f.'s for the queue length at embedded epochs per queue. This system of equations is solved by determining a recursive re-

lation representing the queue-length evolution during the visit (see Eq. (1.2)) along a methodology similar to the one introduced by Leung (see Sect. 1.3.3.2). Finally, we provide a number of extensions for the basic polling system and indicate how these can be incorporated in the analysis. These extensions broaden the applicability of the analysis to more general mobile ad hoc networks.

In Chapter 5, we consider the pure and exhaustive time-limited polling system extended with customer routing. Particularly, we present an alternative exact analysis for the recursive relation obtained in Chapter 4 representing the queue-length evolution during the visit. The analysis of the pure time-limited discipline builds on results from the transient analysis of the M/G/1 queue. Thus, we obtain a direct, non-recursive, relation, which resembles the form of Eq. (1.8), that describes the queue-length evolution during a visit. A similar approach is applied to analyze the exhaustive time-limited discipline. To this end, several novel results for the transient queue-length of the M/G/1 busy period are derived. The final expression for the exhaustive time-limited discipline extends the results of [31] with customer routing. The interpretation of our results suggests that for any branching-type service discipline restricted by an exponential time-limit the queue-length evolution during a visit can be expressed in a similar simple form.

The computation of the joint queue-length distribution along the techniques described in Chapter 4 becomes less attractive as the load or the number of queues in the polling system grows large. Moreover, the sojourn time of a customer may not readily be derived from the queue-length distribution when routing of customers is allowed. Hence, in Chapter 6, we will present two approximations: a joint queue-length approximation for the basic polling model and a sojourn time approximation for a specific MANET application. This queue-length approximation is a product-form approximation for the conditional distribution which is based on the presumably small correlation between the queue lengths of the various queues. First, we investigate the range of parameters for which this hypothesis holds indeed true. Subsequently, we present the approximation which is based on the analysis of an unreliable-server model. Finally, we validate the approximation results with the exact solution along the measure of total variation distance. The results may be used to approximate performance measures for complex multi-queue models by analyzing a simple single-queue model only. The second approximation regards an approximation for the sojourn time in a simple network model for a novel mobile ad hoc networking paradigm. This small ad hoc network comprises two fixed stations and one mobile relay station. Using matrix-geometric methods, we construct an approximation for the Laplace-Stieltjes Transform of the sojourn time at the mobile relay station. The approximation has been validated for a wide range of scenarios. Additional numerical results discuss the insensitivity of the mean end-to-end sojourn time to the switch-over time distribution and the optimization of the mean sojourn time under power control.

1.4.3 Part III: Multi-server polling models

In the third part, we study multi-server polling systems operating under the pure time-limited discipline. The analysis is presented for the case of two servers, but most of the presented techniques readily carry over to systems with more than two servers. We will concentrate on the derivation of an expression similar to Eq. (1.2), but now for a two-dimensional server visit process. Essentially, we distinguish two cases in the analysis, viz., (i) both servers are at the same queue, and (ii) the servers are at different queues. These cases provide a unified framework to analyze multi-server polling systems capturing both the coupled and the individual-server strategy. The analysis is carried out under the assumption of exponential service times.

In Chapter 7, we present a complete framework to analyze the steady-state queue-length distribution for the basic two-server polling model with customer routing similar to the framework of Chapter 4. The key relation within this approach, which describes the queue-length evolution during a period in which the servers do not switch, is constructed in a recursive fashion for both cases separately. Also, we include two examples to illustrate the applicability of the analysis.

Finally, we study a direct solution of the key relation for the basic two-server polling system in Chapter 8. This is done according to a transient analysis using similar ideas as in Chapter 5. When the servers are at different queues, the analysis boils down to evaluating non-trivial complex integrals. These integrals must be solved numerically. When the servers are at the same queue, we may apply results of the transient analysis of the M/M/2 queue to analyze the queue-length process. This leads to an explicit, direct relation between the queue-length p.g.f.'s at the start and the end of such a period. Moreover, these results suggest that a direct relation may indeed be found for the basic multi-server polling system with any finite number of coupled servers.

Part I

Network capacity and stability

CHAPTER

2

Network capacity under optimal multi-path routing

2.1 Introduction

In this chapter, we study the network capacity of ad hoc networks. More precisely, we analyze the capacity of finite networks with an arbitrary topology for a stationary scenario. A key ingredient of the analysis is multi-path routing (see [76]). Multi-path routing is an enhanced version of traditional single-path routing which uses only one path from source to destination for data transmission. Conversely, multi-path routing uses multiple routes for transmission, hereby offering on the one hand more opportunities for distributing the traffic over the network and on the other hand inducing interference between the different paths. Thus, we will focus on the problem of optimizing the network performance under multi-path routing while explicitly considering the interference between the stations (and thus also paths). More specifically, the trade-off between the capacity gain using multiple paths and the loss of capacity due to the additional interference is investigated. This trade-off is illustrated by the following example.

Example 2.1. *Consider the network presented in Fig. 2.1. In these figures, data is generated at a source station s and destined for destination d. The solid lines indicate transmission links, while the dashed ones indicate interference between stations. The*

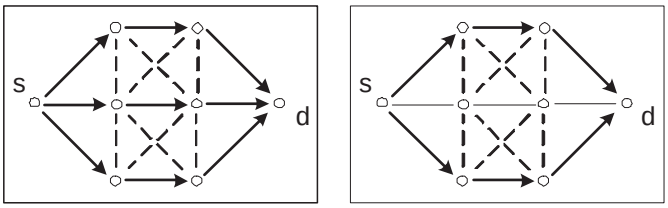

Figure 2.1: Greedy routing (left) vs. interference-aware routing (right).

network comprises three station-independent paths, i.e., paths that do not comprise common stations apart from the source and destination. It can be shown (using the techniques proposed in this chapter) that in such a dense network greedily routing traffic over all available paths (left figure) would yield a network capacity of only 1/3 (i.e, the destination station receives data a fraction 1/3 of the time). However, an interference-aware routing approach (right figure) could attain a capacity of 1/2 of the link capacity by using only two of the available paths.

The concept of multi-path routing is still relatively immature in the context of ad hoc networking. So far, multi-path routing protocols are mainly applied for two reasons: (i) to have backup paths available in case of path failures (see, e.g., [63, 78]), or (ii) to employ multiple paths to spread traffic as to increase the effective bandwidth (see, e.g., [64, 86]). However, there are also a number of drawbacks of employing multiple paths. In single-path routing, normally the shortest path is selected; hence, any additional path will typically be longer, so that the average number of hops to the destination will increase. This may not be harmful when considering capacity questions, but it definitely is when regarding the transfer delay of a packet. Another, frequently underexposed, drawback, which occurs typically in wireless environments, is that stations located on nearby paths may interfere. Interference here refers to the situation of stations in the network overhearing multiple radio signals and consequently observe a useless distorted incoming signal. This leads to unsuccessful packet transmissions and a huge decrease of capacity. Hence, the actual performance gain of using multiple paths over using a single path is unclear.

Although the research focus has been on the back-up application for years, recently the possible bandwidth gain has received more attention. The research effort in this area mainly boils down to the development of novel routing protocols. These protocols aim at finding link- or station-independent paths and do not explicitly take signal interference that occurs between such paths into account (see, e.g., [101]). A plausible explanation is that the notion of interference is very hard to quantify. As a matter of fact, it is still an on-going debate how one can appropriately measure the amount of interference in a network despite of the several metrics that have been suggested. Wu and Harms [107] model the interference (between two paths) metric as the number of links connecting the two paths. Their simulation results do not show significant performance changes for different values of this metric by

choosing alternative paths. Contrary, Pearlman et al. [84] show by using simulation that interference can have a large impact on the network performance. They define an interference metric (coupling) as the average number of stations that are unable to receive data on one path when a single station in the other path is transmitting. Although it is shown that interference affects the performance, a clear relationship between those variables is not derived. These studies reveal that the quantification of interference is indeed non-trivial and that the impact of interference may also strongly depend on the metric defined.

Several interesting papers have then appeared which analytically assess the network capacity by taking interference into account (see [47, 51]). These works follow a graph-theoretical approach for a given finite network instance rather than applying asymptotic techniques conform the pioneering work of Gupta and Kumar [46]. In [51], a multi-commodity flow problem is formulated and extended by interference-related constraints in order to find lower and upper bounds for the capacity. These additional constraints follow by regarding cliques and independent sets in the so-called conflict graph of the network. The approach assumes a central scheduling entity and also extensive computations are required, even for small networks. Later work of Gupta et al. [47] follows a similar approach but aims at a more distributed manner to control the traffic streams in the network. They develop a low-complexity algorithm to find approximate cliques and use this to find lower and upper bounds for the capacity.

In the present chapter, which is an extended version of [H6], we first describe the ad hoc network model and then we present a general stochastic framework in which many performance metrics of interest can be investigated. The analysis within this framework is partly based on the MAC layer IEEE 802.11 protocol [61] for which it has been shown that single-hop ad hoc networks can successfully be modelled by a Processor Sharing queue [70]. We formulate the problem of capacity optimization as a nonlinear programming problem which is then analyzed in order to obtain structural relations in the network. Hence, in contrast to most of the work that appeared in this area which focuses primarily on the capacity value itself, our interest is mainly in the underlying fundamental aspects as to get structural insights into the network performance under interference. At a later stage these insights may be applied in the development of routing protocols. However, the design of such protocols is outside the scope of this thesis.

The chapter is organized as follows. In Sect. 2.2, we will describe the ad hoc network model, state several assumptions, present the mathematical framework and formulate network optimization as a programming problem. The solution techniques for this particular problem are discussed in Sect. 2.3. Next, in Sect. 2.4, we discuss the impact of the interference on the network capacity by virtue of a number of illustrative example topologies. We discuss our model assumptions in Sect. 2.5 and wrap up this chapter in Sect. 2.6.

2.2 Model

2.2.1 Ad hoc network model

Consider an ad hoc network in which all stations are equipped with an identical packet radio (with omnidirectional antenna) operating in half-duplex mode, and transmit over a common channel at identical (maximum) power. Half-duplex mode means that the packet radios are able to transmit and receive data but cannot do so at the same time. A station may transmit data to stations that are within its *transmission range*. During data reception at a station, all stations within its *interference range* must be silent for the reception to be successful. Typically, the interference range exceeds the transmission range, since any small distortion of the radio signal readily excludes a successful data packet reception. Links (or connections) between stations are assumed error free. However, link errors may easily be incorporated in our model as will be discussed in Sect. 2.5. We assume a distributed transmission scheduling mechanism that mimics the IEEE 802.11 MAC protocol, which aims to prevent data-packet collisions. Data transmission in the network is between source-destination pairs (SD-pairs).

2.2.2 Mathematical framework

We will move from the practical setting to a more abstract, mathematical setting. One implication is that we will talk of *nodes* rather than stations. We consider a network consisting of a set of nodes $\mathcal{N} = \{1, \ldots, N\}$. This set comprises a collection of source nodes $\mathcal{S} = \{s_1, \ldots, s_F\}$ and destination nodes $\mathcal{D} = \{d_1, \ldots, d_F\}$, where F denotes the total number of SD-pairs. Thus, s_f and d_f will respectively denote the source and destination of SD-pair f. The remaining nodes can be seen as (pure) relay nodes, but we note that source and destination nodes can also relay traffic. For $j \in \mathcal{N}$, let $\mathcal{N}_{\mathcal{T}}(j) \subset \mathcal{N}$ denote the transmission neighborhood of node j, that is, the set of nodes (possibly sources or destinations) that node j can successfully transmit packets to, and let $\mathcal{N}_{\mathcal{I}}(j) \subset \mathcal{N}$ denote the interference neighborhood of node j, that is, the set of nodes that must be quiet for successful packet reception at node j. The neighborhood relation is not necessarily a symmetrical relation as is well-known from experimental studies (see, e.g., [83]). Thus, e.g., $n \in \mathcal{N}_T(n')$ does not imply $n' \in \mathcal{N}_T(n)$. When $n' \in \mathcal{N}_T(n)$, we say that the network contains a *link* from node n to node n'. Paths in the network consist of a number of links starting at a source node and ending at a destination node. On a path from source s to destination d consisting of $\ell + 2$ links via nodes $n_1, \ldots, n_\ell$, the nodes must be such that $n_1 \in \mathcal{N}_{\mathcal{T}}(s)$, $n_j \in \mathcal{N}_{\mathcal{T}}(n_{j-1})$, $j = 2, \ldots, \ell$, $d \in \mathcal{N}_{\mathcal{T}}(n_\ell)$.

Source s_f generates data flows according to a Poisson process at rate $\lambda^{(f)}$ for SD-pair f. A flow consists of a series of data packets. Let $\beta^{(f)}$ denote the mean number of packets per flow between SD-pair f. Then $\alpha_j^{(f)} = \lambda^{(f)}\beta^{(f)}$, for $j = s_f$, is the mean rate at which source s_f generates packets for SD-pair f; moreover,

$\alpha_j^{(f)} = 0,\ \forall_{j\in\mathcal{N}:j\neq s_f,d_f}$, and we set $\alpha_j^{(f)} = -\lambda^{(f)}\beta^{(f)}$, for $j = d_f$. We assume that ordering of packets at the destination can be handled without loss of information. Thus, a flow may be split and its packets may be transferred over different paths selected according to a suitable network optimization mechanism.

Let node j forward a fraction $p_{jk}^{(f)}$ of its incoming packets for SD-pair f to node k. Denote by $\rho_{t,j}^{(f)}$ and $\rho_{r,j}^{(f)}$ the average number of packets transmitted and received per time unit by node j for SD-pair f. It is assumed here that for any node transmitting a packet takes exactly one unit of time. We may thus also interpret $\rho_{t,j}^{(f)}$ as the average fraction of time node j is transmitting packets for SD-pair f. Then, for the network to sustain all these packet transmissions, the following relations must hold:

$$\rho_{t,j}^{(f)} = \alpha_j^{(f)} + \rho_{r,j}^{(f)},\ \forall_{j\in\mathcal{N}},\ \forall_{f\in\mathcal{F}}, \tag{2.1}$$

$$\rho_{r,j}^{(f)} = \sum_{k\in\mathcal{N}} \rho_{t,k}^{(f)} p_{kj}^{(f)},\ \forall_{j\in\mathcal{N}},\ \forall_{f\in\mathcal{F}}. \tag{2.2}$$

Equation (2.1) equates the amount of data entering and leaving a node per SD-pair; hereby correcting for any generated or absorbed data at the particular node. The other equation, Eq. (2.2), indicates that the amount of data received at a node for a specific SD-pair should have been sent by some neighbor of the particular node. Further, we define the total average fraction of time of transmission and reception for node j by $\rho_{t,j} := \sum_f \rho_{t,j}^{(f)}$ and $\rho_{r,j} := \sum_f \rho_{r,j}^{(f)}$. Notice that for nodes j which are pure relay nodes $\rho_{t,j}^{(f)} = \rho_{r,j}^{(f)},\ \forall_{f\in\mathcal{F}}$. Optimal network design then corresponds to the selection of the fractions $p_{kj}^{(f)}$ in (2.2), so as to maximize a network performance criterion. These fractions determine the optimal path selection for data flows in the network.

The number of packets arriving at a node determines the workload of the node. As a consequence, from the flow perspective, flows share the transmission capacity of the nodes. A single node can handle packets originating from different flows and these packets will be transferred in order of arrival. As nodes are identical, each node transmits packets at a normalized unit rate in the absence of signal interference. Clearly, interference among neighboring nodes reduces the amount of time a node is allowed to transmit packets. In a coordinated network, when neighboring nodes each have packets to transmit, these nodes will share the resources. This is achieved, for example, by the MAC layer protocol of IEEE 802.11 [61], where each node uses its fair share of the medium. It is demonstrated in [70] that the PS queue is an adequate model for MAC layer sharing among multiple nodes. The overhead of the MAC layer protocol results in a reduced data rate, e.g., for a single-cell WLAN operating under IEEE 802.11 with RTS/CTS the MAC layer achieves roughly 85% efficiency (see, e.g., [7]). As a consequence, we may model all nodes in the interference neighborhood (which include the nodes that transmit packets to j) of node j as a single PS queue. We normalize this maximum transmission rate of the nodes at 1. The resulting interference restriction under which node j can still receive data successfully is

$$\sum_{m\in\mathcal{N}_I(j)} \rho_{t,m} + \rho_{t,j} \leq 1,\ \forall_{j\in\mathcal{N}}. \tag{2.3}$$

Notice that this restriction is conservative: when none of the multiple transmissions overheard at a node are directed to this node, then those are unnecessarily prohibited. Moreover, the capacity restriction need not be imposed when node j is not a recipient of any data in the optimal design. This is incorporated via the following modification of Eq. (2.3):

$$\rho_{r,j}\left(\sum_{m\in\mathcal{N}_I(j)}\rho_{t,m}+\rho_{t,j}\right)\leq\rho_{r,j},\ \forall_{j\in\mathcal{N}}. \tag{2.4}$$

Our aim is to investigate the performance trade-off between single-path and multi-path routing. In particular, we investigate the maximum data rate (i.e., capacity) that can be sustained by this optimal path selection under interference (cf. Eq. (2.4)). Thus, we consider capacity optimization of the network in equilibrium. However, our framework also allows to consider alternative performance measures such as the maximal delay for a given traffic load.

2.2.3 Network optimization formulation

Optimal design of paths in the ad hoc network comes down to determining the $p_{jk}^{(f)}$ as the fraction of SD-flow f routed from node j to node k, or equivalently as the probability of routing traffic of SD-flow f from node j to node k. Define the matrices $R_t=(\rho_{t,j}^{(f)})_{j\in\mathcal{N},f\in\mathcal{F}}$ and $P=(p_{jk}^{(f)})_{j,k\in\mathcal{N},f\in\mathcal{F}}$. Network optimization can then be formulated as a nonlinear programming problem:

$$\max \quad h(R_t,P) \tag{2.5}$$

$$\begin{aligned}
\text{s.t.}\quad \rho_{t,j}^{(f)}-\rho_{r,j}^{(f)} &= \alpha_j^{(f)},\ \forall_{j\in\mathcal{N}},\ \forall_{f\in\mathcal{F}}, && (2.6)\\
\rho_{r,j}^{(f)}-\textstyle\sum_{k:j\in\mathcal{N}_T(k)}\rho_{t,k}^{(f)}p_{kj}^{(f)} &= 0,\ \forall_{j\in\mathcal{N}},\ \forall_{f\in\mathcal{F}}, && (2.7)\\
\rho_{r,j}\cdot\left(\textstyle\sum_{m\in\mathcal{N}_I(j)}\rho_{t,m}+\rho_{t,j}\right)-\rho_{r,j} &\leq 0,\ \forall_{j\in\mathcal{N}}, && (2.8)\\
1-\textstyle\sum_{k\in\mathcal{N}_T(j)}p_{jk}^{(f)} &= 0,\ \forall_{j\in\mathcal{N}:j\neq d_f},\forall_{f\in\mathcal{F}}, && (2.9)\\
\rho_{t,j}-\textstyle\sum_f\rho_{t,j}^{(f)} &= 0,\ \forall_{j\in\mathcal{N}}, && (2.10)\\
\rho_{r,j}-\textstyle\sum_f\rho_{r,j}^{(f)} &= 0,\ \forall_{j\in\mathcal{N}}, && (2.11)\\
\rho_{t,j}^{(f)},\ \rho_{r,j}^{(f)} &\geq 0,\ \forall_{j\in\mathcal{N}},\ \forall_{f\in\mathcal{F}}, && (2.12)\\
p_{jk}^{(f)} &\geq 0,\ \forall_{j,k\in\mathcal{N}},\forall_{f\in\mathcal{F}}, && (2.13)
\end{aligned}$$

where $h(R_t,P)$ is a general objective function. For instance, for $h(R_t,P):=\sum_{i=1}^{F}w_i\times\rho_{t,s_i}^{(i)}$ with weights $w_i\geq 0,\ i=1,\ldots,F$, the total weighted capacity in the network is optimized, and for $h(R_t,P):=-\sum_{j=1}^{N}\frac{\rho_{t,j}}{1-\rho_{t,j}}$ the mean number of packets in the network can be minimized for a given set of flows $\alpha_{s_1}^1,\ldots,\alpha_{s_F}^F$. The first set of constraints (2.6)–(2.7) refers to the traffic equations (2.1) and (2.2), and describes

flow conservation; Eq. (2.8) is the interference constraint; Eq. (2.9) indicates that the fractions $p_{jk}^{(f)}$ must sum to one.

The feasible region (2.6)–(2.13) shows that we are dealing with a nonlinear programming problem in the unknowns $\rho_{t,j}^{(f)}$ and $p_{jk}^{(f)}$. The interference assumptions yield constraints that are quadratic in $\rho_{t,j}^{(f)}$ (consider Eqs. (2.7), (2.8) and (2.11)) and moreover interacting terms of $p_{jk}^{(f)}$ and $\rho_{t,j}^{(f)}$ show up in Eq. (2.7). These latter interactions terms can, however, conveniently be eliminated by introducing new variables referring to *link* flows $\lambda_{jk}^{(f)}$, i.e.,

$$\lambda_{jk}^{(f)} := \rho_{t,j}^{(f)} \cdot p_{jk}^{(f)}, \ \forall_{j,k \in \mathcal{N}}, \ \forall_{f \in F},$$

which represent the amount of traffic going from node j to node k per time unit for SD-pair f. Noting that $\rho_{t,j}^{(f)} = \sum_k \lambda_{j,k}(f)$, it is readily observed that all constraints except (2.8) become linear in $\lambda_{jk}^{(f)}$.

In the remainder of this chapter, we will focus on the linear objective of network capacity optimization. We note that this optimization problem for a single SD-pair resembles the well-known max-flow problem in discrete optimization. For multiple SD-pairs, it resembles the multi-commodity flow problem (see, e.g., [91]). Unfortunately, due to the nonlinear interference constraints, our optimization problem cannot be recast in the framework of these problems.

2.3 Solution techniques

We approach the program of (2.5) by exact and approximative solution techniques which are based on the following crucial observation. If the problem is formulated in terms of link flows $\lambda_{jk}^{(f)}$, then the program of (2.5) becomes linear, except for the interference constraints in Eq. (2.8). Furthermore, these quadratic constraints can be replaced by linear ones if the nodes which do not receive any data packets in the optimal traffic distribution (i.e., the nodes with $\rho_{r,j} = 0$) are known. More explicitly, when $\rho_{r,j} = 0$, the constraint is always satisfied, and when $\rho_{r,j} > 0$, we can divide by $\rho_{r,j}$ to get the equivalent constraint (2.3).

2.3.1 Exact approach

The above observation shows that any feasible solution of the nonlinear programming problem characterizes a linear programming (LP) problem. The global optimum is also a feasible solution and therefore we could solve our nonlinear programming problem by consecutively solving 2^N linear problems. However, in practice such an approach is not computationally feasible.

Common techniques (see, e.g., [79, p.12]) would then define the interference constraints by means of functions which indicate whether a node receives data or not.

Instead of Eq. (2.8), we can write:

$$\begin{aligned} \textstyle\sum_{m\in\mathcal{N}_I(j)}\rho_{t,m} + \rho_{t,j} &\leq 1 + N(1-r_j),\ \forall_{j\in\mathcal{N}},\\ \rho_{r,j} &\leq r_j,\ \forall_{j\in\mathcal{N}},\\ r_j &\in \{0,1\},\ \forall_{j\in\mathcal{N}}. \end{aligned}$$

The introduction of the indicator function r_j transforms the quadratic programming problem into a mixed integer programming problem that is linear in $\lambda_{jk}^{(f)}$. Although such a mixed problem is commonly NP-hard, the advantage of this formulation is that standard solvers for this class of programming problems are widely available. Such standard solvers often embed branch-and-bound techniques to reduce the number of LP problems to be solved. We note that the efficiency of such a technique heavily depends on the upper and lower bounds that may be found. Lower bounds follow immediately from feasible solutions. Initially, trivial lower bounds can be obtained by using all nodes in the network or using only a single path for each SD-pair f. For the case of a single SD-pair in the network, a good upper bound can be derived by looking at minimal cuts in the network, since this number is an upper bound on the maximum number of interference-independent paths in the network and for such a path the capacity can easily be determined. An even simpler upper bound can also be found by using the number of outgoing transmission links of s, i.e., $|\mathcal{N}_T(s)|$, and assuming that the paths between source and destination are all node-independent, i.e., the paths do not have any nodes in common. Then, by defining m as $m := |\mathcal{N}_T(s_f)|$, an upper bound of $\frac{m}{m+2}$ for the capacity can readily be derived. Nonetheless, even though such branch-and-bound techniques provide the optimal solution, unfortunately no guarantees on the number of programs to be solved can be given.

2.3.2 Greedy approximation approach

For the evaluation of large networks, a more efficient technique than inspecting all LP problems or applying a branch-and-bound technique will be required. To this end, we introduce an approximate greedy algorithm which works linearly in N, the number of nodes. The greedy algorithm is defined as follows.

Initially, assume that all nodes in the network receive data (i.e., $\rho_{r,j} > 0,\ \forall_{j\in\mathcal{N}}$) and then solve the linear program (i.e., with all interference constraints included) and its corresponding dual. In each following step, a node j^* is eliminated from the network and the resulting (linear) program is then analyzed again. That means, the program with ρ_{r,j^*} set to zero and thus with one interference constraint removed. This iteration process is continued until the optimum finally decreases after a node elimination.

The key element of the algorithm is the elimination step. This elimination takes place based on the values of the dual variables related to the interference constraints, since these dual values indicate the importance of the primal constraints (which have

a one-to-one correspondence to the nodes). Therefore, in each step we eliminate the node with the greatest dual value (which is strictly positive for a connected network), which means that its constraint is definitely tight by appealing to the complementary slackness conditions. Notice that removing a node corresponding to a primal constraint that is not tight (i.e., its dual value is zero) would never lead to an improvement of the objective function. In the case of a tie, i.e., multiple nodes attain the highest dual value, then we select one of the nodes (by solving the primal problem) with the lowest fraction of transmitted packets per unit time. The rationale behind this is that a smaller fraction can more easily be diverted via other paths.

The great advantage of this greedy approach is that since the set of removable nodes is of size N, if the greedy algorithm is used only order N of the 2^N LP problems have to be solved.

2.4 Numerical results

We evaluate the impact of interference on the network capacity for various scenarios. First, we discuss a number of basic topologies that provide structural insights. Next, we move to more general topologies to assess the performance of our proposed greedy algorithm. Finally, capacity results for multiple SD-pairs and for extended interference and transmission ranges will be discussed. Recall that the network capacity is defined in terms of the fractions of time that the sources are transmitting and is thus a dimensionless metric.

2.4.1 Basic scenarios

First, we present results for scenarios with a single source-destination pair and identical one-hop transmission and interference ranges. We will discuss the impact of interference on the network capacity by virtue of a number of small topologies that can be solved optimally within a limited amount of computation time. For these small topologies, we provide structural insights and also indicate their practical importance. Subsequently, we move to larger topologies for which the optimization problem grows too large to find the optimal solution quickly. For those larger topologies, we compare the performance of our (fast) greedy algorithm with an algorithm that would use node-independent paths only, as the latter is common practice in many ad hoc routing protocols. We will show that the capacity values obtained for those algorithms are quite similar in most cases. Recall further that in this case of only a single SD-pair, we do not impose interference constraints on s and node d as these are for sure not relaying any data.

2.4.1.1 Small topologies

We restrict ourselves to topologies consisting of at most two node-independent paths. In the topology figures presented, two nodes can communicate,

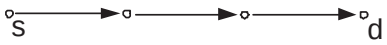

Figure 2.2: Single path.

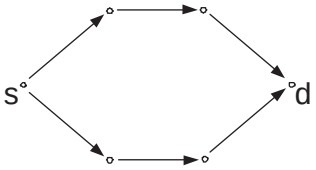

Figure 2.3: Two independent paths.

hence interfere, if and only if they are connected by an edge or arrow. These arrows indicate the paths that may be utilized. The given paths can be seen as provided by a multi-path routing protocol, which typically aims at selecting node-independent paths. All the numerical values for the network capacity are computed by solving the program of (2.5) optimally using a Branch-and-Bound implementation. We note that also the greedy algorithm would give the optimal solution in the presented cases. However, rather than simply stating these capacity values, we attempt throughout to provide structural insights which might carry over to larger networks.

Basic examples

Single path For convenience, let us first explain the situation for the simple topology as shown in Fig. 2.2. In this network structure, there exists merely a single path from s to d. Hence, the only type of interference that matters is the self-interference within the path. Obviously, in order to transmit any data from source s to destination d, all nodes on the path must cooperate to forward the data. Moreover, the nodes will need to be utilized in a identical fashion for capacity optimization. The network capacity equals $1/3$ (i.e., the source transmits data $1/3$ of the time) for such a single-path topology, since nodes are not allowed to transmit simultaneously with neighboring nodes. This capacity result also holds for longer paths consisting of more than three transmitting nodes. In the case that the network reduces to a chain with only one or two transmitting nodes, the capacity would become 1 and $1/2$, respectively.

Independent paths Independent paths are paths from s to d that do not interfere with each other except that they share the resources at node s and finally come together in d. An example of such is given in Fig. 2.3. The capacity of the network is equal to the sum of the flows over all individual paths from s to d. If we choose to use only one of the paths, then the network capacity clearly equals $1/3$. In the case that we use both paths, it readily follows that a capacity of $1/2$ can be obtained. This is well below two times this single path capacity, but clearly more than the

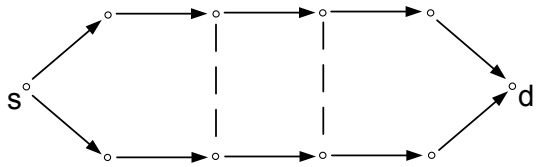

Figure 2.4: Bridges between paths.

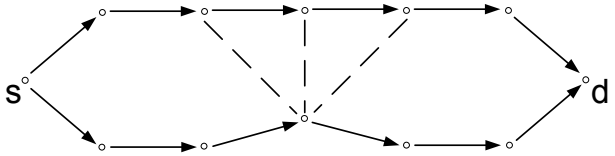

Figure 2.5: Claw structure.

1/3 result for the single path. Hence, in a situation with interference only at the endpoints of the paths, it is definitely favorable to split the traffic at the source and spread it over multiple paths as to optimize the network capacity.

Bridge structures Even for paths that are not independent (cf. the description in the previous paragraph), a capacity of 1/2 may be attained. A topology comprising several of these so-called "bridges" between the paths is presented in Fig. 2.4. In this example the bridges incur interference, but these cannot be employed to switch data packets from one path to the other. This may occur in situations where the routing protocol selects node-independent paths or where the interference range of a packet radio is larger than its transmission range. The fact that a capacity of 1/2 is still feasible is most easily seen by assuming equal flows per path. As each node interferes with at most three other nodes, this directly implies that each node is able to handle a capacity of 1/4. Hence, also each individual path can support a flow of 1/4, and thus the total capacity will equal 1/2. This example shows that interference between paths not necessarily has a negative impact on the capacity. In particular, this demonstrates that the interference metric as defined in [107] may not be an appropriate metric to quantify the impact of interference.

Claw structure Although little interference between paths need not be harmful, there are also situations in which interference has a nefarious effect on the network capacity. More specifically, there exist specific structures for which the capacity is equal to the single path capacity of 1/3. That means that utilizing an additional path is not beneficial with respect to the network capacity. This occurs when (at least) one node interferes with three (or even more) nodes on the other path; this situation is depicted in Fig. 2.5. As one node interferes with three nodes on its path (including itself) and also three nodes on the other path, it readily follows from the observation that under optimality nodes on the same path have identical utilization,

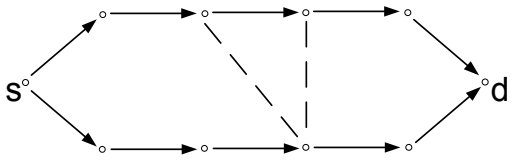

Figure 2.6: Triangular structure.

Scenario	Capacity
Single path	1/3
Indep. paths	1/2
Bridge	1/2
Claw	1/3
Triangle	3/7

Table 2.1: Network capacity for small topologies.

that the capacity is limited to 1/3. Hence, in a situation of two paths with at least one heavily interfering node, employing an extra path is not meaningful with regard to the capacity.

Triangular structure We have sketched above the extreme cases of (nearly) independent paths and paths that interfere so heavily that using them becomes useless. Of course, there remain some intermediate interference cases to investigate. Consider the situation that a single node is interfering with two nodes on the other path, leading to a triangular structure between the paths as shown in Fig. 2.6. The "bridge" and the "claw" example constitute capacity bounds for the "triangular" situation described here. Another observation is that this structure takes the symmetry of the topology away, so that symmetrical flows are not necessarily optimal anymore. To see this, suppose that we would assume symmetry, then since there exists one node that interferes with in total five other nodes (see Fig.2.6), the flow per path would be smaller or equal to 1/5. As a consequence, the network capacity would equal at most 2/5. Conversely, the (asymmetric) solution obtained by solving the program (2.5) shows that sending 2/7 over the upper path and 1/7 over the lower path is optimal and yields a capacity of 3/7 ($> 2/5$).

Combination of structures The previous topologies do not encompass a complete study of all possible structures between paths. An overview of their capacity values is given in Table 2.1. In particular, these investigated structures were all considered in isolation, while typically several of those structures might arise simultaneously between paths.

In the following example, we show that, contrary to what most multi-path routing protocols pursue, link-dependent (and thus also node-dependent) paths should sometimes be selected in order to obtain the optimal network capacity. Consider

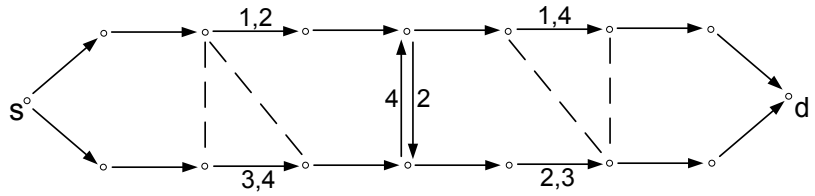

Figure 2.7: Dependent paths with mix of structures.

the topology of Fig. 2.7 comprising two triangular structures, one bridge as shown in Fig. 2.7, and paths numbered 1, 2, 3 and 4. Based on the investigation of the isolated triangular structure, we have an upper bound for the capacity of 3/7. Besides, we could symmetrically transmit 1/5 over the upper and the lower path (1 and 3, respectively) and obtain a capacity of 2/5; this would give us a lower bound. However, if additionally link-dependent paths would be used (by exploiting the bridge), then the capacity would indeed reach the upper bound, viz., 3/7. This is obtained by transmitting 1/7 over the paths 1, 3 and 4 and nothing over path 2. This at first sight paradoxical result follows from the fact that the nodes at the ends of the bridge are not fully utilized in the case of link-independent paths and hence this leaves some space for capacity improvements by deviating traffic over the bridge.

This example demonstrates that the typical procedure of multi-path routing protocols selecting independent paths needs not necessarily be optimal from a capacity point of view. Another interesting conclusion that can be drawn is that in certain cases it is optimal to split a traffic flow into several smaller subflows rather than maintaining all traffic together, even at intermediate nodes in the network.

Applied examples

Grids Rather than the illustrative but arbitrary graphs discussed in previous sections, let us consider here a network with a more general structure: a grid (or mesh) network (see Fig. 2.8). Assume that each of the intersections corresponds to a node and the edges indicate that node pairs can communicate. Typically for a grid network there exist many equal-length paths between source s and destination d. Employing (two) independent shortest paths would yield the optimal capacity of 1/2; however, one could argue that paths that more closely mimic the straight line between s and d should be preferred as those would interfere less with possible neighboring SD-pairs. In that case, we see (Fig. 2.8) that several basic structures, namely triangles (notice that due to the grid structure those show up here as squares) and bridges will appear between those paths. Hence, for a grid network the capacity will be dictated by the underlying structures between the paths.

Honeycomb graphs Another regular type of network arises in the situation in which nodes are distributed uniformly in a worst-case fashion with respect to interference. The network is then typically modelled as a collection of honeycombs as

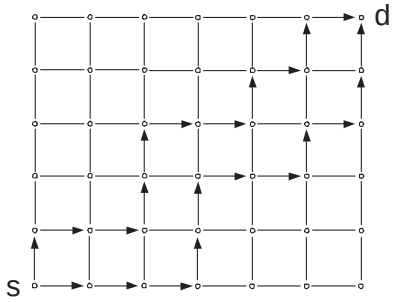

Figure 2.8: Network with grid structure.

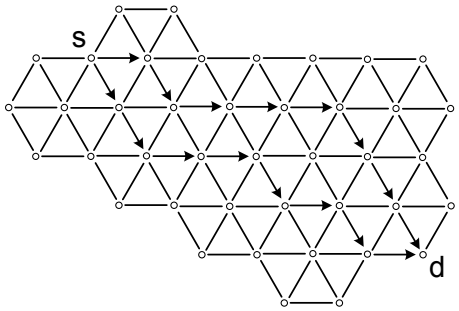

Figure 2.9: Network with honeycomb structure.

shown in Fig. 2.9. Using the same arguments as for the grid structure, let us skip the independent paths and focus on the situation of dependent paths. In such a situation, besides bridges and triangles, also claws can appear (see Fig. 2.9) in which case the capacity reduces to the single-path capacity, and using multiple paths is no longer meaningful. This poor performance can readily be explained as this is a worst-case interference model, since each node interferes with as much as six neighbors whereas in the grid structure nodes interfere with four neighboring nodes only.

2.4.1.2 General topologies

To assess the quality of our greedy algorithm, we have studied several general topologies. More precisely, we have constructed topologies comprising equal-length, parallel, node-independent paths and randomly generated links between nodes on adjacent paths according to independent Bernoulli experiments with success probability p. A sample scenario for such a topology is provided in Fig. 2.10 where the solid lines indicate the links on the node-independent paths and the dashed lines the randomly generated links between these paths. The interference and transmission range are

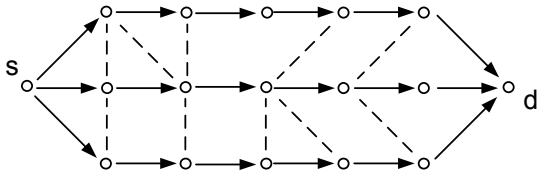

Figure 2.10: Example of a topology for 3 node-independent paths of 5 nodes each.

Scenario:	Opt.	Greedy	Err.>5%	Err.>10%	#LP
(paths, nodes, prob.)					
(3, 5, 0.40)	0.518	0.516	0%	0%	4.50
(3, 5, 0.80)	0.500	0.500	0%	0%	7.70
(4, 4, 0.40)	0.568	0.566	0%	0%	4.66
(4, 4, 0.80)	0.512	0.512	2%	0%	6.38
(5, 3, 0.40)	0.630	0.630	0%	0%	4.80
(5, 3, 0.80)	0.602	0.588	18%	10%	6.84

Table 2.2: Network capacity results for moderate-size topologies.

set to be equal, i.e., any link in the network induces interference, but can also be used for transmission. For all $j \in \mathcal{N}$, the neighborhoods $\mathcal{N}_{\mathcal{T}}(j)$ and $\mathcal{N}_{\mathcal{I}}(j)$ follow directly from this link set and we have $\mathcal{N}_{\mathcal{T}}(j) = \mathcal{N}_{\mathcal{I}}(j)$. For each scenario, we study the performance of the greedy algorithm by carrying out 50 runs of different random link configurations between the paths.

For moderate-size topologies (i.e., fewer than 20 nodes), we are still able to attain the optimal capacity by means of standard solvers. In Table 2.2, the average capacity values are presented for several scenarios and compared with the outcomes achieved by the greedy algorithm. Also included are the percentages of runs for which the greedy algorithm deviates more than 5% or 10% from the optimum and the average number of LPs solved per run. In the final scenario (5 paths of 3 nodes, $p = 0.80$), the greedy algorithm deviates more from the optimum, because it only succeeds in finding two independent paths while three such paths are present. However, the results show that, on average, the greedy algorithm performs close to optimal, that its solution rarely deviates far from the optimal capacity, while it requires only a limited computational effort (see the last column).

Scenario		p=0.4		p=0.8	
(paths, nodes)	Total nr. of nodes	All used	Greedy	All used	Greedy
(4, 6)	26	0.564	0.568	0.502	0.512
(5, 6)	32	0.616	0.624	0.522	0.590
(6, 6)	38	0.654	0.660	0.568	0.612
(7, 6)	44	0.684	0.692	0.612	0.646

Table 2.3: Network capacity results for large topologies.

For larger topologies (i.e., greater or equal than 20 nodes), exact solution approaches are no longer computationally feasible. On the contrary, we can rely on the greedy algorithm to find a fast approximation for the capacity. In Table 2.3, we compare the greedy approximation with the solution for the case that all interference constraints (see Eq. (2.8)) are taken into account and only a single LP problem is to be solved; this corresponds to replacing Eq. (2.8) by Eq. (2.3). The comparison shows that for $p = 0.8$ our approach, which comprises the more sophisticated program, yields a much better approximation for the capacity, while for $p = 0.4$ the approximations are similar. We can conclude that application of the greedy algorithm is definitely valuable in situations with heavy interference. Another observation from our experiments is that for high values of p, the source forwards traffic via fewer nodes than for low values of p. This suggests that in the case of heavy interference only a few paths must be used.

2.4.2 Advanced scenarios

2.4.2.1 Multiple source-destination pairs

Many networks are employed in areas where typically more than a single subject wants to transmit data. This case of multiple SD-pairs is not essentially different from a modelling perspective. However, for each link in the network one should record for which SD-pair it is transferring data and hence the number of variables increases linearly in the number of SD-pairs. Therefore, we restrict ourselves to the situation of two (intersecting) SD-pairs. Of course, in the situation where the pairs are far apart they act independently, and therefore we focus on the scenarios in which the pairs really have to share the medium to get their data through (see, e.g., Fig. 2.11). The objective function $h(R_t, P)$ differs now from the single SD-pair case. We do not want merely to optimize the total capacity (as one pair may consume all capacity, while the other pair may starve), but a function of the individual capacities per SD-pair. More specifically, we let $h(R_t, P) = \rho^{(1)}_{t,s_1} + \rho^{(2)}_{t,s_2}$, and impose $\rho^{(2)}_{t,s_2} = \frac{1-c}{c} \cdot \rho^{(1)}_{t,s_1},\ 0 \leq c \leq 1$. That means that, e.g., for $c = 1/2$, both flows are equally important, while for $c > 1/2$, SD-pair 1 is prioritized. In this way, one may differentiate between various priority classes of traffic.

We consider three scenarios: two single-path SD-pairs, one single-path and one multi-path SD-pair, and two multi-path SD-pairs; the latter one is depicted in Fig. 2.11. In all cases the paths are chosen long enough to avoid boundary effects as our focus is more on structural insights. All results are determined only using the approximate greedy algorithm as the experiments in the previous section have proven its good performance.

The capacity of the first scenario is not hard to assess. The central node can still be active at most a fraction 1/3 of the time, so that the capacity equals 1/3 independent of the value for c.

The second scenario refers to the case with three node-independent paths available for the first pair and one for the second. Obviously, for this scenario both the value

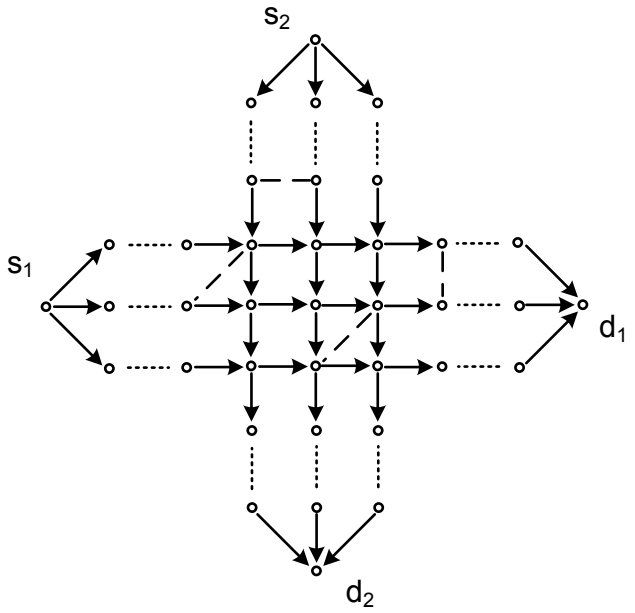

Figure 2.11: Two multi-paths.

of c and of the link probability p will play a role. The role of p is the same as in the previous subsection, so that links are added according to a Bernoulli experiment between adjacent nodes at the paths per SD-pair. That means that we could differ this value per SD-pair, however here we take these values equal for both SD-pairs. We discuss here only the extreme cases $p = 0$ and $p = 1$. The capacity results are presented in Table 2.4 (left). It appears that if the multi-path gets higher priority, then the capacity increases. This makes sense as in isolation the multi-path SD-pair would attain a capacity of 3/5 (for $p = 0$) or 1/2 (for $p = 1$), whereas the single path would only have a capacity of 1/3.

The last scenario we present comprises two interacting multi-paths as depicted in

c	Network capacity	
	p=0	p=1
1/4	0.400	0.381
1/2	0.476	0.444
3/4	0.571	0.533

c	Network capacity	
	p=0	p=1
1/4	0.666	0.593
1/2	0.739	0.480
3/4	0.666	0.593

Table 2.4: Network capacity results for the case of one single-path and one multi-path SD-pair (left) and two multi-path SD-pairs (right).

Scenario	Neighborhoods	1 path	3 paths		5 paths	
		p=0/1	p=0	p=1	p=0	p=1
Basic scenario	$\mathcal{N}_T(j) = \mathcal{N}_I(j) = \mathcal{N}_{0/1}(j)$	0.333	0.600	0.500	0.714	0.583
Ext. interference	$\mathcal{N}_T(j) = \mathcal{N}_{0/1}(j)$, $\mathcal{N}_I(j) = \mathcal{N}^*_{0/1}(j)$	0.200	0.333	0.250	0.429	0.382
Ext. power	$\mathcal{N}_T(j) = \mathcal{N}_I(j) = \mathcal{N}^*_{0/1}(j)$	0.333	0.500	0.400	0.600	0.529

Table 2.5: Network capacity results for different neighborhoods.

Fig. 2.11. Also in this case the value of c influences the capacity, however this takes now place in a symmetrical fashion. The capacity results for this case are presented in Table 2.4 (right). Under this scenario the extra links do not appear very helpful as the capacity decreases for $p = 1$. Besides, favoring one flow at the cost of the other is not good from a capacity viewpoint as for $p = 0$ the capacity ultimately will reach the single SD-pair capacity of 3/5 (for $c = 0$ and $c = 1$).

2.4.2.2 Multiple ranges

As a final illustration, in accordance with more realistic applications, we analyze different interference and transmission neighborhoods. We construct scenarios which conform to the approach in Sect. 2.4.1.2 with link probabilities $p = 0$ and $p = 1$ (for a single SD-pair). For these choices of p, the neighborhoods $\mathcal{N}_T(j)$ and $\mathcal{N}_I(j)$ are fixed and identical for a given $j \in \mathcal{N}$. Let us denote these neighborhoods by $\mathcal{N}_0(j)$ and $\mathcal{N}_1(j)$, respectively. Next, we define extended neighborhoods as follows. For $p = 0$, let $\mathcal{N}^*_0(j)$ extend $\mathcal{N}_0(j)$ with the three parallel nodes on the two nearest paths and with the two adjacent nodes on the same path. For $p = 1$, let $\mathcal{N}^*_1(j)$ extend $\mathcal{N}^*_0(j)$ with the three parallel nodes of the paths located next to the adjacent paths (whenever present). Thus, we have that $\mathcal{N}^*_0(j) \supseteq \mathcal{N}_0(j)$ and $\mathcal{N}^*_1(j) \supseteq \mathcal{N}_1(j)$. We stress that the exact structure of the neighborhoods is not especially important here. Our intention is to gain some understanding of what might happen in realistic cases. In practice, the exact neighborhoods could be constructed based on network measurements.

We have evaluated scenarios with 1, 3 and 5 node-independent paths consisting of six "single-links" from source to destination. In Table 2.5, the capacity results (obtained via our greedy algorithm) for the different neighborhoods and link probabilities are presented. There are two main conclusions that can be drawn from our experiments. First, increasing the interference range for a fixed transmission range has a clear negative effect on the capacity, and this is shown to hold for all scenarios. Second, also increasing both ranges (in an equal fashion) has a (minor) negative effect on the capacity, except for the case of a single path. Further, we observe that the solutions found comprise fewer paths when the interference range is extended; this indicates again that the impact of interference on path selection cannot be ignored.

2.5 Discussion

A key assumption we made is that communication links in the network are error-free. However, in a wireless environment links are almost never free of errors. Thus, this is a valid assumption only if lossy links are excluded in advance (cf. the approach in, e.g., [82]). Otherwise, one should follow another approach which exploits the fact that the network is likely to operate in a stable fashion for a certain period of time. It is then useful to initially test all the links between the nodes in the network. Subsequently, only links with sufficiently low error-rate should be included in the topology used for capacity optimization. Then, these links will still not be fully error-free, but at least they can be assumed to have all a low error rate. Although the remaining errors still lead to retransmissions which costs an amount of capacity, this can now readily be incorporated in our model by adjusting the transmission rates per node. These adjusted rates can again be normalized.

Another important element of our modelling approach is the construction of interference and transmission neighborhoods. Our model facilitates to consider any neighborhood of interest; in particular, the neighborhoods do not need to depend on the distance between nodes. This means that by performing some network measurements before the actual network operation starts, the neighborhoods can accurately be defined. Subsequently, the capacity of the constructed network can be determined using the techniques presented in this chapter.

2.6 Concluding remarks

Signal interference plays an important part in the performance analysis of mobile ad hoc networks. Despite of this, the impact of interference remains often underexposed in the development of novel multi-path protocols. The focus in this development is rather on selecting independent paths than on a more profound investigation of the network structure. As in practice the number of independent paths available is often limited or may yield great diversity in path lengths, it is of the utmost importance to study the network performance in situations where independence cannot simply be exploited.

In this chapter, we presented a general stochastic framework for the analysis of network performance under interference for a multi-hop ad hoc network in the absence of mobility. We specialized to a mathematical model to assess the network capacity in a multi-path environment. A nonlinear programming problem formulation incorporating interference was introduced and it was shown that the global optimum can be obtained by solving a number of linear programming problems. To this end, efficient solution techniques were proposed so that the computational effort to actually obtain this optimum remains limited. Then, we have given a classification of interference situations in a network consisting of two paths by presenting several illustrative examples. In addition, several results were presented for the case of two source-destination pairs and for the case of multiple ranges.

These examples demonstrate that paths do not require to be fully independent in order to attain the optimal capacity under interference. Moreover, using multiple paths that moderately interfere appears favorable over using only a single independent path. Only in situations with heavily interfering nodes, it becomes pointless to use multiple paths, at least from a capacity optimization point of view. It is further enunciated that there even exist situations in which it is optimal to exert additional link-dependent paths to reach the optimal capacity. This makes clear that many protocols which consider only independent paths yield merely suboptimal capacity values. The numerical results indicate that the capacity may drastically reduce in scenarios for which the interference neighborhood is much larger than the transmission neighborhood. More surprisingly, when both interference and transmission range are increased, the network capacity decreases significantly compared to the original situation.

CHAPTER

3

Stability of two exponential time-limited polling models

3.1 Introduction

In this chapter, we will state and prove the stability conditions for the pure and exhaustive service discipline in the context of polling models. The pure time-limited discipline states that the server visits a queue exactly for a random amount of time, while according to the exhaustive time-limited discipline the server already leaves a queue as soon as it becomes empty. Both service disciplines are preemptive disciplines meaning that any on-going service at the time limit will be interrupted and must be restarted at a next visit. The stability conditions prescribe limits on the amount of traffic that can be sustained by the system. Exceeding these limits leads to instable behavior, but operation of the system just below these limits will already lead to large buffers and long transfer delays.

Stability conditions for a large class of polling systems have been proven by Fricker and Jaïbi [37]. In particular, the authors considered periodic polling systems under non-preemptive and work-conserving service disciplines. The necessary and sufficient condition for stability reads:

$$\text{System is stable} \iff \rho + \max_{1 \leq i \leq M} \left(\frac{\lambda_i}{\mathbb{E}[G_i^*]} \right) \cdot c_T < 1,$$

where ρ is the total offered load to the system, c_T is the mean total switch-over time during a cycle and $\mathbb{E}[G_i^*]$ denotes the mean number of served customers at Q_i during a cycle when Q_i is saturated. Saturation in this context means that there is an unlimited number of customers waiting to be served at a polling instant to Q_i. For the exhaustive and gated service discipline, this condition readily simplifies to:

$$\text{System is stable} \iff \rho < 1,$$

since the number of served customers may grow to infinity in these cases, i.e., $\mathbb{E}[G_i^*] = \infty$. Later, the same authors have also proven a similar stability condition for the Markovian polling strategy [38]. However, stability results for preemptive service disciplines, such as the pure and exhaustive time-limited disciplines, are not available in the literature.

To close this gap, we will prove stability conditions for both time-limited disciplines. This will be done for the periodic polling strategy and an exponential time limit. For the pure-time limited discipline, the stability question can be resolved by studying each queue in isolation. For the exhaustive time-limited discipline, stability must be considered for the system as a whole and we will prove the stability condition by adopting the line of proof of [37].

This chapter is organized as follows. The model is formally described in Sect. 3.2. Next, the stability conditions for the pure and the exhaustive time-limited disciplines are given in Sects. 3.3 and 3.4, respectively. We conclude the chapter in Sect. 3.5.

3.2 Model

Let us consider the basic polling system of M queues with Poisson arrivals and generally distributed service and switch-over times. The server visits the queues according to the periodic polling strategy. Without loss of generality (w.l.o.g.) we define a cycle as the time period between two consecutive polling instants at the 1st stage (or visit) of the cycle. A cycle consists of a stages and we denote by $t(j),\ j = 1, \ldots, a$, the queue served during stage j of the cycle. Further, the number of times Q_i is visited during a cycle is denoted by $a_i,\ i = 1, \ldots, M$, with $a_i \geq 1$ and $\sum_{i=1}^{M} a_i = a$.

The service discipline assumed in Sect. 3.3 is the pure time-limited discipline, whereas in Sect. 3.4 the exhaustive time-limited discipline is assumed. The time limit at Q_i for both disciplines is exponentially distributed with parameter ξ_i. Due to the time limit, service will be preempted at the timer expiration and in such case a service time will be redrawn from the original distribution at the start of the next visit; thus, we assume the so-called *preemptive-repeat with resampling* strategy.

3.3 Pure exponential time-limited discipline

For the pure time-limited discipline, the queues in the system are independent from a stability perspective as service capacity cannot be exchanged among the queues. The polling system is stable if and only if there exists a stationary regime in which each customer in the system can be served in a finite period of time. We will say that the system is stable if and only if all the queues in the system are stable.

A necessary and sufficient condition for the stability of a polling system with the server operating under the pure exponential time-limited discipline is given in the following theorem.

Theorem 3.1 (Pure exponential time-limited discipline)**.**

$$\textit{System is stable} \iff \rho_i < \kappa_i, \ \forall_{i\in\{1,\ldots,M\}},$$

where

$$\begin{aligned} \rho_i &= \lambda_i \cdot \frac{1-\tilde{X}_i(\xi_i)}{\xi_i \cdot \tilde{X}_i(\xi_i)}, \\ \kappa_i &= \frac{a_i/\xi_i}{\sum_{j=1}^{M} a_j/\xi_j + \sum_{k=1}^{a} c_{t(k),t(k+1)}}, \end{aligned}$$

where $c_{t(k),t(k+1)}$ is the mean switch-over time from the queue visited in stage k to the queue visited in stage $k+1$.

Proof. It is well-known that for a single queue the nonsaturation condition is both a necessary and sufficient condition for stability, i.e.,

$$Q_i \text{ is stable} \iff \rho_i < \kappa_i, \ i = 1, \ldots, M,$$

where ρ_i is the mean effective amount of work arriving per time unit to Q_i and κ_i is the availability fraction of the server at Q_i.

Consider first the mean effective amount of work arriving per time unit to Q_i. This amount is determined by the total number of customers arriving per time unit λ_i and the mean effective amount of work each individual brings for the server, denoted by $\tilde{\sigma}_i$, as follows:

$$\rho_i = \lambda_i \cdot \tilde{\sigma}_i.$$

The quantity $\tilde{\sigma}_i$ is in fact the mean total time the server spends on serving a customer at Q_i including any interrupted services. Noting that the number of interruptions per customer is geometrically distributed, it can be found via simple calculus that:

$$\tilde{\sigma}_i = \frac{1-\tilde{X}_i(\xi_i)}{\xi_i \cdot \tilde{X}_i(\xi_i)}.$$

The availability fraction of the server κ_i is fully specified by the mean visit times, the visit frequencies and the switch-over times between the queues. Notice that a complete cycle consists of a_i visits to Q_i, $i = 1, \ldots, M$, and the switch-over times between the queues. It then readily follows for the availability fraction of the server at Q_i:

$$\kappa_i = \frac{a_i/\xi_i}{\sum_{j=1}^{M} a_j/\xi_j + \sum_{k=1}^{a} c_{t(k),t(k+1)}}.$$

It is good to notice that the fraction κ_i is independent of the load at the queues. The observation that the system is stable if and only if all the queues in the system are stable completes the proof. □

3.4 Exhaustive exponential time-limited discipline

For the exhaustive time-limited discipline, service capacity can be exchanged between the queues. This suggests that stability must be considered for the system as a whole. However, as the visit time to each queue is bounded by the timer, the occupancy of individual queues also plays a role. The polling system is considered stable if there exists a stationary regime in which each customer in the system can be served in a finite period of time.

A necessary and sufficient condition for the stability of a polling system with the server operating under the exhaustive exponential time-limited discipline is given in the following theorem.

Theorem 3.2 (Exhaustive exponential time-limited discipline)**.**

$$\textit{System is stable} \iff \rho + \max_{1 \le i \le M} \left(\frac{\lambda_i}{\mathbb{E}[G_i^{*-}]} \right) \cdot c_T < 1,$$

where

$$\mathbb{E}[G_i^{*-}] = \frac{a_i \cdot \tilde{X}_i(\xi_i)}{1 - \tilde{X}_i(\xi_i)},$$

denotes the mean number of served customers at Q_i during a cycle when Q_i is saturated and c_T is the mean total switch-over time during a cycle.

We will prove the theorem adopting the approach of Fricker and Jaïbi [37]. To this end, we will often stick to their notation whenever it does not lead to ambiguity. We should emphasize that the authors of [37] considered only work-conserving service disciplines. The exhaustive time-limited (E-TL) discipline allows for preemption of service and thus is definitely not work conserving.

The organization of the proof is as follows. In Sects. 3.4.1 and 3.4.2, we state several preliminary and monotonicity results which are analogous to the results in

[37]. Essentially, we need to introduce notation to account for the preemption of the service, but the line of proof remains similar. Hence, the lemmas and theorems corresponding directly to the ones provided in [37] will be given without proof. In Sect. 3.4.3, we present several novel results for the visit time to the queues during a cycle and also give the proofs. These results are then incorporated in the final necessary and sufficiency proofs of [37], so that these account for preemptive service.

3.4.1 Preliminaries and stochastic monotonicity

The general service disciplines for which stability is proven in [37] should satisfy four properties. Property 1 and 3 refer to the independence of the service discipline on the history of the service process and on the independence of the customer selection. These properties are readily seen to be satisfied for the E-TL discipline. Property 2 deals with the work conservation and is not satisfied for this discipline since work is created due to preemptions. However, during the course of a visit the server is always working and does not idle. Finally, Property 4 is the so-called stochastic monotonicity property and is defined as follows [37]: "*As the queue size grows, the number of customers served during one stage (visit) grows stochastically, but such that the number of customers left at the end of the stage (visit) grows stochastically as well.*" This latter property plays a crucial role in the proof.

Let us w.l.o.g. consider an arbitrary queue in the polling system. First, we define an independent and identically distributed (i.i.d.) sequence $(\sigma^m)_{m=1,2,\ldots}$, as the modified service times of a customer, with mean σ, and being distributed as $\min(X, V)$, where X is distributed according to the original service time distribution, and V is exponentially distributed and independent of X. That is, the modified service times can be seen as the duration of a service attempt (which can either be successful or interrupted). For non-preemptive service disciplines, the number of customers taken into service is equal to the number of customers served during a visit. However, this is not always true for the preemptive discipline that we consider here. Hence, we will define also the following quantities for a visit with x customers present at the start (i.e., $t = 0$):

- $f^+(x)$: the number of customers that is taken into service during the visit;
- $f^-(x)$: the number of customers that is actually served during the visit;
- $v(x)$: the duration of the visit;
- $\phi(x)$: the number of customers at the end of the visit.

The quantities $f^+(x)$ and $f^-(x)$ are given by:

$$\begin{aligned} f^+(x) &= \min(N^0(x), N^*), \\ f^-(x) &= \min(N^0(x), N^* - 1), \end{aligned} \tag{3.1}$$

where $N^0(x)$ refers to the number of served customers during a visit which started with x customers and which ends due to the queue becoming empty, and N^* refers to the number of customers taken into service during a visit which ends due to the expiration of the time limit. Notice that, due to the exponential visit times, N^* is in fact a geometrically distributed random variable independent of x.

Let us denote by $N(a,b]$ the number of arrivals to the queue during the interval $(a,b]$. Thus, we may write the following relations between $f^+(x)$, $f^-(x)$, $v(x)$ and $\phi(x)$:

$$v(x) = \sum_{m=1}^{f^+(x)} \sigma^m, \tag{3.2}$$

$$\phi(x) = x - f^-(x) + N(0, v(x)], \tag{3.3}$$

with $f^+(0) = f^-(0) = v(0) = 0$. It is good to notice that σ^m in Eq. (3.2) refers to the duration of an arbitrary service attempt rather than an original service time of an arbitrary customer.

Let us next recall the definitions of $\leq$-monotonicity and $\leq_d$-monotonicity as given in [37]:

Definition 3.3. *($\leq$-monotonicity)*
A real function h defined on $\mathbb{R}^n$ is called $\leq$-monotone when:

$$x \leq y \Rightarrow h(x) \leq h(y).$$

Definition 3.4. *($\leq_d$-monotonicity)*
Two (cumulative) distributions functions P_1 and P_2 on $\mathbb{R}^n$ satisfy $P_1 \leq_d P_2$ when:

$$\int h \; dP_1 \leq \int h \; dP_2,$$

for any $\leq$-monotone function h such that the integrals are well defined.
Two random vectors X_1 and X_2 satisfy $X_1 \leq_d X_2$ if their distributions satisfy $P_1 \leq_d P_2$.

Hence, the monotonicity property for the E-TL discipline is that $(f^+(x), f^-(x), \phi(x))$ is $\leq_d$-monotone in x. It follows immediately from Eq. (3.2) that $\leq_d$-monotonicity of $f^+(x)$ implies $\leq_d$-monotonicity of $v(x)$, and that $\leq_d$-monotonicity of $f^-(x)$ does not imply that of $\phi(x)$.

Next, we embed the queue into the polling system. Let the nth visit to the queue start at stopping time T_n (with respect to the complete history of the system) with N_n customers waiting. Define the following quantities:

- F_n^+: the number of customers that is taken into service during visit n;
- F_n^-: the number of customers that is actually served during visit n;

- V_n: the duration of visit n;
- Φ_n: the number of customers at the end of visit n.

Let us introduce the tuple (f^+, f^-, v, ϕ) which represents the service discipline. It can readily be argued (cf. [37, p.215]) that for each n:

$$(F_n^+, F_n^-, V_n, \Phi_n) =_d (f^+(N_n), f^-(N_n), v(N_n), \phi(N_n)).$$

Along the single-queue equations, Eqs. (3.2) and (3.3), we find that for any n, V_n and Φ_n are related to F_n^+ and F_n^- as follows.

$$\begin{aligned} V_n &= \sum_{i=D_n+1}^{D_n+F_n^+} \sigma^i, \\ \Phi_n &= N_n - F_n^- + N(T_n, T_n + V_n], \end{aligned}$$

where D_n denotes the number of service attempts performed up to T_n. Since $(F_n^+, F_n^-, V_n, \Phi_n)$ is independent of future service attempt durations, i.e., $(\sigma^i)_{i>D_n+F_n^+}$, and the future customer arrival process, i.e., $N(T_n + V_n, T_n + V_n + \cdot]$, we may apply Wald's equation and obtain:

$$\mathbb{E}[V_n] = \mathbb{E}[F_n^+] \cdot \sigma, \tag{3.4}$$

$$\mathbb{E}[N(T_n, T_n + V_n]] = \mathbb{E}[F_n^+] \cdot \lambda \cdot \sigma. \tag{3.5}$$

Notice that the expectations in (3.4) and (3.5) are finite, since the visit duration is always bounded by the exponential timer.

Let F^{*+} (F^{*-}) be the number of customers that are taken into service (served) during a visit if there are infinitely many customers waiting in the queue, and V^* the duration of such a visit, i.e.,

$$\begin{aligned} 0 < \lim_{x\to\infty} \mathbb{E}[f^+(x)] = \mathbb{E}[F^{*+}] &< \infty, \\ 0 < \lim_{x\to\infty} \mathbb{E}[f^-(x)] = \mathbb{E}[F^{*-}] &< \infty, \end{aligned}$$

and also,

$$\lim_{x\to\infty} \mathbb{E}[v(x)] = \mathbb{E}[V^*] = \mathbb{E}[F^{*+}] \cdot \sigma < \infty.$$

Next, we present a lemma which will be needed in the final part of the proof. This lemma substitutes in fact Lemma 1 of [37].

Lemma 3.5. *Let $(N_n)_n$ be a sequence of random variables converging in distribution to a, possibly degenerate, integer-valued random variable N. Let (f^+, f^-, v, ϕ) be induced by the E-TL service discipline and be independent of $(N, (N_n)_n)$, i.e., N^* is independent of $(N, (N_n)_n)$ (see, Eq.(3.1)). The sequence $(N_n, f^+(N_n), f^-(N_n), v(N_n), \phi(N_n))_n$ converges in distribution to $(N, f^+(N), f^-(N), v(N), \phi(N))$, and*

(i) when $\mathbb{E}[F^{*-}] < \infty$*, and if* N *has a defective distribution, then so is the limiting distribution of* $N_n - f^-(N_n)$*;*
(ii) when $\mathbb{E}[F^{*-}] < \infty$*,* $\mathbb{E}[F^-(N)] < \mathbb{E}[F^{*-}]$ *if and only if there exists a* $y < \infty$ *such that* $\mathbb{P}(N \leq y) > 0$ *and* $\mathbb{E}[f^-(y)] < \mathbb{E}[F^{*-}]$*.*
In both cases, if $(N_n)_n$ *is* $\leq_d$*-monotone,* $\lim_{n\to\infty} \mathbb{E}[F^-(N_n)] = \mathbb{E}[f^-(N)]$ *and* $\lim_{n\to\infty} \mathbb{E}[v(N_n)] = \mathbb{E}[v(N)]$*.*

Proof. The proof is immediate from the proof of Lemma 1 in [37]. □

Remark 3.6 (Number of customers taken into service)**.** *We have defined Lemma 3.5 in terms of the number of customers served. Analogously, this lemma can be defined for the number of customers taken into service.*

Finally, the following lemma is essential for the remainder of the proof and refers to a specific property of the service discipline addressed in the beginning of this section.

Lemma 3.7. *The E-TL discipline satisfies the stochastic monotonicity property.*

Proof. The proof follows by sample-path arguments and is immediate from the proof of Lemma 2 in [37]. □

3.4.2 Monotonicity

The stochastic monotonicity property plays a key role in the stability proof. Therefore, we will state several monotonicity results [37] which are valid for service disciplines satisfying this property.

To this end, we describe the system by the queue lengths at the polling instants and define $M(t)$ as follows:

$$M(t) = (N_1(t), \ldots, N_M(t)),\ t \geq 0.$$

Recall that $t(i)$ denotes the queue served at visit i of a cycle. We denote by visit (n, i) the ith visit in the nth cycle and let visit $(1, 1)$ start at time $t = 0$. Let $T_{n,i}$ denote the time of the polling instant of visit (n, i), so that we have:

$$0 = T_{1,1} \leq T_{1,2} \leq \ldots \leq T_{1,a} \leq T_{2,1} \leq \ldots\ .$$

For convenience, we write $M_{n,i}$ for $M(T_{n,i})$ and $N_{n,i}$ for $N_{t(i)}(T_{n,i})$. Hence, we can describe the Markovian behavior of the system as follows.

Proposition 3.8. *(Prop. 1 of [37]) The sequence* $(M_{n,i})_{n,i}$ *is a Markov chain. For each* i *fixed in* $\{1, \ldots, a\}$*, the Markov chain* $(M_{n,i})_n$ *is homogeneous, aperiodic and irreducible on (a subset of)* $\mathbb{N}^M$*.*

Proof. See [37]. □

Let us define by π_i the transition operator at visit i, $1 \leq i \leq a$, of the Markov chain $(M_{n,i})_{n,i}$ as follows:

$$\pi_i h(\mathbf{m}) = \mathbb{E}[h(M_{n,i+1})|M_{n,i} = \mathbf{m}],$$

for any $\mathbf{m} = (m_1, \ldots, m_M)$ and any real function h defined on $\mathbb{N}^M$ for which the expectation exists. Besides, we let $\tilde{\pi}$ be the transition operator of the Markov chain $(M_{n,i})_n$. An operator π is said to be $\leq_d$-monotone if for all distributions $P_1 \leq_d P_2$, $\pi P_1 \leq_d \pi P_2$. This holds if πh is $\leq$-monotone when h is.

Lemma 3.9. *(Lemma 3 of [37]) For all i, π_i and $\tilde{\pi}_i$ are $\leq_d$-monotone.*

Proof. See [37]. □

Let us next define the following quantities:

- $F^+_{n,i}$: the number of customers taken into service during visit (n, i);
- $F^-_{n,i}$: the number of customers served during visit (n, i);
- $V_{n,i}$: the duration of visit (n, i).

An immediate consequence of Lemma 3.9 is the monotonicity property of the state process.

Proposition 3.10. *Suppose $M_{1,1} = (0, \ldots, 0)$. Then, for each i, $(M_{n,i})_n$ and $(F^+_{n,i}, F^-_{n,i}, V_{n,i})$ are $\leq_d$-monotone.*

Proof. The proof is immediate from the proof of Proposition 2 in [37]. □

Next, we turn to dominance relations between polling systems. In particular, we compare systems with a different number of saturated queues. Here, saturation means that at a polling instant of a queue there is an infinite number of customers waiting. The saturation of a queue implies that the server serves the queue up to the time limit and then leaves. From the viewpoint of the other queues in the system, such a visit to a saturated queue is merely an additional switch-over time. Let $\mathcal{S}$ be the initial polling system with queues $1, \ldots, M$. For $e \in \{0, \ldots, M\}$, we define the subsystem $\mathcal{S}^e$ as the polling system consisting of the queues $1, \ldots, e$, resulting from the saturation of the queues $e+1, \ldots, M$, and served according to the same periodic schedule as the original system. We emphasize that if $t(i) > e$ then in $\mathcal{S}^e$ no queue is served but the server becomes unavailable for a duration of $V^*_{t(i)}$, which is defined as the stationary duration of a visit to queue $t(i)$ with an infinite number of customers waiting at the start of the visit. Let us define σ_j as the mean duration of a service attempt at Q_j, i.e., $\sigma_j := \mathbb{E}[min(X_j, V_j)]$, where X_j refers to the original service time of a customer at Q_j and V_j to the visit time of the server to Q_j. Further, denote by $\mathbb{E}[G^{*+}_j]$ the expected number of customers taken into service at Q_j during a cycle

when Q_j is saturated. Then, the mean total switch-over time in the subsystem c_T^e can be written as:

$$c_T^e = c_T + \sum_{j=e+1}^{M} \sigma_j \mathbb{E}[G_j^{*+}] = c_T + \sum_{j=e+1}^{M} a_j/\xi_j.$$

The state space of the subsystem $\mathcal{S}^e$ is given by the sequence $M_{n,i}^e = (N_1^e(T_{n,i}^e), \ldots, N_e^e(T_{n,i}^e))$ at the polling instants $T_{n,i}^e$. For each visit i, $(M_{n,i}^e)_n$ is a Markov chain and is $\leq_d$-monotone if the initial state is the empty state. The subsystem $\mathcal{S}^e$ is similar to the original system $\mathcal{S}$ in the sense that all previous results apply to it. Let denote by $M^{g|e}$ the e first components of a vector M^g having $g > e$ components. Then, the subsystems $\mathcal{S}^e$ satisfy the following dominance property.

Lemma 3.11. *(Lemma 4 of [37]) For $e < g$ both in $\{0, \ldots, M\}$, $\mathcal{S}^e$ dominates $\mathcal{S}^g$ in the sense that if $M_{1,1}^{g|e} \leq_d M_{1,1}^e$ then $M_{n,i}^{g|e} \leq_d M_{n,i}^e$ for all (n, i).*

Proof. See [37]. □

3.4.3 Stability proof

The polling system is said to be stable if:

- there exists a proper stationary joint-distribution for the queue lengths at the polling instants at stage k, for all $k = 1, \ldots, a$;
- the stationary cycle time is finite.

3.4.3.1 Proof: Sufficient condition

We assume w.l.o.g. that the system is empty at time 0 as the stationary distribution of the Markov chain does not depend on the initial distribution. For convenience, let us introduce several definitions for the number of customers at a specific queue, viz.,

- H_k^- : number of customers actually served at Q_k during a visit to Q_k;
- H_k^+ : number of customers taken into service at Q_k during a visit to Q_k;
- H_k^{*-} : number of customers actually served at Q_k during a visit when Q_k is saturated;
- H_k^{*+} : number of customers taken into service at Q_k during a visit when Q_k is saturated.

Notice that these definitions resemble the definitions of F_n^-, F_n^+, F_n^{*-} and F_n^{*+}. However, the latter quantities refer to the number of customers at the nth visit rather than to the number at a specific queue.

W.l.o.g. we consider the cycle from $T_{n,1}$ to $T_{n+1,1}$. Then, we may similarly define the counterparts G_k^-, G_k^+, G_k^{*-} and G_k^{*+} which count the same quantities but over a complete cycle. Hence, we may then also write:

$$\begin{aligned}
\mathbb{E}[G_k^-] &:= \mathbb{E}[H_{k,1}^-] + \cdots + \mathbb{E}[H_{k,a_k}^-],\\
\mathbb{E}[G_k^+] &:= \mathbb{E}[H_{k,1}^+] + \cdots + \mathbb{E}[H_{k,a_k}^+],\\
\mathbb{E}[G_k^{*-}] &:= \mathbb{E}[H_{k,1}^{*-}] + \cdots + \mathbb{E}[H_{k,a_k}^{*-}],\\
\mathbb{E}[G_k^{*+}] &:= \mathbb{E}[H_{k,1}^{*+}] + \cdots + \mathbb{E}[H_{k,a_k}^{*+}],
\end{aligned}$$

where $\mathbb{E}[H_{k,i}^-]$ is the mean number of customers served at Q_k during the ith visit to Q_i in a cycle, and $\mathbb{E}[H_{k,i}^+]$, $\mathbb{E}[H_{k,i}^{*-}]$, and $\mathbb{E}[H_{k,i}^{*+}]$ are defined similarly. Besides, we define $\tilde{\sigma}_k$ as the mean effective service time of a customer at Q_k. The effective service time refers to the total time spent by the server on serving a customer (including interrupted service attempts, but excluding the periods that the server is not present at the queue) and it is in fact a geometric sum of service attempt durations. Thus, using Wald's equation, we may write for its mean:

$$\tilde{\sigma}_k = \mathbb{E}\left[\sum_{n=1}^{N} \min(X_{k,n}, V_{k,n})\right] = \sigma_k / \tilde{X}_k(\xi_k), \tag{3.6}$$

where N is geometrically distributed with success probability $p = \tilde{X}_k(\xi_k)$, and $\{X_{k,n}\}_{n\geq 1}$ and $\{V_{k,n}\}_{n\geq 1}$ are two independent families of independent random variables distributed as X_k and V_k, respectively.

Let us denote by $\mathbb{E}[V_k^c]$ the mean total visit time to Q_k during a cycle, i.e.,

$$\mathbb{E}[V_k^c] = \mathbb{E}[V_{k,1}] + \cdots + \mathbb{E}[V_{k,a_k}],$$

where $\mathbb{E}[V_{k,j}]$ stands for the mean visit time during the jth visit to Q_k of a cycle. Then, we are ready to present the following lemma:

Lemma 3.12.

$$\mathbb{E}[V_k^c] = \mathbb{E}[G_k^-] \cdot \tilde{\sigma}_k,\ k = 1, \ldots, M. \tag{3.7}$$

The proof of the lemma will be given below. However, we will derive several intermediate results first.

Clearly, when Q_k is saturated, there is always exactly one interrupted service. Thus, we have the following property:

Property 3.13.

$$H_k^{*+} = H_k^{*-} + 1,\ k = 1, \ldots, M,$$

and since $\mathbb{E}[H_k^{*+}] < \infty$ *also:*

$$\mathbb{E}[H_k^{*+}] = \mathbb{E}[H_k^{*-}] + 1,\ k = 1, \ldots, M.$$

Besides, there is a less obvious relation between the quantities H_k^+, H_k^-, H_k^{*+} and H_k^{*-}. However, before we get to this relation, we give a lemma and present some useful properties for H_k^+ and H_k^-.

Lemma 3.14. *Let H be a geometrically distributed random variable and let W be a non-negative discrete random variable independent of H. Then, the following assertion holds:*

$$\mathbb{E}[H\,\mathbf{1}_{\{W\geq H\}}] + \mathbb{E}[W\,\mathbf{1}_{\{W<H\}}] = \mathbb{E}[H]\cdot\mathbb{E}[\mathbf{1}_{\{W\geq H\}}].$$

Proof.

$$\begin{aligned}\mathbb{E}[H] &= \mathbb{E}[H\mathbf{1}_{\{W\geq H\}}] + \mathbb{E}[H\mathbf{1}_{\{W<H\}}]\\ &= \mathbb{E}[H\mathbf{1}_{\{W\geq H\}}] + \mathbb{E}[W\mathbf{1}_{\{W<H\}}] + \mathbb{E}[(H-W)\mathbf{1}_{\{W<H\}}].\end{aligned}$$

Next, we may use the fact that H is a geometric and thus memoryless random variable, i.e., $H - W|_{H>W} =_d H$, so that:

$$\mathbb{E}[H] = \mathbb{E}[H\mathbf{1}_{\{W\geq H\}}] + \mathbb{E}[W\mathbf{1}_{\{W<H\}}] + \mathbb{E}[H]\cdot\mathbb{E}[\mathbf{1}_{\{W<H\}}].$$

This completes the proof. □

Denote by N_k^0 the number of customers served until Q_k would become empty for the first time if there were no timer. The following properties are readily verified:

Property 3.15.

$$\begin{aligned}H_k^+ &= \min(N_k^0, H_k^{*+}) = N_k^0\,\mathbf{1}_{\{N_k^0\leq H_k^{*+}\}} + H_k^{*+}\,\mathbf{1}_{\{N_k^0>H_k^{*+}\}},\\ H_k^- &= \min(N_k^0, H_k^{*-}) = N_k^0\,\mathbf{1}_{\{N_k^0\leq H_k^{*-}\}} + H_k^{*-}\,\mathbf{1}_{\{N_k^0>H_k^{*-}\}}.\end{aligned}$$

These properties imply that if the server leaves Q_k because it is empty, then $H_k^+ = H_k^-$ and $H_k^+ = H_k^- + 1$, otherwise.

The following lemma demonstrates that the ratio of mean number of served customers and mean number of customers taken into service is equal both for a saturated and a non-saturated queue.

Lemma 3.16.

$$\frac{\mathbb{E}[H_k^{*-}]}{\mathbb{E}[H_k^{*+}]} = \frac{\mathbb{E}[H_k^-]}{\mathbb{E}[H_k^+]},\ k=1,\ldots,M.$$

Proof. Note that H_k^{*+} is a geometrically distributed random variable (with success probability $p = 1 - \tilde{X}_k(\xi_k)$, since an interruption is seen as a success). Then,

$$\begin{aligned}
\mathbb{E}[H_k^+] \cdot \mathbb{E}[H_k^{*-}] &= \left(\mathbb{E}[N_k^0 \mathbf{1}_{\{N_k^0 < H_k^{*+}\}}] + \mathbb{E}[H_k^{*+} \mathbf{1}_{\{N_k^0 \geq H_k^{*+}\}}]\right) \left(\mathbb{E}[H_k^{*+}] - 1\right) \\
&= \left(\mathbb{E}[N_k^0 \mathbf{1}_{\{N_k^0 < H_k^{*+}\}}] + \mathbb{E}[H_k^{*+} \mathbf{1}_{\{N_k^0 \geq H_k^{*+}\}}]\right) \cdot \mathbb{E}[H_k^{*+}] \\
&\quad - \left(\mathbb{E}[N_k^0 \mathbf{1}_{\{N_k^0 < H_k^{*+}\}}] + \mathbb{E}[H_k^{*+} \mathbf{1}_{\{N_k^0 \geq H_k^{*+}\}}]\right) \\
&= \left(\mathbb{E}[N_k^0 \mathbf{1}_{\{N_k^0 < H_k^{*+}\}}] + \mathbb{E}[H_k^{*+} \mathbf{1}_{\{N_k^0 \geq H_k^{*+}\}}]\right) \cdot \mathbb{E}[H_k^{*+}] \\
&\quad - \mathbb{E}[\mathbf{1}_{\{N_k^0 \geq H_k^{*+}\}}] \cdot \mathbb{E}[H_k^{*+}] \\
&= \mathbb{E}[H_k^{*+}] \cdot \left(\mathbb{E}[N_k^0 \mathbf{1}_{\{N_k^0 < H_k^{*+}\}}] + \mathbb{E}[(H_k^{*+} - 1) \mathbf{1}_{\{N_k^0 \geq H_k^{*+}\}}]\right),
\end{aligned}$$

where for the third equality sign we used Lemma 3.14. Finally, observe that $\{N_k^0 < H_k^{*+}\} = \{N_k^0 \leq H_k^{*-}\}$ (since all variables are discrete), so that we may write:

$$\begin{aligned}
&\mathbb{E}[N_k^0 \mathbf{1}_{\{N_k^0 < H_k^{*+}\}}] + \mathbb{E}[(H_k^{*+} - 1) \mathbf{1}_{\{N_k^0 \geq H_k^{*+}\}}] \\
&= \mathbb{E}[N_k^0 \mathbf{1}_{\{N_k^0 \leq H_k^{*-}\}}] + \mathbb{E}[(H_k^{*+} - 1) \mathbf{1}_{\{N_k^0 > H_k^{*-}\}}] \\
&= \mathbb{E}[N_k^0 \mathbf{1}_{\{N_k^0 \leq H_k^{*-}\}}] + \mathbb{E}[H_k^{*-} \mathbf{1}_{\{N_k^0 > H_k^{*-}\}}] \\
&= \mathbb{E}[H_k^-],
\end{aligned}$$

where we used Proposition 3.15 in the final step. □

Remark 3.17 (Independence of N_k^0)**.** *It is important to notice that the equivalence of the ratios does not depend on the distribution of N_k^0. In particular, we have made no assumptions whatsoever on the number of customers present at the start of a visit in the unsaturated case. So, the time of the previous polling instant of the queue does not impact the ratio of the unsaturated case (while the ratio of the saturated case is obviously fixed).*

Recall that $\mathbb{E}[V_k^*]$ denotes the mean visit time of the server to Q_k when Q_k is saturated. This quantity satisfies the following relation.

Lemma 3.18.

$$\mathbb{E}[V_k^*] = \mathbb{E}[H_k^{*+}] \cdot \sigma_k \;=\; \mathbb{E}[H_k^{*-}] \cdot \tilde{\sigma}_k, \;\; k = 1, \ldots, M.$$

Proof. Consider the saturated case. We write: $V_k^* = \sum_{j=1}^{H_k^{*+}} Y_{k,j}$, where $Y_{k,j}$, $j = 1, 2, \ldots$, are i.i.d. random variables distributed as $min(X_k, V_k)$ and with mean σ_k. Let further $V_{k,j}$ and $X_{k,j}$, $j = 1, 2, \ldots$, be i.i.d. random variables distributed as the generic random variables X_k and V_k, with X_k and V_k independent. Notice that $H_k^{*+} = \min\{j : X_{k,j} > V_{k,j}\}$. Therefore, H_k^{*+} is a stopping time for $Y_{k,j}$, $j =$

$1, 2, \ldots$, so that we may apply Wald's equation yielding: $\mathbb{E}[V_k^*] = \mathbb{E}[H_k^{*+}] \cdot \sigma_k$. Next, consider a period Z comprising a single visit of length V_k^* to Q_k extended with the (service) time needed to complete the service of the customer that was interrupted at the end of the visit. That is, Z is the time needed to complete all (residual) services that were started during V_k^* (in particular, we include a possible residual service time, which is in fact distributed as an effective service time due to the geometric nature of the effective service time). Thus, $\mathbb{E}[Z] = \mathbb{E}[H_k^{*+}] \cdot \tilde{\sigma}_k$, but also $\mathbb{E}[Z] = \mathbb{E}[V_k^*] + \tilde{\sigma}_k$, since there is always an interrupted service with a mean residual service time identical to the original mean effective service time. Hence, it follows that: $\mathbb{E}[V_k^*] = (\mathbb{E}[H_k^{*+}] - 1) \cdot \tilde{\sigma}_k = \mathbb{E}[H_k^{*-}] \cdot \tilde{\sigma}_k$. □

Combining Lemma 3.18 with Eq. (3.6) gives after some manipulations:

Corollary 3.19.

$$\mathbb{E}[G_k^{*-}] = a_k \cdot \mathbb{E}[H_k^{*-}] = \frac{a_k \cdot \tilde{X}_k(\xi_k)}{1 - \tilde{X}_k(\xi_k)}.$$

Proof. (Proof of Lemma 3.12) It is readily seen that to prove Eq. (3.7) it is sufficient to show:

$$\mathbb{E}[V_{k,j}] = \mathbb{E}[H_{k,j}^-] \cdot \tilde{\sigma}_k, \;\; j = 1, \ldots, a_k.$$

W.l.o.g. we consider the first visit to Q_k in a cycle and leave out the subscript 1. Thus, we need to prove the following:

$$\mathbb{E}[V_k] = \mathbb{E}[H_k^-] \cdot \tilde{\sigma}_k.$$

Analogously to the proof of Lemma 3.18, we write $V_k = \sum_{j=1}^{H_k^+} X_{k,j}$ for the unsaturated case. By arguing that H_k^+ is a stopping time for the sequence $\{X_{k,j}\}_j$, it immediately follows via Wald that: $\mathbb{E}[V_k] = \mathbb{E}[H_k^+] \cdot \sigma_k$. The proof is then completed by appealing to Lemma 3.16 and Lemma 3.18. □

Let us define $\hat{\rho}_k, \;\; k = 1, \ldots, M$ as follows:

$$\hat{\rho}_k := \sum_{j=1}^{k} \rho_j = \sum_{j=1}^{k} \lambda_j \tilde{\sigma}_j.$$

Next, we define a stability condition for the complete system and for the subsystems $\mathcal{S}^e$ with $e \in \{0, \ldots, M\}$:

Definition 3.20. *(Condition $\mathcal{C}^M$)*

$$\mathcal{C}^M : \hat{\rho}_M + \max_{1 \le j \le M} (\lambda_j / \mathbb{E}[G_j^{*-}]) c_T < 1.$$

Definition 3.21. *(Condition $\mathcal{C}^e$)*

$$\mathcal{C}^e : \hat{\rho}_e + \max_{1\leq j\leq e}(\lambda_j/\mathbb{E}[G_j^{*-}])c_T^e < 1.$$

We number the queues according to the ratio $\lambda_j/\mathbb{E}[G_j^{*-}]$ in non-decreasing order. Hence, we have that:

$$\mathcal{C}^e : \hat{\rho}_e + (\lambda_e/\mathbb{E}[G_e^{*-}])c_T^e < 1.$$

Further, we note that it can verified by simple calculations that $\mathcal{C}^{e+1}$ implies $\mathcal{C}^e$.

We are now ready to present the following lemma (cf. Lemma 6 of [37]) which forms a crucial link in the proof:

Lemma 3.22. *If condition $\mathcal{C}^e$ holds, then*

$$\mathbb{E}[G_k^{e-}] < \mathbb{E}[G_k^{*-}],\ 1 \leq k \leq e.$$

Proof. Let us consider the mean duration of a cycle. W.l.o.g. we say that the nth cycle starts at time $T_{n,1}$ and ends at time $T_{n+1,1}$. A cycle consists of the visits to the queues and the switch-over times, so that we may write (cf. Lemma 3.12):

$$\mathbb{E}[T_{n+1,1} - T_{n,1}] = \sum_{j=1}^{M} \tilde{\sigma}_j \mathbb{E}[G_{n,j}^-] + c_T,\ n = 0, 1, \dots\ .$$

Hence, for the change in number of customers at Q_k during this cycle, we readily have for $k = 1, \dots, M,\ n = 0, 1, \dots$:

$$\mathbb{E}[N_k(T_{n+1,1}) - N_k(T_{n,1})] = \lambda_k \cdot \left(\sum_{j=1}^{M} \tilde{\sigma}_j \mathbb{E}[G_{n,j}^-] + c_T \right) - \mathbb{E}[G_{n,k}^-].$$

Suppose w.l.o.g. that the system is empty at time 0. Using the $\leq_d$-monotonicity for each given visit, it follows that the expectations of the queue lengths at the polling times are non-decreasing, i.e.,

$$\mathbb{E}[N_k(T_{n+1,1}) - N_k(T_{n,1})] \geq 0,\ k = 1, \dots, M,\ n = 0, 1, \dots,$$

which provides us immediately with the following system of equations:

$$\mathbb{E}[G_{n,k}^-] \leq \lambda_k \cdot \left(\sum_{j=1}^{M} \tilde{\sigma}_j \mathbb{E}[G_{n,j}^-] + c_T \right),\ k = 1, \dots, M,\ n = 0, 1, \dots\ . \tag{3.8}$$

Observe that $\mathbb{E}[G_{n,k}^-]$ and $\mathbb{E}[G_{n,k}^+]$ are non-decreasing in n and are bounded from above by $\mathbb{E}[G_k^{*+}] < \infty$. Thus, we may write for the following limits, $k = 1, \ldots, M$:

$$\begin{aligned} \lim_{n\to\infty} \mathbb{E}[G_{n,k}^-] &= \mathbb{E}[G_k^-], \\ \lim_{n\to\infty} \mathbb{E}[G_{n,k}^+] &= \mathbb{E}[G_k^+]. \end{aligned}$$

Let us consider next Eq. (3.8) for $k = 1$ and let $n \to \infty$:

$$\mathbb{E}[G_1^-] \leq \lambda_1 \cdot \left(\sum_{j=1}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + c_T \right).$$

This can readily be rewritten to:

$$\mathbb{E}[G_1^-] \cdot (1 - \hat{\rho}_1) \leq \lambda_1 \cdot \left(\sum_{j=2}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + c_T \right).$$

Applying a triangularization procedure (see Appendix 3.A), we may obtain for $1 \leq k \leq M$:

$$\mathbb{E}[G_k^-] \cdot (1 - \hat{\rho}_k) \leq \lambda_k \cdot \left(\sum_{j=k+1}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + c_T \right). \tag{3.9}$$

The latter result does also hold if we consider the system with queues $e + 1$ up to M being saturated. From the point of view of the first e queues only the return time of the server will change while the behavior of the server during a visit remains identical. Denoting the quantities in this modified system by adding the superscript e, we may write:

$$\mathbb{E}[G_k^{e-}] \leq \frac{\lambda_k}{1 - \hat{\rho}_k} \cdot \left(\sum_{j=k+1}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^{e-}] + c_T^e \right), \; k = 1, \ldots, e,$$

where $c_T^e = c_T + \sum_{j=e+1}^{M} \mathbb{E}[V_j^*]$. Since, $\mathbb{E}[G_j^{e-}] \leq \mathbb{E}[G_j^{*-}], \; j = 1, \ldots, M$, we obtain:

$$\begin{aligned} \mathbb{E}[G_k^{e-}] &\leq \frac{\lambda_k}{1 - \hat{\rho}_k} \cdot \left(\sum_{j=k+1}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^{*-}] + c_T \right) \\ &= \frac{\lambda_k}{1 - \hat{\rho}_k} \cdot c_T^k, \; k = 1, \ldots, e. \end{aligned}$$

On the other hand, the condition $\mathcal{C}^k, \; k = 1, \ldots, e$, which is implied by $\mathcal{C}^e$, reads:

$$\mathcal{C}^k : \hat{\rho}_k + \max_{1 \leq j \leq k} (\lambda_j / \mathbb{E}[G_j^{*-}]) \cdot c_T^k < 1.$$

Under the assumption that the ratios $\lambda_j/\mathbb{E}[G_j^{*-}]$ are ordered non-decreasingly, it is readily found that $\mathcal{C}^k$ implies:

$$\mathbb{E}[G_k^{*-}] > \frac{\lambda_k}{1-\hat{\rho}_k} \cdot c_T^k,$$

which completes the proof. □

The remainder of the proof is along the lines of [37]. Recall that we want to show here that if condition $\mathcal{C}^M$ is satisfied, then the system is stable. An equivalent definition of stability (see [37]) is that there exists a proper stationary joint queue-length distribution at the polling instants such that the expectation of the stationary cycle time is finite. A sufficient condition for the stationary distribution to exist is that the multi-dimensional Markov chain $(M_{n,1}^e)$ is ergodic. The ergodicity of this chain $(M_{n,1}^e)$ is equivalent to the existence of $\mathbf{m}^e$ such that the limit

$$\lim_{n\to\infty} \mathbb{P}(M_{n,1}^e \leq \mathbf{m}^e) \geq 1 - \sum_{k=1}^{e} \lim_{n\to\infty} \mathbb{P}(N_k(T_{n,1}^e) \geq \mathbf{m}_k), \tag{3.10}$$

is strictly positive. We note that the system will become empty once the chain enters some state $\leq \mathbf{m}^e$ with a strictly positive probability (due to no arrivals for a specific time). Since the positiveness of the limit implies that you return infinitely often to some state $\leq \mathbf{m}^e$, it follows that with probability one you will reach the empty state in a finite amount of time; in other words, it excludes transient or null-recurrent behaviour of the chain. Thus, to have ergodicity, we need the sum on the right-hand side of Eq. (3.10) to be strictly smaller than one. This can be established if for one k the limiting distribution $(N_k(T_{n,1}^e))_n$ is not concentrated at infinity, i.e, $\mathbb{P}(N_k < \infty) > 0$ and all other limiting distributions are proper, i.e., $\mathbb{P}(N_k < \infty) = 1$.

We will prove this by induction starting with the subsystem $\mathcal{S}^0$. This system $\mathcal{S}^0$ is readily seen to be stable. Next, we suppose $\mathcal{S}^{e-1}$ is stable, and consider $\mathcal{S}^e$, $1 \leq e \leq M$. We note since $\mathcal{S}^{e-1}$ is stable, the Markov chain $(M_{n,1}^{e-1})$ is ergodic and in particular $(N_k(T_{n,1}^{e-1}))_n$, $1 \leq k \leq e-1$ has a proper distribution. Also, $(M_{n,i}^{e-1})_n$, $i = 1, ..., a$ has a proper limiting distribution and by Lemma 3.11, $M_{n,i}^{e|e-1} \leq_d M_{n,i}^{e-1}$ for all n. Thus, $(M_{n,i}^{e|e-1})_n$ has a proper limiting distribution. Moreover, from Lemma 3.12, we have that $\mathbb{E}[G_e^{e-}] < \mathbb{E}[G_e^{*-}]$. Hence, there exists a visit r such that $\lim_{n\to\infty} \mathbb{E}[F_{n,r}^{e-}] < \mathbb{E}[F_r^{*-}]$. Then, by Lemma 3.5-ii there exists a y such that $\lim_{n\to\infty} \mathbb{P}(N_{n,r}^e \leq y) > 0$, i.e., the limiting distribution of the last component $N_{n,r}^e = N_e^e(T_{n,r})$ of $M_{n,r}^e$ is not concentrated at infinity. Thus, the chain $(M_{n,r}^e)_n$ is ergodic. The observation that the expectation of the cycle time is finite completes the proof.

□

3.4.3.2 Proof: Necessary condition

Suppose the polling system $\mathcal{S}$ is stable. Let us define F^-_{n,k_l} as the mean number of customers served during the k_l-th stage of the nth cycle, where the k_l-th stage corresponds to exactly the lth visit to Q_k in the cycle. We let for each visit i the initial distribution of $(M_{n,i})_n$ be its stationary distribution. Since $\mathcal{S}$ is stable, these chains are stationary with positive-recurrent states. As a result, $\mathbb{P}(N_k(T_{n,i}) = 0) > 0$ for all k and (n, i). Further, as the expected cycle time is finite, $\mathbb{E}[G^-_k] = \sum_{l=1}^{a_k} \mathbb{E}[F^-_{n,k_l}]$ does not depend on n and is finite for all k. It follows by Lemma 3.5 that $\mathbb{E}[G^-_k] < \mathbb{E}[G^{*-}_k]$ for $1 \leq k \leq M$ and in particular that $\mathbb{E}[G^-_M] < \mathbb{E}[G^{*-}_M]$.

On the other hand, it can readily be seen that:

$$N_k(T_{2,1}) - N_k(T_{1,1}) = N_k(T_{1,1}, T_{2,1}] - \sum_{l=1}^{a_k} F^-_{1,k_l}.$$

Hence, we can bound $N_k(T_{2,1}) - N_k(T_{1,1})$ as follows:

$$-\sum_{l=1}^{a_k} F^-_{1,k_l} \leq N_k(T_{2,1}) - N_k(T_{1,1}) \leq N_k(T_{1,1}, T_{2,1}].$$

Both the lower and upper bound have finite expectation, such that for all k (see [37, Lemma 7]):

$$\mathbb{E}[N_k(T_{2,1}) - N_k(T_{1,1})] = 0,$$

and in general for $n \geq 1$:

$$\mathbb{E}[N_k(T_{n+1,1}) - N_k(T_{n,1})] = 0.$$

This leads to (cf. Eq. (3.8)) the following system of equalities:

$$\mathbb{E}[G^-_k] = \lambda_k \cdot \left(\sum_{j=1}^{M} \tilde{\sigma}_j \mathbb{E}[G^-_j] + c_T \right), \ 1 \leq k \leq M,$$

and along the lines of deriving Eq. (3.9), we obtain:

$$\mathbb{E}[G^-_k] \cdot (1 - \hat{\rho}_k) = \lambda_k \cdot \left(\sum_{j=k+1}^{M} \tilde{\sigma}_j \mathbb{E}[G^-_j] + c_T \right), \ 1 \leq k \leq M.$$

Specifically, for $k = M$ this implies:

$$\mathbb{E}[G^-_M] \cdot (1 - \hat{\rho}_M) = \lambda_M \cdot S.$$

Together with the observation above, $\mathbb{E}[G^-_M] < \mathbb{E}[G^{*-}_M]$, it follows that condition $\mathcal{C}^M$ holds.

3.5 Concluding remarks

We have proven the stability conditions for two polling models with time-limited service and periodic polling. The proof for the pure time-limited discipline is straightforward, since the queues can in fact be decoupled and thus studied in isolation. For the proof of the exhaustive time-limited discipline, we have relied largely on the rigorous stability proof of Fricker and Jaïbi [37] for a class of service disciplines. Unfortunately, this class covers only non-preemptive and work-conserving service disciplines. Though, the main ideas of their proof could still be used to prove stability here.

A logical next step would be to extend the results to Markovian polling of the server. The fixed cycle of periodic polling will then become a random cycle. One would typically consider per-queue cycles, i.e., a cycle starting and ending at consecutive polling instants of a specific queue. Consequently, the number of visits to the other queues during a cycle are random variables. For the pure time-limited discipline, such an extension can readily be incorporated by appropriately adjusting the availability fraction κ_i. For the exhaustive time-limited, it might require some more work to prove this extension. However, we strongly believe this could also be done using similar techniques as the ones presented in this chapter.

3.A Triangularization

Let us explain below the triangularization method that we apply. We depart from the following set of equalities:

$$\mathbb{E}[G_k^-] \leq \lambda_k \cdot \left(\sum_{j=1}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right), \ k = 1, \ldots, M. \tag{3.11}$$

Rearranging the equality for $k = 1$, we obtain:

$$(1 - \hat{\rho}_1) \cdot \mathbb{E}[G_1^-] \leq \lambda_1 \cdot \left(\sum_{j=2}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right). \tag{3.12}$$

Next, we will show that also for $2 \leq k \leq M$ we may write:

$$(1 - \hat{\rho}_k) \cdot \mathbb{E}[G_k^-] \leq \lambda_k \cdot \left(\sum_{j=k+1}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right).$$

This will be done by proving the following inequalities by induction.

$$\sum_{j=1}^{k} \tilde{\sigma}_j \mathbb{E}[G_j^-] \leq \frac{\hat{\rho}_k}{1 - \hat{\rho}_k} \cdot \left(\sum_{j=k+1}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right), \ k = 1, \ldots, M, \tag{3.13}$$

$$(1 - \hat{\rho}_{k+1}) \cdot \mathbb{E}[G_{k+1}^-] \leq \lambda_{k+1} \cdot \left(\sum_{j=k+2}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right), \ k = 1, \ldots, M-1. \tag{3.14}$$

First, notice that for $k = 1$ Eq. (3.13) has been shown above, while Eq. (3.14) reads as follows:

$$(1 - \hat{\rho}_2) \cdot \mathbb{E}[G_2^-] \leq \lambda_2 \cdot \left(\sum_{j=3}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right).$$

This inequality can be proven from Eq. (3.11) and taking $k = 2$. First, we take all terms $\mathbb{E}[G_2^-]$ to the left-hand side, second we apply Eq. (3.12), and finally some simple manipulations provide us with the desired result. Next, we show that once these inequalities hold for l these also hold for $l + 1$. First, consider Eq. (3.13) for

$l+1$:

$$\begin{aligned}
\sum_{j=1}^{l+1} \tilde{\sigma}_j \mathbb{E}[G_j^-] &= \sum_{j=1}^{l} \tilde{\sigma}_j \mathbb{E}[G_j^-] + \tilde{\sigma}_{l+1} \mathbb{E}[G_{l+1}^-] \\
&\leq \frac{\hat{\rho}_l}{1-\hat{\rho}_l} \cdot \left(\sum_{j=l+2}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right) + \frac{1}{1-\hat{\rho}_l} \cdot \tilde{\sigma}_{l+1} \mathbb{E}[G_{l+1}^-] \\
&\leq \left(\frac{\hat{\rho}_l}{1-\hat{\rho}_l} + \frac{\rho_{l+1}}{(1-\hat{\rho}_l)(1-\hat{\rho}_{l+1})} \right) \cdot \left(\sum_{j=l+2}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right) \\
&= \frac{\hat{\rho}_{l+1}}{1-\hat{\rho}_{l+1}} \cdot \left(\sum_{j=l+2}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right).
\end{aligned}$$

Second, we have to prove Eq. (3.14) for $l+1$, i.e.,

$$(1-\hat{\rho}_{l+2}) \cdot \mathbb{E}[G_{l+2}^-] \leq \lambda_{l+2} \cdot \left(\sum_{j=l+3}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right). \tag{3.15}$$

To this end, we depart from Eq. (3.11) for $l+2$:

$$\begin{aligned}
\mathbb{E}[G_{l+2}^-] &\leq \lambda_{l+2} \cdot \left(\sum_{j=1}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right) \\
&= \lambda_{l+2} \cdot \sum_{j=1}^{l+1} \tilde{\sigma}_j \mathbb{E}[G_j^-] + \lambda_{l+2} \cdot \left(\sum_{j=l+2}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right) \\
&\leq \lambda_{l+2} \cdot \frac{\hat{\rho}_{l+1}}{1-\hat{\rho}_{l+1}} \cdot \left(\sum_{j=l+2}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right) \\
&\quad + \lambda_{l+2} \cdot \left(\sum_{j=l+2}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right) \\
&= \frac{\lambda_{l+2}}{1-\hat{\rho}_{l+1}} \cdot \left(\sum_{j=l+3}^{M} \tilde{\sigma}_j \mathbb{E}[G_j^-] + S \right) + \frac{\lambda_{l+2}}{1-\hat{\rho}_{l+1}} \cdot \tilde{\sigma}_{l+2} \mathbb{E}[G_{l+2}^-].
\end{aligned}$$

Hence, moving all the terms $\mathbb{E}[G_{l+2}^-]$ to the left-hand side and performing some rearrangements yields Eq. (3.15).

Part II

Single-server polling models

CHAPTER

4

Analysis of the basic polling model

4.1 Introduction

Single-server polling systems operating under the pure exponential time-limited service discipline can be considered as a building block towards more sophisticated performance models for data communication in ad hoc networks with mobile stations. The time-limited aspect of this service discipline captures explicitly the randomness of the lifetime of a wireless communication link. Moreover, the preemptive nature of the service discipline grasps the interruption of a data transmission in such networks.

Despite the vast amount of polling literature, there exist hardly any analytical studies on the class of the pure time-limited service disciplines. In fact, the only work that analyzes this discipline in the context of polling systems is [108]. This work considers the workload process for a pure time-limited polling model with deterministic visit times and a cyclic visit schedule. Due to the deterministic nature of the model, the queue lengths at the different queues can be decoupled and each queue is modelled as an M/G/1 queue with server vacations. Using an approximate analysis, the mean workload and mean message delay are studied.

A well-studied class of disciplines that is closely related to the pure time-limited discipline is the class of *exhaustive* time-limited service disciplines (see, e.g., [95, 39, 68, 99]). According to the exhaustive discipline, the server leaves the queues as soon as it becomes empty. Leung [68] analyzes the exhaustive time-limited discipline

with an exponential time-limit and non-preemptive service, whereas Eliazar and Yechiali [31] cover the preemptive case. The exhaustive time-limited discipline with deterministic timer and preemption is considered by De Souza e Silva et al. [95] for Poisson arrivals and by Frigui and Alfa [39] for Markovian Arrival Processes. A specific application of an exhaustive time-limited polling system to a timed-token protocol can be found in [99].

This impatient property of the server, i.e., the server moves to a next queue once the queue becomes empty, as witnessed above is a common assumption in the analysis of polling models. Nonetheless, several analytical papers treat queueing models with a server that remains at a queue even when it becomes empty. Such models are often referred to as patient server or stopping server models. For instance, Eisenberg [28] and Borst [13] analyze several stopping strategies for the server once the complete system becomes empty as to optimize specific performance measures. More recently, Boxma et al. considered a single-queue vacation model [16] and a two-queue polling model [17] in which the server upon arriving to an empty queue waits patiently for a customer arrival until a pre-specified timer expires. We note that in the latter two-queue polling model (contrary to the models in [13] and [28]) there is no notion of work conservation anymore, since the server may wait patiently at one queue while the other queue is nonempty. Thus, these approaches implicitly relate the patience of the server to the number of customers in the system, which is not the case for the pure time-limited discipline.

The present chapter is an elaborate version of [H7]. We will analyze the joint queue-length distribution of the basic polling system in a stable environment. In particular, we assume the pure time-limited service discipline with an exponential time limit. Thus, the main characteristics of the model are that the server visits a queue for an exponential amount of time (irrespective of the number of customers present at a queue) and that the service strategy is preemptive-repeat with resampling.

For the case of a single queue (i.e., $M = 1$), the specific polling model becomes in fact an unreliable-server model (USM). Gaver [42] analyzed the queue-length distribution for such a model. The extension of the analysis to a two-queue polling model appears basically feasible along the approach of [22] or [34]. This approach requires to solve a boundary value problem. Unfortunately, this solution method becomes an extremely difficult task for the two-queue model already, while for three or more queues analytical solutions along this direction are not anticipated. For the multi-queue case, we will concentrate on the joint queue-length distribution at various embedded epochs. Our analytical approach builds on the work of Eisenberg [29]. We set-up a system of equations which relates the queue-length distributions at specific instants. The solution of this system is obtained via the explicit determination of the distribution at visit completion instants using an iterative approach. This approach is similar to the one introduced by Leung for probabilistically-limited polling models [67]. Finally, we discuss several extensions for the basic polling model, viz., customer routing, Markovian polling and non-exponential visit times of the server.

We should emphasize that the queue-length processes at the different queues

in the system are not independent. For instance, the number of arrivals to the queues depend on the realizations of the random visit times at the other queues. Therefore, even for zero switch-over times, the queue-length processes are definitely not independent. However, we note that if the interest would only be in mean performance measures, then the queues could be considered in isolation.

This chapter is organized as follows. In Sect. 4.2, we describe the polling model in detail. The analysis for the single-queue model and multi-queue model are given in Sect. 4.3 and Sect. 4.4, respectively. In Sect. 4.5, we discuss three extensions of the basic polling model. The chapter is concluded in Sect. 4.6.

4.2 Model

Let us consider the basic polling system of $M \geq 1$ queues with Poisson arrivals and generally distributed service and switch-over times. The server visits the queues according to the cyclic polling strategy. The service discipline assumed is the pure time-limited discipline. The time limit at Q_i is exponentially distributed with parameter ξ_i. Service will be preempted at the expiration of the timer and a new service time will be redrawn from the original distribution at the start of the next visit; that is, we assume the preemptive-repeat with resampling strategy.

We recall that the random variables $I_i,\ X_i,\ C_{i-1,i},$ and $Y_i,\ i = 1, \ldots, M$ refer to the interarrival time of customers, the service time of customers, the switch-over time of the server and visit time (or time limit) of the server at a queue. These random variables are assumed independent and identically distributed and also assumed to be mutually independent.

4.3 Analysis of the single-queue model

The single-queue model comprises a single queue, say Q_1, fed by a Poisson arrival process of customers. Each customer brings a generally distributed amount of work to the system. The server visits the queue for an exponential period during which customers may be served and then leaves for a random period of time. Let these intervisit times consist of the sum of the visit times to the other queues plus the total switch-over times of the multi-queue model. Then, the queue-length distribution of this single-queue model corresponds to the marginal queue-length distribution of the queue in the polling system. In the literature, such a single-queue model is known as an unreliable-server model or, alternatively, as a vacation model with preemptive service. The first to study this specific model was Gaver [42] by analyzing it as a priority queueing model with high (i.e., interruptions) and low priority customers (i.e., common arrivals). Below, we describe the unreliable-server model, present the expression for the queue-length distribution and give a direct proof of this expression (which to our opinion is somewhat more insightful than the proof in [42]).

Consider a sequence of alternating *processing* and *non-processing periods*. During a processing period, there are some customers at the queue and one of these is being served. During a non-processing period no customers are present. The server may break down (and thus needs repair) at random points in time both during processing and non-processing periods. The *repair periods* have a duration B which follows a distribution $B(t)$ with Laplace-Stieltjes Transform (LST) $\tilde{B}(s)$ and mean $\mathbb{E}[B]$, and correspond to the intervisit times at Q_1 in our polling system. Thus, within the setting of the polling system, we have that $B = C_{M,1} + \sum_{i \neq 1}(C_{i-1,i} + Y_i)$. We note that in this section, since only a single queue is considered, we will drop the subscript 1 whenever this does not lead to ambiguity. The periods between consecutive repairs, the so-called *availability periods*, are assumed exponentially distributed with mean $1/\xi$ and these periods correspond to the visit times in the polling model. Customers arrive to the system according to a Poisson process with rate λ. We assume that a preemptive-repeat servicing strategy with resampling is followed, i.e., if a service is interrupted, then at the start of the next availability period the service requirement is redrawn from the original service-time distribution.

Let $\tilde{X}_G(s)$ and $\mathbb{E}[X_G]$ denote the LST and the mean of the *generalized service time* of a customer, respectively. The latter period of time is defined as the period which starts when a customer receives service for the first time and ends when the customer leaves the system. We denote by $\hat{U}(z)$ the probability generating function (p.g.f.) of the number of customers that arrive during the generalized service time of a customer which arrives to an empty system. Such a latter service (customer) will be referred to as an *exceptional first service (customer)*. This service is exceptional in the sense that it may include the residual repair time of the server. Finally, let $\mathbb{E}[K]$ refer to the mean number of customers served during a processing period and define ρ_G as the generalized load of the queue, i.e., $\rho_G = \lambda \mathbb{E}[X_G]$. Notice that it follows from our model assumptions that repair periods, availability periods and service times are independent.

Let us denote the queue-length distribution of the number of customers left behind by a departing customer by d_n, $n = 0, 1, 2, \ldots$. Then, the probability generating function $P_{L_d}(z)$ of the queue-length distribution at customer departure instants is known (see, e.g., [42]) and given by the following theorem.

Theorem 4.1 (Queue-length distribution of the unreliable-server model)**.**

$$P_{L_d}(z) = \frac{1}{\mathbb{E}[K]} \cdot \left(1 + \frac{z(1 - \hat{U}(z))}{\tilde{X}_G(\lambda(1-z)) - z}\right), \tag{4.1}$$

where

$$\begin{aligned}\tilde{X}_G(s) &= \frac{\tilde{X}(\xi+s)\cdot(\xi+s)}{(\xi+s)-\xi\cdot(1-\tilde{X}(\xi+s))\cdot\tilde{B}(s)},\\ \hat{U}(z) &= \tilde{X}_G(\lambda(1-z))\cdot\frac{\lambda z+\xi\cdot(\tilde{B}(\lambda(1-z))-\tilde{B}(\lambda))}{z\cdot(\lambda+\xi(1-\tilde{B}(\lambda)))},\\ \mathbb{E}[K] &= \frac{1}{1-\rho_G}\cdot\frac{\lambda(1+\xi\cdot\mathbb{E}[B])}{\lambda+\xi\cdot(1-\tilde{B}(\lambda))}.\end{aligned}$$

Next, we will present several lemmas and defer the proof of the theorem until the end of this section.

Remark 4.2 (Steady-state queue-length distribution)**.** *Notice that the departure distribution equals the steady-state queue-length distribution. This can be argued by first using an up- and downcrossings argument and next appealing to the well-known Poisson Arrivals See Time Averages (PASTA) property [106].*

Let us denote by V^* the processing time given that the service is interrupted. Further, we denote by X^* the service time given that the service is successful. Let V_j^* be i.i.d. copies of V^*, B_j i.i.d. copies of B, and N a random variable denoting the number of interruptions during a service. Notice that B, X^*, V_j^* and N are independent random variables. Then, the generalized service time X_G satisfies:

$$X_G = X^* + \sum_{j=1}^{N}(V_j^* + B_j).$$

Lemma 4.3.

$$\tilde{X}_G(s) = \mathbb{E}[e^{-sX_G}] = \frac{\tilde{X}(\xi+s)\cdot(\xi+s)}{(\xi+s)-\xi(1-\tilde{X}(\xi+s))\tilde{B}(s)}. \tag{4.2}$$

Proof. The random variable N is geometrically distributed with success probability $\tilde{X}(\xi)$. The result for $\tilde{X}_G(s)$ follows by conditioning on N and some elementary calculus. □

The service time U of an exceptional first customer is given by

$$U = X_G + B_R\cdot\mathbf{1}_{\{B_R\}}, \tag{4.3}$$

where B_R denotes the residual repair time as seen by the first customer which arrives during a repair period, and $\mathbf{1}_{\{B_R\}}$ is the indicator function of the event that a customer which arrives during a repair time arrives to an empty system. It should be noted that X_G and B_R are independent. Let us introduce $N(T)$ to refer to the number of arrivals to the queue during a random period T, so that we can present the following lemma.

Lemma 4.4.

$$\hat{U}(z) = \mathbb{E}[z^{N(U)}] = \mathbb{E}[z^{N(X_G)}] \cdot \mathbb{E}[z^{N(B_R \mathbf{1}_{\{B_R\}})}],$$

where

$$\begin{aligned} \mathbb{E}[z^{N(X_G)}] &= \tilde{X}_G(\lambda(1-z)), \\ \mathbb{E}[z^{N(B_R \mathbf{1}_{\{B_R\}})}] &= 1 - (1 - \mathbb{E}[z^{N(B_R)}]) \cdot \frac{\xi \cdot (1 - \tilde{B}(\lambda))}{(\lambda + \xi) - \xi \cdot \tilde{B}(\lambda)}. \end{aligned}$$

Proof. By Eq. (4.3) and using that Poisson arrivals in disjoint intervals are independent, we have:

$$\hat{U}(z) = \mathbb{E}[z^{N(U)}] = \mathbb{E}[z^{N(X_G)+N(B_R \mathbf{1}_{\{B_R\}})}] = \mathbb{E}[z^{N(X_G)}] \cdot \mathbb{E}[z^{N(B_R \mathbf{1}_{\{B_R\}})}].$$

Also due to the Poisson assumption, we have:

$$\mathbb{E}[z^{N(X_G)}] = \tilde{X}_G(\lambda(1-z)).$$

Let us denote by $\mathbb{P}(\text{XFS})$ the probability that an arbitrary arriving customer has an exceptional first service (XFS). Then, we can write:

$$\begin{aligned} \mathbb{E}[z^{N(B_R \mathbf{1}_{\{B_R\}})}] &= \mathbb{E}[z^{N(B_R)}] \cdot \mathbb{P}(\text{XFS}) + 1 \cdot (1 - \mathbb{P}(\text{XFS})) \\ &= 1 - (1 - \mathbb{E}[z^{N(B_R)}]) \cdot \mathbb{P}(\text{XFS}) \ . \end{aligned}$$

The p.g.f. $\mathbb{E}[z^{N(B_R)}]$ can be found by conditioning on the event of at least one arrival during the repair time:

$$\mathbb{E}[z^{N(B_R)}] = \frac{\mathbb{E}[z^{N(B)} \mid N(B) \geq 1]}{z} = \frac{\tilde{B}(\lambda(1-z)) - \tilde{B}(\lambda)}{z(1 - \tilde{B}(\lambda))}.$$

The probability $\mathbb{P}(\text{XFS})$ is obtained by considering its counterpart $\mathbb{P}(\overline{XFS}) = 1 - \mathbb{P}(\text{XFS})$. The sequence of instants at which the queue becomes empty forms a renewal process. Note that the queue becomes empty only during an availability period and that the residual availability period is again exponentially distributed. Thus, by considering the first customer arriving after a renewal epoch, we can write a recursive relation for $\mathbb{P}(\overline{XFS})$:

$$\begin{aligned} \mathbb{P}(\overline{XFS}) &= \mathbb{P}(\{\text{arrival in availability period}\}) \\ &\quad + (1 - \mathbb{P}(\{\text{arrival in availability period}\})) \\ &\quad \cdot \mathbb{P}(\{\text{no arrival in the following repair period}\}) \cdot \mathbb{P}(\overline{XFS}). \end{aligned}$$

It follows that:

$$\mathbb{P}(\overline{XFS}) = \frac{\lambda}{\lambda + \xi} + \frac{\xi}{\lambda + \xi} \cdot \tilde{B}(\lambda) \cdot \mathbb{P}(\overline{XFS}) = \frac{\lambda}{\lambda + \xi \cdot (1 - \tilde{B}(\lambda))}.$$

□

Proof of Theorem 4.1. Equation (4.1) can readily be obtained by studying the embedded Markov chain at service completion instants (see, e.g., [81]). The explicit expressions that show up were derived in Lemmas 4.3 and 4.4. Finally, the term $\mathbb{E}[K]$ follows by inserting $z = 1$ into Eq. (4.1) and applying L'Hospital's rule, yielding:

$$\mathbb{E}[K] = \frac{1}{1-\rho_G} \cdot \frac{\lambda(1+\xi \cdot \mathbb{E}[B])}{\lambda + \xi(1-\tilde{B}(\lambda))}.$$

□

4.4 Analysis of the multi-queue model

We have shown that for the polling system under consideration the marginal queue-length distributions can be obtained by analyzing each queue in isolation. However, the joint queue-length distribution cannot be obtained in this way due to the stochastics in the visit times of the server. Our analysis of the multi-queue model builds on the work of Eisenberg [29] which considers a polling model with an impatient server and non-preemptive service. For this model, the queue-length distribution is determined at visit beginning, visit completion, service beginning, and service completion instants by studying the embedded Markov chains defined at these instants. The fundamental relation in the analysis is the relation that counts the number of events with state $\mathbf{n}$ that occurred until time t [29, Eq.(4)]. In our work, we extend this relation for the polling model under consideration and we use this as a building block for obtaining the queue-length distribution at specific instants.

We will state the stability conditions of the system in Sect. 4.4.1. Next, in Sect. 4.4.2, we treat the extended counting relation in more detail. This counting relation is not sufficient to determine the queue-length distribution at all instants. To this end, we derive additional relations between the random variables in Sect. 4.4.3. However, even with these additional relations we still do not have enough information to solve our model completely. We will resolve this problem by deriving an explicit expression for the queue-length distribution at visit completion instants (see Sect. 4.4.4). This latter approach is based on work of Leung [67] for a probabilistically-limited polling model. Finally, we present the steady-state probabilities for the multi-queue model in Sect. 4.4.5.

4.4.1 Stability condition

The specific necessary and sufficient condition for stability of this cycling polling system follows immediately from Thm. 3.1 in Chapter 3, i.e.,

Theorem 4.5. *(Stability condition for a cyclic polling model with pure exponential time-limited service)*

$$\textit{System is stable} \iff \rho_i < \kappa_i, \ \forall_{i \in \{1,\ldots,M\}},$$

where

$$\rho_i = \lambda_i \cdot \frac{1-\tilde{X}_i(\xi_i)}{\xi_i \cdot \tilde{X}_i(\xi_i)},$$

$$\kappa_i = \frac{1/\xi_i}{\sum_{j=1}^{M}(1/\xi_j + c_{j-1,j})}.$$

4.4.2 A relation for the queue-length distribution at specific instants

We set up a relation for the number of occurrences of specific events. Apart from the events defined in [29], we define a number of additional events. We introduce events related to the start and the completion of an idle period. These events do not appear in Eisenberg's model as in his model the server leaves a queue as soon as it becomes empty. Moreover, we introduce events related to the interruption of a service or idle period due to the end of a server visit. Let us denote by n_i the number of customers at Q_i. Next, we can define the following variables which all refer to the number of the given events with state $\mathbf{n} = (n_1, \ldots, n_M)$ that occur in $(0, t)$ at Q_i:

$\omega^i(t;\mathbf{n})$, service beginnings;

$\pi^i(t;\mathbf{n})$, successful service completions (i.e., server does not switch during a service);

$\pi^i_*(t;\mathbf{n})$, interrupted services (i.e., server switches during service);

$\alpha^i(t;\mathbf{n})$, visit beginnings;

$\beta^i(t;\mathbf{n})$, visit completions;

$a^i(t;\mathbf{n})$, idle period beginnings;

$b^i(t;\mathbf{n})$, idle period completions (i.e., server does not switch during an idle period);

$b^i_*(t;\mathbf{n})$, interrupted idle periods (i.e., server switches during an idle period).

We note that $\mathbf{n}$ refers to the number of customers present in the system (either waiting or in service) immediately after the specific event occurred. These variables are related in the following way for all $\mathbf{n} \in \mathbb{N}^M$ and $t \geq 0$:

$$\pi^i(t;\mathbf{n}) + \pi^i_*(t;\mathbf{n}) + \alpha^i(t;\mathbf{n}) + b^i(t;\mathbf{n}) + b^i_*(t;\mathbf{n}) = \omega^i(t;\mathbf{n}) + \beta^i(t;\mathbf{n}) + a^i(t;\mathbf{n}). \quad (4.4)$$

This counting relation should be read as follows. At each instant that one of the events present at the l.h.s. of Eq. (4.4) with state $\mathbf{n}$ occurs, also exactly one event with the same state $\mathbf{n}$ at the r.h.s. occurs. For instance, a visit beginning event at Q_i at time t with state $\mathbf{n}$', $\pi^i_*(t;\mathbf{n}')$, always coincides exactly either with a service beginning event at Q_i with the same state $\mathbf{n}'$, $\omega^i(t;\mathbf{n}')$ (if there are some customers present upon the server's arrival) or with an idle period beginning at Q_i with the same state $\mathbf{n}'$, $a^i(t;\mathbf{n}')$ (if no customers are present upon the server's arrival).

We note that the end of a server visit always corresponds to an interruption event and vice versa. Therefore, we can isolate these events and break up Eq. (4.4) into:

$$\pi^i_*(t;\mathbf{n}) + b^i_*(t;\mathbf{n}) = \beta^i(t;\mathbf{n}), \tag{4.5}$$
$$\pi^i(t;\mathbf{n}) + \alpha^i(t;\mathbf{n}) + b^i(t;\mathbf{n}) = \omega^i(t;\mathbf{n}) + a^i(t;\mathbf{n}). \tag{4.6}$$

Let us define embedded Markov chains each corresponding to instants at which one of the counting processes increases. Each state in a Markov chain is uniquely defined by the position i of the server ($i = 1, \ldots, M$) and $\mathbf{n} = (n_1, \ldots, n_M)$, the number of customers present in the system. We define the steady-state probabilities for each event type by dividing the number of events with state $\mathbf{n}$ that occurred until t by the total number of the events until t, and then taking the limit for t to infinity, yielding:

$$\begin{aligned}
\alpha^i_{\mathbf{n}} &= \lim_{t\to\infty}[\alpha^i(t;\mathbf{n})/\alpha^i(t)], & \beta^i_{\mathbf{n}} &= \lim_{t\to\infty}[\beta^i(t;\mathbf{n})/\beta^i(t)],\\
b^i_{\mathbf{n}} &= \lim_{t\to\infty}[b^i(t;\mathbf{n})/b^i(t)], & b^i_{*,\mathbf{n}} &= \lim_{t\to\infty}[b^i_*(t;\mathbf{n})/b^i_*(t)],\\
a^i_{\mathbf{n}} &= \lim_{t\to\infty}[a^i(t;\mathbf{n})/a^i(t)], & \omega^i_{\mathbf{n}} &= \lim_{t\to\infty}[\omega^i(t;\mathbf{n})/\omega(t)],\\
\pi^i_{\mathbf{n}} &= \lim_{t\to\infty}[\pi^i(t;\mathbf{n})/\pi(t)], & \pi^i_{*,\mathbf{n}} &= \lim_{t\to\infty}[\pi^i_*(t;\mathbf{n})/\pi_*(t)],
\end{aligned}$$

where

$$\begin{aligned}
\alpha^i(t) &= \textstyle\sum_{\mathbf{n}}\alpha^i(t;\mathbf{n}), & \beta^i(t) &= \textstyle\sum_{\mathbf{n}}\beta^i(t;\mathbf{n}),\\
b^i(t) &= \textstyle\sum_{\mathbf{n}}b^i(t;\mathbf{n}), & b^i_*(t) &= \textstyle\sum_{\mathbf{n}}b^i_*(t;\mathbf{n}),\\
a^i(t) &= \textstyle\sum_{\mathbf{n}}a^i(t;\mathbf{n}), & \omega(t) &= \textstyle\sum_i\sum_{\mathbf{n}}\omega^i(t;\mathbf{n}),\\
\pi(t) &= \textstyle\sum_i\sum_{\mathbf{n}}\pi^i(t;\mathbf{n}), & \pi_*(t) &= \textstyle\sum_i\sum_{\mathbf{n}}\pi^i_*(t;\mathbf{n}).
\end{aligned}$$

It can be seen that for a stable system all these limits exist with probability one by using renewal theory arguments. For instance, consider $\pi^i_{\mathbf{n}}$ as given above. We have that for given i and $\mathbf{n}$, $\{\pi^i(t;\mathbf{n}),\ t \geq 0\}$ forms a renewal process, thus $\lim_{t\to\infty}[\pi^i(t;\mathbf{n})/t]$ exists with probability one (see, e.g., [89]). Moreover, it follows that under stability, $\lim_{t\to\infty}[\pi(t)/t]$ exists, and thus we can conclude that the ratio of these latter limits exists. The other probabilities can be argued analogously and are thus also correctly defined.

Notice that (hereby following [29]) we have that all probabilities are conditioned on Q_i except for $\omega^i_{\mathbf{n}}$, $\pi^i_{\mathbf{n}}$ and $\pi^i_{\mathbf{n},*}$. Along with the steady-state probabilities, let us also define the corresponding p.g.f.'s as follows:

$$\begin{aligned}
\alpha^i(\mathbf{z}) &= \textstyle\sum_{\mathbf{n}}\alpha^i_{\mathbf{n}}\cdot\mathbf{z}^{\mathbf{n}}, & \beta^i(\mathbf{z}) &= \textstyle\sum_{\mathbf{n}}\beta^i_{\mathbf{n}}\cdot\mathbf{z}^{\mathbf{n}}, & b^i(\mathbf{z}) &= \textstyle\sum_{\mathbf{n}}b^i_{\mathbf{n}}\cdot\mathbf{z}^{\mathbf{n}},\\
b^i_*(\mathbf{z}) &= \textstyle\sum_{\mathbf{n}}b^i_{*,\mathbf{n}}\cdot\mathbf{z}^{\mathbf{n}}, & a^i(\mathbf{z}) &= \textstyle\sum_{\mathbf{n}}a^i_{\mathbf{n}}\cdot\mathbf{z}^{\mathbf{n}}, & \omega(\mathbf{z}) &= \textstyle\sum_{\mathbf{n}}\omega^i_{\mathbf{n}}\cdot\mathbf{z}^{\mathbf{n}},\\
\pi(\mathbf{z}) &= \textstyle\sum_{\mathbf{n}}\pi^i_{\mathbf{n}}\cdot\mathbf{z}^{\mathbf{n}}, & \pi_*(\mathbf{z}) &= \textstyle\sum_{\mathbf{n}}\pi^i_{*,\mathbf{n}}\cdot\mathbf{z}^{\mathbf{n}},
\end{aligned}$$

where $\mathbf{z}^{\mathbf{n}} := z_1^{n_1}\cdots z_M^{n_M}$.

Next, we divide Eqs. (4.5) and (4.6) by $\pi(t)$ and take the limit of $t \to \infty$, yielding:

$$\pi^i_{*,\mathbf{n}} \lim_{t\to\infty}[\pi_*(t)/\pi(t)] + b^i_{*,\mathbf{n}} \lim_{t\to\infty}[b^i_*(t)/\pi(t)] = \beta^i_{\mathbf{n}} \lim_{t\to\infty}[\beta^i(t)/\pi(t)], \tag{4.7}$$

$$\pi^i_{\mathbf{n}} \lim_{t\to\infty}[\pi(t)/\pi(t)] + \alpha^i_{\mathbf{n}} \lim_{t\to\infty}[\alpha^i(t)/\pi(t)] + b^i_{\mathbf{n}} \lim_{t\to\infty}[b^i(t)/\pi(t)] = \omega^i_{\mathbf{n}} \lim_{t\to\infty}[\omega(t)/\pi(t)] + a^i_{\mathbf{n}} \lim_{t\to\infty}[a^i(t)/\pi(t)]. \tag{4.8}$$

Using similar renewal arguments as above, it is readily verified that under our model assumptions all these limits indeed exist with probability one if the stability conditions are satisfied.

Let us introduce some additional notation. We denote by $p_{pr,X}$ the probability of an arbitrary service at some queue in the system being preempted and by $p^i_{pr,I}$ the probability of an idle period at Q_i being preempted (i.e., the server switches before the next customer arrives to the queue). The mean cycle time of the server will be denoted by $\mathbb{E}[C]$. Finally, we define $\kappa_i := \lim_{t\to\infty}[a^i(t)/\pi(t)]$. This enables us to present the following theorem.

Theorem 4.6. *The p.g.f.'s of the queue-length distribution at Q_i, $i = 1, \ldots, M$, at specific embedded instants in a polling model operating under the pure exponential time-limited discipline read as follows:*

$$\begin{aligned}
\frac{p_{pr,X}}{1-p_{pr,X}} \cdot \pi^i_*(\boldsymbol{z}) + \vartheta_i \cdot p^i_{pr,I} \cdot b^i_*(\boldsymbol{z}) &= \gamma \cdot \beta^i(\boldsymbol{z}), \\
\pi^i(\boldsymbol{z}) + \gamma \cdot \alpha^i(\boldsymbol{z}) + \vartheta_i \cdot (1-p^i_{pr,I}) \cdot b^i(\boldsymbol{z}) &= \frac{\omega^i(\boldsymbol{z})}{1-p_{pr,X}} + \vartheta_i \cdot a^i(\boldsymbol{z}),
\end{aligned}$$

where

$$\begin{aligned}
p_{pr,X} &= 1 - \frac{\sum_j \lambda_j}{\sum_j \lambda_j / \tilde{X}_j(\xi_j)}, \\
p^i_{pr,I} &= 1 - \tilde{I}_i(\xi_i), \ i = 1, \ldots, M, \\
\vartheta_i &= \frac{1}{p^i_{pr,I}} \cdot \left(\gamma - \frac{\lambda_i}{\sum_j \lambda_j} \cdot \frac{1 - \tilde{X}_i(\xi_i)}{\tilde{X}_i(\xi_i)} \right), \ i = 1, \ldots, M, \\
\gamma &= \frac{1}{\sum_j \lambda_j \mathbb{E}[C]}.
\end{aligned}$$

It should be noted that $1/X_i(\xi_i)$ refers to the mean number of required server visits to serve a single customer. Next, we will present several lemmas and defer the proof of the theorem until the end of this section.

Lemma 4.7.

$$\lim_{t\to\infty}[\alpha^i(t)/\pi(t)] = \lim_{t\to\infty}[\beta^i(t)/\pi(t)] = \frac{1}{\sum_j \lambda_j \mathbb{E}[C]}, \ i = 1, \ldots, M.$$

Proof. First, notice that the number of visit completions, $\beta^i(t)$, differs at most one from the number of visit beginnings, $\alpha^i(t)$, for any $t \geq 0$. Therefore, we have that:

$$\lim_{t\to\infty}[\alpha^i(t)/\beta^i(t)] = 1.$$

Second, the number of visit beginnings at Q_i per cycle is exactly one. Hence, it is follows that:

$$\lim_{t\to\infty}[\alpha^i(t)/t] = \frac{1}{\mathbb{E}[C]},$$

where for $\mathbb{E}[C]$, the mean cycle time, we have:

$$\mathbb{E}[C] = \sum_j \left(\frac{1}{\xi_j} + c_{j-1,j}\right)$$

Third, the average total number of service completions per cycle is equal to the average total number of arrivals per cycle (assuming a stable system). This implies that:

$$\lim_{t\to\infty}[\pi(t)/t] = \sum_j \lambda_j.$$

Combining these three limits yields the desired result. □

Lemma 4.8.

$$\begin{aligned}\lim_{t\to\infty}[\omega(t)/\pi(t)] &= \frac{1}{1-p_{pr,X}},\\ \lim_{t\to\infty}[\pi_*(t)/\pi(t)] &= \frac{p_{pr,X}}{1-p_{pr,X}}.\end{aligned}$$

Proof. The $\lim_{t\to\infty}[\omega(t)/\pi(t)]$ is defined as the limit of the ratio of the total number of service beginnings and the total number of (successful) service completions. The numerator and denominator are related via the probability of an arbitrary service being preempted, $p_{pr,X}$. More precisely,

$$\lim_{t\to\infty}[\pi(t)/\omega(t)] = 1 - p_{pr,X}.$$

Similar to the relation between $\alpha^i(t)$ and $\beta^i(t)$, we note that $\omega(t)$ and $\pi(t)+\pi_*(t)$ differ at most one for $t \geq 0$. Therefore, we can write:

$$\lim_{t\to\infty}[\pi_*(t)/\pi(t)] = \lim_{t\to\infty}[(\omega(t)-\pi(t))/\pi(t)] = \frac{p_{pr,X}}{1-p_{pr,X}}.$$

□

Lemma 4.9.

$$\begin{aligned}\lim_{t\to\infty}[b^i(t)/\pi(t)] &= \vartheta_i \cdot (1-p^i_{pr,I}),\ i=1,\ldots,M,\\ \lim_{t\to\infty}[b^i_*(t)/\pi(t)] &= \vartheta_i \cdot p^i_{pr,I},\ i=1,\ldots,M\ .\end{aligned}$$

Proof. Recall that we set $\lim_{t\to\infty}[a^i(t)/\pi(t)] =: \vartheta_i$, where ϑ_i is a constant yet to be determined. These limits do not have a simple interpretation, but we can relate them to limits for other events. The number of events $a^i(t)$ and $b^i(t)$ are related as follows:

$$\lim_{t\to\infty}[b^i(t)/a^i(t)] = 1-p^i_{pr,I},$$

where $p^i_{pr,I}$, the probability that an idle period at Q_i is preempted, depends on i, and is given by:

$$p^i_{pr,I} = 1-\tilde{I}_i(\xi_i).$$

Analogously, $a^i(t)$ and $b^i_*(t)$ are related via:

$$\lim_{t\to\infty}[b^i_*(t)/a^i(t)] = p^i_{pr,I},$$

which completes the proof. □

Proof of Theorem 4.6. The presented equations follow by first evaluating the limit expressions in Eqs. (4.7) and (4.8). The limit expressions are derived in the lemmas above. However, these expressions still contain the unknowns $p_{pr,X}$ and ϑ_i, $i = 1,\ldots,M$.

For the service preemption probability $p_{pr,X}$, we obtain:

$$\begin{aligned}p_{pr,X} &= \sum_j \mathbb{P}(\text{service is preempted} \mid \text{s.b. at } Q_j)\\ &\qquad \times \mathbb{P}(\text{s.b. at } Q_j \mid \text{s.b. at some queue})\\ &= \sum_j (1-\tilde{X}_j(\xi_j)) \cdot \mathbb{P}(\text{s.b. at } Q_j \mid \text{s.b. at some queue})\\ &= \frac{\sum_j (1-\tilde{X}_j(\xi_j))\lambda_j/\tilde{X}_j(\xi_j)}{\sum_k \lambda_k/\tilde{X}_k(\xi_k)} = 1-\frac{\sum_j \lambda_j}{\sum_k \lambda_k/\tilde{X}_k(\xi_k)},\end{aligned}$$

where we use that:

$$\begin{aligned}&\mathbb{P}(\text{s.b. at } Q_i \mid \text{s.b. at some queue})\\ &= \frac{\lambda_i/(1-\mathbb{P}(\text{serv. at } Q_i \text{ is preempted} \mid \text{s.b. at } Q_i))}{\sum_j \lambda_j/(1-\mathbb{P}(\text{serv. at } Q_j \text{ is preempted} \mid \text{s.b. at } Q_j))}\\ &= \frac{\lambda_i/\tilde{X}_i(\xi_i)}{\sum_j \lambda_j/\tilde{X}_j(\xi_j)},\end{aligned} \tag{4.9}$$

where the condition *s.b. at* Q_i refers to the fact that we consider the embedded Markov chain of all service beginning instants. Notice that multiple service beginning events may correspond to a single customer.

The unknown $\vartheta_i,\ i = 1, \ldots, M$, can be found from Eq. (4.8) (or alternatively from Eq. (4.7)) by inserting all the limit expressions and summing both sides over $\mathbf{n}$. After several rearrangements and using that

$$\sum_{\mathbf{n}} \pi^i_{*,\mathbf{n}} = \mathbb{P}(\text{s.i. at } Q_i \mid \text{s.i. at some queue}) = \frac{\lambda_i/\tilde{X}_i(\xi_i) - \lambda_i}{\sum_j \left(\lambda_j/\tilde{X}_j(\xi_j) - \lambda_j\right)}, \quad (4.10)$$

where we use *s.i.* as short for service interruption, we eventually obtain:

$$\vartheta_i = \frac{1}{p^i_{pr,I}} \left(\gamma - \frac{\lambda_i}{\sum_j \lambda_j} \cdot \frac{1 - \tilde{X}_i(\xi_i)}{\tilde{X}_i(\xi_i)} \right).$$

The final step is to write these equations in terms of p.g.f.'s by multiplication with $\mathbf{z}^{\mathbf{n}}$ and summation over $\mathbf{n}$. □

4.4.3 Additional relations for the queue-length distributions at specific instants

We need additional relations to obtain the queue-length distributions at the different instants defined. Eisenberg [29] presents relations between $\pi^i(\mathbf{z})$ and $\omega^i(\mathbf{z})$ for the non-patient server model with non-preemptive services. We show that with a minor modification this relation can be used to relate both $\pi^i(\mathbf{z})$ and $\omega^i(\mathbf{z})$ and $\pi^i_*(\mathbf{z})$ and $\omega^i(\mathbf{z})$ in our model. Moreover, relations between $a^i(\mathbf{z})$ and $b^i(\mathbf{z})$ and between $a^i(\mathbf{z})$ and $b^i_*(\mathbf{z})$ can be established in a similar fashion. Finally, for completeness we repeat the relation from [29] between $\alpha^i(\mathbf{z})$ and $\beta^{i-1}(\mathbf{z})$.

Recall that $\omega^i(\mathbf{z}),\ \pi^i_*(\mathbf{z})$ and $\pi^i(\mathbf{z})$ refer to the number of customers at all queues at instants of service beginning, service interruption and successful service completion, respectively. The relations between these quantities are given in the following lemma.

Lemma 4.10.

$$\pi_i(\boldsymbol{z}) = \frac{\tilde{X}_i(\xi_i) \cdot (\sum_j \lambda_j/\tilde{X}_j(\xi_j))}{\sum_j \lambda_j} \cdot \hat{X}'_i(\boldsymbol{z}) \cdot \frac{\omega_i(\boldsymbol{z})}{z_i}, \quad (4.11)$$

$$\pi^i_*(\boldsymbol{z}) = (1 - \tilde{X}_i(\xi_i)) \cdot (\textstyle\sum_j \lambda_j/\tilde{X}_j(\xi_j) - \lambda_j) \cdot \hat{X}^*_i(\boldsymbol{z}) \cdot \omega_i(\boldsymbol{z}), \quad (4.12)$$

where

$$\hat{X}'_i(\boldsymbol{z}) = \frac{\tilde{X}_i(\xi_i + \sum_j \lambda_j(1 - z_j))}{\tilde{X}_i(\xi_i)},$$

$$\hat{X}^*_i(\boldsymbol{z}) = \frac{\xi_i}{\xi_i + \sum_j \lambda_j(1 - z_j)} \cdot \frac{1 - \tilde{X}_i(\xi_i + \sum_j \lambda_j(1 - z_j))}{1 - \tilde{X}_i(\xi_i)}.$$

Proof. Let us first consider Eq. (4.11), i.e., the relation between $\omega^i(\mathbf{z})$ and $\pi^i(\mathbf{z})$. Unfortunately, we cannot relate $\omega^i(\mathbf{z})$ and $\pi^i(\mathbf{z})$ in the straightforward manner as was done by Eisenberg [29]. In our case, every successful service completion instant has exactly one corresponding service beginning instant, while the correspondence the other way round is not true due to preemption (which is caused by the exogenously determined visit times of the server). This also implies that the long-term fraction of all service beginnings that occur at Q_i and the long-term fraction of all service completions that occur at Q_i are no longer equal for all parameter settings. Or, in mathematical terms, the relation $\omega^i(\mathbf{z}) \,|_{\mathbf{z}=\mathbf{1}} = \pi^i(\mathbf{z}) \,|_{\mathbf{z}=\mathbf{1}}$, $\forall_i$, is no longer necessarily true. Thus, the derivation of Eq. (4.11) requires a bit more work here.

Recall first the definitions of $\omega^i(\mathbf{z})$ and $\pi^i(\mathbf{z})$:

$$\begin{aligned}
\omega^i(\mathbf{z}) &= \sum_{n_1} \cdots \sum_{n_M} z_1^{n_1} \cdots z_M^{n_M} \\
&\qquad \times \mathbb{P}(\{\mathbf{N} = \mathbf{n}\} \cap \{\text{s.b. at } Q_i\} \mid \text{s.b. at some queue}), \\
\pi^i(\mathbf{z}) &= \sum_{n_1} \cdots \sum_{n_M} z_1^{n_1} \cdots z_M^{n_M} \\
&\qquad \times \mathbb{P}(\{\mathbf{N} = \mathbf{n}\} \cap \{\text{s.c. at } Q_i\} \mid \text{s.c. at some queue}),
\end{aligned}$$

where *s.c.* is used as short for service completion. Then, to circumvent the use of these unconditional p.g.f.'s, we define $\omega_c^i(\mathbf{z})$ and $\pi_c^i(\mathbf{z})$ as follows.

$$\begin{aligned}
\omega^i(\mathbf{z}) &= \sum_{n_1} \cdots \sum_{n_M} z_1^{n_1} \cdots z_M^{n_M} \cdot \mathbb{P}(\mathbf{N} = \mathbf{n} \mid \text{s.b. at } Q_i) \\
&\qquad \times \mathbb{P}(\text{s.b. at } Q_i \mid \text{s.b. at some queue}) \\
&=: \omega_c^i(\mathbf{z}) \cdot \mathbb{P}(\text{s.b. at } Q_i \mid \text{s.b. at some queue}), \\
\pi^i(\mathbf{z}) &= \sum_{n_1} \cdots \sum_{n_M} z_1^{n_1} \cdots z_M^{n_M} \cdot \mathbb{P}(\mathbf{N} = \mathbf{n} \mid \text{s.c. at } Q_i) \\
&\qquad \times \mathbb{P}(\text{s.c. at } Q_i \mid \text{s.c. at some queue}) \\
&=: \pi_c^i(\mathbf{z}) \cdot \mathbb{P}(\text{s.c. at } Q_i \mid \text{s.c. at some queue}),
\end{aligned}$$

where

$$\mathbb{P}(\text{s.c. at } Q_i \mid \text{s.c. at some queue}) = \frac{\lambda_i}{\sum_j \lambda_j}. \tag{4.13}$$

The latter equation follows immediately by the observation that the number of arriving customers per time unit is equal to the number of served customers per time unit for a system in equilibrium. Further, notice that the conditional probability $\mathbb{P}(\text{s.b. at } Q_i \mid \text{s.b. at some queue})$ was already given in Eq. (4.9).

Using that the distribution of the number of customers at the start of a service is independent of the success of the service, we can relate the conditional p.g.f.'s in the following manner:

$$\pi_c^i(\mathbf{z}) = \frac{\hat{X}_i'(\mathbf{z})}{z_i} \cdot \omega_c^i(\mathbf{z}), \tag{4.14}$$

where the term $1/z_i$ is due to the fact that the number of customers at Q_i at a service completion instant is exactly one less than at the service beginning instant and $\hat{X}_i'(\mathbf{z})$ is the p.g.f. of the number of customers that arrive at all queues during a service at Q_i given that is indeed completed. The latter is given by:

$$\begin{aligned} \hat{X}_i'(\mathbf{z}) &:= \mathbb{E}[\mathbf{z}^{N(X_i)} \mid X_i < Y_i] = \frac{\mathbb{E}[\mathbf{z}^{N(X_i)} \mathbf{1}_{\{X_i<Y_i\}}]}{\mathbb{P}(X_i < Y_i)} \\ &= \frac{\tilde{X}_i(\xi_i + \sum_j \lambda_j(1-z_j))}{\tilde{X}_i(\xi_i)}, \end{aligned} \tag{4.15}$$

where we used the notation $N(T)$ to refer to the number of arrivals to all queues during a random period T. The final equality sign follows from first conditioning on X_i and Y_i and next using that $\mathbb{E}[\mathbf{z}^{N(x)}]$ is Poisson distributed with parameter $\sum_j \lambda_j(1-z_j)x$ for a given x. Combining the definitions of the conditional p.g.f.'s and Eq. (4.14), we obtain:

$$\pi_i(\mathbf{z}) = \frac{\mathbb{P}(\text{s.c. at } Q_i \mid \text{s.c. at some queue})}{\mathbb{P}(\text{s.b. at } Q_i \mid \text{s.b. at some queue})} \cdot \hat{X}_i'(\mathbf{z}) \cdot \frac{\omega_i(\mathbf{z})}{z_i}. \tag{4.16}$$

Equation (4.12), which relates $\pi_*^i(\mathbf{z})$ and $\omega_i(\mathbf{z})$, is derived analogously and it resembles Eq. (4.16):

$$\pi_*^i(\mathbf{z}) = \frac{\mathbb{P}(\text{s.i. at } Q_i \mid \text{s.i. at some queue})}{\mathbb{P}(\text{s.b. at } Q_i \mid \text{s.b. at some queue})} \cdot \hat{X}_i^*(\mathbf{z}) \cdot \omega_i(\mathbf{z}), \tag{4.17}$$

where

$$\begin{aligned} \hat{X}_i^*(\mathbf{z}) &:= \mathbb{E}[\mathbf{z}^{N(Y_i)} \mid X_i > Y_i] \\ &= \frac{\xi_i}{\xi_i + \sum_j \lambda_j(1-z_j)} \cdot \frac{1 - \tilde{X}_i(\xi_i + \sum_j \lambda_j(1-z_j))}{1 - \tilde{X}_i(\xi_i)}. \end{aligned}$$

The derivation of $\hat{X}_i^*(\mathbf{z})$ is done analogously to the derivation of $\hat{X}_i'(\mathbf{z})$. Notice that the term $1/z_i$ is absent in Eq. (4.17), since no customer departs from the queue. The final step of the proof is to substitute the conditional probabilities of Eqs. (4.9), (4.10) and (4.13) into Eqs. (4.16) and (4.17). □

Remark 4.11 (Non-preemptive service)**.** *We note that for non-preemptive service the first ratio on the r.h.s. of Eq. (4.16) equals one as a service beginning corresponds uniquely to a service completion. Further, in this case, we have that the term* $\hat{X}_i'(\boldsymbol{z})$ *equals* $\mathbb{E}[\mathbf{z}^{N(X_i)}]$*, so that we obtain Eq. (17) of [29].*

Recall that $a^i(\mathbf{z})$, $b^i_*(\mathbf{z})$ and $b^i(\mathbf{z})$ refer to the number of customers at instants of idle period beginning, idle period interruption and idle period completion at Q_i, respectively. The relations between these quantities are given in the following lemma.

Lemma 4.12.

$$\begin{aligned} b^i(\boldsymbol{z}) &= \hat{I}'_i(\boldsymbol{z}) \cdot z_i \cdot a^i(\boldsymbol{z}), \\ b^i_*(\boldsymbol{z}) &= \hat{I}'_i(\boldsymbol{z}) \cdot a^i(\boldsymbol{z}), \end{aligned}$$

where

$$\hat{I}'_i(\boldsymbol{z}) = \frac{\tilde{I}_i(\xi_i + \sum_{j \neq i} \lambda_j (1 - z_j))}{\tilde{I}_i(\xi_i)}.$$

Proof. Let us first consider the relation between $a^i(\mathbf{z})$ and $b^i(\mathbf{z})$. We note that every idle period completion instant has a corresponding idle period beginning instant, while the correspondence the other way round is not true. This is due to the exponential visit time of the server. Whether the idle period gets interrupted depends on the arrival process and on the distribution of the visit time of the server only. In particular, it does not depend on the queue-length distribution at the start of an idle period. Thus, we can relate the generating functions $a^i(\mathbf{z})$ and $b^i(\mathbf{z})$ by the following observations. The p.g.f. of the number of customers that arrive at all queues different from Q_i during an idle period given that it is completed with an arrival is given by:

$$\hat{I}'_i(\mathbf{z}) := \mathbb{E}[\mathbf{z}^{N(I_i)} \mid I_i < Y_i] = \frac{\tilde{I}_i(\xi_i + \sum_{j \neq i} \lambda_j (1 - z_j))}{\tilde{I}_i(\xi_i)}.$$

This expression can be derived in a similar fashion as Eq. (4.15). Further, we note that exactly one customer arrives at Q_i at the end of the idle period. Together, this yields the following relation between $a_i(\mathbf{z})$ and $b_i(\mathbf{z})$:

$$b^i(\mathbf{z}) = \hat{I}'_i(\mathbf{z}) \cdot z_i \cdot a^i(\mathbf{z}).$$

In the same manner, the relation between $b^i_*(\mathbf{z})$ and $a^i(\mathbf{z})$ can be established:

$$b^i_*(\mathbf{z}) = \hat{I}'_i(\mathbf{z}) \cdot a^i(\mathbf{z}).$$

Notice that $\mathbb{E}[\mathbf{z}^{N(Y_i)} \mid I_i > Y_i] = \mathbb{E}[\mathbf{z}^{N(I_i)} \mid I_i < Y_i] = \hat{I}'_i(\mathbf{z})$, since both Y_i and I_i are assumed exponentially distributed. □

Recall that $\alpha^i(\mathbf{z})$ and $\beta^i(\mathbf{z})$ refer to the number of customers at visit beginning instants and visit completion instants at Q_i, respectively. The relations between these quantities are given in the following lemma.

Lemma 4.13.

$$\alpha^i(\boldsymbol{z}) = \hat{C}_{i-1,i}(\boldsymbol{z}) \cdot \beta^{i-1}(\boldsymbol{z}), \tag{4.18}$$

where

$$\hat{C}_{i-1,i}(\boldsymbol{z}) = \tilde{C}_{i-1,i}\left(\sum_j \lambda_j(1-z_j)\right).$$

Proof. There exists a well-known relation (see, e.g., [29]) between the number of customers that the server leaves behind in the system at departure from Q_{i-1} and the number of customers in the system that the server finds upon arrival to Q_i. This difference is characterized by the number of arriving customers during a switch-over time from Q_{i-1} to Q_i. We denote by $\hat{C}_{i-1,i}(\mathbf{z})$ the p.g.f. of this number, which is given by:

$$\hat{C}_{i-1,i}(\mathbf{z}) = \tilde{C}_{i-1,i}\left(\sum_j \lambda_j(1-z_j)\right),$$

since arrivals to the queues are according to a Poisson process. Hence, we immediately obtain Eq. (4.18). □

Altogether, we have derived $7 \cdot M$ relations between the $8 \cdot M$ p.g.f's of our interest. By combining the equations from Lemmas 4.10, 4.12 and 4.13 with the ones of Theorem 4.6, we are able to fully specify all the p.g.f.'s in terms of $\beta^i(\mathbf{z})$. It can then be shown that it is sufficient to determine the M p.g.f.'s for $\beta^i(\mathbf{z})$, $i = 1, \ldots, M$, explicitly; the latter will be done below.

4.4.4 Queue-length probabilities at visit completion instants

We shall determine the p.g.f. of the queue-length distribution at visit completion instants, $\beta^i(\mathbf{z})$, explicitly. This part of the analysis is based on work by Leung [67] for the study of a probabilistically-limited polling model, which was later extended in [68] to a time-limited polling model, and involves setting up an iterative scheme. A key role in this iterative scheme is played by the (auxiliary) p.g.f.'s $\phi_k(\mathbf{z})$ and $\phi_k^s(\mathbf{z})$, which will be explained below. In the final step of the iteration scheme, $\beta^i(\mathbf{z})$ is obtained as a simple function of $\phi_k^s(\mathbf{z})$.

We consider a tagged queue i and we will leave out the subscript and superscript i whenever it does not lead to ambiguity. We will recursively relate the number of customers present at the end of a server visit to the tagged queue to the number present at the beginning. To this end, we partition a server visit by means of so-called *marked events*. We mark three types of events that may occur during a server visit

viz., a customer arrival to an empty Q_i, a service completion at Q_i, and a departure of a server from Q_i. Further, we define for $k \geq 0$, $N_k^i := (N_{k,1}^i, \ldots, N_{k,M}^i)$, where $N_{k,j}^i$, $j = 1, \ldots, M$, denotes the number of customers at Q_j just after the epoch of the k-th marked event at Q_i. We let N_0^i refer to the number of customers present at the arrival of the server to Q_i. Finally, we denote by the random variable $\Upsilon_i \geq 1$ the number of marked events that occurs during a visit time of Q_i. Due to the Poisson arrival stream, it is clear that the sequence $\{N_k^i\}_{k=0}^{\infty}$ is a Markov chain. We let $\phi_k^i(\mathbf{z})$ be the joint p.g.f. of the number of customers at all queues at the k-th marked event epoch at Q_i and marked event k is not the final marked event during the visit (i.e., marked event $k+1$ will occur). Similarly, we introduce $\phi_k^{s,i}(\mathbf{z})$ as the joint p.g.f. of the number of customers at all queues at the k-th marked event epoch at Q_i and k is the final marked event (i.e., marked event k is a server departure event, and marked event $k+1$ will not occur). Thus, we formally define for $k \geq 1$

$$\begin{aligned} \phi_k^i(\mathbf{z}) &:= \mathbb{E}[\mathbf{z}^{N_k^i} \mathbf{1}_{\{\Upsilon_i > k\}}], \\ \phi_k^{s,i}(\mathbf{z}) &:= \mathbb{E}[\mathbf{z}^{N_k^i} \mathbf{1}_{\{\Upsilon_i = k\}}], \end{aligned}$$

where $\mathbf{1}_{\{A\}}$ is the indicator function of event A ($\mathbf{1}_{\{A\}} = 1$, if A is true, and 0 otherwise). Let $N(T)$ be the number of arrivals during a random period T, I be the (exponential) interarrival time of customers at Q_i, and $\hat{C}_{i-1,i}(\mathbf{z})$ be the p.g.f. of the number of arrivals during a switch-over time from Q_{i-1} to Q_i. Then, we can present the explicit expression for $\beta^i(\mathbf{z})$ in the following lemma.

Lemma 4.14.

$$\beta^i(\boldsymbol{z}) = \sum_{k=1}^{\infty} \phi_k^{s,i}(\boldsymbol{z}),$$

where

$$\begin{aligned} \phi_k^i(\boldsymbol{z}) &= \phi_{k-1}^i(\boldsymbol{z}) \mid_{z_i=0} \cdot \mathbb{E}[\boldsymbol{z}^{N(I)} \boldsymbol{1}_{\{Y>I\}}] \cdot z_i \\ &\quad + \left(\phi_{k-1}^i(\boldsymbol{z}) - \phi_{k-1}^i(\boldsymbol{z}) \mid_{z_i=0}\right) \cdot \mathbb{E}[\boldsymbol{z}^{N(X)} \boldsymbol{1}_{\{Y>X\}}]/z_i, \qquad (4.19) \\ \phi_k^{s,i}(\boldsymbol{z}) &= \phi_{k-1}^i(\boldsymbol{z}) \mid_{z_i=0} \cdot \mathbb{E}[\boldsymbol{z}^{N(Y)} \boldsymbol{1}_{\{Y<I\}}] \\ &\quad + \left(\phi_{k-1}^i(\boldsymbol{z}) - \phi_{k-1}^i(\boldsymbol{z}) \mid_{z_i=0}\right) \cdot \mathbb{E}[\boldsymbol{z}^{N(Y)} \boldsymbol{1}_{\{Y<X\}}], \qquad (4.20) \end{aligned}$$

where

$$\phi_0^i(\boldsymbol{z}) = \hat{C}_{i-1,i}(\boldsymbol{z})\beta^{i-1}(\boldsymbol{z}), \tag{4.21}$$

$$\mathbb{E}[\boldsymbol{z}^{N(I)}\mathbf{1}_{\{Y>I\}}] = \tilde{I}_i(\xi_i + \sum_{j\neq i}\lambda_j(1-z_j)), \tag{4.22}$$

$$\mathbb{E}[\boldsymbol{z}^{N(X)}\mathbf{1}_{\{Y>X\}}] = \tilde{X}_i(\xi_i + \sum_{j}\lambda_j(1-z_j)), \tag{4.23}$$

$$\mathbb{E}[\boldsymbol{z}^{N(Y)}\mathbf{1}_{\{Y<I\}}] = \frac{\xi_i}{\xi_i + \sum_{j\neq i}\lambda_j(1-z_j)} \cdot (1 - \tilde{I}_i(\xi_i + \sum_{j\neq i}\lambda_j(1-z_j))), \tag{4.24}$$

$$\mathbb{E}[\boldsymbol{z}^{N(Y)}\mathbf{1}_{\{Y<X\}}] = \frac{\xi_i}{\xi_i + \sum_{j}\lambda_j(1-z_j)} \cdot (1 - \tilde{X}_i(\xi_i + \sum_{j}\lambda_j(1-z_j))). \tag{4.25}$$

Proof. Equation (4.19) follows from the following observations. First, the time between the $(k-1)$-th and the k-th marked event epoch (and thus also the number of arriving customers) depends on whether at least one customer was present at the $(k-1)$-th marked event epoch. This explains why the equation consists of two parts. Second, the number of customers at all queues at a marked event epoch is equal to the number present at the previous marked event epoch adjusted for the arrivals and departures in the meantime. Similarly, Eq. (4.20) can be explained. By definition $\phi_0^i(\mathbf{z}) = \alpha^i(\mathbf{z})$, so that Eq. (4.21) directly follows from Eq. (4.18). It is also worthwhile to note that $\phi_0^i(\mathbf{1}) = 1$, while $\phi_k^i(\mathbf{1}) < 1$, for all $k = 1, 2, \ldots$, since the $(k+1)$th service period need not occur at all during a visit to Q_i. The expressions in Eqs. (4.22), (4.23), (4.24) and (4.25) follow from simple probabilistic calculations.

Notice that the final marked event is always a departure of the server from Q_i. Therefore, we can write for the number of customers at the queues at the end of a server visit to Q_i:

$$\beta^i(\mathbf{z}) = \sum_{k=1}^{\infty}\phi_k^{s,i}(\mathbf{z}), \tag{4.26}$$

which completes the proof. □

We set up an iterative scheme to compute $\beta^i(\mathbf{z})$ numerically. The scheme is constructed in terms of Discrete Fourier Transforms (DFTs) as described in Sect. 1.3.3.2. The pseudo-code of the iterative scheme is presented in Algorithm 4.16. The common values that have been used for the convergence parameters are $\epsilon = 10^{-6}$ and $\delta = 10^{-9}$. Finally, via the Inverse Fourier Transform, the steady-state probabilities $P_{\beta^i}(\mathbf{n})$ are obtained. It is good to keep in mind that this approach is mainly applicable to systems with a light to moderate load.

Remark 4.15 (Queue-length at service completion epochs)**.** *The p.g.f. $\pi^i(\mathbf{z})$, which refers to the queue-length at service completion instants, can now be obtained using the derived relations (see Sect. 4.4.2-4.4.3) and the explicit computation of $\beta^i(\mathbf{z})$.*

However, $\pi^i(\mathbf{z})$ can also directly be expressed in terms of the introduced auxiliary p.g.f. $\phi_k^i(\mathbf{z})$ using arguments from the theory of regenerative processes and some simple manipulations. Eventually, this yields:

$$\pi^i(\mathbf{z}) = \frac{\mathbb{P}(s.c.\ at\ Q_i \mid s.c.)}{\mathbb{E}[\#\ s.c.\ per\ visit\ to\ Q_i]} \cdot \sum_{k=1}^{\infty} \phi_k^i(\mathbf{z}),$$

where

$$\mathbb{E}[\#\ s.c.\ per\ visit\ to\ Q_i] = \sum_{k=1}^{\infty} \phi_k^i(\mathbf{1}) = \lambda_i \cdot \mathbb{E}[C],$$

and where $\mathbb{P}(s.c.\ at\ Q_i \mid s.c.)$ is given in Eq. (4.13).

Algorithm 4.16. *Pseudo-code of the iterative scheme for determining $\check{\beta}^i(\boldsymbol{k}), \forall_i, \forall_{\boldsymbol{k}}$.*

$\check{\beta}^{i_0}(\boldsymbol{k}) = 1,\ \forall_{i_0},\ \forall_{\boldsymbol{k}}$; *(start with an empty system)*
FOR $i_1 = 1, \ldots, M$
set $i_2 := i_1$;
REPEAT
$\bar{\beta}^{i_2}(\boldsymbol{k}) = \check{\beta}^{i_2}(\boldsymbol{k}),\ \forall_{\boldsymbol{k}}$;
set $j := 0$;
set $\phi_0^i(\boldsymbol{k}) = \check{\beta}^{i_2-1}(\boldsymbol{k}) \cdot \check{C}_{i_2-1,i_2}(\boldsymbol{k})$;
REPEAT
set $j := j + 1$;
compute $\phi_j^{i_2}(\boldsymbol{k}),\ \forall_{\boldsymbol{k}}$, *using Eq. (4.19);*
compute $\phi_j^{s,i_2}(\boldsymbol{k}),\ \forall_{\boldsymbol{k}}$, *using Eq. (4.20);*
compute $\check{\beta}^{i_2}(\boldsymbol{k}) = \sum_{l=1}^{j} \phi_l^{s,i_2}(\boldsymbol{k}),\ \forall_{\boldsymbol{k}}$;
UNTIL $1 - Re(\check{\beta}^{i_2}(\boldsymbol{0})) < \delta$
set $i_2 := MOD(i_2, M) + 1$;
UNTIL $\lvert Re(\check{\beta}^{i_1}(\boldsymbol{k})) - Re(\bar{\beta}^{i_1}(\boldsymbol{k}))\rvert < \epsilon,\ \forall_{\boldsymbol{k}}$
END

Remark 4.17 (Relation to the work of Leung [67, 68])**.** *In our model, interruptions can occur both during services and during idle periods, while in Leung's time-limited model (see [68]) only services can be interrupted. The latter is due to the fact that in Leung's model the server moves to the next queue if there are no customers present anymore. Due to the additional event of idle period interruptions in our model, the probability of $\psi_i(j) \geq 1$ (i.e., one or more customers present at Q_i after j services)*

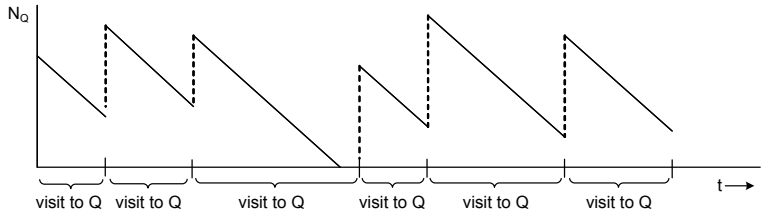

Figure 4.1: Sketch of the induced queue-length process at queue Q.

of Eq.(9) of [67] which is conditioned on the event that no interruption occurs during the jth service is no longer equal to the unconditional probability. Nevertheless, we strongly believe that for our model the approach of [67] could still be followed to find $\beta^i(\mathbf{z})$. However, the expressions will become quite involved, so that we proposed an unconditional approach here.

4.4.5 Steady-state queue-length probabilities and sojourn times

The exponential visit times allow us to obtain the steady-state queue-length probabilities. To see this, consider the full queue-length process and condition on the server being at some tagged queue Q. Next, remove from this full process the time periods that the server is not at Q and concatenate the remaining parts. This induced process then consists of a series of exponentially distributed periods with jumps between each two periods (see Fig. 4.1). These jumps in fact constitute a Poisson batch arrival process and reflect the arrivals to the system when the server is not at Q. Due to the PASTA property, these batches see time average behavior upon arrival. Moreover, the system observed by the arriving batches is exactly the system as observed by the server when it departs from Q. Thus, we have that a departing server observes the system in steady-state conditioned on the position of the server.

Let us denote the p.g.f. of the number of customers present at a random instant during a switch-over time from Q_{i-1} to Q_i by $\hat{C}^R_{i-1,i}(\mathbf{z})$. It is well-known that this p.g.f. should satisfy:

$$\hat{C}^R_{i-1,i}(\mathbf{z}) = \beta^i(\mathbf{z}) \cdot \frac{1 - \tilde{C}_{i-1,i}\left(\sum_j \lambda_j(1-z_j)\right)}{c_{i-1,i} \cdot \left(\sum_j \lambda_j(1-z_j)\right)}.$$

Hence, by conditioning on the position of the server (either it is visiting a queue or switching between two queues), we may write for $P(\mathbf{z}) := \mathbb{E}[\mathbf{z}^{\mathbf{N}}]$, the joint p.g.f. of the steady-state queue lengths,

$$P(\mathbf{z}) = \frac{1}{\mathbb{E}[C]} \cdot \sum_{i=1}^{M} \left(\beta^i(\mathbf{z}) \cdot \frac{1}{\xi_i} + \hat{C}^R_{i-1,i}(\mathbf{z}) \cdot c_{i-1,i} \right). \tag{4.27}$$

It should be noted that in the discussion above the size of the arriving batches depends on the realizations of the random visit times at the other queues. Therefore, even for zero switch-over times, the queue-length processes at the different queues are definitely not independent. Besides, it is good to notice that the acquirement of these probabilities is not a very common result in the polling literature. For most polling models that have been analyzed no steady-state queue-length probabilities are known.

Let us next turn to the marginal distribution. We denote by $P_i(z)$ the marginal queue-length distribution for Q_i, which readily follows from $P(\mathbf{z})$, i.e.,

$$P_i(z) = P(z_1, \ldots, z_M)|_{z_j=1,\ j\neq i,\ z_i=z}.$$

Also, the marginal distribution $P_i(z)$ could be obtained using the techniques discussed in Sect. 4.3. Alternatively, these marginal probabilities can also be obtained via $\pi^i(\mathbf{z})$ (see Remark 4.15). Using an up- and downcrossings argument and the PASTA property [106], we may write:

$$P_i(z) = \frac{\pi^i(z_1, \ldots, z_M)|_{z_j=1,\ j\neq i,\ z_i=z}}{\pi^i(1, \ldots, 1)}.$$

We note that this latter method does not rely on the exponentiality of the visit times and it may also be applied for the analysis of other service disciplines.

From the marginal queue-length distribution, the LST of the sojourn time (or delay) of a customer, which we denote by $\tilde{D}_i(s)$, can be obtained using the distributional form of Little's law (see [55]). In particular, we have that:

$$\tilde{D}_i(s) = P_i(z)\ |_{z=1-s/\lambda_i}\ .$$

Thus, all moments of the sojourn time at each queue can be determined.

4.5 Model extensions

The analysis of the basic polling model may be extended in various directions. Extending the basic model increases the range of applications that can be modelled while other applications may be modelled in greater detail. Moreover, such extensions are also interesting from a theoretical point of view. The specific extensions that we discuss here are the following: customer routing, Markovian polling of the server, and non-exponential server visit times. We will first outline in detail how to incorporate the first two extensions and finally comment on relaxing the exponential server visit times in the model.

4.5.1 Customer routing

The basic model assumes that customers upon being served leave the system without requiring any additional work. We drop this assumption and allow customers to join

any other queue in a probabilistic fashion upon service completion. More formally, customers who have completed their service at Q_i will join Q_j, $i, j = 1, \ldots, M$ with probability $r_{ij} \geq 0$ and with probability $r_{i0} \geq 0$ they will leave the system. Clearly, these routing probabilities r_{ij} must satisfy $\sum_{j=0}^{M} r_{ij} = 1$, $i = 1, \ldots, M$. We will assume in the sequel that $r_{ii} = 0$, i.e., no self loops are allowed. However, we note that $r_{ii} > 0$ might also be incorporated in the model (see Remark 4.18). Finally, we let $r^i(\mathbf{z})$ denote the p.g.f. of the number of arrivals to all queues generated by a single departing customer at Q_i, i.e., $r^i(\mathbf{z}) = r_{i0} + \sum_j r_{ij} z_j$.

Incorporating customer routing into the framework for the computation of the joint queue-length probabilities at visit completion instants (see Sect. 4.4.4) is then readily done. Specifically, the expression for $\phi_k^i(\mathbf{z})$ in Eq. (4.19) should simply be adjusted for $k = 1, 2, \ldots$, to:

$$\begin{aligned} \phi_k^i(\mathbf{z}) &= \phi_{k-1}^i(\mathbf{z}) \mid_{z_i=0} \cdot \mathbb{E}[\mathbf{z}^{N(I)} \mathbf{1}_{\{Y>I\}}] \cdot z_i \\ &\quad + \left(\phi_{k-1}^i(\mathbf{z}) - \phi_{k-1}^i(\mathbf{z}) \mid_{z_i=0}\right) \cdot \mathbb{E}[\mathbf{z}^{N(X)} \mathbf{1}_{\{Y>X\}}] \cdot \frac{r^i(\mathbf{z})}{z_i}. \end{aligned}$$

All other expressions involved in the computation of $\beta^i(\mathbf{z})$ remain unchanged.

Remark 4.18 (Self-loops in customer routing)**.** *The case $r_{ii} > 0$ may be incorporated in the model by appropriately scaling the service rates and the routing probabilities at a queue. To be precise, the service time should be scaled such that its mean, denoted by $1/\mu_i'$, equals $1/(\mu_i(1 - r_{ii}))$. The scaled routing probabilities, r_{ij}', should be set to $r_{ij}/(1 - r_{ii})$, $j \neq i$, while r_{ii}' should be set to zero. In this modified system, the server serves each arriving customer only once, but as each brings more work to the queue the total effective amount of work arriving per time unit to the queue remains the same as for the original system. Finally, using a sample-path comparison, it can readily be seen that the queue-length distribution of the modified system is equal to the one of the original system.*

4.5.2 Markovian polling of the server

We have assumed until now that the server polls the queues according to a fixed cyclic schedule. To allow for more general polling schedules, the routing (or polling) of the server is taken to follow a Markovian pattern (see [18, 69]). The polling of the server being Markovian signifies that upon the end of a visit to a queue the next queue to be served is selected in a probabilistic manner. More specifically, the probability of choosing the next queue upon a visit completion depends only the queue left behind by the server. Thus, we let $s_{i,j} \geq 0$, $j = 1, \ldots, M$, denote the probability upon departure from Q_i that the next queue that will be visited is Q_j. We note that this next queue could also be Q_i again. This feature is particularly meaningful in the case of nonzero switch-over times, whereas in the case of zero switch-over times one may simply set $s_{i,i} = 0$, $i = 1, \ldots, M$. For instance, in an applied setting of wireless communication with overlapping transmission ranges (and

thus zero switch-over times), the server can not leave Q_i without visiting some other $Q_j,\ j \neq i$, so that subsequent visits to the same queue could not occur. However, if transmission ranges do not overlap and there exists some "free space" (and thus switch-over times are nonzero), the server may leave Q_i's range, enter the free space and return to Q_i again without visiting any other queue in the meanwhile.

Let us describe the embedded process of visits to the queues (thus neglecting switch-over times for the moment) by a discrete-time Markov chain $X_n \in \{1, \ldots, M\}$, $n \geq 0$, driven by the transition probability matrix $S = \{s_{i,j}\}_{i,j=1,\ldots,M}$. We assume that this Markov chain (or jump chain) has an equilibrium distribution which we denote by $\tau_i,\ i = 1, \ldots, M$. Our interest is in the fraction of time that the random process is at a state i (denoted by κ_i) and also in the fraction that the process is switching from state i to j (denoted by $\Delta_{i,j}$). Let us define the cycle time of Q_i by C_i as the time between two consecutive polling instants of the server at Q_i. This means that a cycle C_i comprises exactly one visit time to Q_i. The mean cycle time can then be expressed as a weighted sum of mean visit and mean switch-over times as follows:

$$\mathbb{E}[C_i] = \sum_k \frac{\tau_k}{\tau_i} \left(\mathbb{E}[Y_k] + \sum_j s_{k,j} \cdot c_{k,j} \right),\ i = 1, \ldots, M,$$

where $\mathbb{E}[Y_k]$ denotes the mean visit time to Q_k and $c_{k,j}$ the mean switch-over time from Q_k to Q_j. Notice that the ratio τ_k/τ_i equals the mean number of visits to Q_k per visit to Q_i. There exists a simple relation between the mean cycle times of the different queues in the sense that the product $\mathbb{E}[C_i] \cdot \tau_i$ is constant for all queues. This relation immediately explains why under Markovian polling, in contrast to the cyclic routing case (for which $\tau_i = 1/M,\ i = 1, \ldots, M$), the mean cycle times per state are not necessarily equal for all states.

It then readily follows for $\kappa_i,\ i = 1, \ldots, M$, which is in fact the long-term fraction of time the server is available at Q_i, that:

$$\begin{aligned} \kappa_i &= \frac{\mathbb{E}[Y_i]}{\mathbb{E}[C_i]} = \frac{\mathbb{E}[Y_i]}{\sum_k \frac{\tau_k}{\tau_i}(\mathbb{E}[Y_k] + \sum_j s_{k,j} \cdot c_{k,j})} \\ &= \frac{\tau_i \mathbb{E}[Y_i]}{\sum_k \tau_k (\mathbb{E}[Y_k] + \sum_j s_{k,j} \cdot c_{k,j})}, \end{aligned} \tag{4.28}$$

where it is should be noticed that the denominator in the rightmost ratio of Eq. (4.28) does not depend on Q_i. Similarly, for $\Delta_{i,j}$ we find that:

$$\Delta_{i,j} = \frac{\tau_i \cdot s_{i,j} \cdot c_{i,j}}{\sum_k \tau_k (\mathbb{E}[Y_k] + \sum_j s_{k,j} \cdot c_{k,j})}.$$

Again, the adjustments in our modelling framework for the joint queue-length probabilities are relatively simple. We note that routing of the server solely plays a role in the relation between the queue length at a visit beginning instant and the

queue length at the preceding visit completion instant. Since this preceding queue that was served is now random, the expression for $\alpha^i(\mathbf{z})$ becomes:

$$\alpha^i(\mathbf{z}) = \sum_j q_{j,i} \hat{C}_{j,i}(\mathbf{z}) \beta^j(\mathbf{z}),$$

where $q_{j,i}$ is the probability that given that the server is at Q_i the preceding queue has been Q_j, and is given by (see, e.g., [56, p.28]):

$$q_{j,i} = \frac{\tau_j \cdot s_{j,i}}{\tau_i}.$$

We note that this affects only the initial expression for $\phi_0^i(\mathbf{z})$ in the iterative computation scheme (see Eq. (4.21)). Finally, the p.g.f. of the steady-state joint queue-length probabilities then yields:

$$P(\mathbf{z}) = \sum_{i=1}^{M} \left(\beta^i(\mathbf{z}) \cdot \kappa_i + \sum_{j=1}^{M} \hat{C}_{i,j}^R(\mathbf{z}) \cdot \Delta_{i,j} \right),$$

where

$$\hat{C}_{i,j}^R(\mathbf{z}) = \beta^i(\mathbf{z}) \cdot \frac{1 - \tilde{C}_{i,j}(\sum_k \lambda_k(1 - z_k))}{c_{i,j} \cdot \sum_k \lambda_k(1 - z_k)}.$$

4.5.3 Non-exponential visit times of the server

It is assumed in the basic polling model that visit times (or time limits) of the server are exponentially distributed. However, in practice these visit times distributions might as well range from deterministic to even heavy-tailed distributions. Deterministic time limits appear for instance in the area of token bus networks (see [95] and references therein). In the context of mobile ad hoc networking, the time limit appears as contact time in the experimental analysis of delay-tolerant and opportunistic networking. More precisely, the contact time is then defined as the contiguous period of time during which two (wireless) devices can communicate. Research on simple analytic mobility models has shown (see, e.g., [77]) that these contact times are exponentially distributed. Conversely, traces of real-world experiments for mobile users in WLAN campus networks have illustrated that the contact times exhibit power-law behavior [4, 48, 49]. More recently proposed mobility models based on social-network theory (see, e.g., [11, 77]), but also models that mimic commuters networks [30] show power-law behavior for the contact times.

Hence, it is certainly useful to consider visit-time distributions beyond the exponential one. Let us consider first the family of phase-type distributions [80], since by using these phase-type distributions any (light-tailed) visit-time distribution can be approximated arbitrary close. Replacing the exponential distribution for the server

visit time by a phase-type distribution may readily be done. This is due to the fact that each phase in a phase-type distribution is still exponentially distributed and we have shown that we can relate the queue length at the start and end of such a period. For example, let us consider a polling model consisting of two queues, zero switch-over times and with a cyclic server operating under the pure time-limited discipline. The visit times to Q_1 are taken exponentially distributed. The visit times to Q_2 are assumed Coxian-2 distributed with parameter b, i.e., with probability b the visit time to Q_2 consist of two exponential phases and with probability $1-b$ of only one exponential phase. We may incorporate this in our model by defining (cf. [29]) visit stage m_1 which corresponds to the visit to Q_1. Besides, we define stages m_2 and m_3 which refer to the first phase and the second phase of the visit to Q_2. This means that we should organize the server routing matrix S in terms of stages and assign the nonzero probabilities $s_{m_1,m_2} = 1$, $s_{m_2,m_3} = b$ and $s_{m_2,m_1} = 1-b$. This leads to the following relations between the different stages:

$$\alpha^{m_1}(\mathbf{z}) = (1-b)\cdot\beta^{m_2}(\mathbf{z}) + b\cdot\beta^{m_3}(\mathbf{z}), \quad \alpha^{m_2}(\mathbf{z}) = \beta^{m_1}(\mathbf{z}), \quad \alpha^{m_3}(\mathbf{z}) = \beta^{m_2}(\mathbf{z}).$$

Unfortunately, the extreme case of heavy-tailed distributed visit times may not be analyzed within our framework. Though, in the other extreme case of deterministic visit times, our framework may be used as an approximation. We note that in this case the queues in the polling system become independent when the server follows a cyclic visit schedule. Hence, the analysis of the system boils down to analyzing a specific single-server M/G/1 vacation queue with preemptive service. Although this may seem simple at first sight, the analysis turns out to be rather complex (see [108]). In particular, the probability of the queue being empty during a visit is not tractable and thus [108] follows an approximation approach. Alternatively, we could apply our framework and approximate the visit times by an Erlang-k distribution as discussed above. Moreover, this approach remains applicable even if the server follows a non-cyclic (e.g., Markovian) polling strategy.

4.6 Concluding remarks

Polling models operating under the pure exponential time-limited discipline may arise quite naturally as a performance model in the context of mobile wireless technologies. We have analyzed this novel service discipline in great detail. First, we have related the p.g.f.'s of the joint queue-length distribution at specific embedded epochs by extending the beautiful approach introduced by Eisenberg [29] for the exhaustive and gated service discipline. Next, we have presented an exact (numerical) analysis to find several of these p.g.f.'s explicitly. In particular, we have coupled the p.g.f. of the joint queue length at visit completion instants, $\beta^i(\mathbf{z})$, to the p.g.f. at visit beginning instants, $\alpha^i(\mathbf{z})$, in a recursive fashion by segmenting the visit according to specific intermediate events. Thus, the queue-length distribution can be obtained at all embedded epochs defined. Moreover, owing to the exponential structure of the visit

times, the steady-state joint queue-length distribution has been found. Finally, we have discussed several enhancements of the basic polling model. These enhancements greatly expand the applicability of the techniques presented in this chapter with regard to the development of accurate performance models for MANETs.

CHAPTER

5

Transient analysis for exponential time-limited polling models

5.1 Introduction

Already in the introductory chapter, Chapter 1, we have highlighted the celebrated approach to analyze polling systems based on constructing Markov chains at specific embedded epochs and subsequently relating the state space at these epochs (see also [29]). The key relation within this approach relates the joint queue length at the end of a server visit to Q_i to the joint queue length at the start of the visit to Q_i and can be written in the following general form:

$$\beta^i(\mathbf{z}) = \mathcal{F}(\alpha^i)(\mathbf{z}), \tag{5.1}$$

where $\beta^i(\mathbf{z})$ is the p.g.f. of the joint queue length at the end of a server visit to Q_i, $\alpha^i(\mathbf{z})$ is the p.g.f. of the joint queue length at the start of a server visit to Q_i and $\mathcal{F}$ is an operator representing the mapping between the queue lengths at these epochs and depends on the assumed service discipline.

In the previous chapter, we have established this key relation for the pure exponential time-limited discipline with preemptive service in an indirect, recursive manner (see Sect. 4.4.4). Under the assumption of exponential service times, Al Hanbali et al. [H3] derived a direct relation between $\beta^i(\mathbf{z})$ and $\alpha^i(\mathbf{z})$ for this discipline using a

matrix geometric approach. In the same article, the authors also rederived a result of Eliazar and Yechiali [31] for the exhaustive exponential time-limited discipline for the special case of exponential service times. The latter authors studied the exhaustive exponential time-limited discipline for preemptive service [31, 32]. Observing that upon successful service completion at a queue the busy period in fact regenerates, the authors could obtain a closed-form relation between the joint queue length at the end and the start of a server visit of the following form:

$$\beta^i(\mathbf{z}) = c(\mathbf{z}) \cdot (\alpha^i(\mathbf{z}) - \alpha^i(\mathbf{z}_i^*)) + \alpha^i(\mathbf{z}_i^*),$$

where $\alpha^i(\mathbf{z}_i^*) := \alpha^i(z_1, \ldots, z_{i-1}, k_i(\mathbf{z}), z_{i+1}, \ldots, z_M)$, and $c(\mathbf{z})$ and $k_i(\mathbf{z})$ are functions of $\mathbf{z}$ with $k_i(\mathbf{z})$ being related to the LST of the busy period of a customer at Q_i. Leung [68] analyzed the key relation for the exhaustive exponential time-limited discipline and non-preemptive service. This was done in a recursive manner by conditioning on specific intermediate events during a server visit. Ultimately, De Souza e Silva et al. [95] studied the key relation, Eq. (5.1), for the exhaustive deterministic time-limited discipline both for preemptive and non-preemptive service. Under the assumption of exponential service times, the authors analyze the transient behavior of the system by applying uniformization techniques as to find the joint queue-length distribution $\beta^i(\mathbf{z})$.

In the present chapter, we will derive a direct relation for the evolution of the joint queue-length during the course of a server visit, i.e., we relate $\beta^i(\mathbf{z})$ and $\alpha^i(\mathbf{z})$ in a direct manner. This will be done both for the pure and the exhaustive exponential time-limited discipline for general service time requirements and preemptive service. More specifically, service of individual customers is according to the *preemptive-repeat-random* strategy, i.e., if a service is interrupted, then at the next server visit a new service time will be drawn from the original service-time distribution. Moreover, we incorporate customer routing in our analysis, such that it may be applied to a large variety of queueing networks with a single server operating under one of the before-mentioned time-limited service disciplines. The analysis of the pure time-limited discipline builds on several known results for the transient analysis of the M/G/1 queue. Besides, for the analysis of the exhaustive discipline, we will derive several new results for the transient analysis of the M/G/1 queue during a busy period. The final expressions (both for the exhaustive and pure case) that we obtain for the key relations are of the form:

$$\beta^i(\mathbf{z}) = d_1(\mathbf{z}) \cdot (\alpha^i(\mathbf{z}) - \alpha^i(\mathbf{z}_i^*)) + d_2(\mathbf{z}) \cdot \alpha^i(\mathbf{z}_i^*),$$

where $\alpha^i(\mathbf{z}_i^*) := \alpha^i(z_1, \ldots, z_{i-1}, l_i(\mathbf{z}), z_{i+1}, \ldots, z_M)$, $d_1(\mathbf{z})$ and $d_2(\mathbf{z})$ are functions of $\mathbf{z} = (z_1, \ldots, z_M)$ which are largely determined by the LST of the service-time distribution, and $l_i(\mathbf{z})$ is related to the length of the busy period of a customer at Q_i. These relations generalize previous results by incorporating customer routing ([32] and [H3]) and by relaxing the exponentiality assumption on the service times [H3]. Finally, based on the interpretation of these key relations, we formulate a

conjecture for the key relation for any branching-type service discipline operating under an exponential time-limit.

Remember that complementary to these key relations, there exists a relation between the p.g.f.'s $\beta^i(\mathbf{z})$ and $\alpha^j(\mathbf{z})$ which couples the queue length at the start of a visit to Q_j to the queue length at the end of a visit to Q_i (see Chapter 4). Together, these relations for all queues in the system give rise to a system of equations which may be solved numerically in an iterative fashion. In this respect, the key relation for the time-limited discipline derived in this chapter does not only present a more elegant counterpart of the recursive relation obtained in Chapter 4, but also offers a computationally more attractive alternative.

The rest of this chapter is organized as follows. We describe the model and the notation in Sect. 5.2. The key relations for the pure and the exhaustive exponential time-limited discipline are presented in Sects. 5.3 and 5.4, respectively. We wrap up with a discussion on the final results for the key relations in Sect. 5.5. In Appendix 5.A, the transient analysis for the M/G/1 queue during a busy period is given. The complete proofs of the key relations are given in Appendices 5.B and 5.C.

5.2 Model and notation

Let us consider the basic polling system of $M \geq 1$ queues with Poisson arrivals and generally distributed service times extended with customer routing (cf. Sect. 4.5.1). We will consider in this chapter two service disciplines, viz.,

- pure exponential time-limited discipline;
- exhaustive exponential time-limited discipline.

The time limit at Q_i is exponentially distributed with parameter ξ_i. Finally, we assume the preemptive-repeat with resampling servicing strategy.

We recall that the random variables $I_i,\ X_i,$ and $Y_i,\ i = 1, \ldots, M$, refer to the interarrival time of customers, the service time of customers, and the visit time (or time limit) of the server at Q_i. These random variables $I_i,\ X_i$, and Y_i are assumed independent and identically distributed. Besides, these random variables are assumed to be mutually independent.

The switch-over times and the polling strategy of the server will be excluded from the analysis in this chapter. We purely focus on the key relation, Eq. (5.1), in this chapter. In particular, we refer to Chapter 4 for the incorporation of this relation into the framework for the computation of the queue-length probabilities. Further, we assume that the polling system operates in a stable environment, i.e., the stability conditions as presented in Chapter 3 are satisfied.

Below, we introduce the notation that will be used throughout.

- Q; an arbitrary queue in the system;
- x_t; number of customers at time t at Q;

- $z_{(n)}$; number of customers left behind by the nth departing customer from Q;
- r'_n; time of the nth departure from Q;
- $D(t)$; number of departures from Q in $[0,t)$;
- $A(t)$; number of arrivals to Q in $[0,t)$;
- I; exponentially distributed random variable with parameter λ denoting the interarrival time to Q;
- X; generally distributed random variable denoting the service time at Q;
- $\mathbf{1}_{\{A\}}$; indicator function of event A;
- $\tilde{X}(\cdot)$; LST of random variable X;
- $\hat{\mu}(s,y)$; root x with the smallest absolute value less than one of $x = y \cdot \tilde{X}(s + \lambda(1-x))$;
- $\mathbf{N}_i^s$; number of customers at all queues at the start of a server visit to Q_i;
- $\mathbf{N}_i^e$; number of customers at all queues at the end of a server visit to Q_i;
- $N_{i,j}(t)$; number of customers at Q_j at time t during a server visit to Q_i;
- $\alpha^i(\mathbf{z})$; p.g.f. of $\mathbf{N}_i^s$;
- $\beta^i(\mathbf{z})$; p.g.f. of $\mathbf{N}_i^e$.

5.3 Analysis of the pure time-limited service discipline

In this section, we analyze the pure time-limited discipline. Under this discipline, the server will only depart from the queue when the time limit has been reached. It should be stressed that the server will not leave the queue when it becomes empty. We will derive an expression for $\beta^i(\mathbf{z})$, the p.g.f. of the number of customers at all queues at the instant that the server leaves Q_i, in terms of the number present at the start of the visit, $\alpha^i(\mathbf{z})$. Here, we present only the essential analytical steps and the main result. The proofs will be given in Appendix 5.B.

Consider a visit of the server to an arbitrary queue Q. During such a visit, the queue-length process at Q is a birth-and-death process, while the queue-length process at the other queues is a pure birth-process. Notice also that arrivals to these other queues may be both exogenous and endogenous (from Q). Our interest is in the number of customers at time t given an initial number of customers at the queues. Moreover, to include customer routing in the analysis, we need to keep track of the

number of departures during a visit. Notice that to record this number of departures, it is not sufficient to know the number of customers at Q at the beginning and the end of a visit. Therefore, we will focus on the transient probabilities $p_{hk}^{(n)}(t)$ which are defined as follows:

$$p_{hk}^{(n)}(t) := \begin{cases} \mathbb{P}(x_t = k,\ D(t) = n | x_0 = h), & h,k,n = 0,1,\ldots, \\ 0, & \text{otherwise}, \end{cases}$$

where x_t refers to the number of customers at Q at time t and $D(t)$ refers to the number of departures from Q until time t. We relate these probabilities to the transient probabilities for the standard M/G/1 queue, which we denote by $P_{hj}^{(n)}(t)$. These time-dependent conditional probabilities which incorporate also the number of departures until time t are defined for $n = 1,2,\ldots,\ h,j = 0,1,\ldots$, and $t > 0$ as [23, p.239]:

$$P_{hj}^{(n)}(t) := \mathbb{P}(z_{(n)} = j,\ r'_n \le t | z_{(0)} = h),$$

where $z_{(n)}$ refers to the number of customers left behind by the nth departure, and r'_n to the epoch of the nth departure. Further, it is assumed that at time $t = 0$ the 0-th customer left the queue. We consider the function $\pi_h(r,s,y)$ which is defined in terms of $P_{hj}^{(n)}(t)$ as follows:

$$\pi_h(r,s,y) := \sum_{n=1}^{\infty} y^n \sum_{j=0}^{\infty} r^j \int_0^{\infty} e^{-st} dP_{hj}^{(n)}(t),\ h = 0,1,\ldots,$$

and which is explicitly provided in Cohen [23, p.240] for $h = 0,1,\ldots$, as

$$\pi_h(r,s,y) = \frac{y \cdot \tilde{X}(s+\lambda(1-r))}{r - y \cdot \tilde{X}(s+\lambda(1-r))} \cdot \left\{ r^h - \frac{\lambda(1-r)+s}{\lambda(1-\hat{\mu}(s,y))+s} \cdot \hat{\mu}^h(s,y) \right\}, \quad (5.2)$$

where $\hat{\mu}(s,y)$ is the root x with the smallest absolute value less than one of $x = y \cdot \tilde{X}(s+\lambda(1-x))$. Notice that $\hat{\mu}(s,y)$ is the joint generating function of the busy period and the number of customers served during this period.

To take advantage of this explicit result, we will first present an explicit expression for the transient probabilities $p_{hk}^{(n)}(t)$ in terms of $P_{hj}^{(n)}(t)$. For convenience, we define:

$$\begin{aligned} F_k^{(0)}(t) &= \mathbf{1}_{\{k=0\}} \mathbb{P}(A(t) = 0,\ I > t) \\ &\quad + \mathbf{1}_{\{k \ge 1\}} \mathbb{P}(A(t) = k,\ I + X > t),\ k = 0,1,\ldots, \\ F_k(t) &= \mathbb{P}(A(t) = k,\ X > t),\ k = 0,1,\ldots\ . \end{aligned}$$

That is, $F_k(t)$ refers to k exogenous arrivals to Q during a period which is shorter than a service time X (thus assuming a nonempty queue at $t = 0$) meaning that a service is interrupted. In the special case of an empty queue at $t = 0$ (see $F_k^{(0)}(t)$), we need to account for the fact that first an arrival should occur before any service may start at all. Then, we can relate $p_{hk}^{(n)}(t)$ to $P_{hj}^{(n)}(t)$ for $n = 1,2,\ldots,\ h,k = 0,1,\ldots$, and $t > 0$ as follows:

Lemma 5.1.

$$p_{hk}^{(n)}(t) = \int_{u=0}^{t} F_k^{(0)}(t-u) dP_{h0}^{(n)}(u) + \sum_{j=1}^{k} \int_{u=0}^{t} F_{k-j}(t-u) dP_{hj}^{(n)}(u).$$

To retrieve the terms $\pi_h(r,s,y)$, we first take the LST of $p_{hk}^{(n)}(t)$ (see Remark 5.5 below). Next, we take the generating function of this expression with respect to the number of customers at the end of a server visit. Notice that our interest here is specifically in this number rather than in the number at the time of the nth departure, since the server only leaves upon expiration of the timer. In a final step, we take the generating function with respect to the number of departures until time t as to obtain an expression for $p_{hk}^{(n)}(t)$ in terms of $\pi_h(r,s,y)$. These consecutive steps provide us with the following result for $h = 0, 1, \ldots$.

Lemma 5.2.

$$\begin{aligned} &\sum_{n=1}^{\infty} y^n \sum_{k=0}^{\infty} r^k \int_{t=0}^{\infty} e^{-st} dp_{hk}^{(n)}(t) \qquad (5.3)\\ &= \frac{s}{\lambda(1-r)+s} \cdot \frac{\lambda(1 - r \cdot \tilde{X}(\lambda(1-r)+s)) + s}{\lambda + s} \cdot \pi_{h0}(s,y) \\ &\quad + \frac{s}{\lambda(1-r)+s} \cdot (1 - \tilde{X}(\lambda(1-r)+s)) \cdot (\pi_h(r,s,y) - \pi_{h0}(s,y)), \end{aligned}$$

where the terms $\pi_{h0}(s,y)$ are given by (see [23, p.240]),

$$\pi_{00}(s,y) = \frac{\lambda}{\lambda(1-\hat{\mu}(s,y))+s} \cdot \hat{\mu}(s,y), \qquad (5.4)$$

$$\pi_{h0}(s,y) = \frac{\lambda+s}{\lambda(1-\hat{\mu}(s,y))+s} \cdot \hat{\mu}^h(s,y), \ h = 1, 2, \ldots . \qquad (5.5)$$

The right-hand side of Eq. (5.3) can be interpreted as follows. The first part refers to the case that upon the nth departure zero customers are left behind, while the second part refers to a strictly positive number left behind by the nth departing customer. Moreover, the second part can be decomposed in two independent components: $\pi_h(r,s,y) - \pi_{h0}(s,y)$ accounts for the queue-length evolution until nth departure and the other component for the queue-length evolution during the final, interrupted service. A similar reasoning holds for the first part.

Thus, we have related the transient probabilities of our interest to known results for the M/G/1 queue. To incorporate these results in the polling model, we refer to a specific queue Q_i by adding an index i to the generic variables. Next, by unconditioning on the system state at the start of a visit and incorporating the expressions above into the definition of $\beta^i(\mathbf{z})$, we obtain the main result of this section for the p.g.f. of the joint queue-length at the end of a server visit under the pure exponential time-limited discipline.

Theorem 5.3 (Pure exponential time-limited discipline)**.**

$$\beta^i(\boldsymbol{z}) = d_1^P(\boldsymbol{z}) \cdot (\alpha^i(\boldsymbol{z}) - \alpha^i(\boldsymbol{z}_i^*)) + d_2^P(\boldsymbol{z}) \cdot \alpha^i(\boldsymbol{z}_i^*), \tag{5.6}$$

where

$$d_1^P(\boldsymbol{z}) = \frac{\xi_i}{z_i - r^i(\boldsymbol{z}) \cdot \tilde{X}_i(\lambda_i(1-z_i) + \xi_i^*)} \cdot \frac{z_i \cdot (1 - \tilde{X}_i(\lambda_i(1-z_i) + \xi_i^*))}{\lambda_i(1-z_i) + \xi_i^*}, \tag{5.7}$$

$$\begin{aligned} d_2^P(\boldsymbol{z}) &= d_1^P(\boldsymbol{z}) + \frac{\xi_i}{z_i - r^i(\boldsymbol{z}) \cdot \tilde{X}_i(\lambda_i(1-z_i) + \xi_i^*)} \\ &\quad \times \frac{(z_i - r^i(\boldsymbol{z})) \cdot \tilde{X}_i(\lambda_i(1-z_i) + \xi_i^*)}{\lambda_i(1 - \hat{\mu}_i(\xi_i^*, r^i(\boldsymbol{z}))) + \xi_i^*}, \\ \xi_i^* &= \xi_i + \sum_{j \neq i} \lambda_j(1 - z_j), \end{aligned}$$

and $\alpha^i(\boldsymbol{z}_i^*) := \alpha^i(z_1, \ldots, z_{i-1}, \hat{\mu}_i(\xi_i^*, r^i(\boldsymbol{z})), z_{i+1}, \ldots, z_M)$.

Remark 5.4 (Exponential service times)**.** *For the case of exponential service times at* Q_i *(with rate* μ_i*), it can be shown that Eq. (5.6) can be rewritten to:*

$$\beta^i(\boldsymbol{z}) = \frac{\xi_i \cdot z_i}{V_i(\boldsymbol{z})} \cdot (\alpha^i(\boldsymbol{z}) - \alpha^i(\boldsymbol{z}_i^*)) + \frac{\xi_i \cdot r^i(\boldsymbol{z}) \cdot (\hat{\mu}_i(\xi_i^*, r^i(\boldsymbol{z})) - z_i)}{V_i(\boldsymbol{z}) \cdot (\hat{\mu}_i(\xi_i^*, r^i(\boldsymbol{z})) - r^i(\boldsymbol{z}))} \cdot \alpha^i(\boldsymbol{z}_i^*), \tag{5.8}$$

where

$$V_i(\boldsymbol{z}) = -\lambda_i z_i^2 + (\mu_i + \lambda_i + \sum_{j \neq i} \lambda_j(1 - z_j) + \xi_i) z_i - r^i(\boldsymbol{z}) \cdot \mu_i. \tag{5.9}$$

We note that Eq. (5.8) generalizes the result for the special case $r^i(\boldsymbol{z}) = 1$ *(i.e., no customer routing) given in [H3].*

Remark 5.5 (Exponential time limit)**.** *The step of taking the LST of* $p_{hk}^{(n)}(t)$ *corresponds in fact to unconditioning over the exponentially distributed visit time. This shows that our assumption on the visit time plays a crucial role in the analysis.*

5.4 Analysis of the exhaustive time-limited service discipline

Let us next consider the exhaustive time-limited discipline. Notice that under this discipline the server will depart from the queue when it becomes empty or when the time limit has been reached, whichever occurs first. Again, we will derive an expression for $\beta^i(\mathbf{z})$, the p.g.f. of the number of customers at all queues at the instant

that the server leaves Q_i. This will be done in terms of the number present at the start of the visit, $\alpha^i(\mathbf{z})$. As in the previous section, we present here only the main analytical steps and the final result. The proofs will be given in Appendix 5.C.

Under the exhaustive time-limited discipline, the server may leave a queue for two reasons, viz., the server departs due to the queue being empty or due to the timer expiring. Let $\{empty\}$ and $\{timer\}$ denote the corresponding server events. Recall that $\mathbf{N}_i^s$ and $\mathbf{N}_i^e$ denote the multi-dimensional random variable of the number of customers at all queues at the start and the end of a visit to Q_i, respectively. The p.g.f. of $\mathbf{N}_i^e$, $\beta^i(\mathbf{z})$, can be decomposed in two parts depending on the reason of a server departure as the server departs only if the queue is empty or if the timer expires. Moreover, these events are readily seen to be mutually exclusive (the service-time distribution and timer distribution are both continuous distributions, so that the probability of the given events occurring simultaneously is zero). Hence, the p.g.f. of the number of customers at the end of a visit period to Q_i satisfies,

$$\beta^i(\mathbf{z}) = \mathbb{E}[\mathbf{z}^{\mathbf{N}_i^e}] = \mathbb{E}[\mathbf{z}^{\mathbf{N}_i^e}\mathbf{1}_{\{empty\}}] + \mathbb{E}[\mathbf{z}^{\mathbf{N}_i^e}\mathbf{1}_{\{timer\}}].$$

Next, in the Sects. 5.4.1 and 5.4.2, we will derive the conditional p.g.f.'s $\mathbb{E}[\mathbf{z}^{\mathbf{N}_i^e}\mathbf{1}_{\{empty\}}|\mathbf{N}_i^s = \mathbf{n}]$ and $\mathbb{E}[\mathbf{z}^{\mathbf{N}_i^e}\mathbf{1}_{\{empty\}}|\mathbf{N}_i^s = \mathbf{n}]$, where $\mathbf{n}$ denotes the vector $(n_1, \ldots, n_M)$. Finally, we will uncondition these expressions to get our main result in Sect. 5.4.3.

5.4.1 $\mathbb{E}[\mathbf{z}^{\mathbf{N}_i^e}\mathbf{1}_{\{empty\}}|\mathbf{N}_i^s = \mathbf{n}]$

We note that in case the $\{empty\}$ event occurs the queue may be empty upon arrival of the server or become empty upon departure of a customer. If the server finds an empty queue upon arrival, then clearly $\mathbf{N}_i^e = \mathbf{N}_i^s$. Else, if the queue is nonempty, then the evolution of queue-length process during the visit is strongly related to the length of a busy period in a standard M/G/1 queue. This is formalized in the following proposition.

Proposition 5.6. *The joint conditional p.g.f. of the number of customers at the end of a visit period to Q_i and the server departs due to the queue being empty satisfies,*

$$\mathbb{E}[\boldsymbol{z}^{\boldsymbol{N}_i^e}\boldsymbol{1}_{\{empty\}}|\boldsymbol{N}_i^s = \boldsymbol{n}] = \hat{\mu}_i^{n_i}(\xi_i^*, r^i(\boldsymbol{z})) \cdot \prod_{j\neq i} z_j^{n_j}, \tag{5.10}$$

where

$$\xi_i^* = \xi_i + \sum_{j\neq i} \lambda_j(1 - z_j).$$

5.4.2 $\mathbb{E}[\mathbf{z}^{\mathbf{N}_i^e}\mathbf{1}_{\{timer\}}|\mathbf{N}_i^s = \mathbf{n}]$

We note that in case the $\{timer\}$ event occurs the queue must be nonempty upon arrival of the server, then it remains nonempty during the course of the visit and it

is still nonempty at the expiration of the timer. The analysis of this case builds on the work of Cohen for the transient analysis of the M/G/1 queue. However, contrary to the analysis for the pure time-limited discipline, we cannot directly apply the formulae derived in [23]. This is due to the fact that we need specifically to account for not entering the state with zero customers at Q_i during the course of a server visit. Below, we state the transient probabilities of interest and several related expressions. Next, using these expressions, we will derive $\mathbb{E}[\mathbf{z}^{\mathbf{N}_i^e}\mathbf{1}_{\{timer\}}|\mathbf{N}_i^s = \mathbf{n}]$.

We consider first the conditional joint queue-length distribution at time $t > 0$ given an initial number of customers at time $t = 0$ and given that the server is at an arbitrary queue Q. It is good to notice that during a server visit to Q the queue-length process at the other queues is simply a pure-birth process. Hence, we neglect the other queues for the moment and concentrate on the marginal queue-length probabilities for Q, denoted by $q_{hk}^{(n)}(t)$, which we define as:

$$q_{hk}^{(n)}(t) := \begin{cases} \mathbb{P}(x_t = k,\ D(t) = n,\ x_v > 0,\ 0 < v < t | x_0 = h), \\ \qquad\qquad\qquad\qquad\qquad\qquad n = 0, 1, \ldots, \quad h, k = 1, 2, \ldots, \\ 0, \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{otherwise.} \end{cases}$$

For completeness, let us recall the definition of the probabilities $P_{hj}^{(n)}(t)$, for $n = 1, 2, \ldots,\ h, j = 0, 1, \ldots$ and $t > 0$,

$$P_{hj}^{(n)}(t) := \mathbb{P}(z_{(n)} = j,\ r'_n < t | z_{(0)} = h).$$

Analogously, we define $R_{hj}^{(n)}(t)$, for $h, j, n = 1, 2, \ldots$, and $t > 0$,

$$R_{hj}^{(n)}(t) := \mathbb{P}(z_{(n)} = j,\ r'_n < t,\ z_{(k)} > 0,\ 0 < k < n | z_{(0)} = h),$$

where it is assumed that at time $t = 0$ a new service starts. We note that $R_{hj}^{(n)}(t)$ is only defined for $h, j = 1, 2, \ldots$. This is due to the fact that the event of a server arriving to an empty queue (i.e., $h = 0$) and the event of the nth customer leaving an empty queue behind (i.e., $j = 0$) are never considered as $\{timer\}$ events, but always as $\{empty\}$ events.

We consider the function $\gamma_h(r, s, y)$ which is defined in terms of $R_{hj}^{(n)}(t)$ as follows:

$$\gamma_h(r, s, y) := \sum_{n=1}^{\infty} y^n \sum_{j=1}^{\infty} r^j \int_0^{\infty} e^{-st} dR_{hj}^{(n)}(t),\ h = 1, 2, \ldots,$$

and which is explicitly given (see Sect. 5.A for the derivation) for $h = 1, 2, \ldots$, as

$$\gamma_h(r, s, y) = \frac{r\left(-\hat{\mu}^h(s, y) + y \cdot \tilde{X}\left(\lambda(1 - r) + s\right) \cdot r^{h-1}\right)}{r - y \cdot \tilde{X}\left(\lambda(1 - r) + s\right)}.$$

Analogously to the approach in the previous section, we intend to utilize the explicit expressions for $\gamma_h(r,s,y)$. To this end, we will start by relating the transient probabilities $q_{hk}^{(n)}(t)$ to the time-dependent probabilities $R_{hj}^{(n)}(t)$ at embedded epochs of service completion. For convenience, we recall that:

$$F_k(t) = \mathbb{P}(A(t) = k,\ X > t),\ k = 0, 1, \ldots,$$

that is, $F_k(t)$ refers to the number of arrivals to Q during a period which duration is shorter than a service time X. The specific relation between $q_{hk}^{(n)}(t)$ and $R_{hj}^{(n)}(t)$ is then given in the following lemma.

Lemma 5.7.

$$q_{hk}^{(n)}(t) = \int_{u=0}^{t} \sum_{j=1}^{k} F_{k-j}(t-u) dR_{hj}^{(n)}(u),\ n = 1, 2, \ldots,\ h, k = 1, 2, \ldots\ .$$

Again, to obtain the terms $\gamma_h(r,s,y)$, we take the LST of $q_{hk}^{(n)}(t)$ (see Remark 5.5). Next, we take the generating function with respect to the number of customers at the end of the server visit of the resulting expression and finally we take the generating function with respect to the number of departures. Hence, we obtain the following result.

Lemma 5.8.

$$\begin{aligned} &\sum_{n=1}^{\infty} y^n \sum_{k=1}^{\infty} r^k \int_{t=0}^{\infty} e^{-st} dq_{hk}^{(n)}(t) \\ &= \gamma_h(r,s,y) \cdot \frac{s}{\lambda(1-r)+s} \cdot (1 - \tilde{X}(\lambda(1-r)+s)),\ h = 1, 2, \ldots\ . \end{aligned} \tag{5.11}$$

The right-hand side of Eq. (5.11) can be recognized as a convolution of two independent parts. The first part, $\gamma_h(r,s,y)$, refers to the queue length at the instant of the final (successful) service completion during the visit, while the other part refers to the number of arrivals during an interrupted service.

Next, we can present the explicit expression for the joint conditional p.g.f. of the number of customers at all queues at the end of a visit to a specific queue Q_i when the server departure is due to the timer expiration. The condition is on the number of customers present at the start of the visit.

Proposition 5.9.

$$\begin{aligned} &\mathbb{E}[\boldsymbol{z}^{\boldsymbol{N}_i^e} \boldsymbol{1}_{\{timer\}} | \boldsymbol{N}_i^s = \boldsymbol{n}] \\ &= \frac{\xi_i \cdot z_i \cdot (1 - \tilde{X}_i(\lambda_i(1-z_i) + \xi_i^*)) \left(z_i^{n_i} - \hat{\mu}_i^{n_i}(\xi_i^*, r^i(\boldsymbol{z}))\right)}{[\lambda_i(1-z_i) + \xi_i^*] \cdot [z_i - r^i(\boldsymbol{z}) \cdot \tilde{X}_i(\lambda_i(1-z_i) + \xi_i^*)]} \cdot \prod_{j \neq i} z_j^{n_j} \end{aligned} \tag{5.12}$$

where

$$\xi_i^* = \xi_i + \sum_{j \neq i} \lambda_j (1 - z_j)\ .$$

5.4.3 $\mathbb{E}[\mathbf{z}^{\mathbf{N}_i^e}]$

Combining the two conditional results of Eqs. (5.10) and (5.12), we obtain our main result of this section for the exhaustive exponential time-limited service discipline.

Theorem 5.10 (Exhaustive exponential time-limited discipline)**.**

$$\beta^i(\boldsymbol{z}) = d_1^E(\boldsymbol{z}) \cdot (\alpha^i(\boldsymbol{z}) - \alpha^i(\boldsymbol{z}_i^*)) + d_2^E(\boldsymbol{z}) \cdot \alpha^i(\boldsymbol{z}_i^*), \tag{5.13}$$

where

$$\begin{aligned} d_1^E(\boldsymbol{z}) &= d_1^P(\boldsymbol{z}), \\ d_2^E(\boldsymbol{z}) &= 1, \\ \xi_i^* &= \xi_i + \sum_{j \neq i} \lambda_j (1 - z_j), \end{aligned}$$

and $\alpha^i(\boldsymbol{z}_i^*) := \alpha^i(z_1, \ldots, z_{i-1}, \hat{\mu}_i(\xi_i^*, r^i(\boldsymbol{z})), z_{i+1}, \ldots, z_M)$.

We note that Eq. (5.13) generalizes the result for the special case $r^i(\mathbf{z}) = 1$ (i.e., no customer routing) given in [32].

Remark 5.11 (Exponential service times)**.** *For the case of exponential service times at Q_i (with rate μ_i), it can be shown that Eq. (5.13) can be rewritten to:*

$$\beta^i(\boldsymbol{z}) = \frac{\xi_i \cdot z_i}{V_i(\boldsymbol{z})} \cdot (\alpha^i(\boldsymbol{z}) - \alpha^i(\boldsymbol{z}_i^*)) + \alpha^i(\boldsymbol{z}_i^*), \tag{5.14}$$

where $V_i(\boldsymbol{z})$ is given in Eq. (5.9). We note that thus Eq. (5.14) generalizes the result for the special case $r^i(\boldsymbol{z}) = 1$ (i.e., no customer routing) given in [H3].

Remark 5.12 (Exhaustive service discipline)**.** *We note that in the limit case of $\xi_i \downarrow 0$ the time limit is of infinite length. Hence, in this case (assuming a stable queue), the server will always depart due to Q_i being empty. It can readily be found that for $\lim \xi_i \downarrow 0$ and $r^i(\boldsymbol{z}) = 1$ the following expression for $\beta^i(\boldsymbol{z})$ is obtained:*

$$\beta^i(\boldsymbol{z}) = \alpha^i(\boldsymbol{z}_i^*),$$

where $\alpha^i(\boldsymbol{z}_i^*) := \alpha^i(z_1, \ldots, z_{i-1}, \hat{\mu}_i(\sum_{j \neq i} \lambda_j(1 - z_j), 1), z_{i+1}, \ldots, z_M)$. This result matches the well-known result for the exhaustive service discipline.

5.5 Discussion

The final results for the pure exponential time-limited discipline (P-TL) and the exhaustive exponential time-limited discipline (E-TL) are similar. More specifically, these results can be written in the following form:

$$\text{P-TL: } \beta^i(\mathbf{z}) = d_1^P(\mathbf{z}) \cdot (\alpha^i(\mathbf{z}) - \alpha^i(\mathbf{z}_i^*)) + d_2^P(\mathbf{z}) \cdot \alpha^i(\mathbf{z}_i^*) \tag{5.15}$$

$$\text{E-TL: } \beta^i(\mathbf{z}) = d_1^E(\mathbf{z}) \cdot (\alpha^i(\mathbf{z}) - \alpha^i(\mathbf{z}_i^*)) + d_2^E(\mathbf{z}) \cdot \alpha^i(\mathbf{z}_i^*), \tag{5.16}$$

where $d_1^E(\mathbf{z}) = d_1^P(\mathbf{z})$ is given in Eq. (5.7), $d_2^E(\mathbf{z}) = 1$ and $d_2^P(\mathbf{z})$ is given by

$$\begin{aligned} d_2^P(\mathbf{z}) &= \frac{\xi_i}{z_i - r^i(\mathbf{z}) \cdot \tilde{X}_i(\lambda_i(1-z_i) + \xi_i^*)} \\ &\quad \times \Bigg\{ \frac{\tilde{X}_i(\lambda_i(1-z_i) + \xi_i^*) \cdot (z_i - r^i(\mathbf{z}))}{\lambda_i(1 - \hat{\mu}_i(\xi_i^*, r^i(\mathbf{z}))) + \xi_i^*} \\ &\quad + \frac{z_i(1 - \tilde{X}_i(\lambda_i(1-z_i) + \xi_i^*))}{\lambda_i(1-z_i) + \xi_i^*} \Bigg\}. \end{aligned} \tag{5.17}$$

Equations (5.15) and (5.16) can be interpreted as follows. Consider a visit of the server to Q_i. Regarding the timer, it may occur that (i) the timer expires before Q_i gets empty for the first time, or (ii) the timer expires only after Q_i becoming empty for the first time. It is readily seen that the queue-length process is identical for both service disciplines in the first case. This is reflected in the equivalence of the terms $d_1^P(\mathbf{z}) \cdot (\alpha^i(\mathbf{z}) - \alpha^i(\mathbf{z}_i^*))$ and $d_1^E(\mathbf{z}) \cdot (\alpha^i(\mathbf{z}) - \alpha^i(\mathbf{z}_i^*))$. However, in the second case, the queue length process is different for each discipline. Under the exhaustive time-limited discipline, the server immediately leaves upon the queue becoming empty, say this occurs at time t_0. Conversely, under the pure time-limited discipline, at time t_0 the server will remain at the queue and a sequence of idle and busy periods will follow until eventually the timer expires. This latter contribution (after t_0) to the queue-length process is represented in the term $d_2^P(\mathbf{z})$.

Hence, $d_2^P(\mathbf{z})$ reflects the p.g.f. of the number of customers at all queues at the end of a server visit process which runs for an exponential amount of time and which starts from an empty queue. This function can be analyzed as follows. First, observe that the timer will interrupt the visit process either during an idle or a busy period. Second, observe that this process is regenerative in the sense that if the timer does not expire before the end of the first busy period, then the process starts like anew at that specific time instant. Recall that I_i denotes the length of an idle period at Q_i and Y_i the exponential visit time of the server to Q_i. Let further U_i denote the length of a busy period at Q_i starting with a single customer. Then, we may write the following relation for $d_2^P(\mathbf{z})$:

$$\begin{aligned} d_2^P(\mathbf{z}) &= \mathbb{E}[\mathbf{z}^{\mathbf{N}_i(Y_i)} \mathbf{1}_{\{I_i > Y_i\}} | \mathbf{N}_i(0) = \mathbf{n},\ N_{i,i}(0) = 0] \\ &\quad + \mathbb{E}[\mathbf{z}^{\mathbf{N}_i(Y_i)} \mathbf{1}_{\{I_i < Y_i, I_i + U_i > Y_i\}} | \mathbf{N}_i(0) = \mathbf{n},\ N_{i,i}(0) = 0] \\ &\quad + \mathbb{E}[\mathbf{z}^{\mathbf{N}_i(I_i+U_i)} \mathbf{1}_{\{I_i + U_i < Y_i\}} | \mathbf{N}_i(0) = \mathbf{n},\ N_{i,i}(0) = 0] \cdot d_2^P(\mathbf{z}) \\ &= \frac{\xi_i}{\lambda_i + \xi_i^*} + \frac{\lambda_i}{\lambda_i + \xi_i^*} \\ &\quad \times \mathbb{E}[\mathbf{z}^{\mathbf{N}_i(Y_i)} \mathbf{1}_{\{timer\}} | \mathbf{N}_i(0) = (n_1, \ldots, n_{i-1}, 1, n_{i+1}, \ldots, n_M)] \\ &\quad + \frac{\lambda_i}{\lambda_i + \xi_i^*} \cdot \hat{\mu}_i(\xi_i^*, r^i(\mathbf{z})) \cdot d_2^P(\mathbf{z}), \end{aligned}$$

where $\mathbb{E}[\mathbf{z}^{\mathbf{N}_i(Y_i)} \mathbf{1}_{\{timer\}} | \mathbf{N}_i(0) = \mathbf{n}]$ is provided in the analysis of the exhaustive time-limited discipline (see Proposition 5.9). Then, inserting this result of Proposition 5.9

and reorganizing the terms appropriately, we obtain:

$$\begin{aligned} d_2^P(\mathbf{z}) &= \frac{\xi_i}{z - r^i(\mathbf{z}) \cdot \tilde{X}_i(\lambda_i(1-z_i)+\xi_i^*)} \\ &\quad \times \left\{ \frac{(\lambda_i(1-z_i)+\xi_i^*)(z - r^i(\mathbf{z}) \cdot \tilde{X}_i(\lambda_i(1-z_i)+\xi_i^*))}{(\lambda_i(1-z_i)+\xi_i^*)(\lambda_i(1-\hat{\mu}_i(\xi_i^*, r^i(\mathbf{z})))+\xi_i^*)} \right. \\ &\quad \left. + \frac{\lambda_i \cdot z_i \cdot (1-\tilde{X}_i(\lambda_i(1-z_i)+\xi_i^*))(z_i - \hat{\mu}_i(\xi_i^*, r^i(\mathbf{z})))}{(\lambda_i(1-z_i)+\xi_i^*)(\lambda_i(1-\hat{\mu}_i(\xi_i^*, r^i(\mathbf{z})))+\xi_i^*)} \right\}, \end{aligned}$$

where

$$\xi_i^* = \xi_i + \sum_{j \neq i} \lambda_j (1 - z_j).$$

It can readily be verified that the latter expression is indeed equal to Eq. (5.17).

This result confirms indeed the interpretation given above. We remarked above that for the common exhaustive discipline (E), i.e., the time-limited version with the time limit set to infinity, the term $d_1^E(\mathbf{z})$ vanishes. Thus, yielding:

$$\text{E: } \beta^i(\mathbf{z}) = \alpha^i(\mathbf{z}_i^*), \tag{5.18}$$

Notice that in fact all service disciplines satisfying the branching property can be cast in the form of Eq. (5.18), while for other disciplines it is unlikely that such a result exists. This suggests, together with the interpretation above, that similarly to the exhaustive time-limited discipline key relations for $\beta^i(\mathbf{z})$ may be found for any branching-property satisfying discipline operating under an exponential time limit. Indeed for the gated time-limited discipline (G-TL), we may readily find:

$$\text{G-TL: } \beta^i(\mathbf{z}) = d_1^G(\mathbf{z}) \cdot (\alpha^i(\mathbf{z}) - \alpha^i(\mathbf{z}_i^\bullet)) + d_2^G(\mathbf{z}) \cdot \alpha^i(\mathbf{z}_i^\bullet), \tag{5.19}$$

where $\alpha^i(\mathbf{z}_i^\bullet) := \alpha^i(z_1, \ldots, z_{i-1}, r^i(\mathbf{z}) \cdot \tilde{X}_i(\xi_i^*), z_{i+1}, \ldots, z_M)$, $d_1^G(\mathbf{z}) = d_1^E(\mathbf{z})$ and $d_2^G(\mathbf{z}) = d_2^E(\mathbf{z})$. This result follows by differentiating between the server departing due to having served all customers that were present at the start of the visit or due to the timer expiration. The former case readily gives the term $\alpha^i(\mathbf{z}_i^\bullet)$, since each customer is essentially replaced in an i.i.d. manner during the visit. In the latter case, it can be observed that the number of customers served during the visit is geometrically distributed. The stated expression is then readily obtained via some simple calculus. In particular, we note that the property of the gated discipline that each customer served during a visit was already present at the start of the visit renders the analysis straightforward.

We do believe that these results indeed carry over to any branching-property satisfying service discipline [40, 88]. According to such a discipline, customers at Q_i will effectively be replaced in an i.i.d. manner by a random population during the course of a server visit. Let us denote by $h_i^*(\mathbf{z})$ the corresponding p.g.f. which accounts for these replacements under the exponential time-limit. Then, we conclude this discussion with the following conjecture.

Conjecture 5.13. *For a FCFS single-server polling system with no customer routing and with the server at Q_i operating under a branching-property satisfying service discipline with replacement p.g.f. $h_i^*(\boldsymbol{z})$ which is restricted by an exponentially distributed time limit, the queue-length evolution during a server visit to Q_i can be described as follows:*

$$\beta^i(\boldsymbol{z}) = d(\boldsymbol{z}) \cdot (\alpha^i(\boldsymbol{z}) - \alpha^i(\boldsymbol{z}_i^\star)) + \alpha^i(\boldsymbol{z}_i^\star),$$

where $\alpha^i(\boldsymbol{z}_i^\star) := \alpha^i(z_1, \ldots, z_{i-1}, h_i^*(\boldsymbol{z}), z_{i+1}, \ldots, z_M)$, *and*

$$d(\boldsymbol{z}) = \frac{\xi_i}{z_i - r^i(\boldsymbol{z}) \cdot \tilde{X}_i(\lambda_i(1-z_i) + \xi_i^*)} \cdot \frac{z_i \cdot (1 - \tilde{X}_i(\lambda_i(1-z_i) + \xi_i^*))}{\lambda_i(1-z_i) + \xi_i^*}.$$

Remark 5.14 (Customer routing)**.** *The inclusion of customer routing in the analysis depends on the service discipline (see, Eqs. (5.16) and (5.19)). Although it is anticipated that it can readily be included for any branching-type discipline, it is intentionally left out here as to preserve a general formulation.*

5.6 Concluding remarks

In this chapter, we have studied two time-limited service disciplines for polling systems, viz., the pure and the exhaustive exponential time-limited discipline with preemptive service. Specifically, we have obtained relations between the p.g.f.'s of the joint queue-length distribution at visit beginning and visit completion epochs. To this end, we have used both known (see Sect. 5.3) and novel (see Sect. 5.4) results for the transient behavior of the M/G/1 queue. Our final expressions for the key relation, Eq. (5.1), enhance previous results for the pure [H3] and the exhaustive time-limited discipline [31] by including customer routing. These relations can be used to obtain the joint queue-length distribution at these embedded epochs, e.g., along the framework presented in Chapter 4.

Let us also confront the specific relations for the pure time-limited discipline with our earlier analysis in Chapter 4. In that chapter, we found for the key relation:

$$\beta^i(\mathbf{z}) = \sum_{k=1}^{\infty} \phi_k^{s,i}(\mathbf{z}), \tag{5.20}$$

where $\phi_k^{s,i}(\mathbf{z})$ is related in a recursive manner to $\alpha^i(\mathbf{z})$. Contrary, by exploiting known results on the transient analysis of the M/G/1 queue, we obtained the following direct relation in this chapter (see Thm. 5.3):

$$\beta^i(\mathbf{z}) \quad = \quad d_1^P(\mathbf{z}) \cdot (\alpha^i(\mathbf{z}) - \alpha^i(\mathbf{z}_i^*)) + d_2^P(\mathbf{z}) \cdot \alpha^i(\mathbf{z}_i^*). \tag{5.21}$$

Hence, the infinite sum of Eq. (5.20), which is a consequence of conditioning on intermediate epochs during a visit, is now replaced by a single expression as shown

in Eq. (5.21). This does not only provide us with a more elegant result, but it has also shown to significantly alleviate the computational efforts required to compute the joint queue-length distribution of the polling system along the framework of Chapter 4.

Finally, we have been able to give a clear interpretation of the final expressions for the key relations, Eqs. (5.15) and (5.16). This interpretation has given rise to a conjecture for a direct expression of the key relation, Eq. (5.1), for any branching-type service discipline operating under an exponential timer.

5.A Transient analysis of the M/G/1 queue during a busy period

In this section, we analyze the transient behavior of the M/G/1 queue during a busy period. We follow a similar approach as Cohen [23] used to study the transient behavior of the full queue-length process of the M/G/1 queue. To this end, we consider a single queue served by a single server. Customer arrive to the queue according to a Poisson process with rate λ. The service requirement X of a customer is generally distributed with mean μ.

Our interest is in the queue-length process during a busy period with some initial number of customers. Moreover, we keep track of the number of departures until time t. Therefore, similar to the transient transition probabilities $P_{hj}^{(n)}(t)$ that were defined in [23, p.239], we define the transient probabilities $R_{hj}^{(n)}(t)$ which specifically account for the fact that the system is nonempty from time 0 up to time t. More precisely, the transient probabilities $R_{hj}^{(n)}(t)$ are defined for $h, j, n = 1, 2, \ldots,$ and $t > 0$ as:

$$R_{hj}^{(n)}(t) := \mathbb{P}(z_{(n)} = j,\ r'_n < t,\ z_{(k)} > 0,\ 0 < k < n | z_{(0)} = h),$$

where it is assumed that at time $t = 0$ a new service starts. Notice that $R_{hj}^{(n)}(t)$ is only defined for $h, j \geq 1$. Our objective is to find an explicit expression for $\gamma_h(r, s, y)$ which is defined as:

$$\gamma_h(r, s, y) := \sum_{n=1}^{\infty} y^n \sum_{j=0}^{\infty} r^j \int_0^{\infty} e^{-st} dR_{hj}^{(n)}(t),\ h = 1, 2, \ldots\ .$$

From the definition of $R_{hj}^{(n)}(t)$, it follows immediately that:

$$\begin{aligned}
R_{1j}^{(1)}(t) &= \int_{\tau=0}^{t} e^{-\lambda\tau} \frac{(\lambda\tau)^j}{j!} dX(\tau),\ j = 1, 2, \ldots, \\
R_{hj}^{(1)}(t) &= \int_{\tau=0}^{t} e^{-\lambda\tau} \frac{(\lambda\tau)^{j+1-h}}{(j+1-h)!} dX(\tau),\ j = h-1, h, \ldots,\ h = 2, 3, \ldots, \\
R_{hj}^{(1)}(t) &= 0, \text{ otherwise.}
\end{aligned}$$

Also, analogously to Eq. (4.20) of [23], we have the following recursive relation for $R_{hj}^{(n)}(t)$ for $t > 0$, $h, j = 1, 2, \ldots,$ $n = 2, 3, \ldots,$

$$R_{hj}^{(n)}(t) = \sum_{l=1}^{\infty} \int_{u=0}^{t} R_{hl}^{(n-1)}(t-u) d_u R_{lj}^{(1)}(u). \tag{5.22}$$

The following definitions will be used in the sequel:

$$\begin{aligned}
\gamma_{hj}^{(n)}(s) &:= \int_0^\infty e^{-st} dR_{hj}^{(n)}(t),\ h,j,n = 1,2,\ldots, \\
\gamma_h^{(n)}(r,s) &:= \sum_{j=1}^{\infty} r^j \gamma_{hj}^{(n)}(s),\ h,n = 1,2,\ldots, \\
\gamma_{hj}(s,y) &:= \sum_{n=1}^{\infty} y^n \gamma_{hj}^{(n)}(s),\ h,j = 1,2,\ldots, \\
\gamma_h(r,s,y) &:= \sum_{n=1}^{\infty} y^n \gamma_h^{(n)}(r,s),\ h = 1,2,\ldots\ .
\end{aligned}$$

As an immediate consequence of Eq. (5.22), we obtain the following result.

Lemma 5.15.

$$\gamma_h^{(n)}(r,s) = \sum_{l=1}^{\infty} \gamma_{hl}^{(n-1)}(s) \cdot \gamma_l^{(1)}(r,s),\ h = 1,2,\ldots,\ n = 2,3,\ldots,. \tag{5.23}$$

The final term in the right-hand side of Eq. (5.23), $\gamma_l^{(1)}(r,s)$, refers to the number of arrivals during a service time starting with l customers. We have to distinguish between starting with one or with two or more customers, since in the former case the queue might be empty upon service completion and this situation should be excluded. A closed-form expression for this term is then given in the following lemma.

Lemma 5.16.

$$\gamma_1^{(1)}(r,s) = \tilde{X}\left(\lambda(1-r)+s\right) - \tilde{X}\left(\lambda+s\right),$$

and for $h \geq 2$,

$$\gamma_h^{(1)}(r,s) = r^{h-1} \cdot \tilde{X}\left(\lambda(1-r)+s\right).$$

Proof. Let us consider first the case $h \geq 2$:

$$\begin{aligned}
\gamma_h^{(1)}(r,s) &= \sum_{j=1}^{\infty} r^j \gamma_{hj}^{(1)}(s) \\
&= \sum_{j=h-1}^{\infty} r^j \gamma_{hj}^{(1)}(s) \\
&= \sum_{j=h-1}^{\infty} r^j \int_{t=0}^{\infty} e^{-st} dR_{hj}^{(1)}(t) \\
&= \int_{t=0}^{\infty} s e^{-st} \sum_{j=h-1}^{\infty} r^j R_{hj}^{(1)}(t) dt \\
&= \int_{t=0}^{\infty} s e^{-st} \int_{\tau=0}^{t} e^{-\lambda\tau} \sum_{j=h-1}^{\infty} r^{h-1} \cdot \frac{(r\lambda\tau)^{j+1-h}}{(j+1-h)!}\, dX(\tau)\, dt \\
&= r^{h-1} \int_{\tau=0}^{\infty} e^{-\lambda\tau(1-r)} \int_{t=\tau}^{\infty} s \cdot e^{-st} dt\, dX(\tau) \\
&= r^{h-1} \cdot \tilde{X}\left(\lambda(1-r)+s\right).
\end{aligned}$$

In case $h = 1$, we should have at least one arrival before the first departure, otherwise the queue would become empty. Hence, in the derivation of $\gamma_1^{(1)}(r,s)$, we do not encounter the complete power series representation of the exponential function, so that the final expression will consist of two parts. More precisely,

$$\begin{aligned}
\gamma_1^{(1)}(r,s) &= \sum_{j=1}^{\infty} r^j \gamma_{hj}^{(n)}(s) \\
&= \ldots = \int_{t=0}^{\infty} s e^{-st} \int_{\tau=0}^{t} e^{-\lambda\tau} \sum_{j=1}^{\infty} \frac{(r\lambda\tau)^j}{j!} dX(\tau)\, dt \\
&= \int_{\tau=0}^{\infty} e^{-\lambda\tau} \cdot (e^{-\lambda\tau r} - 1) \int_{t=\tau}^{\infty} s e^{-st} dt\, dX(\tau) \\
&= \tilde{X}\left(\lambda(1-r)+s\right) - \tilde{X}\left(\lambda+s\right).
\end{aligned}$$

□

Next, we are ready to present our main result of this section, i.e., a closed-form expression for $\gamma_h(r,s,y)$.

Theorem 5.17.

$$\begin{aligned}
\gamma_h(r,s,y) &= \frac{r}{r - y \cdot \tilde{X}\left(\lambda(1-r)+s\right)} \\
&\quad \times \left(-\hat{\mu}^h(s,y) + y \cdot \tilde{X}\left(\lambda(1-r)+s\right) \cdot r^{h-1}\right), \; h = 1,2,\ldots,
\end{aligned}$$

where $\hat{\mu}(s,y)$ is the smallest root of the function $x = y \cdot \tilde{X}\left(\lambda(1-x)+s\right)$ in x with the absolute value smaller than one.

Proof. Starting from the definition of $\gamma_h(r,s,y)$ and applying Lemmas 5.15 and 5.16, we obtain the following relations after some manipulations:

$$\begin{aligned}\gamma_1(r,s,y) \cdot \left(1 - \frac{y}{r} \cdot \tilde{X}\left(\lambda(1-r)+s\right)\right) \\ = \quad y \cdot \left(\tilde{X}\left(\lambda(1-r)+s\right) - \tilde{X}\left(\lambda+s\right) \cdot \left(1+\gamma_{11}(s,y)\right)\right),\end{aligned} \tag{5.24}$$

and for $h = 2,3,\ldots$:

$$\begin{aligned}\gamma_h(r,s,y) \cdot \left(1 - \frac{y}{r} \cdot \tilde{X}\left(\lambda(1-r)+s\right)\right) \\ = \quad y \cdot \left(\tilde{X}\left(\lambda(1-r)+s\right) \cdot r^{h-1} - \tilde{X}\left(\lambda+s\right) \cdot \gamma_{h1}(s,y)\right).\end{aligned} \tag{5.25}$$

Denote by $\hat{\mu}(s,y)$ the smallest root of the function $x = y \cdot \tilde{X}\left(\lambda(1-x)+s\right)$ in x with the absolute value smaller than one. Since the functions $\gamma_h(r,s,y)$ should be analytic for $|r| \leq 1$, it follows that $\hat{\mu}(s,y)$ is a zero of the right-hand side of the expressions above. Thus, we immediately obtain for $\gamma_{h1}(s,y)$:

$$\begin{aligned}\gamma_{11}(s,y) &= \frac{\hat{\mu}(s,y) - y \cdot \tilde{X}\left(\lambda+s\right)}{y \cdot \tilde{X}\left(\lambda+s\right)}, \\ \gamma_{h1}(s,y) &= \frac{\hat{\mu}^h(s,y)}{y \cdot \tilde{X}\left(\lambda+s\right)}, \ h = 2,3,\ldots \ .\end{aligned}$$

Notice that inserting $h = 1$ in the latter expression, which we denote by $(\gamma_{h1}(s,y))|_{h=1}$, shows that: $\gamma_{11}(s,y) + 1 = (\gamma_{h1}(s,y))|_{h=1}$. Finally, inserting these expressions into Eqs. (5.24) and (5.25) completes the proof. □

5.B Proofs of results Section 5.3

For convenience, let us recall the following definitions for $t > 0$:

$$\begin{aligned}p_{hk}^{(n)}(t) &:= \begin{cases} \mathbb{P}(x_t = k,\ D(t) = n | x_0 = h), & h,k,n = 0,1,\ldots, \\ 0, & \text{otherwise}, \end{cases} \\ P_{hj}^{(n)}(t) &:= \mathbb{P}(z_{(n)} = j,\ r'_n < t | z_{(0)} = h),\ n = 1,2,\ldots,\ h,j = 0,1,\ldots, \\ F_k^{(0)}(t) &:= \mathbf{1}_{\{k=0\}}\mathbb{P}(A(t) = 0,\ I > t) \\ &\quad + \mathbf{1}_{\{k\geq 1\}}\mathbb{P}(A(t) = k,\ I + X > t),\ k = 0,1,\ldots, \\ F_k(t) &:= \mathbb{P}(A(t) = k,\ X > t),\ k = 0,1,\ldots \ .\end{aligned}$$

5.B.1 Proof of Lemma 5.1

The proof of the lemma is carried out as follows. First, we rewrite the event $D(t) = n$ and use the assumption that at time 0 the 0-th customer departed from the queue, so that we obtain Eq. (5.26) below. Next, we condition on the number of customers present at the nth departure, $z_{(n)}$, and on the time this departure occurs, r'_n, which leads to Eq. (5.27). Finally, observing that $r_{n+1},\ n = 0, 1, \ldots,$ depends in fact only on r_n and $z_{(n)}$, using that the arrival process is stationary and applying the definitions of $F_k^{(0)}(t),\ F_k(t)$ and $P_{hj}^{(n)}(t)$ provides us with Eq. (5.28).

$$\begin{aligned}
p_{hk}^{(n)}(t) &:= \mathbb{P}(x_t = k, D(t) = n | x_0 = h) \\
&= \mathbb{P}(x_t = k, r'_n \le t, r'_{n+1} > t | z_{(0)} = h) \qquad (5.26) \\
&= \int_{u=0}^{t} \sum_{j=0}^{k} \mathbb{P}(x_t = k,\ r'_{n+1} > t |\ r'_n = u,\ z_{(0)} = h,\ z_{(n)} = j) \\
&\quad \times d_u \mathbb{P}(\ r'_n \le u,\ z_{(n)} = j | z_{(0)} = h) \qquad (5.27) \\
&= \int_{u=0}^{t} F_k^{(0)}(t-u) dP_{h0}^{(n)}(u) + \sum_{j=1}^{k} \int_{u=0}^{t} F_{k-j}(t-u) dP_{hj}^{(n)}(u). \qquad (5.28)
\end{aligned}$$

Let us define the following LSTs:

$$\begin{aligned}
\tilde{F}_k^{(0)}(s) &:= \int_{0-}^{\infty} e^{-st} dF_k^{(0)}(t),\ k = 0, 1, \ldots, \\
\tilde{F}_k(s) &:= \int_{0-}^{\infty} e^{-st} dF_k(t),\ k = 0, 1, \ldots, \\
\pi_{hj}^{(n)}(s) &:= \int_{0-}^{\infty} e^{-st} dP_{hj}^{(n)}(t),\ n = 1, 2, \ldots,\ h, j = 0, 1, \ldots\ .
\end{aligned}$$

Then, we may present the following result as an immediate consequence of Lemma 5.1:

Corollary 5.18.

$$\int_{t=0-}^{\infty} e^{-st} dp_{hk}^{(n)}(t) = \tilde{F}_k^{(0)}(s) \pi_{h0}^{(n)}(s) + \sum_{j=1}^{k} \tilde{F}_{k-j}(s) \pi_{hj}^{(n)}(s).$$

5.B.2 Proof of Lemma 5.2

Before we get to the actual proof of Lemma 5.2, we present another lemma. Let us introduce the auxiliary functions $G^{(0)}(r, s)$ and $G(r, s)$. These functions refer to the number of customers that arrive to the system during a period which starts at

a service completion instant at an arbitrary queue Q and ends at a timer expiration which occurs before a next service is completed. More specifically, the function $G^{(0)}(r,s)$ refers to the case with zero customers present after a service completion, while $G(r,s)$ refers to the case with a strictly positive number of customers present at a service completion instant.

Lemma 5.19.

$$\begin{aligned} G^{(0)}(r,s) &:= \sum_{k=0}^{\infty} r^k \tilde{F}_k^{(0)}(s) \\ &= \frac{s}{\lambda(1-r)+s} \cdot \frac{\lambda(1-r\cdot\tilde{X}(\lambda(1-r)+s))+s}{\lambda+s}, \\ G(r,s) &:= \sum_{k=0}^{\infty} r^k \tilde{F}_k(s) \\ &= \frac{s}{\lambda(1-r)+s} \cdot \Big(1-\tilde{X}(\lambda(1-r)+s)\Big). \end{aligned}$$

Proof. First, we will prove the expression for $G^{(0)}(r,s)$. We separate the terms for $k=0$ and $k\geq 1$, insert the expression for $\tilde{F}_k^{(0)}(s)$ and perform some simple calculations yielding Eq. (5.29). Next, we condition on the interarrival time (for the case $k\geq 1$) and use the fact that for a given time t the events $\{A(t)=k\}$ and $\{X>t\}$ are independent. The final expression, Eq. (5.30), then readily follows from the Poisson arrival assumption and some simple manipulations.

$$\begin{aligned} G^{(0)}(r,s) &:= \sum_{k=0}^{\infty} r^k \tilde{F}_k^{(0)}(s) \\ &= s\cdot\int_{t=0}^{\infty} e^{-st}\mathbb{P}(A(t)=0)dt \\ &\quad + r\cdot\sum_{k=1}^{\infty} r^{k-1}\cdot s\cdot\int_{t=0}^{\infty} e^{-st}\mathbb{P}(A(t)=k,\ I+X>t)dt \qquad (5.29) \\ &= \frac{s}{\lambda(1-r)+s}\cdot\frac{\lambda(1-r\cdot\tilde{X}(\lambda(1-r)+s))+s}{\lambda+s}. \qquad (5.30) \end{aligned}$$

Analogously, we find for $G(r,s)$:

$$\begin{aligned} G(r,s) &:= \sum_{k=0}^{\infty} r^k \tilde{F}_k(s) \\ &= \sum_{k=0}^{\infty} r^k\cdot s\int_{t=0}^{\infty} e^{-st}\mathbb{P}(A(t)=k,\ X>t)dt \qquad (5.31) \\ &= \frac{s}{\lambda(1-r)+s}\cdot\Big(1-\tilde{X}(\lambda(1-r)+s)\Big). \end{aligned}$$

□

Let us give several definitions which will be used in the proof of Lemma 5.2:

$$\begin{aligned}
\pi_h^{(n)}(r,s) &:= \sum_{j=0}^{\infty} r^j \pi_{hj}^{(n)}(s),\ h=0,1,\ldots,\ n=1,2,\ldots,\\
\pi_{h0}(s,y) &:= \sum_{n=1}^{\infty} y^n \pi_{h0}^{(n)}(s),\ h=0,1,\ldots,\\
\pi_h(r,s,y) &:= \sum_{n=1}^{\infty} y^n \pi_h^{(n)}(r,s),\ h=0,1,\ldots\ .
\end{aligned}$$

Proof of Lemma 5.2. The proof of Lemma 5.2 consists in fact of three main steps. In the first step, we substitute the result of Corollary 5.18 into Eq. (5.32) leading to Eq. (5.33). Next, we work out the generating function with respect to the number of customers at the end of a visit. After some manipulations and using the definitions of $G^{(0)}(r,s)$, $G(r,s)$, $\pi_{h0}^{(n)}(s)$ and $\pi_h^{(n)}(r,s)$, we arrive at Eq. (5.34). In the final step, we use the definitions of $\pi_h(r,s,y)$ and $\pi_{h0}(s,y)$ and insert the explicit expressions for $G^{(0)}(r,s)$ and $G(r,s)$ which were derived in Lemma 5.19.

$$\sum_{n=1}^{\infty} y^n \sum_{k=0}^{\infty} r^k \int_{t=0}^{\infty} e^{-st} dp_{hk}^{(n)}(t) \tag{5.32}$$

$$= \sum_{n=1}^{\infty} y^n \sum_{k=0}^{\infty} r^k \left(\tilde{F}_k^{(0)}(s)\pi_{h0}^{(n)}(s) + \sum_{j=1}^{k} \tilde{F}_{k-j}(s)\pi_{hj}^{(n)}(s) \right) \tag{5.33}$$

$$= \sum_{n=1}^{\infty} y^n \left(G^{(0)}(r,s)\cdot \pi_{h0}^{(n)}(s) + G(r,s)\left(\pi_h^{(n)}(r,s) - \pi_{h0}^{(n)}(s) \right) \right) \tag{5.34}$$

$$\begin{aligned}
&= \frac{s}{\lambda(1-r)+s} \cdot \frac{\lambda(1-r\cdot\tilde{X}(\lambda(1-r)+s))+s}{\lambda+s} \cdot \pi_{h0}(s,y)\\
&\quad + \frac{s}{\lambda(1-r)+s} \cdot (1-\tilde{X}(\lambda(1-r)+s)) \cdot (\pi_h(r,s,y) - \pi_{h0}(s,y)).
\end{aligned}$$

□

5.B.3 Proof of Theorem 5.3

We prove the expression for $\beta^i(\mathbf{z})$ for a specific queue Q_i as given in Thm. 5.3 by first deriving the conditional p.g.f. $\beta_{\mathbf{n}}^i(\mathbf{z}) := \mathbb{E}[\mathbf{z}^{\mathbf{N}_i^e} | \mathbf{N}_i^s = \mathbf{n}]$ and then unconditioning on $\mathbf{N}_i^s$, the number of customers present at the start of a visit to Q_i. For convenience, let us define ξ_i^* as follows.

$$\xi_i^* := \xi_i + \sum_{j\neq i} \lambda_j (1-z_j).$$

We recall that we refer to a specific queue Q_i by adding an index i to a generic variable. Next, $\beta^i_{\mathbf{n}}(\mathbf{z})$ can be expressed as follows.

Lemma 5.20.

$$\begin{aligned}\beta^i_{\boldsymbol{n}}(\boldsymbol{z}) &= \frac{\xi_i}{\xi_i^*}\cdot\Big(G_i^{(0)}(z_i,\xi_i^*)\cdot\big(\pi_{n_i,0}(\xi_i^*,r^i(\boldsymbol{z}))+\mathbf{1}_{\{n_i=0\}}\big)\\ &\quad+G_i(z_i,\xi_i^*)\cdot\big(\pi_{n_i}(z_i,\xi_i^*,r^i(\boldsymbol{z}))+z_i^{n_i}\\ &\quad-\pi_{n_i,0}(\xi_i^*,r^i(\boldsymbol{z}))-\mathbf{1}_{\{n_i=0\}}\big)\Big)\cdot\prod_{j\neq i}z_j^{n_j}. \qquad (5.35)\end{aligned}$$

Proof. Let $A_{i,j}(t)$ denote the number of arrivals to Q_j (both external and internal arrivals) during a visit to Q_i. Recall further that $D_i(t)$ denotes the number of departures at Q_i from time 0 to t. Starting from the definition of the p.g.f., we condition on the timer Y_i and introduce the number of departures from Q_i until time t:

$$\begin{aligned}\beta^i_{\mathbf{n}}(\mathbf{z}) &= \sum_{m_1=0}^{\infty}\cdots\sum_{m_M=0}^{\infty}z_1^{m_1}\cdots z_M^{m_M}\mathbb{P}(\mathbf{N}_i^e=\mathbf{m}|\mathbf{N}_i^s=\mathbf{n})\\ &=\int_0^{\infty}\xi_i e^{-\xi_i t}\sum_{m_1=0}^{\infty}\cdots\sum_{m_M=0}^{\infty}z_1^{m_1}\cdots z_M^{m_M}\\ &\qquad\times\sum_{n}\mathbb{P}(\mathbf{N}_i(t)=\mathbf{m},\ D_i(t)=n|\mathbf{N}_i(0)=\mathbf{n})dt.\end{aligned}$$

After some simple rearrangements and using that given t and $D_i(t)$ the queue-length process at Q_i is independent of the aggregate arrival process to the other queues, we obtain the following:

$$\begin{aligned}&\int_0^{\infty}\xi_i e^{-\xi_i t}\sum_{n}\sum_{m_1=0}^{\infty}\cdots\sum_{m_M=0}^{\infty}z_1^{m_1-n_1}\cdots z_M^{m_M-n_M}\\ &\quad\times\ \mathbb{P}(\{A_{i,j}(t)=m_j-n_j,\ \forall_{j\neq i}\}|D_i(t)=n,\ \mathbf{N}_i(0)=\mathbf{n})\\ &\quad\times\ \sum_{m_i}z_i^{m_i}\mathbb{P}(N_{i,i}(t)=m_i|D_i(t)=n,\ \mathbf{N}_i(0)=\mathbf{n})\\ &\quad\times\ \mathbb{P}(D_i(t)=n|\mathbf{N}_i(0)=\mathbf{n})\ dt\cdot\prod_{j\neq i}z_j^{n_j}.\end{aligned}$$

These aggregate arrivals to $Q_j,\ j\neq i$, can be decomposed in two independent parts, viz., a first part referring to external arrivals at each queue and a second part referring to customers that were served at Q_i and routed to some other queue. The latter is represented by the term $(r^i(\mathbf{z}))^n$. Also noting that $N_{i,i}(t)$ depends only on $\mathbf{N}_i(0)$

through $N_{i,i}(0)$, we retrieve $p^{(n)}_{n_i m_i}(t)$ and eventually find that:

$$\beta^i_{\mathbf{n}}(\mathbf{z}) = \int_0^\infty \xi_i e^{-\xi_i^* t} \sum_{n=0}^\infty \sum_{m_i=0}^\infty z_i^{m_i} (r^i(\mathbf{z}))^n p^{(n)}_{n_i m_i}(t) dt \cdot \prod_{j \neq i} z_j^{n_j}. \tag{5.36}$$

Then, we can apply Lemma 5.2 for $n \geq 1$, while for $n = 0$ we use:

$$\begin{aligned}
&\sum_{m_i=0}^\infty z_i^{m_i} \int_0^\infty \xi_i e^{-\xi_i^* t} p^{(0)}_{n_i m_i}(t) dt \\
&= \mathbf{1}_{\{n_i=0\}} \cdot \sum_{m_i=0}^\infty z_i^{m_i} \int_0^\infty \xi_i e^{-\xi_i^* t} \mathbb{P}(A_i(t) = m_i,\ I_i + X_i > t) dt \\
&\qquad + \mathbf{1}_{\{n_i \geq 1\}} \cdot \sum_{m_i=0}^\infty z_i^{m_i} \int_0^\infty \xi_i e^{-\xi_i^* t} \mathbb{P}(A_i(t) = m_i - n_i,\ X_i > t) dt \\
&= \frac{\xi_i}{\xi_i^*} \cdot \left(\mathbf{1}_{\{n_i=0\}} \cdot G_i^{(0)}(z_i, \xi_i^*) + \mathbf{1}_{\{n_i \geq 1\}} \cdot z_i^{n_i} \cdot G_i(z_i, \xi_i^*) \right).
\end{aligned}$$

The final expression for $\beta^i_{\mathbf{n}}(\mathbf{z})$ follows from inserting this result together with the result from Lemma 5.2 into Eq. (5.36) and some simple manipulations. □

Proof of Theorem 5.3. Essentially, the proof follows immediately by unconditioning $\beta^i_{\mathbf{n}}(\mathbf{z})$ on the state $\mathbf{n} = (n_1, \ldots, n_M)$ at the start of the visit. The result of this operation is shown below. Equation (5.37) follows by substitution of Eq. (5.35) into the definition of $\beta^i(\mathbf{z})$. We note that the final expression, Eq. (5.38), follows from inserting the explicit expressions for $G_i^{(0)}(r, s)$ and $G_i(r, s)$ (see Lemma 5.19), inserting the expressions for $\pi_h(z_i, \xi_i^*, r^i(\mathbf{z}))$ and $\pi_{h0}(\xi_i^*, r^i(\mathbf{z}))$, $h \geq 0$, which are

given in Eqs. (5.2), (5.4) and (5.5), and some simple manipulations.

$$\begin{aligned}
\beta^i(\mathbf{z}) &= \sum_{n_1=0}^{\infty} \cdots \sum_{n_M=0}^{\infty} \beta^i_{\mathbf{n}}(\mathbf{z}) \mathbb{P}(\mathbf{N}^s_i = \mathbf{n}) \\
&= \sum_{n_1=0}^{\infty} \cdots \sum_{n_M=0}^{\infty} \mathbb{P}(\mathbf{N}^s_i = \mathbf{n}) \cdot \prod_{j \neq i} z_j^{n_j} \cdot \frac{\xi_i}{\xi_i^*} \\
&\quad \times \Big(G_i(z_i, \xi_i^*) \cdot \big(\pi_{n_i}(z_i, \xi_i^*, r^i(\mathbf{z})) + z_i^{n_i} - \pi_{n_i,0}(\xi_i^*, r^i(\mathbf{z})) - \mathbf{1}_{\{n_i=0\}} \big) \\
&\qquad + G_i^{(0)}(z_i, \xi_i^*) \cdot \big(\pi_{n_i,0}(\xi_i^*, r^i(\mathbf{z})) + \mathbf{1}_{\{n_i=0\}} \big) \Big) \qquad (5.37) \\
&= \frac{\xi_i}{z_i - r^i(\mathbf{z}) \cdot \tilde{X}_i(\xi_i + \sum_j \lambda_j(1 - z_j))} \\
&\quad \times \left(\frac{\tilde{X}_i(\xi_i + \sum_j \lambda_j(1 - z_j)) \cdot (z_i - r^i(\mathbf{z}))}{\lambda_i(1 - \hat{\mu}_i(\xi_i, r^i(\mathbf{z}))) + \xi_i^*} \cdot \alpha^i(\mathbf{z}_i^*) \right. \\
&\qquad \left. + \frac{(1 - \tilde{X}_i(\xi_i + \sum_j \lambda_j(1 - z_j))) \cdot z_i}{\lambda_i(1 - z_i) + \xi_i^*} \cdot \alpha^i(\mathbf{z}) \right), \qquad (5.38)
\end{aligned}$$

where $\alpha^i(\mathbf{z}_i^*) := \mathbb{E}[z_1^{N_1^s} \cdots \hat{\mu}_i(\xi_i^*, r^i(\mathbf{z}))^{N_i^s} \cdots z_M^{N_M^s}]$ and $\hat{\mu}_i(\xi_i^*, r^i(\mathbf{z}))$ is the root x with the smallest absolute value less than one of $x = r^i(\mathbf{z}) \cdot \tilde{X}_i(\xi_i^* + \lambda_i(1 - x))$. □

5.C Proofs of results Section 5.4

For convenience, let us recall the following definitions for $t > 0$:

$$\begin{aligned}
q^{(n)}_{hk}(t) &:= \begin{cases} \mathbb{P}(x_t = k,\ D(t) = n,\ x_v > 0,\ 0 < v < t | x_0 = h), \\ \qquad\qquad n = 0, 1, \ldots, \quad h, k = 1, 2, \ldots, \\ 0, \\ \qquad\qquad \text{otherwise}, \end{cases} \\
R^{(n)}_{hj}(t) &:= \mathbb{P}(z_{(n)} = j,\ r'_n < t,\ z_{(k)} > 0,\ 0 < k < n | z_{(0)} = h),\ h, j, n = 1, 2, \ldots, \\
F_k(t) &:= \mathbb{P}(A(t) = k,\ X > t),\ k = 0, 1, \ldots\ .
\end{aligned}$$

5.C.1 Proof of Proposition 5.6

The first observation is that each customer at Q_j, $j \neq i$, will still be present at the end of the visit, which is accounted for in the term $\prod_{j \neq i} z_j^{n_j}$. Second, each customer present at the start of the visit at Q_i will effectively be replaced by a random population during the course of the visit in an identical fashion. In particular, the size of this population is given by $\hat{\mu}_i(\xi_i^*, r^i(\mathbf{z}))$. To see this, recall that $\hat{\mu}_i(s, y)$ refers to the joint generating function of the busy period and the number of customers

served during this period. The term ξ_i^* in $\hat{\mu}_i(\xi_i^*, r^i(\mathbf{z}))$ accounts for the exogenous arrivals to the other queues in the system during a busy period which ends before the timer expires. Similarly, the term $r^i(\mathbf{z})$ in $\hat{\mu}_i(\xi_i^*, r^i(\mathbf{z}))$ accounts for the internal arrivals to the other queues (from Q_i) during this period. As initially there are n_i identical customers present at Q_i, this leads to n_i independent contributions which are recognized in the power of $\hat{\mu}_i(\xi_i^*, r^i(\mathbf{z}))$.

5.C.2 Proof of Lemma 5.7

Lemma 5.7 is readily proven by using similar arguments as in the proof of Lemma 5.1:

$$\begin{aligned} q_{hk}^{(n)}(t) &= \mathbb{P}(x_t = k,\ D(t) = n,\ x_v > 0,\ 0 < v < t | x_0 = h) \\ &= \mathbb{P}(x_t = k,\ r'_n \le t,\ r'_{n+1} > t,\ x_v > 0,\ 0 < v < t | z_{(0)} = h) \\ &= \int_{u=0}^{t} \sum_{j=1}^{k} \mathbb{P}(x_t = k, r'_{n+1} > t | r'_n = u, z_{(0)} = h, z_{(m)} > 0, 0 \le m \le n, \\ &\qquad z_{(n)} = j) \cdot d_u \mathbb{P}(\ r'_n \le u,\ z_{(n)} = j,\ z_{(m)} > 0,\ 0 < m < n | z_{(0)} = h) \\ &= \int_{u=0}^{t} \sum_{j=1}^{k} F_{k-j}(t-u) dR_{hj}^{(n)}(u). \end{aligned}$$

Let us define the following LSTs.

$$\begin{aligned} \tilde{F}_k(s) &:= \int_{0-}^{\infty} e^{-st} dF_k(t),\ k = 0, 1, \dots, \\ \gamma_{hj}^{(n)}(s) &:= \int_{0-}^{\infty} e^{-st} dR_{hj}^{(n)}(t),\ h, j, n = 1, 2, \dots\ . \end{aligned}$$

Then, a direct consequence of Lemma 5.7 is:

Corollary 5.21.

$$\int_{t=0}^{\infty} e^{-st} dq_{hk}^{(n)}(t) = \sum_{j=1}^{k} \gamma_{hj}^{(n)}(s) \tilde{F}_k(s),\ h, k, n = 1, 2, \dots\ . \tag{5.39}$$

5.C.3 Proof of Lemma 5.8

Let us give several definitions which will be used in the proof of Lemma 5.8:

$$\begin{aligned} \gamma_h^{(n)}(r, s) &:= \sum_{j=0}^{\infty} r^j \gamma_{hj}^{(n)}(s),\ h, n = 1, 2, \dots, \\ \gamma_h(r, s, y) &:= \sum_{n=1}^{\infty} y^n \gamma_h^{(n)}(r, s),\ h = 1, 2, \dots\ . \end{aligned}$$

Proof of Lemma 5.8. The proof consists of three consecutive steps similar to the proof of Lemma 5.2:

$$\sum_{n=1}^{\infty} y^n \sum_{k=1}^{\infty} r^k \int_{t=0}^{\infty} e^{-st} dq_{hk}^{(n)}(t) \tag{5.40}$$

$$= \sum_{n=1}^{\infty} y^n \sum_{k=1}^{\infty} r^k \sum_{j=1}^{k} \gamma_{hj}^{(n)}(s) \tilde{F}_{k-j}(s)$$

$$= \sum_{n=1}^{\infty} y^n \gamma_h^{(n)}(r,s) G(r,s) \tag{5.41}$$

$$= \gamma_h(r,s,y) \cdot \frac{s}{\lambda(1-r)+s} \cdot (1 - \tilde{X}(\lambda(1-r)+s)).$$

First, we substitute Eq. (5.39) into Eq. (5.40). Next, using the definitions of $\gamma_h^{(n)}(r,s)$ and $G(r,s)$ (see Lemma 5.19) immediately yields Eq. (5.41). The final step follows from the definition of $\gamma_h(r,s,y)$ and the substitution of the explicit expression for $G(r,s)$. □

5.C.4 Proof of Proposition 5.9

As a preliminary to proving Proposition 5.9, we present the following result for $h = 1, 2, \ldots$, for the special case of $D(t) = 0$, i.e., no departures occur before the timer expires.

Lemma 5.22.

$$\sum_{k=1}^{\infty} r^k \int_{t=0}^{\infty} e^{-st} dq_{hk}^{(0)}(t) = r^h \cdot \frac{s}{\lambda(1-r)+s} \cdot (1 - \tilde{X}(\lambda(1-r)+s)). \tag{5.42}$$

Proof.

$$\sum_{k=1}^{\infty} r^k \int_{t=0}^{\infty} e^{-st} dq_{hk}^{(0)}(t)$$

$$= r^h \cdot \int_{t=0}^{\infty} s e^{-st} \sum_{k=h}^{\infty} r^{k-h} \mathbb{P}(A(t) = k-h,\ X > t) dt \tag{5.43}$$

$$= r^h \cdot \frac{s}{\lambda(1-r)+s} \cdot (1 - \tilde{X}(\lambda(1-r)+s)). \tag{5.44}$$

Elaborating on the definition of $q_{hk}^{(0)}(t)$, we may obtain Eq. (5.43) after some simple manipulations. Equation (5.44) then follows directly from the earlier derivation of $G(r,s)$ (see Lemma 5.19). □

Proof of Proposition 5.9. To consider a specific queue Q_i, we will add here again an index i to the generic variables. Let $A_{i,j}(t)$ denote the number of arrivals to Q_j (both external and internal arrivals) during a visit to Q_i. Recall further that $D_i(t)$ denotes the number of departures at Q_i from time 0 to t. Starting from the definition of the p.g.f., we condition on the timer Y_i and introduce the number of departures from Q_i until time t:

$$\begin{aligned}
&\mathbb{E}[\mathbf{z}^{\mathbf{N}_i^e}\mathbf{1}_{\{timer\}}|\mathbf{N}_i^s=\mathbf{n}]\\
&= \sum_{m_1=0}^{\infty}\cdots\sum_{m_M=0}^{\infty} z_1^{m_1}\cdots z_M^{m_M}\,\mathbb{P}(\mathbf{N}_i^e=\mathbf{m},\ \{timer\}|\mathbf{N}_i^s=\mathbf{n})\\
&= \int_0^{\infty}\xi_i e^{-\xi_i t}\sum_{m_1=0}^{\infty}\cdots\sum_{m_M=0}^{\infty} z_1^{m_1}\cdots z_M^{m_M}\\
&\quad\times\sum_n \mathbb{P}(\mathbf{N}_i(t)=\mathbf{m},\ \{timer\},\ D_i(t)=n|\mathbf{N}_i(0)=\mathbf{n})dt.
\end{aligned}$$

Using that given t and $D_i(t)$ the queue-length process at Q_i is independent of the aggregate arrival process to the other queues and working out the event $\{timer\}$, we obtain:

$$\begin{aligned}
&\int_0^{\infty}\xi_i e^{-\xi_i t}\sum_n\sum_{m_1=0}^{\infty}\cdots\sum_{m_M=0}^{\infty} z_1^{m_1-n_1}\cdots z_M^{m_M-n_M}\\
&\quad\times\mathbb{P}(\{A_{i,j}(t)=m_j-n_j,\ \forall_{j\neq i}\}|D_i(t)=n,\ \mathbf{N}_i(0)=\mathbf{n})\\
&\quad\times\sum_{m_i} z_i^{m_i}\mathbb{P}(N_{i,i}(t)=m_i,\quad N_{i,i}(v)>0,\ 0<v<t|D_i(t)=n,\ \mathbf{N}_i(0)=\mathbf{n})\\
&\quad\times\mathbb{P}(D_i(t)=n|\mathbf{N}_i(0)=\mathbf{n})dt\cdot\prod_{j\neq i} z_j^{n_j}.
\end{aligned}$$

Exactly following the same reasoning that led to Eq. (5.36), we have that:

$$\mathbb{E}[\mathbf{z}^{\mathbf{N}_i^e}\mathbf{1}_{\{timer\}}|\mathbf{N}_i^s=\mathbf{n}] = \int_{t=0}^{\infty}\xi_i e^{-\xi_i^* t}\sum_{n=0}^{\infty}\sum_{m_i=1}^{\infty} z_i^{m_i}(r^i(\mathbf{z}))^n q_{n_i m_i}^{(n)}(t)dt\cdot\prod_{j\neq i} z_j^{n_j}.$$

Then, we may apply Lemma 5.8 for $n\geq 1$ and Lemma 5.22 for $n=0$. This leads after substituting the explicit expressions for $G_i(r,s)$ (see Lemma 5.19) and $\gamma_h(r,s,y)$ (see Eq. (5.2)) and performing some simple manipulations to the final expression, viz.:

$$\begin{aligned}
&\mathbb{E}[\mathbf{z}^{\mathbf{N}_i^e}\mathbf{1}_{\{timer\}}|\mathbf{N}_i^s=\mathbf{n}]\\
&= \frac{\xi_i\cdot z_i\cdot(1-\tilde{X}_i(\lambda_i(1-z_i)+\xi_i^*))\left(z_i^{n_i}-\hat{\mu}_i^{n_i}(\xi_i^*,r^i(\mathbf{z}))\right)}{[\lambda_i(1-z_i)+\xi_i^*]\cdot[z_i-r^i(\mathbf{z})\cdot\tilde{X}_i(\lambda_i(1-z_i)+\xi_i^*)]}\cdot\prod_{j\neq i} z_j^{n_j}.
\end{aligned}$$

□

5.C.5 Proof of Theorem 5.10

The final result for $\beta^i(\mathbf{z})$ is obtained by unconditioning the conditional p.g.f.'s of the previous propositions and then merging these outcomes. Let us define $\beta_e^i(\mathbf{z}) := \mathbb{E}[\mathbf{z}^{\mathbf{N}_i^e}\mathbf{1}_{\{empty\}}]$, $\beta_t^i(\mathbf{z}) := \mathbb{E}[\mathbf{z}^{\mathbf{N}_i^e}\mathbf{1}_{\{timer\}}]$ and $\alpha^i(\mathbf{z}_i^*) := \alpha^i(z_1, \ldots, z_{i-1}, \hat{\mu}_i(\xi_i^*, r^i(\mathbf{z})), z_{i+1}, \ldots, z_M)$. First, $\beta_e^i(\mathbf{z})$ and $\beta_t^i(\mathbf{z})$ are given in the following two lemmas which are immediate from unconditioning the expressions in Propositions 5.6 and 5.9.

Lemma 5.23.

$$\beta_e^i(\boldsymbol{z}) = \alpha^i(\boldsymbol{z}_i^*).$$

Lemma 5.24.

$$\beta_t^i(\boldsymbol{z}) = \frac{\xi_i \cdot z_i \cdot (1 - \tilde{X}_i(\lambda_i(1-z_i) + \xi_i^*))}{(\lambda_i(1-z_i) + \xi_i^*)(z_i - r^i(\boldsymbol{z}) \cdot \tilde{X}_i(\lambda_i(1-z_i) + \xi_i^*))} \cdot (\alpha^i(\boldsymbol{z}) - \alpha^i(\boldsymbol{z}_i^*)).$$

Proof of Theorem 5.10. The proof follows directly from the two final lemmas above:

$$\begin{aligned}
\beta^i(\mathbf{z}) &= \beta_e^i(\mathbf{z}) + \beta_t^i(\mathbf{z}) \\
&= \left(1 - \frac{\xi_i \cdot z_i \cdot (1 - \tilde{X}_i(\lambda_i(1-z_i) + \xi_i^*))}{(\lambda_i(1-z_i) + \xi_i^*)(z_i - r^i(\mathbf{z}) \cdot \tilde{X}_i(\lambda_i(1-z_i) + \xi_i^*))}\right) \cdot \alpha^i(\mathbf{z}_i^*) \\
&\quad + \frac{\xi_i \cdot z_i \cdot (1 - \tilde{X}_i(\lambda_i(1-z_i) + \xi_i^*))}{(\lambda_i(1-z_i) + \xi_i^*)(z_i - r^i(\mathbf{z}) \cdot \tilde{X}_i(\lambda_i(1-z_i) + \xi_i^*))} \cdot \alpha^i(\mathbf{z}).
\end{aligned}$$

□

CHAPTER

6

Approximations for the basic polling model

6.1 Introduction

Closed-form expressions for performance measures in polling systems are rather scarce. Even for measures such as the mean number of customers, except for some particular cases, there exist no such results for polling systems with a server operating under the well-studied exhaustive or gated disciplines. Nevertheless, there do exist efficient numerical solution methods to obtain exact performance results under these service disciplines, such as the Buffer Occupancy Method (see, e.g., [25, 29]), the Descendent Set Approach [59] or Mean Value Analysis [105]. It appears then also unlikely that explicit (i.e., non-recursive) closed-form *distributional* results for polling systems may be found. For this reason, several researchers have explored the numerical computation of the complete distribution of polling systems (see, e.g., [9, 67]). Such a numerical approach is particularly meaningful for the analysis of polling systems operating under non-branching type service disciplines for which generally the common methods for moment calculation cannot be applied. Unfortunately, such numerical methods break down when the number of stations grows large. As our proposed solution method for computation of the queue-length distribution of the basic polling model (see Chapter 4) is largely based on the approach of [67],

we suffer identical problems. It is thus meaningful to consider alternative analytical tools. Hence, we will resort to approximations in this chapter.

More specifically, we present two distinct approximations, viz.,

- a queue-length approximation for the basic polling model;
- a sojourn-time approximation for a mobile ad hoc networking application.

This first approximation is a product-form approximation for the joint queue-length distribution of the basic polling model. In particular, we will consider the conditional distribution where the condition is on the position of the server. The approximation is based on the marginal queue-length distributions as the correlation in queue lengths among the various queues appears small. First, we investigate the range of parameters for which this hypothesis holds indeed true. Subsequently, we present the approximation which is based on the unreliable-server model. Finally, we compare the queue-length results of the approximation and the exact solution along the measure of total variation distance.

The second approximation is an approximation for the marginal sojourn time at a mobile relay queue in a two-queue tandem network. This tandem network represents a model for a novel wireless communication paradigm and it can be seen as the basic polling model extended with customer routing. We note that the mean sojourn time could be derived from the exact analysis of the queue-length distribution. However, in a practical setting, a robust approximation might be preferred over an exact, but time-consuming, solution. Thus, we develop an analytical approach that provides an approximate expression for the LST of the sojourn time which enables a fast and accurate computation of the moments. The approximation is designed for exponential service requirements and builds on absorbing Markov chain analysis.

The organization of this chapter is as follows. First, we discuss the product-form approximation for the basic polling model in Sect. 6.2. Next, we present the approximation for the sojourn time in the two-queue tandem network in Sect. 6.3. We conclude this chapter in Sect. 6.4.

6.2 Queue-length approximation for the basic polling model

The numerical approach to obtain the joint queue-length probabilities (see Sect. 4.4.4) suffers from the state-space expansion. Hence, particularly for highly loaded systems, but also in case of light to moderate load, it is meaningful to consider an alternative analysis. As the queue lengths at the various queues in the polling system interact only via the common server, we will consider an approximation based on the assumption of the queues operating independently. That means, we will assume that each queue in the system can be seen as an unreliable-server model (see Sect. 4.3).

First, we study the correlation between the queue lengths in Sect. 6.2.1. The outcomes of this study have led to a proposed approximation which is discussed

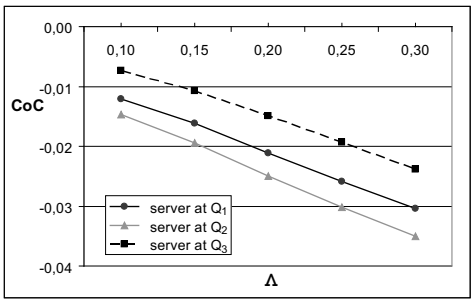

Figure 6.1: The coefficient of correlation (CoC) as function of Λ for $\mu = 1.0$ and $\xi = 1.0$ (exponential service times).

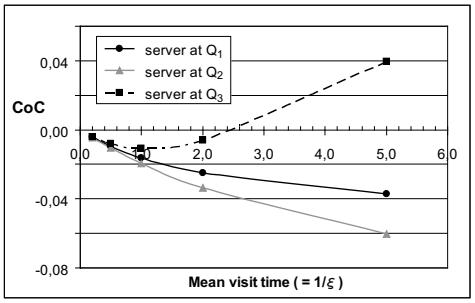

Figure 6.2: The coefficient of correlation as function of the mean visit time ($1/\xi$) for $\Lambda = 0.15$ and $\mu = 1.0$ (exponential service times).

in Sect. 6.2.2. In Sect. 6.2.3, we perform a numerical study on the quality of the approximation and we will conclude in Sect. 6.2.4.

6.2.1 Queue-length correlation

We have investigated the correlations among the conditional queue lengths for the basic polling model with three symmetric queues for a wide range of parameter values. It is assumed that the server strategy is cyclic, service times of the customers are exponential and that switch-over times have zero duration. We should emphasize that we only consider conditional queue lengths, i.e, conditional on the position of the server. This is because the system state generally depends heavily on the position of the server, so that it is more meaningful to compare conditional probabilities. Also, if we would take a snapshot of the system state at a random instant in time, then we do not expect it to be in line with the unconditional steady-state probabilities. For ease of presentation, we define $\Lambda := \sum_j \lambda_j,\ \mu := \mu_i,\ \xi := \xi_i$, and denote by N_i the steady-state queue length at $Q_i,\ i = 1, \ldots, M$.

Let us introduce first the performance measure for the amount of correlation between the queue lengths. We will consider the coefficient of correlation, $\rho_{1,2|Q_j},\ j = 1, 2, 3$, for the conditional queue length at Q_1 and Q_2 as function of Λ and ξ, where

$\rho_{1,2|Q_j}$ is defined as follows:

$$\begin{aligned}\rho_{1,2|Q_j} &:= \frac{\text{Cov}(N_1\ ,\ N_2 \mid \text{server at } Q_j)}{\sqrt{\text{Var}(N_1 \mid \text{server at } Q_j)\text{Var}(N_2 \mid \text{server at } Q_j)}} \\ &= \frac{\mathbb{E}[N_1, N_2 \mid \text{server at } Q_j]}{\sqrt{\text{Var}(N_1 \mid \text{server at } Q_j)\text{Var}(N_2 \mid \text{server at } Q_j)}} \\ &\quad - \frac{\mathbb{E}[N_1 \mid \text{server at } Q_j] \cdot \mathbb{E}[N_2 \mid \text{server at } Q_j]}{\sqrt{\text{Var}(N_1 \mid \text{server at } Q_j)\text{Var}(N_2 \mid \text{server at } Q_j)}}.\end{aligned}$$

The results for $\rho_{1,2|Q_j}$ are computed using the exact analysis of Chapter 4.

In Fig. 6.1, we plot $\rho_{1,2|Q_j}$ as function of the total arrival rate Λ for $\mu = 1.0$ and $\xi = 1.0$. It is shown that the correlation between the queues is quite small (for all server's positions), although it increases (in absolute sense) slightly in Λ. Figure 6.2 shows the impact of the mean visit time to a queue ($= 1/\xi$) on $\rho_{1,2|Q_j}$ for $\Lambda = 0.15$ and $\mu = 1.0$. The plot shows that the coefficient of correlation is small for short visit times, but that it may drift rapidly away from zero when the visit times grow large. This is in accordance with the fact that for $1/\xi \downarrow 0$ the queue lengths indeed become independent (under zero switch-over times) yielding a coefficient of correlation equal to zero. In fact, in this limit case, each queue operates as an M/M/1 queue with arrival rate λ_i and rescaled service rate $\mu_i^* := \mu_i/\xi_i/\mathbb{E}[C]$ (here, $\mu_i^* = \mu/M$), where $\mathbb{E}[C]$ is the mean cycle time. We note that this specific regime is attained only for zero switch-over times and exponential service times.

We have also generated results for many other parameter settings for this symmetric three-queue polling system. These results demonstrate that for a wide range of settings the coefficient of correlation is indeed quite small which indicates little dependence between the queue lengths at the different queues.

6.2.2 Approximation

A natural next step is to propose an approximation for the conditional joint queue-length distribution of the polling model exploiting the observed "quasi-independence" of the queues. Such approximations are of great value since we have experienced that the computation time for the exact joint queue-length probabilities in the polling model may grow quite large, while for a single queue the results are immediate. Thus, we base the approximation for the conditional joint queue-length distribution on the marginal distributions. These marginal distributions can be computed almost directly via the unreliable-server model (USM) (see Sect. 4.3). Specifically, the approximation reads as follows:

$$\mathbb{P}(N_1 = n_1, \ldots, N_M = n_M | \text{server at } Q_j) \approx \prod_{i=1}^{M} \mathbb{P}(N_i = n_i | \text{server at } Q_j). \quad (6.1)$$

To assess the quality of this approximation, we will compute the terms on the r.h.s. of Eq. (6.1) via the analysis of the USM. As we have not analyzed these conditional terms yet, this will be done next.

Let us consider the unreliable-server model as described in Sect. 4.3 with arrival rate λ, service rate μ, exponentially distributed availability periods with parameter ξ. We denote the Erlang$_{M-1}(\xi)$ distributed repair period by B, and its LST by $\tilde{B}(\cdot)$. Individual (exponential) repair stages are denoted by $B^j,\ j = 1, \ldots, M-1$, with LST $\tilde{B}^j(\cdot)$. W.l.o.g. we consider the p.g.f. $\hat{N}_{1j}(z) = \mathbb{E}[z^{N(Q_1)}|\text{server at } Q_j],\ j = 1, \ldots, M$, which refer to the number of customers at Q_1 given that the server is either at the queue ($j = 1$) or at stage $j-1$ of the repair period ($j \neq 1$).

Let us first consider $\hat{N}_{11}(z)$ in more detail. Notice that (due to exponentially distributed availability periods) $\hat{N}_{11}(z)$ in fact refers to the p.g.f. of the number of customers present at an arbitrary instant of the availability period. Denote further by $\hat{N}_{1B}(z)$ the p.g.f. of the number of customers present at an arbitrary instant of the repair period. These quantities are related to $P_{L_d}(z)$ (see Eq. (4.1)) as follows:

$$P_{L_d}(z) = \kappa \cdot \hat{N}_{11}(z) + (1-\kappa) \cdot \hat{N}_{1B}(z),$$

where κ $(= 1/M)$ and $1-\kappa$ are the long-term fractions that the server is available and being repaired, respectively. Observe that $\hat{N}_{11}(z)$ and $\hat{N}_{1B}(z)$ are also related via:

$$\hat{N}_{1B}(z) = \hat{N}_{11}(z) \cdot \hat{B}_A(z),$$

where $\hat{B}_A(z)$ is the p.g.f. of the number of arrivals from the start of the repair period until an arbitrary instant of that period, and satisfies, using simple regenerative processes theory (see, e.g., [41]):

$$\hat{B}_A(z) = \frac{1-\hat{B}(z)}{\hat{B}'(1)(1-z)},$$

where $\hat{B}(z)$ $(= \tilde{B}(\lambda(1-z)))$ is the p.g.f. of the number of arrivals during the complete repair period.

Hence, it follows that:

$$\hat{N}_{11}(z) = \frac{P_{L_d}(z)}{\kappa + (1-\kappa) \cdot \hat{B}_A(z)}.$$

Next, we note that $\hat{N}_{1j}(z),\ j \neq 1$, can be decomposed in three independent parts. The first part refers to the number of customers present at the end of an availability period. The second part accounts for the arrivals during the already completed repair stages. Finally, the last part represents the number of arrivals from the beginning of repair stage $j-1$ until a random instant during this stage. In terms of p.g.f.'s this leads to:

$$\hat{N}_{1j}(z) = \hat{N}_{11}(z) \cdot \prod_{k=1}^{j-2} \hat{B}^k(z) \cdot \hat{B}_A^j(z),\ j = 2, \ldots, M,$$

where $\hat{B}^k(z)$ refers to the arrivals during the k-th stage of the repair period and is given by:

$$\hat{B}^k(z) = \tilde{B}^k(\lambda(1-z)),\ k = 1, \ldots, M-2,$$

and $\hat{B}_A^j(z)$ (cf., $\hat{B}_A(z)$) is given by:

$$\hat{B}_A^j(z) = \frac{1-\hat{B}^j(z)}{\hat{B}'^j(1)(1-z)},\ j = 2, \ldots, M.$$

Finally, the probabilities $\mathbb{P}(N_i = n_i|\text{server at } Q_j)$ are obtained from $\hat{N}_{1j}(z)$ in a numerical fashion using Discrete Fourier Transform techniques (see also Sect. 1.3.3.2).

We have outlined above the proposed approximation for the conditional joint queue-length distribution. The approximation is anticipated to work well in situations where the individual queues behave "almost" independently. Let us emphasize that our objective here is not to perform an exhaustive numerical study for all system parameters and service time distributions. The underlying idea of the approximation is that if the queues in the system would turn out to be "almost" independent, then the results of a much simpler single-queue model can be used as a good approximation for a complex multi-queue polling model. Therefore, our purpose is mainly to gain preliminary insight in the parameter ranges for which the approximation works well.

Remark 6.1 (Asymmetric systems)**.** *Notice that this approximation approach also works for asymmetric polling systems. However, in that case all steps above have to be performed for each queue separately.*

6.2.3 Numerical evaluation

We present numerical results from experiments for a symmetric three-queue polling model for both exponentially and deterministically distributed service times with the mean service time $1/\mu$ set equal to one. It follows that for these distributions the (effective) load at Q_i is given by (see Thm. 3.1):

$$\begin{aligned} \rho_i &= \lambda_i && \text{(exponential service)}, \\ \rho_i &= \lambda_i \cdot \frac{1-e^{-\xi_i}}{\xi_i \cdot e^{-\xi_i}} && \text{(deterministic service)}. \end{aligned}$$

The total load of the system equals $M \cdot \rho_i$. The performance measure that we adopt to assess the quality of the distributional approximation is the measure of total variation distance [35, p.286] for the queue-length distribution conditional on the

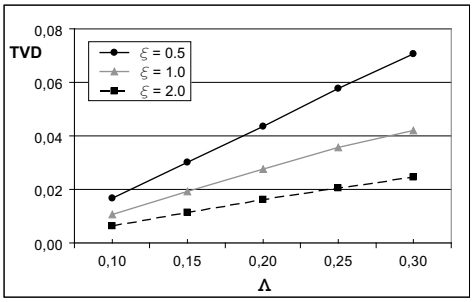

Figure 6.3: The total variation distance (TVD) as function of Λ (exponential service times).

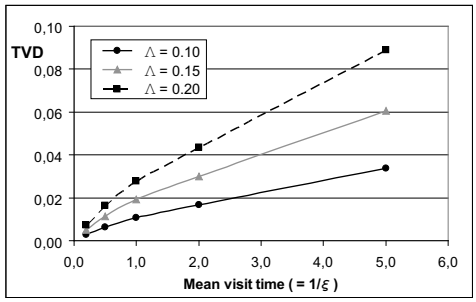

Figure 6.4: The total variation distance (TVD) as function of the mean visit time ($1/\xi$) (exponential service times).

position of the server, denoted by $\theta^p_{cond,j}$:

$$\theta^p_{cond,j} := \sum_{\mathbf{n}} \Big| \mathbb{P}(N_1 = n_1, \ldots, N_M = n_M \mid \text{server at } Q_j) - \prod_{i=1}^{M} \mathbb{P}(N_i = n_i \mid \text{server at } Q_j) \Big|.$$

For ease of presentation, we define $\theta^p_{cond} := \theta^p_{cond,j}$, for $j = 1, \ldots, M$. This measure quantifies the difference between the exact and the approximate distribution. Clearly, if the approximation is exact (e.g., when the queue lengths are indeed independent) θ^p_{cond} equals zero, and it is strictly positive otherwise.

The total variation distance in the exponential case is presented in Figs. 6.3 and 6.4. First, consider Fig. 6.3 in which θ^p_{cond} is plotted as function of Λ for various values of ξ. The slopes observed in this figure clearly show that θ^p_{cond} is not insensitive to Λ, but increases about linearly in the arrival rate. Moreover, it can be witnessed that θ^p_{cond} decreases in ξ. This is further illuminated in Fig. 6.4 which shows the impact of the mean visit time (i.e., $1/\xi$) on θ^p_{cond} for various values of Λ. It is shown that θ^p_{cond} is quite small for short visit times and increases linearly in $1/\xi$ for longer visit times. This former observation follows readily from the fact that the approximation is exact for the limit case $\xi \to \infty$ as discussed before.

The results for deterministic service times are presented in Figs. 6.5 and 6.6. Figure 6.5 shows θ^p_{cond} as function of Λ for various values of ξ. Again, as for the exponential case, θ^p_{cond} increases linearly in Λ. The impact of ξ on θ^p_{cond} appears small. This is confirmed by the plot of Fig. 6.5 which shows the total variation distance as function of the mean visit time for various values of Λ. However, an important difference with respect to the exponential case is that θ^p_{cond} stays away from zero when the mean visit time decreases as depicted in Fig. 6.6. The latter occurs since the effective load for the deterministic case increases in ξ (due to the

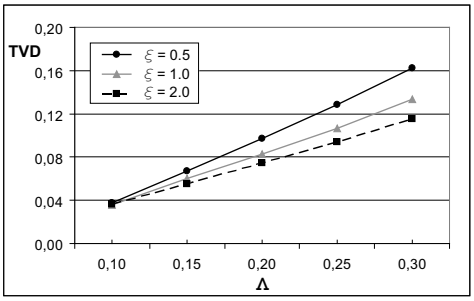

Figure 6.5: The total variation distance (TVD) as function of Λ (deterministic service times).

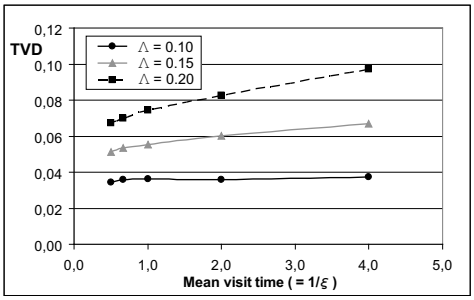

Figure 6.6: The total variation distance (TVD) as function of the mean visit time (deterministic service times).

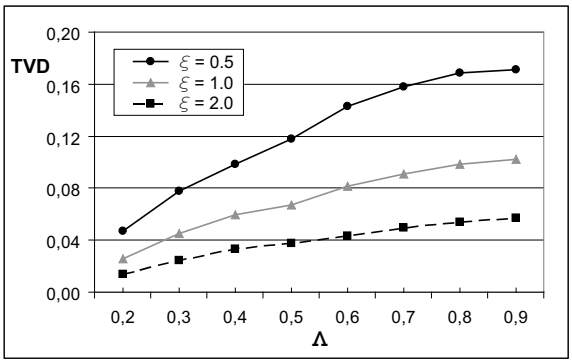

Figure 6.7: The total variation distance (TVD) as function of Λ for the two-queue system (exponential service times).

increasing number of preemptions), so that the queue lengths will not approach independence for $1/\xi \downarrow 0$ under the stable regime.

The figures reveal the complete behaviour of the TVD as function of the mean visit time. Contrary, the behaviour of the TVD for high values of Λ (i.e., for heavy load) remains unrevealed, while in fact this range of input values may often be interesting from an operational point of view. Unfortunately, we are not able to push the load in the three-queue system much further due to the curse of the rapidly growing state space. Therefore, we have performed additional experiments for a two-queue polling system with exponential service times instead, so that we can get insight in the performance measure under the heavy load regime. These results for a symmetric two-queue polling system are presented in Fig. 6.7.

Figure 6.7 shows that the relation between the total arrival rate Λ (here, also the system load) and TVD is not linear over the whole load range, but assumes in fact

a sublinear shape. This delineates that the quality of the approximation does not degrade very fast as the load increases, so that our approximation appears a quite useful approach to assess the joint distribution when other analytical methods fail.

6.2.4 Concluding remarks on the queue-length approximation

The exact computation of the queue-length distribution of the basic polling model suffers from the state-space expansion. This problem renders the computational process slow and thus impracticable. To resolve this problem, we have proposed a product-form approximation which neglects the small correlation in the queue lengths that is present. We have compared the results of our approach with the exact results mainly for three-queue polling systems along the measure of total variation distance, θ^p_{cond}. The main observations in our experiments can be summarized as follows: (i) θ^p_{cond} is positively correlated to the arrival rate Λ and the two-queue results indicate that the relation is even sublinear; (ii) θ^p_{cond} also increases in the mean visit time; though, it decreases rapidly toward zero when the mean visit time is shortened for exponential service times, while for deterministic services θ^p_{cond} seems to decrease to an asymptotic value strictly larger than zero.

We have seen that for a wide range of parameter settings the approximation for the three-queue polling system works quite well. Though, it is not clear to what extent the experimental results will carry over to systems with many more than three queues. Besides, the approximation appears less applicable to systems with both extremely large visit times and heavily loaded queues. For such situations, it might be worthwhile to consider specific heavy-traffic approximations.

6.3 Sojourn-time approximation for a two-queue tandem model

Communication in a wireless ad hoc environment is typically multi-hop, i.e., an end-to-end transmission consists of multiple (wireless) links. The performance in terms of delay and buffer sizes of single queues is well-studied, but for networks of multiple queues fewer analytical results are known. Moreover, exact numerical methods may break down for large networks or heavily loaded networks as witnessed in Chapter 4. Hence, in this section, we present an approximate analysis for the performance of a mobile ad hoc networking application. This analysis is largely based on [H2].

In particular, we will study a model for the paradigm known as opportunistic, or also delay tolerant, networking (see, e.g., [26, 85]). This novel networking approach targets at providing end-to-end quality of service in networks with highly volatile network topologies (due to mobility, station breakdowns, etc.). Our model extends previous analytical work in this area in various directions. We will account, unlike [44, 49, 94], for the fact that the transmission of packets may fail due to the short period of link availability and a retransmission is required. Also, we assume

that the source station has a stream of packet arrivals instead of only one packet, like it was considered in [72, 94, 110]. Besides, we are interested in what happens in a more practical case of small, finite-size networks, rather than in asymptotic cases (see, e.g., [45, 110]). To this end, we adopt the network scenario of a fixed source and destination station and a single mobile station which acts as a relaying device. Although it is a small model, it contains the main characteristics of an opportunistic network and it is also non-trivial from an analytical perspective.

The network model of our interest is reminiscent of a two-queue tandem model with a single alternating server. Such a tandem model has been analyzed under various servicing strategies (see, e.g., [104]) which are typically based on the assumption that the server can be controlled. However, in the mobility-driven network model of our interest, the server is autonomous and there is no possibility to control its movements. The research efforts on models with exhaustive time-limited service periods are also closely related to our work. In a two-queue setting, [22] analyzes the model via boundary value techniques. Exhaustive time-limited service models have also been studied in the context of polling systems (see, e.g., [39, 68]). However, also in these models, there exists a notion of server control, since it is assumed that whenever a queue becomes empty the server moves to another queue.

It is good to notice here that the delay at the source station can easily be found from the analysis of an unreliable-server model. Hence, our main interest is to assess the sojourn time or delay at the relay station in the opportunistic network described above. We study this tandem queueing model at the packet level by considering it as an extended version of the basic polling model as presented in Chapter 4. That is, a two-queue polling system with the server operating under the pure exponential time-limited discipline and with customers of one queue being routed to the next queue after being served. We note that the mean sojourn times at the queues could be obtained via the exact analytical framework (see Chapter 4). However, the computation time of the joint queue-length probabilities (and thus also the mean sojourn times) may grow rapidly when the load in the system increases. Hence, we perform an approximative analysis for the LST of the delay at the mobile relay station for the case of exponential service requirements at all stations. This is done by analyzing the queue-length process at the relay station in isolation as a workload process with Poisson batch arrivals. This Poisson process is inherited directly from the exponential time-limited discipline. The key element of the analysis is to accurately approximate the batch-size distribution. Numerical experiments for the mean sojourn time at the relay queue demonstrate the excellent performance of the approximation for a broad range of parameter settings. The experiments further show that the mean end-to-end sojourn time is insensitive to third and higher moments of the switch-over times. Finally, several guidelines are given for sojourn time optimization by power control.

The rest of this section is organized as follows. In Sect. 6.3.1, we give the description of the opportunistic network model and show the mapping to the polling model. For completeness, in Sect. 6.3.2, we discuss the exact analysis for the queue-

length distribution and the mean sojourn time at the relay station for this specific model. In Sect. 6.3.3, we present the analytic approximation for the sojourn time of a customer at the relay queue. Numerical results are given in Sect. 6.3.4 to validate the approximation, to assess the impact of the switch-over time distribution, and to consider optimization of the mean end-to-end sojourn time by tuning specific model parameters. We wrap up this approximation in Sect. 6.3.5.

6.3.1 Model

We consider a tandem model consisting of three first-in-first-out (FIFO) single-server queues with unlimited buffer size. The third queue functions merely as a sink station and will not be included in the rest of the analysis. Customers arrive to Q_1 according to a Poisson process with rate λ and subsequently require service at Q_2 before they arrive at the sink queue. The service requirements at Q_i are according to a generic random variable X_i which follows a general distribution $X_i(\cdot)$, with LST $\tilde{X}_i(\cdot)$, and mean $1/\mu_i$. The special feature of the model is that Q_2 is a mobile queue and it alternates between two positions, viz., L_1 and L_2. We will say that when Q_2 is in position L_1 (see Fig. 6.8(a)) that the server visits Q_1, and when Q_2 is in position L_2 (see Fig. 6.8(b)) the server visits Q_2. It is good to notice that at most one queue (either Q_1 or Q_2) can be served at a time. The movement of Q_2 is autonomous in the sense that Q_2 remains during its n-th visit, $n = 0, 1, \ldots$, an exponential period of time at location L_1 (resp. L_2), which we denote by $E^n_{L_1}$ (resp. $E^n_{L_2}$), before it migrates to position L_2 (resp. L_1). We assume that $E^n_{L_1}$ ($E^n_{L_2}$) is an independent and identically distributed (i.i.d.) sequence of exponentially distributed random variables with rate ξ_1 (ξ_2). When the server moves away from a queue during the service of a customer (i.e., at the end of a visit), this service will be preempted. At the beginning of a next server visit, the service time will be re-sampled according to $X_i(\cdot)$. This discipline is referred to as *preemptive-repeat-random* discipline. We recall that in a wireless environment the transmission rate is dominated by the dynamic channel conditions, so that adopting for instance a preemptive-resume strategy would make little sense here. In addition, Q_2 incurs a switching time from L_i to L_j ($i \neq j,\ j \in \{1, 2\}$) during which no customers are served. Specifically, after the n-th visit to L_1, Q_2 incurs a switch-over time $C^n_{1,2}$ from L_1 to L_2, and similarly a switch-over time $C^n_{2,1}$ after the n-th visit to L_2. We assume that $C^n_{1,2}$ ($C^n_{2,1}$) is an i.i.d. sequence with general distribution $C_{1,2}(\cdot)$ $\big(C_{2,1}(\cdot)\big)$, LST $\tilde{C}_{1,2}(\cdot)$ $\big(\tilde{C}_{2,1}(\cdot)\big)$, and mean $c_{1,2}$ ($c_{2,1}$). Furthermore, we assume $\{E^n_{L_1}, E^n_{L_2}, C^n_{1,2}, C^n_{2,1}\}$ are i.i.d. and mutually independent. Hence, the location of Q_2 can be described by a continuous-time process $\{L(t) : t \geq 0\}$ on the state space $\{-2, -1, 0, 1\}$. More precisely, we let $L(t) = 1$ ($L(t) = 0$) when Q_2 is at L_1 (resp. L_2) at time t, and $L(t) = -1$ ($L(t) = -2$) when Q_2 switches from L_1 to L_2 (L_2 to L_1). Without loss of generality, let $L(0) = 1$. Finally, we let $N_i(t)$ denote the number of customers at Q_i, $i = 1, 2$, at time t and assume that $N_i(0) = 0$, $i = 1, 2$.

The mapping of the two-hop tandem model to the polling model with the server

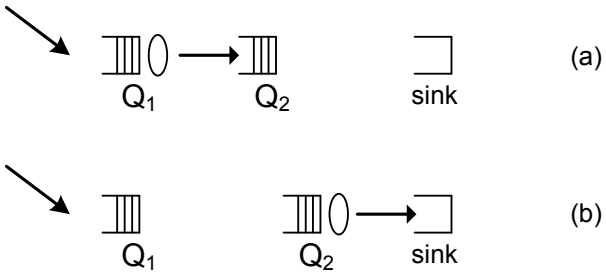

Figure 6.8: Customer flows when (a) Q_2 is in position L_1 and when (b) Q_2 is in position L_2.

operating under the pure time-limited service discipline is as follows. We adopt the basic polling model with $M = 2$ queues and Poisson arrivals to Q_1 with rate λ and to Q_2 with rate 0. For $Q_i,\ i = 1, 2$, the service times and visit times are exponentially distributed with rate μ_i and ξ_i, respectively. The nonzero customer-routing probabilities are $r_{12} = r_{20} = 1$. The switch-over times $C_{1,2}$ and $C_{2,1}$ are generally distributed with mean $c_{1,2}$ and $c_{2,1}$.

The necessary and sufficient condition for stability for $Q_i,\ i = 1, 2$, reads (see Sect. 3.3):

$$Q_i \text{ is stable } \iff \frac{\lambda}{\mu_i} < \frac{1/\xi_i}{\mathbb{E}[C]}, \tag{6.2}$$

where $\mathbb{E}[C]$, the mean cycle time, satisfies:

$$\mathbb{E}[C] = \frac{1}{\xi_1} + \frac{1}{\xi_2} + c_{1,2} + c_{2,1}.$$

The whole system is stable when both Q_1 and Q_2 are stable. Notice that in the stable regime, the mean arrival rates to Q_1 and Q_2 are equal.

In the analysis, we will concentrate on the sojourn time at the relay queue, since the marginal analysis of the source queue boils down to analyzing an unreliable-server model. Specifically, the objective is to approximate the *mean* sojourn time of a customer at the relay queue. However, our final approximate expression for the LST of the sojourn time allows for studying second and higher moments of the sojourn time as well. Finally, it is good to notice that the end-to-end sojourn time can be determined from the exact analysis of the polling system.

6.3.2 Exact analysis

6.3.2.1 Queue-length distribution

The joint queue-length distribution at visit completion instants for the specific polling model under consideration can be found using the exact analytical framework pre-

sented in Chapter 4. The key relation within this framework, Eq. (1.2), can be computed in a recursive fashion using the relations in Sect. 4.4.4 and Sect. 4.5.1 (to account for customer routing), or in a direct fashion using the relations in Sect. 5.3. To account specifically for the customer routing, one should use: $r^1(\mathbf{z}) = z_2$ and $r^2(\mathbf{z}) = 1$. Notice that the server polling strategy is in fact cyclic, so that no modifications have to be made there. Recall that the exact framework of Chapter 4 embeds an iterative scheme which is mainly applicable to systems with a light to moderate load.

6.3.2.2 Sojourn times

For the two-queue tandem model, the joint steady-state queue-length distribution, $P(\mathbf{z})$, follows from Eq. (4.27). The marginal queue-length distribution for Q_1, denoted by $P_1(z)$, and Q_2, denoted by $P_2(z)$, are then given by $P_1(z) = P(z, 1)$ and $P_2(z) = P(1, z)$.

The LST of the end-to-end sojourn time, $\tilde{D}(s)$, is immediate from the distributional form of Little's law (see [55]), i.e.,

$$\tilde{D}(s) = P(z, z) \mid_{z=1-s/\lambda} .$$

However, this distributional form cannot be applied to find the LST (and thus also second and higher moments) for the marginal sojourn time at Q_2, since the arrival process to Q_2 does not satisfy the non-anticipating property [109]. Hence, we provide here only the exact expression for the first moment of the sojourn time at Q_2, $\mathbb{E}[D_2]$, which follows immediately from Little's law [71]:

$$\mathbb{E}[D_2] = \frac{\mathbb{E}[N_2]}{\lambda} = \frac{1}{\lambda} \cdot \frac{d}{dz} P_2(z) \mid_{z=1} . \tag{6.3}$$

6.3.3 Approximation

The exact solution approach becomes computationally less attractive when the load increases. To be able to analyze the system performance also in such scenarios, we present an approximation for $\tilde{D}_2(s)$, the LST of the sojourn time of a customer in Q_2, the mobile queue. This will be done under the assumption that the service times are exponentially distributed at both queues (still with rate μ_1 and μ_2).

The approximation builds on the analysis of the workload process in Q_2 when $L(t) = 0$, i.e., Q_2 is served. It turns out that this process corresponds to the workload process in an $M/M/1$ queue with batch arrivals. The sojourn time of a customer in Q_2 then equals the sum of:

- the sojourn time of a customer in the $M/M/1$ batch-arrival queue;
- the period of time a customer is at Q_2, but Q_2 is not served.

We note that for the case of exponential service times the preemptive-repeat-random and the preemptive-resume disciplines are stochastically identical. For ease of presentation, we will consider the preemptive-resume discipline below.

6.3.3.1 The workload process at the relay queue

To study the workload process at the relay queue, Q_2, we split the time into disjoint intervals which begin at the time instants that the $L(t)$-process jumps from state -2 to 1 (i.e., at the start of a server visit to Q_1). Denote the starting points of these intervals by $\{Z_n,\ n = 1, 2, \ldots\}$ with the convention that $Z_1 = 0$. Let the n-th cycle of $L(t)$ denote the time interval $[Z_n, Z_{n+1})$, with duration $E^n_{L_1} + C^n_{1,2} + E^n_{L_2} + C^n_{2,1}$. Let $V(t)$ denote the workload (i.e., virtual waiting time of a customer) at Q_2 at time t. Without loss of generality, we assume that $V(t)$ is left-continuous, i.e., arrivals are not counted until just after they arrive. A sample path of the simultaneous evolution of $L(t)$ and $V(t)$ over time is shown in Figure 6.9.

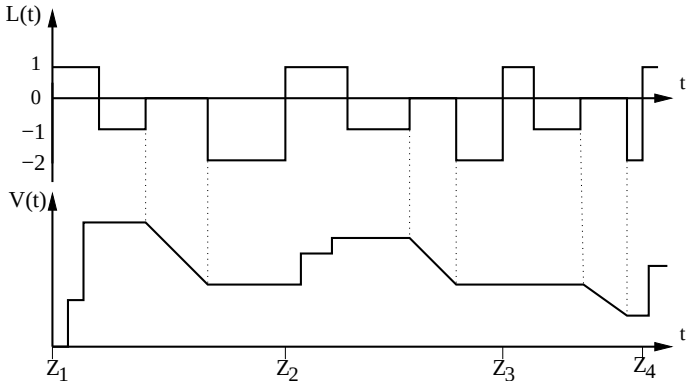

Figure 6.9: Evolution of $L(t)$ and $V(t)$ for Q_2.

Let K_n be the total number of arrivals to Q_2 (or equivalently, the number of departures from Q_1) during $E^n_{L_1}$, $X_{2,i}$ be the service requirement of a customer in Q_2 distributed as X_2, and let W^n_B denote the workload present in Q_2 at time Z_n. Then, based on the evolution of $L(t)$, it is easily seen that:

$$W_B^{n+1} = \left(W_B^n + \sum_{i=1}^{K_n} X_{2,i} - E^n_{L_2}\right)^+, \quad n \geq 0, \tag{6.4}$$

where $(\cdot)^+ = \max(\cdot, 0)$. It is useful to notice that a cycle consists of exactly one period during which arrivals to Q_2 may occur and one period during which the server may serve customers at Q_2. We note that $X_{2,i}$ is independent of $E^n_{L_2}$ and that K_n depends on $N_1(Z_n)$, the number of customers at Q_1 at time Z_n. Therefore,

K_n, $n = 1, 2, \ldots$, are correlated. For the sake of model tractability, we need the following:

Assumption: *K_n, $n = 1, 2, \ldots$, are i.i.d. and independent of $\{E_{L_2}^m : m < n\}$.*

By this assumption, Eq. (6.4) represents the workload seen by the first customer of a batch in a queue with Poisson batch arrivals with rate ξ_2, independent batch size K_n, and exponential service requirements with rate μ_2. The specific distribution of K_n will be considered in the next subsection.

Let $\tilde{G}(s) := \mathbb{E}\left[e^{-s\sum_{i=1}^{K_n} X_{2,i}}\right]$, i.e., $\tilde{G}(s)$ denotes the LST of the service requirement of a batch. It is well known that the batch-arrival queue is stable when $-\xi_2\tilde{G}'(0) = \xi_2\mathbb{E}[K_n]/\mu_2 < 1$. Noting that $\mathbb{E}[K_n] = \lambda \cdot \mathbb{E}[C]$, it can readily be verified that the latter condition is equivalent to the condition in Eq. (6.2) for Q_2. By conditioning on K_n, we find that

$$\tilde{G}(s) \quad = \quad \mathbb{E}\left[\left(\frac{\mu_2}{\mu_2 + s}\right)^{K_n}\right]. \tag{6.5}$$

The LST of the steady-state distribution of W_B^n follows from the well-known Pollaczek-Khinchine formula, yielding:

$$\tilde{W}_B(s) = \left(1 + \xi_2\tilde{G}'(0)\right) \cdot \frac{s}{s - \xi_2(1 - \tilde{G}(s))}.$$

Next, we want to transpose this result for the workload at the embedded points to the workload as seen by an arbitrary arriving customer. To this end, let $\tilde{V}^j(s)$ denote the LST of the workload seen by the j-th customer within a batch upon arrival (including the work brought in by itself). Since the service requirement of arriving customers is independent of the workload present in the queue, it follows that:

$$\tilde{V}^j(s) = \tilde{V}^{j-1}(s) \cdot \frac{\mu_2}{\mu_2 + s}, \quad j = 1, 2, \ldots,$$

with $\tilde{V}^0(s) = \tilde{W}_B(s)$. Moreover, since the K_n are assumed i.i.d., the probability that a customer is the j-th customer within its batch, which we denote by $\mathbb{P}(J = j)$, is equal to the fraction of customers who are the j-th arrival in their own batch. Thus,

$$\mathbb{P}(J = j) = \frac{\mathbb{P}(K_n \geq j)}{\mathbb{E}[K_n]}.$$

Removing the condition on the customer position in a batch, it follows from some simple calculus that the LST of the sojourn time of an arbitrary customer in the batch-arrival queue is given by:

$$\tilde{V}(s) = \tilde{W}_B(s) \cdot \frac{\mu_2(1 - \tilde{G}(s))}{s\mathbb{E}[K_n]}, \tag{6.6}$$

where it should be observed that the ratio on the r.h.s. of this equation can be interpreted as the LST of the residual service requirement of a batch. It remains to compute $\mathbb{E}[z^{K_n}]$ in order to find $\tilde{G}(s)$, and subsequently $\tilde{W}_B(s)$. This will be done next.

6.3.3.2 The p.g.f. of the batch-size distribution

As remarked in the previous subsection, K_n is the total number of departures from Q_1 during the n-th cycle and depends on the queue length of Q_1 at time Z_n. To compute the p.g.f. of K_n, we assume that Q_1 has a finite queue of size $M-1$. This finite version of Q_1 is denoted by Q_1^M. Later, we will drop the assumption and let M tend to infinity to get our final results.

As we require the batch-size distribution of the arrivals in steady state, we will assume that Q_1^M is in steady state at time $t=0$. The probability that there are j customers in Q_1^M at $t=0$ is denoted by $b_M(j)$ and we let $b_M = (b_M(0), \ldots, b_M(M-1))$ denote the steady-state distribution of the finite Q_1^M. Under the assumption that the infinite-sized Q_1 is stable, $\lim_{M\to\infty} b_M(j) = b(j)$. Moreover, we have that $\sum_{j\geq 0} b(j)z^j = F^{\{-2,1\}}(z)$, where $F^{\{-2,1\}}(z)$ is the p.g.f. of the number of customers at the transition of the $L(t)$ process from -2 to 1 (i.e., at the start of a server visit to Q_1). This latter p.g.f. can explicitly be obtained using the analysis of the unreliable-server model (see Sect. 4.3) and several known results for the M/G/1 vacation queue. We denote the duration of a vacation by B and its residual duration by B^R. The LST of the sojourn time of a customer is denoted by $\tilde{D}_1(s)$ and follows from a decomposition argument [54]:

$$\tilde{D}_1(s) = \tilde{W}_1(s)\tilde{X}_G(s), \tag{6.7}$$

where $\tilde{W}_1(s)$ and $\tilde{X}_G(s)$ denote the LST of the waiting time of a customer (until it is taken into service for the first time) and the generalized service time (including possible service interruptions), respectively. The expression for $\tilde{X}_G(s)$ is given in Eq. (4.2), while the LST for the waiting time is given by [54]:

$$\tilde{W}_1(s) = \tilde{W}_{M/G/1}(s)\left(\kappa_1 + (1-\kappa_1)\tilde{B}^R(s)\right), \tag{6.8}$$

where $\tilde{W}_{M/G/1}(s)$ is the LST of the waiting time in the "corresponding" M/G/1 queue with service times according to LST $\tilde{X}_G(s)$. Besides, the p.g.f. of N_1, the number of customers at Q_1, which we denote by $F_1(\cdot)$, can be expressed as function of $\tilde{D}_1(\cdot)$ using the distributional form of Little's law:

$$F_1(z) = \tilde{D}_1\big(\lambda(1-z)\big). \tag{6.9}$$

Next, let us work towards $F^{\{-2,1\}}(z)$. First, using the argument that a departing server observes the system in steady-state conditioned on the position of the server (for more details, see Sect. 4.4.5), we immediately obtain:

$$E[z^{N_1}|\text{server vacation}] = E[z^{N_1}|\text{server visit}]E[z^{\Psi}], \tag{6.10}$$

where Ψ is the number of arrivals to Q_1 during the age of the vacation. As the age of the vacation is equal in distribution to the residual time of a vacation, it readily follows that $E[z^{\Psi}] = \tilde{B}^R(\lambda(1-z))$. We note that the conditional p.g.f.'s of Eq. (6.10) are also related via $F_1(z)$, i.e.

$$F_1(z) = E[z^{N_1}|\text{server visit}] \cdot \kappa_1 + E[z^{N_1}|\text{server vacation}] \cdot (1-\kappa_1), \qquad (6.11)$$

where $\kappa_1 = 1/\xi_1/\mathbb{E}[C]$, the probability that the server is available at Q_1. Next, by inserting Eq. (6.10) into Eq. (6.11), the conditional p.g.f. of N_1 can be written as:

$$\begin{aligned} E[z^{N_1}|\text{server visit}] &= \frac{F_1(z)}{\kappa_1 + (1-\kappa_1)\tilde{B}^R(\lambda(1-z))} \\ &= \tilde{W}_{M/G/1}\big(\lambda(1-z)\big) \cdot \tilde{X}_G\big(\lambda(1-z)\big), \end{aligned} \qquad (6.12)$$

where the last equality follows by appealing successively to Eqs. (6.9), (6.7) and (6.8). The final expression for $F^{\{-2,1\}}(z)$ follows directly from the definition, the arguments used to obtain Eq. (6.10), and Eq. (6.12), yielding:

$$\begin{aligned} F^{\{-2,1\}}(z) &= E[z^{N_1}|\text{server visit}] \cdot \tilde{B}\big(\lambda(1-z)\big) \\ &= \tilde{W}_{M/G/1}\big(\lambda(1-z)\big) \cdot \tilde{X}_G\big(\lambda(1-z)\big) \cdot \tilde{B}\big(\lambda(1-z)\big). \end{aligned} \qquad (6.13)$$

We continue with the derivation of the batch-size p.g.f. of the finite Q_1 along absorbing Markov chain analysis. To this end, let $(N_1(t), D(t))$ denote the two-dimensional continuous-time Markov process living on the state-space $\{0, 1, \ldots, M-1\} \times \{0, 1, \ldots\} \cup \{(M, 0)\}$, where $N_1(t)$ represents the number of customers in Q_1 at time t and $D(t)$ the number of departures from Q_1 until t. The state $(M, 0)$ is an absorbing state. We denote by **AMC** the absorbing Markov chain driven by $(N_1(t), D(t))$. The state-transition diagram of **AMC** is shown in Figure 6.10. The absorption of **AMC** occurs when the server leaves the queue which happens with rate ξ_1. By setting the probability that the initial state (i.e., at $t = 0$) of **AMC** is $(i, 0)$ to $b_M(i)$, the probability that the absorption of **AMC** occurs from one of the states $\{(i, k) : i = 0, 1, \ldots, M-1\}$ equals the steady-state batch-size distribution $\mathbb{P}(K_n = k),\ k = 0, 1, \ldots$.

Let us order the infinite number of **AMC** states as: $(0, 0), \ldots, (M-1, 0)$, $(0, 1), \ldots, (M-1, 1), \ldots$, and ultimately $(M, 0)$. It is easily seen that the transition-rate matrix $\mathbf{P}$ of **AMC** can be written as:

$$\mathbf{P} = \left(\begin{array}{c|c} \mathbf{Q} & \mathbf{R} \\ \hline \mathbf{0} & 0 \end{array} \right),$$

where $\mathbf{Q}$ is an upper-bidiagonal block matrix of infinite dimension, $\mathbf{0}$ is the row vector with all zero entries and $\mathbf{R} = (\xi_1, \ldots, \xi_1)^T$. The blocks of $\mathbf{Q}$'s diagonal are all equal to $\mathbf{A}$, an M-by-M bidiagonal matrix with diagonal $(-\lambda - \xi_1, -\lambda - \xi_1 - \mu_1, \ldots, -\lambda - \xi_1 - \mu_1, -\xi_1 - \mu_1)$ and upper-diagonal $(\lambda, \ldots, \lambda)$. The blocks of $\mathbf{Q}$'s upper-diagonal are all equal to $\mathbf{B}$, an M-by-M lower-diagonal matrix with lower-diagonal $(\mu_1, \ldots, \mu_1)$.

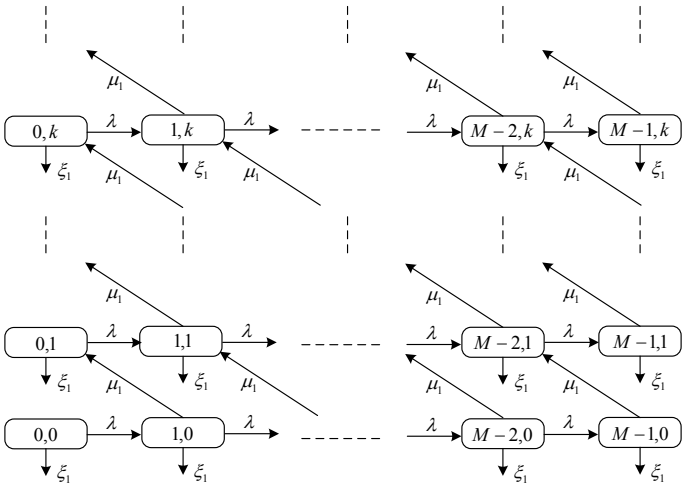

Figure 6.10: Transition state diagram of **AMC**.

Next, we will derive $\mathbb{P}(K_n = k)$ as function of the inverse of $\mathbf{Q}$. Since $\mathbf{Q}$ is an upper-bidiagonal block matrix, $\mathbf{Q}^{-1}$ is an upper-triangular block matrix and is readily obtained as:

$$\mathbf{Q}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} & \mathbf{U}_{0,1} & & \cdots & \\ & \ddots & \ddots & & \\ & & \mathbf{A}^{-1} & \mathbf{U}_{m,m+1} & \cdots \\ & & & \ddots & \ddots \\ & & & & \end{pmatrix}$$

where $\mathbf{U}_{m,l} = \big(-\mathbf{A}^{-1}\mathbf{B}\big)^{l-m}\mathbf{A}^{-1}$ for $m \geq 0$ and $l \geq m$. Note that the matrix $\mathbf{A}$ is invertible since it is upper-bidiagonal with strictly negative diagonal entries.

From the theory of absorbing Markov chains, it is known that given that the initial state vector of **AMC** being b_M, the probability that absorption occurs at one of the states $\{(i,k) : i = 0, 1, \ldots, M-1\}$ is then given by (see, e.g., [43]):

$$P(K_n = k) = -\xi_1 b_M(\mathbf{U}_{0,k})e = -\xi_1 b_M\big(-\mathbf{A}^{-1}\mathbf{B}\big)^k\mathbf{A}^{-1}e,$$

where e denotes the M-dimensional column vector with all entries equal to one. After some algebra, we find:

$$E_M[z^{K_n}] = -\xi_1 b_M(\mathbf{A} + z\mathbf{B})^{-1}e, \quad |z| \leq 1. \tag{6.14}$$

Hence, it remains to find $(\mathbf{A} + z\mathbf{B})^{-1}$.

Now, define $\mathbf{Q}(z) := (\mathbf{A} + z\mathbf{B})$, let $u^T = (1, 0, \ldots, 0)$ and let $v^T = (0, \ldots, 0, 1)$. Observe that $\mathbf{Q}(z) = \mathbf{T}(z) + \mu_1 uu^T + \lambda vv^T$, where $\mathbf{T}(z)$ is an M-by-M tridiagonal Toeplitz matrix, i.e., a matrix with constant entries on each of its diagonals. In this case, the main diagonal entries are equal to $(-\lambda - \mu_1 - \xi_1)$, upper-diagonal entries are equal to λ, and lower-diagonal entries are equal to $z\mu_1$. Let t^*_{ij} denote the (i, j)-entry of $\mathbf{T}^{-1}(z)$. By applying the Sherman-Morrison formula [87, p.76], which allows to express the inverse of a slightly perturbed matrix in terms of the inverse of the unperturbed matrix, we find that the (i, j)-entry of $\mathbf{Q}^{-1}(z)$, $i, j = 1, \ldots, M$, reads:

$$q^*_{ij} = m_{ij} - \lambda \cdot \frac{m_{iM} m_{Mj}}{1 + \lambda m_{MM}}, \quad \text{where} \quad m_{ij} = t^*_{ij} - \mu_1 \cdot \frac{t^*_{i1} t^*_{1j}}{1 + \mu_1 t^*_{11}}. \tag{6.15}$$

The inverse of a tridiagonal Toeplitz matrix has been computed in closed-form (see [27, Sec. 3.1]). Following that same approach, we obtain:

$$t^*_{ij} = \begin{cases} -\dfrac{(r_1(z)^i - r_2(z)^i)(r_1(z)^{M+1-j} - r_2(z)^{M+1-j})}{\lambda(r_1(z) - r_2(z))(r_1(z)^{M+1} - r_2(z)^{M+1})} & , \ i \le j \le M; \\ \dfrac{(r_1(z)^{-j} - r_2(z)^{-j})(r_1(z)^{M+1} r_2(z)^i - r_2(z)^{M+1} r_1(z)^i)}{\lambda(r_1(z) - r_2(z))(r_1(z)^{M+1} - r_2(z)^{M+1})} & , \ j \le i \le M, \end{cases}$$

where $r_1(z)$ and $r_2(z)$ are the roots of the quadratic function $\lambda x^2 - (\lambda + \mu_1 + \xi_1)x + \mu_1 z$ in x, i.e.,

$$r_1(z) = \frac{(\lambda + \mu_1 + \xi_1) - \sqrt{(\lambda + \mu_1 + \xi_1)^2 - 4\lambda\mu_1 z}}{2\lambda}, \tag{6.16}$$

$$r_2(z) = \frac{(\lambda + \mu_1 + \xi_1) + \sqrt{(\lambda + \mu_1 + \xi_1)^2 - 4\lambda\mu_1 z}}{2\lambda}, \tag{6.17}$$

where it is good to notice that $|r_1(z)| < 1 < |r_2(z)|$.

Inserting Eq. (6.15) into Eq. (6.14) yields that:

$$\begin{aligned} E_M[z^{K_n}] &= -\xi_1 \sum_{i=1}^{M} b_M(i-1) \\ &\quad \times \sum_{j=1}^{M} \left[t^*_{ij} - \frac{\mu_1 t^*_{i1} t^*_{1j}}{1 + \mu_1 t^*_{11}} - \frac{\lambda m_{iM}}{1 + \lambda m_{MM}} \left(t_{Mj} - \frac{\mu_1 t^*_{M1} t^*_{1j}}{1 + \mu_1 t^*_{11}} \right) \right]. \end{aligned} \tag{6.18}$$

Thus, it remains to let $M \to \infty$ in Eq. (6.18) to find $\mathbb{E}[z^{K_n}]$. Several of these

limit expressions are quite straightforward, viz.,

$$\begin{aligned}
\lim_{M\to\infty} t^*_{M,M-j} &= -\frac{1}{\lambda}\cdot\frac{r_1(z)^j}{r_2(z)},\\
\lim_{M\to\infty} m_{M-i,M} &= \lim_{M\to\infty} t^*_{M-i,M} = -\frac{1}{\lambda}\cdot r_2(z)^{-(i+1)},\\
\lim_{M\to\infty} t^*_{1j} &= -\frac{1}{\lambda}\cdot r_2(z)^{-j},\\
\lim_{M\to\infty} t^*_{i1} &= -\frac{1}{\lambda}\cdot\frac{r_1(z)^{i-1}}{r_2(z)},
\end{aligned}$$

while it requires some easy but tedious calculus to shows that the following limit vanishes:

$$\lim_{M\to\infty}\xi_1\sum_{i=1}^{M} b_M(i-1)\sum_{j=1}^{M}\frac{\lambda m_{iM}}{1+\lambda m_{MM}}\left(t^*_{Mj}-\frac{\mu_1 t^*_{M1}t^*_{1j}}{1+\mu_1 t^*_{11}}\right).$$

Putting everything together, it follows that:

$$\mathbb{E}[z^{K_n}] = \frac{\xi_1}{\lambda(1-r_1(z))(r_2(z)-1)}\left[1+\mu_1\cdot\frac{1-z}{\lambda r_2(z)-\mu_1}\cdot F^{\{-2,1\}}(r_1(z))\right], \quad (6.19)$$

where $F^{\{-2,1\}}(.)$, $r_1(z)$ and $r_2(z)$ are given in Eqs. (6.13), (6.16) and (6.17), respectively. Inserting $z=\mu_2/(\mu_2+s)$ into Eq. (6.19) readily provides us with $\tilde{G}(s)$, the LST of the service requirement of a batch (see Eq. (6.5)). Finally, the expression for $\tilde{G}(s)$ renders also the expressions for $\tilde{W}_B(s)$ and $\tilde{V}(s)$ explicit.

6.3.3.3 Sojourn time at the relay queue

Let $\{H(t): t\geq 0\}$ denote the residual sojourn time of a customer in the batch-arrival queue if the server would be continuously working at Q_2 from time t onwards. In particular, let $H(0)$ denote the (original) sojourn time of a customer in the batch-arrival queue. It follows that $H(t)$ decreases at rate 1 when $L(t)=0$ (i.e., the server visits Q_1) and $H(t)$ is constant when $L(t)\in\{-2,-1,1\}$ at time t. Let Y denote the number of service interruptions during the sojourn time of a customer. Figure 6.11 displays a sample path of the evolution of $H(t)$ as function of t.

The visit periods have an exponential length with rate ξ_2. Now, given that $H_0=v$, the number of interruptions has a Poisson distribution, i.e.,

$$\mathbb{E}[z^Y|H_0=v] = e^{-\xi_2(1-z)v}.$$

We denote the duration of these interruptions by $(\Xi^n)_{n\geq 1}$ which are i.i.d. copies of the generic random variable $\Xi := C_{2,1}+E_{L_1}+C_{1,2}$. The time it takes before $H(t)$ actually starts decreasing after time 0, which we denote by Ξ^*, satisfies $\Xi^* = E^R_{L_1}+C_{1,2}$, where $E^R_{L_1}$ is the residual of the visit time E_{L_1}. Note that $E^R_{L_1}$ and E_{L_1} are identically

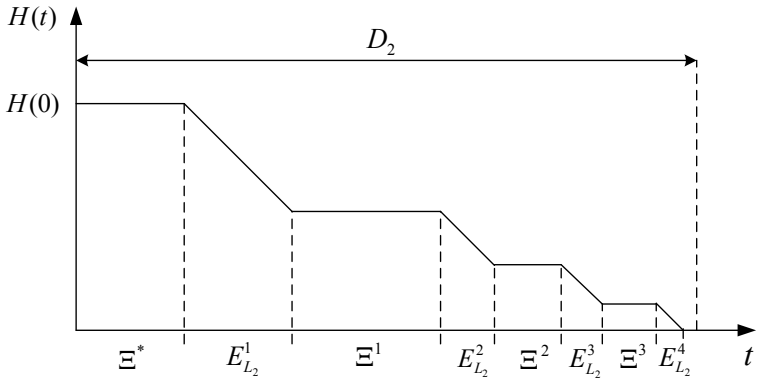

Figure 6.11: Evolution of $H(t)$ as function of t.

distributed due to the memoryless property of the exponential distribution. From Fig. 6.11, it is easily seen that $D_2 = \Xi^* + H_0 + \sum_{i=1}^{Y} \Xi_i$. By conditioning on H_0 and Y, we find for the LST of D_2,

$$\tilde{D}_2(s) = \mathbb{E}\big[e^{-s\Xi^*}\big]\mathbb{E}\big[e^{-s(\sum_{i=1}^{Y}\Xi_i + H_0)}\big] = \mathbb{E}\big[e^{-s\Xi^*}\big]\mathbb{E}\big[e^{-sH_0}e^{-\xi_2 H_0(1-\tilde{\Xi}(s))}\big],$$

where $\tilde{\Xi}(s) = \frac{\xi_1}{\xi_1+s} \cdot \tilde{C}_{1,2}(s) \cdot \tilde{C}_{2,1}(s)$. Since H_0 is the sojourn time in the batch-arrival queue (see, Eq. (6.6)), we readily obtain:

$$\tilde{D}_2(s) \quad = \quad \frac{\xi_1}{\xi_1 + s} \cdot \tilde{C}_{1,2}(s) \cdot \tilde{W}_B\big(\Delta(s)\big) \cdot \frac{\mu_2}{\mathbb{E}[K_n]} \cdot \frac{1 - \tilde{G}\big(\Delta(s)\big)}{\Delta(s)}, \tag{6.20}$$

where $\Delta(s) := s + \xi_2(1 - \tilde{\Xi}(s))$. The mean sojourn time at the relay queue then reads:

$$\mathbb{E}[D_2] = -\frac{d}{ds}\tilde{D}_2(s) \mid_{s=0} . \tag{6.21}$$

6.3.4 Numerical evaluation

The evaluation of the model will be done in three parts. First, we will extensively validate the accuracy of the approximation. Second, we consider the impact of the switch-over time distribution on the mean sojourn time. Notice that the switch-over times determine to a large extent the time between consecutive server visits. Finally, we study the optimization of the end-to-end delay in the two-hop network by adjusting the visit time parameters for a fixed mean cycle length. Throughout this section, service times are assumed exponentially distributed and the switch-over times, $C_{1,2}$ and $C_{2,1}$, are assumed identically distributed with mean $c_{1,2} = c_{2,1}$.

6.3.4.1 Model validation

We validate the approximate model developed in Sect. 6.3.3 for the mean sojourn time at Q_2 for the case of exponentially distributed switch-over times. Recall that the key approximation step in this model was that $(K_n)_{n\geq 1}$, the batch sizes in the batch-arrival queue in the n-th cycle (see, e.g., (6.4)), were assumed to be mutually independent and also independent of $E_{L_2}^m$, $m < n$, the durations of the preceding visits to Q_2. The validation will be done essentially by comparing the results with those of the exact model in Sect. 6.3.2. However, we note that due to the state-space expansion the computation time for the exact joint queue-length probabilities, and thus also the mean sojourn time, may grow large for specific model parameters. Therefore, in these specific cases, we use a simulation program (developed in C++) to determine the sojourn time in Q_2. We have run the simulation program such that the upper and lower bounds of the 95%-confidence interval for the mean sojourn time are always within 1% of the estimation of the mean.

Let $\mathbb{E}[D_2^{exa}]$ (resp. $\mathbb{E}[D_2^{app}]$) denote the mean sojourn time in Q_2 using the exact (resp. approximate) model as expressed in Eq. (6.3) (resp. in Eq. (6.21)). Let R_2 denote the absolute relative difference between the approximate and exact mean sojourn time in Q_2, i.e.,

$$R_2 := \left| 1 - \frac{\mathbb{E}[D_2^{app}]}{\mathbb{E}[D_2^{exa}]} \right|.$$

Further, we note that, by exploiting the exponentiality of the service times and the symmetry of the switch-over times, the *generalized* load ρ_i^G at Q_i (i.e., the load as fraction of the total load that can be sustained by a queue) can be written as:

$$\rho_i^G := \frac{\lambda}{\mu_i} \cdot \xi_i \cdot \left(\frac{\xi_1 + \xi_2}{\xi_1 \xi_2} + 2 \cdot c_{1,2} \right), \; i = 1, 2. \tag{6.22}$$

We note that in this section load and generalized load are used interchangeably, but will always refer to the definition in Eq. (6.22).

The first validation results are depicted in Fig. 6.12 for the scenario with $\mu_2 = 1$, $c_{1,2} = 10$ and $\xi_1 = \xi_2 = 0.1$. This figure shows the relative approximation error, R_2, as function of the load at Q_2 for different values of the load at Q_1. The value of ρ_2^G is adjusted here by changing the value of λ, while the choice for μ_1 follows immediately from ρ_1^G and ρ_2^G. Though, we have also verified that changing ρ_2^G as function of μ_2 (and keeping λ fixed) would give a similar picture. Let us first explain the behavior of R_2 as function of ρ_2^G for a given value of ρ_1^G. For light load at Q_2, the server will leave with high probability (i.e., $1 - \rho_2^G$) an empty queue behind at the end of a visit to Q_2. In the limit case of $\rho_2^G \to 0$, customers served at Q_2 need only to wait for customers that arrived in the same batch of size K_n. As a result, the correlation in $(K_n)_{n\geq 1}$ plays no role and the approximation is exact. When the load at Q_2 is increased, the approximation assumption becomes more important and thus the error increases more than linearly in ρ_2^G. Besides, Fig. 6.12 uncovers the impact

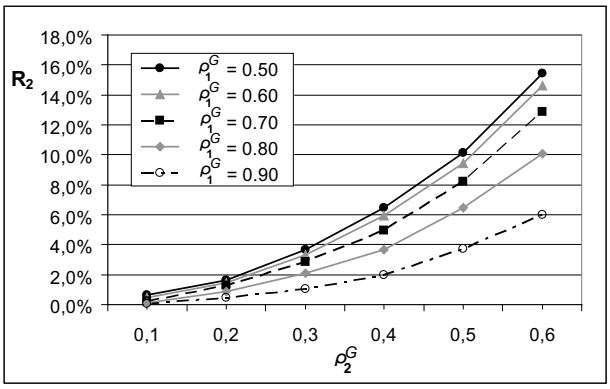

Figure 6.12: R_2 as function of ρ_2^G for different values of ρ_1^G with $\lambda = 0.01$, $\xi_1 = \xi_2 = 0.1$, and $c_{1,2} = 10$.

of ρ_1^G on R_2 for a given value of ρ_2^G. It is shown that the relative error R_2 decreases in ρ_1^G and this can be explained by considering the limit case $\rho_1^G \to 1$. In this situation, the server works almost continuously during its visits to Q_1 and consequently the batch sizes $(K_n)_{n\geq 1}$ will tend to i.i.d. random variables with a geometric distribution living on $\{0, 1, \ldots\}$ with success probability $1 - \tilde{X}_i(\xi_i)$. Thus, the approximation is exact in this limit case and then it deteriorates when ρ_1^G is decreased.

We have witnessed in Fig. 6.12 that the accuracy of the approximation increases in ρ_1^G (for given ρ_2^G) and decreases in ρ_2^G (for given ρ_1^G). Let us next observe what happens when ρ_1^G and ρ_2^G are changed simultaneously as function of λ. Specifically, we consider the relative approximation error when the load at Q_1 and Q_2 are equal ($\rho_1^G = \rho_2^G =: \rho^G$), and with $\mu_1 = \mu_2 = 1$ and $\xi_1 = \xi_2 = 0.1$. The results for R_2 as function of ρ^G are displayed in Figure 6.13 for different values of $c_{1,2}$. Figure 6.13 shows that the approximation error increases in ρ^G (and thus in λ). Apparently, the negative impact of an increase in ρ_2^G on R_2 is stronger than the positive influence of an equal increase in ρ_1^G. Besides, the figure indicates that for a given value of $\rho_1^G = \rho_2^G$ the error R_2 decreases with $c_{1,2}$ (e.g., for $\rho^G = 0.5$, $R_2 = 15\%$ when $c_{1,2} = 1$, while $R_2 = 8\%$ when $c_{1,2} = 20$). This is due to the fact that (for a given load) an increase of $c_{1,2}$ leads to a decrease of the arrival rate λ and thus a more accurate approximation.

Finally, we want to highlight a qualitative observation of our model validation. In Fig. 6.14, we plot the mean sojourn time in Q_2 as function of $\rho^G (= \rho_1^G = \rho_2^G)$. Again, this is done by varying the arrival rate λ for the symmetric scenario with $\mu_1 = \mu_2 = 1$ and $\xi_1 = \xi_2 = 0.1$. The figure shows that the approximation provides always an upper bound for $\mathbb{E}[D_2]$. This observation is consistent with the result in [12] which states that in the correlated $M/G/1$ queue a positive correlation between the service requirement and the last interarrival time reduces the mean sojourn time. We note

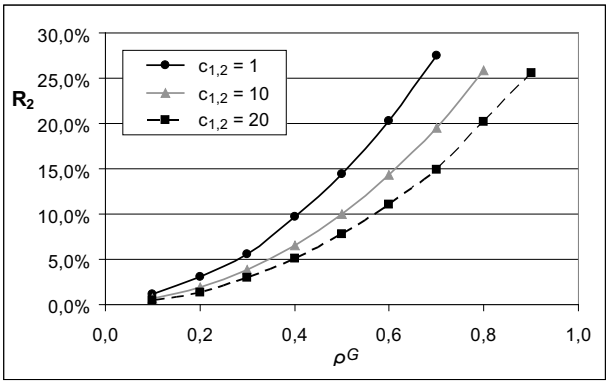

Figure 6.13: R_2 as function of λ for $\mu_1 = \mu_2 = 1$ and $\xi_1 = \xi_2 = 0.1$.

that in our model K_n (the batch size) and the last interarrival time are positively correlated, i.e., an increase of the last interarrival time induces stochastically an increase of K_n (and of the total service requirement of a batch). Thus, since we neglect these correlations in our approximation, the approximation should provide indeed an overestimation of the exact value.

We conclude that the approximate model has the following properties regarding the mean sojourn time:

- It is accurate for *low and moderate* load at Q_1 and Q_2;
- It is accurate for *high* load at Q_1 and *moderate* load at Q_2;
- It gives an upper bound for the sojourn time at Q_2.

6.3.4.2 Impact of the switch-over times distribution on the end-to-end sojourn time

We note that in the analysis the switch-over times were assumed to be arbitrary. In this section, we will study the impact of the distribution of the switch-over times on the end-to-end sojourn time of a customer. This is done by considering three distributions for the switch-over times, viz., a two-phase hyper-exponential, a two-phase Coxian and a Weibull distribution. The parameters of the distributions are chosen such that the first two moments are identical.

Let us denote by $SCOV_s$ and $SCOV_i$ the squared coefficient of variation of the switch-over times and the intervisit times, respectively. That is, $SCOV_s := Var(C_{1,2})/(c_{1,2})^2$, and of $SCOV_i := Var(C_{1,2} + C_{2,1} + E^{L_2})/(2c_{1,2} + 1/\xi_2)^2$. The intervisit times, i.e., the time between two consecutive visits of the server, are included, since in a more practical setting (e.g., for opportunistic networking) this is an important measure.

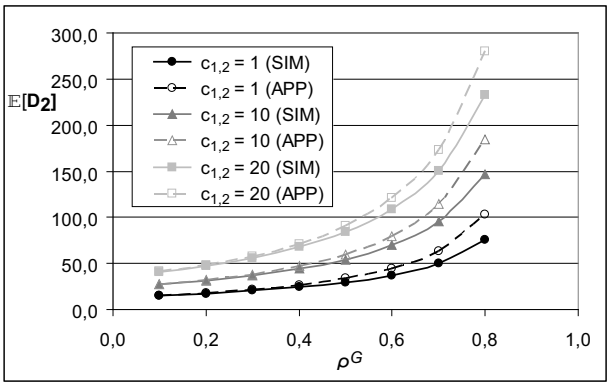

Figure 6.14: Mean sojourn time in Q_2 as function of ρ^G for $\mu_1 = \mu_2 = 1$ and $\xi_1 = \xi_2 = 0.1$.

In Table 6.1, we display the mean sojourn time as function of $SCOV_s$ and of $SCOV_i$ for $\lambda = 0.01$, $\mu_1 = \mu_2 = 1$, $\xi_1 = \xi_2 = 0.1$, and $c_{1,2} = 10$. This is done by using the approximate and exact models for the hyper-exponential and the Coxian distribution. For the Weibull distribution, we used the simulation program, since its LST it is not known in closed form. Observe that in Table 6.1 the mean sojourn time is almost equal for the three different distributions. This suggests that the mean end-to-end sojourn time depends on the distribution of the switch-over times only through their first two moments. We have performed other experiments for various parameter settings which all approved this claim. We note that this conjecture is well known for single-server queues with vacations and also for several polling models, but it is novel in the context of single-server tandem models operating under the pure time-limited discipline.

$SCOV_s$	1	5	10	15	20	30
$SCOV_i$	0.33	1.22	2.33	3.44	4.55	6.78
	Hyper-exponential distribution					
$\mathbb{E}[D^{app}]$	45.07	55.85	69.28	82.69	96.06	122.7
$\mathbb{E}[D^{exa}]$	45.02	55.8	69.25	82.66	95.96	121.4
	Coxian distribution					
$\mathbb{E}[D^{app}]$	45.07	55.85	69.31	82.74	96.15	122.9
$\mathbb{E}[D^{exa}]$	45.02	55.81	69.27	82.71	96.12	122.6
	Weibull distribution					
$\mathbb{E}[D^{sim}]$	45.01	55.85	69.25	82.54	95.88	121.9

Table 6.1: Mean end-to-end sojourn time as function of $SCOV_s$ and $SCOV_i$.

6.3.4.3 Insight in the optimal end-to-end sojourn time

In this section, we study the evolution of ξ_2^{opt}, the optimal value of ξ_2 that yields the minimum value of the mean end-to-end sojourn time in the two-hop tandem network. This will be done under the constraints of zero switch-over time, i.e., $c_{1,2} = 0$, and constant mean cycle length, i.e., $\mathbb{E}[C] = 1/\xi_1 + 1/\xi_2$ is constant. Notice that under these constraints ξ_1 decreases in ξ_2 and vice versa. Since the mean sojourn time at Q_1 decreases with ξ_2 and the mean sojourn time at Q_2 increases with ξ_2, ξ_2^{opt} exists and it is unique. We should note that manually adjusting the parameters ξ_1 and ξ_2 appears incompatible with the autonomous behavior of the server at first sight. However, in practice the adjustment of ξ_1 and ξ_2 can perfectly be done by controlling the transmission power of the stations.

The optimal value ξ_2^{opt} can be computed by applying the numerical optimization package of MAPLE to the approximate mean sojourn time in (6.20). This value was validated afterwards by verifying that the mean sojourn time using the exact model for $\xi_2 = \xi_2^{opt}$ is a local minimum inside $[\xi_2^{opt} - 10^{-3}, \xi_2^{opt} + 10^{-3}]$. For the symmetric case, i.e., $\mu_1 = \mu_2$, it is found that $\xi_2^{opt} = \xi_1 = 2/\mathbb{E}[C]$. For the asymmetric case, i.e., $\mu_1 > \mu_2 = 1$, Table 6.2 presents ξ_2^{opt} as function of μ_1 for $\lambda = 0.10$ and $\mathbb{E}[C] = 10$. Observe that in this case ξ_2^{opt} is smaller than $2/\mathbb{E}[C]$ and that the difference increases with μ_1. Table 6.3 displays ξ_2^{opt} as function of μ_2 for $\mu_2 > \mu_1 = 1$ and $\mathbb{E}[C] = 10$. Contrary to the previous case, we observe that ξ_2^{opt} is greater than $2/\mathbb{E}[C]$. More surprisingly, the values of ξ_2^{opt} and ξ_1 for the two cases are almost exactly exchanged. This result is not anticipated as the arrival processes at the source and the relay queue are essentially different. Finally, we note that the mean end-to-end sojourn times differ significantly in these mirrored cases.

μ_1	1.1	2	3	6	11	16
ξ_2^{opt}	0.197	0.177	0.168	0.158	0.152	0.150
ξ_1	0.203	0.230	0.247	0.272	0.292	0.300
ρ_1^G	0.185	0.115	0.082	0.045	0.027	0.019
ρ_2^G	0.197	0.177	0.168	0.158	0.152	0.150

Table 6.2: ξ_2^{opt} as function of μ_1 for $\mu_2 = 1$, $\lambda = 0.10$, and $\mathbb{E}[C] = 10$.

μ_2	1.1	2	3	6	11	16
ξ_2^{opt}	0.205	0.232	0.248	0.275	0.295	0.304
ξ_1	0.195	0.176	0.168	0.157	0.151	0.149
ρ_1^G	0.195	0.176	0.168	0.157	0.151	0.149
ρ_2^G	0.186	0.116	0.083	0.046	0.027	0.019

Table 6.3: ξ_2^{opt} as function of μ_2 for $\mu_1 = 1$, $\lambda = 0.10$, and $\mathbb{E}[C] = 10$.

It is nice that the optimal value for ξ_2 can be obtained using an optimization

package. However, for practical purposes, it might be more valuable to have a simple rule that provides a value for ξ_2 which yields a mean end-to-end sojourn time close to the optimum. Therefore, we will discuss two alternative, heuristic optimization approaches. The first heuristic selects the values of ξ_1 and ξ_2 such that the load is balanced at both queues, i.e., $\rho_1^G = \rho_2^G$. This gives:

$$\xi_i = (\mu_1 + \mu_2)/(\mu_{3-i} \cdot \mathbb{E}[C]),\ i = 1, 2.$$

The second heuristic chooses ξ_1 and ξ_2 based on the analysis of a tandem model of two M/M/1 queues with shared service capacity. This means that the servers at both queues are always present, but serving at rate ν at Q_1 and at rate $1 - \nu$ at Q_2. Then, the optimal ν, say ν^*, is the one that minimizes the mean end-to-end sojourn time in such a tandem model, which we denote by $\mathbb{E}[D]^{M/M/1}$ and equals simply

$$\mathbb{E}[D]^{M/M/1} = 1/(\mu_1\nu - \lambda) + 1/(\mu_2(1 - \nu) - \lambda).$$

We choose the ratio ξ_1/ξ_2 equal to $(1 - \nu^*)/\nu^*$, such that the fraction of time that the server is at Q_1 in our model equals the optimal rate ν^* in the M/M/1 tandem model.

In Tables 6.4 and 6.5, we present the results of this comparison. Here, ξ_2^{opt}, ξ_2^{LB} and $\xi_2^{M/M/1}$ refer to the choice of ξ_2 in the optimal case, in the load balancing heuristic, and in the M/M/1 tandem heuristic, respectively. Further, we present the relative differences in mean sojourn time using the two heuristics (denoted by ϵ^{LB} and $\epsilon^{M/M/1}$) with respect to the optimal mean sojourn time, $\mathbb{E}[D]^{opt}$. In Table 6.4, we study the performance of those heuristics when μ_1 is increased for the case $\mu_2 = 1.0,\ \lambda = 0.1$ and $\mathbb{E}[C] = 10$. We note that for the symmetric case, $\mu_1 = \mu_2$, the heuristics would also give the optimal solution $\xi_1 = \xi_2$. The performance using load balancing worsens rapidly when μ_1 is increased. Also the M/M/1 tandem heuristic deviates from the optimum, but the relative differences remain small. In Table 6.5, we investigate the performance of the heuristics when the mean cycle time is varied for the case $\mu_1 = 6.0,\ \mu_2 = 1.0$ and $\lambda = 0.1$. The results show that the relative error when using load balancing is almost insensitive to $\mathbb{E}[C]$. We note that in the limit case of the cycle time tending to zero our tandem model approaches the tandem model of two M/M/1 queues. Hence, in this case the M/M/1 tandem heuristic is optimal. This explains why the relative error increases in $\mathbb{E}[C]$. However, notice that the relative error $\epsilon^{M/M/1}$ is still very small for $\mathbb{E}[C] = 20$.

We can conclude that balancing the load is not a good solution for end-to-end sojourn time optimization unless $\mu_1 \approx \mu_2$. However, using an optimization heuristic based on a simple tandem model of two M/M/1 queues will give nearly optimal results for the mean sojourn time under a wide variety of parameter settings.

6.3.5 Concluding remarks on the sojourn-time approximation

We have analyzed here a network consisting of a fixed source station, a fixed destination station, and a mobile relay station. This two-queue tandem model can be seen as

μ_1	1.1	2	3	6	11	16
ξ_2^{opt}	0.194	0.174	0.166	0.156	0.150	0.148
ξ_2^{LB}	0.190	0.150	0.133	0.117	0.109	0.106
$\xi_2^{M/M/1}$	0.195	0.167	0.154	0.138	0.128	0.123
$\mathbb{E}[D]^{opt}$	14.47	12.82	12.14	11.42	11.08	10.94
ϵ^{LB} (%)	<0.1	3.9	8.6	17.2	23.6	26.3
$\epsilon^{M/M/1}$ (%)	<0.1	0.4	0.9	2.3	3.8	4.7

Table 6.4: Comparison of ξ_2 and $\mathbb{E}[D]$ for different optimization approaches for $\mu_2 = 1.0$, $\lambda = 0.1$, and $\mathbb{E}[C] = 10$.

$\mathbb{E}[C]$	1	2	5	10	20
ξ_2^{opt}	1.408	0.719	0.300	0.156	0.080
ξ_2^{LB}	1.167	0.583	0.233	0.117	0.058
$\xi_2^{M/M/1}$	1.375	0.687	0.275	0.137	0.069
$\mathbb{E}[D]^{opt}$	3.15	4.07	6.83	11.42	20.58
ϵ^{LB} (%)	17.3	17.1	17.1	17.2	17.7
$\epsilon^{M/M/1}$(%)	0.2	0.6	1.4	2.3	3.0

Table 6.5: Comparison of ξ_2 and $\mathbb{E}[D]$ for different optimization approaches for $\mu_1 = 6.0$, $\mu_2 = 1.0$, and $\lambda = 0.1$.

a first-step model towards developing analytical models for quantifying the delay in an opportunistic network. More specifically, we proposed an analytical approximation for the mean delay at the relay station. The approximate model has extensively been validated for the mean delay and is fairly accurate. Numerical results on the mean end-to-end delay show that the inter-contact time distribution impacts this metric only through its first two moments. Moreover, load balancing appears not an effective tool for delay optimization under power control, while a simple M/M/1 tandem queue heuristic is nearly optimal.

We emphasize that the sojourn time analysis carried out in this section can readily be extended to model for instance multiple relay stations with single-copy packet approach [96], h-hop ($h \geq 2$) relay routing schemes, or networks with mobile source and destination station. Moreover, the key ideas can be applied to approximate the sojourn time in longer tandem queues with multiple mobile relay stations [H1].

6.4 Concluding remarks

Approximations are often considered as an inferior mathematical tool. After all, explicit exact expressions for system performance appear more elegant and more insightful. Unfortunately, many real-world systems are so complex that properly

describing such systems by tractable mathematical models becomes infeasible. In such situations, approximations may offer a fruitful alternative.

In this chapter, we have presented two approximations for complex multi-queue systems. The first, perhaps more classical, product-form approximation has been used to assess the conditional queue-length distribution in the basic polling system which operates under the pure time-limited discipline. This specific service discipline renders the correlation between the different queues rather small, so that indeed assuming that the queues operate in isolation appears quite reasonable. We have seen that there exists a wide range of parameter settings for which the approximation indeed works quite well, yet the approximation appears still not expedient in scenarios with extremely heavily loaded queues or extremely large visit times. The second approximation has been introduced to efficiently assess the mean sojourn time for a simple opportunistic networking model. By applying techniques from absorbing Markov chain analysis, we have obtained a closed-form approximate expression for the Laplace-Stieltjes Transform which allows for fast numerical evaluation of the sojourn time. The approximation has extensively been validated for the mean sojourn time at the relay station and appears highly accurate. Moreover, the approximation allows to consider also second and higher moments for the sojourn time at the relay station which cannot be found from the exact analysis. Finally, we have performed additional experiments which are especially interesting from a practical viewpoint.

Part III

Multi-server polling models

CHAPTER

7

Recursive analysis for the basic polling system

7.1 Introduction

In the previous part of this thesis, we have studied performance models for mobile ad hoc networks with at most a single data transmission at a time. This scenario occurs typically in networks which are fully connected (i.e., each station is aware of the presence of the other stations) or networks which are small (see, e.g., Sect 6.3). Conversely, ad hoc networks that are not fully connected or even (temporarily) disconnected provide opportunities to sustain multiple transmissions to be successful at the same time. This is advantageous from a performance perspective as it will increase the network capacity. For instance, it was shown for asymptotically large networks [45] that increasing the number of transmission opportunities leads to a higher capacity. However, the specific 2-hop relaying scheme and simple mobility model adopted in that work resulted in unbounded end-to-end transfer delays. Alternative relay schemes and mobility models have then been considered in the same asymptotic setting to study the trade-off between delay and capacity in more detail (see, e.g., [5, 92]). Also for finite-size ad hoc networks (without mobility), multiple simultaneous transmissions will increase the network capacity as demonstrated in Chapter 2. Moreover, the framework presented in that chapter allows also for

studying the (mean) transfer delays in a static setting. However, little is known for the transfer delays when mobility is explicitly included in such finite-size ad hoc networks.

We note that the multiple transmissions correspond in fact to multiple wireless communication links that can be active simultaneously. Together with the randomness of the network topology arising from the mobility of stations, the basic multi-server polling model comes up as a natural model to study the buffer and delay performance in such networks. In Chapter 4, we have illustrated how a polling model with a single server operating under the pure exponential time-limited service discipline can be analyzed. Particularly, we have focussed on the evolution of the queue-length process during a visit by conditioning on intermediate events during a visit. This recursive approach may essentially also be applied to study polling models with multiple servers. The key difference between single and multi-server polling models is that the visit process of the servers becomes multi-dimensional meaning that also multiple queues can be visited simultaneously. Basically, this leads to two different server strategies, viz.,

- coupled servers; i.e., the servers are coupled and move as a group along the queues;
- individual servers; i.e, the servers move individually through the system.

We note that in the individual-server case if a server polls a queue which is already being served by a number of servers, then the server may decide to jump over the queue and move to a next queue.

Both strategies have been studied in the literature along different analytical techniques. Approximation techniques have been used to study the coupled-servers strategy [20] and individual-servers strategy [6, 15, 75]. Also, exact analytical results can be found for both strategies [14, 73, 103]. Most of these studies considered the exhaustive and the gated disciplines. An exception to this rule is [103] which considers the pure time-limited discipline for an infinite-server polling model.

Here, we will also consider this latter time-limited discipline but then for a polling model with a finite number of servers. Specifically, we study the basic multi-server polling model extended with customer routing. We present a unified framework that supports the analysis of both the coupled-servers and the individual-servers case. Although the framework allows for studying any finite number of servers, we will focus here on the two-server case. The analytical approach builds on the single-server analysis (see Chapter 4). The key elements of the analysis are the relations for the queue-length evolution during a visit (cf. Eq. (1.2)). These relations are derived separately for the case of the two servers being at the same queue and the servers being at different queues. This is done in a recursive fashion by conditioning on specific events during a visit. Together, these relations form a system of equations from which the queue-length distribution at these embedded epochs can be solved iteratively. It should be emphasized though that such an iterative approach is generally not feasible, but it works here due to the pure time-limited discipline. Analogously to the

single-server case, the steady-state queue-length distribution can be obtained from the distribution at the embedded epochs. To illustrate the applicability of the analysis, we present two examples. The first example considers a more traditional cyclic polling model with two independent servers. In the second example, we demonstrate how the analysis can be used to study a model for wireless multi-hop communication in the presence of mobile stations. Finally, we delineate how nonzero switch-over times, a larger number of servers, and a limit on the number of servers per queue can be incorporated in our analytical framework.

The remainder of this chapter is organized as follows. First, we give the model description in Sect. 7.2. Next, we state the stability condition and present the queue-length analysis in Sect. 7.3. The examples are given in Sect. 7.4. We conclude this chapter in Sect. 7.5 with a discussion.

7.2 Model description

Let us consider the basic multi-server polling model of $M \geq 2$ queues with $K = 2$ servers as described in Sect. 1.3.5. Thus, customers arrive according to a Poisson process with rate λ_i and require an exponentially distributed amount of service with mean $1/\mu_i$. We allow for customer routing conform the description in Sect. 4.5.1. In particular, we denote by $r^i(\mathbf{z})$ the p.g.f. of the number of arrivals to all queues generated by a single departing customer at Q_i. The switch-over times are assumed to be of zero duration. The service discipline at all queues for both servers is the pure exponential time-limited discipline. Conform the description in Sect. 1.3.4, the server-location process is a two-dimensional process on $\mathcal{L}_1 \times \mathcal{L}_2$ driven by the transition matrix $S = \{s_{l,j}\}_{l,j \in \mathcal{L}_1 \times \mathcal{L}_2}$ with $\mathcal{L}_1, \mathcal{L}_2 \in \{1, \ldots, M\}$. Thus, we let $\xi^l,\ l \in \mathcal{L}_1 \times \mathcal{L}_2$, denote the (total) departure intensity of the servers in server-location state l. For the individual-server case, the per-server departure intensities will be denoted by $\xi_i,\ i = 1, \ldots, M$, so that $\xi^l = \xi_{l_1} + \xi_{l_2},\ l \in \mathcal{L}_1 \times \mathcal{L}_2,\ l_1 \in \mathcal{L}_1,\ l_2 \in \mathcal{L}_2$. Let us denote by $\tau_l,\ l \in \mathcal{L}_1 \times \mathcal{L}_2$, the stationary distribution of the Markov jump chain that is driven by the server-location process. Further, we will denote throughout by N_i the number of customers at Q_i.

For ease of presentation, we have assumed that switch-over times have zero duration and that the number of servers is two. At the end of this chapter, we will indicate how nonzero switch-over times and additional servers may be included in the analysis.

7.3 Analysis

In Sect. 7.3.1, we discuss the stability condition of the two-server polling system. Then, in Sect. 7.3.2, we present our approach to obtain a system of equations for the queue-length distribution at the entrance and departure instants of a server-location state. The essential step here is to recursively relate the queue length at

specific embedded time points, viz., at entrance epochs to a server-location state and at departure epochs to this state. These relations are obtained by conditioning on intermediate events during a specific server-location state. Finally, in Sect. 7.3.3, we discuss how the steady-state joint queue-length distribution may be determined from this set of relations in an iterative manner for this two-server polling system.

7.3.1 Stability condition

The polling system is assumed stable if in steady state each customer in the system can be served in a finite period of time. We must consider stability on a per-queue basis as service capacity cannot be exchanged between the queues.

For an individual queue to be stable, it is well-known that the average amount of work arriving per time unit to the queue must be smaller than the average number of servers present. Let us first denote by γ_i the aggregate customer arrival rate to Q_i. This quantity follows from the set of traffic equations, i.e., $\gamma_i = \lambda_i + \sum_{j=1}^{N} \gamma_j r_{ji},\ i = 1, \ldots, M$, where the r_{ji} refer to the customer routing probabilities. Hence, the load ρ_i at Q_i, i.e., average amount of work arriving per time unit, is given by:

$$\rho_i = \frac{\gamma_i}{\mu_i},\ i = 1, \ldots, M.$$

Next, let us denote by $\nu_i,\ i = 1, \ldots, M$, the mean number of servers present at Q_i. Then, given τ_l, we may write:

$$\nu_i = \sum_{l \in \mathcal{L}_1 \times \mathcal{L}_2} \omega_l \cdot (\mathbf{1}_{\{l_1 = i\}} + \mathbf{1}_{\{l_2 = i\}}),\ i = 1, \ldots, M,$$

where $\mathbf{1}_{\{\cdot\}}$ is the indicator function and ω_l, the average fraction of time that the servers are in state l, is given by:

$$\omega_l = \frac{\tau_l / \xi^l}{\sum_k \tau_k / \xi^k}. \tag{7.1}$$

The necessary and sufficient condition for stability of Q_i then reads:

$$Q_i \text{ is stable } \iff \rho_i < \nu_i,\ i = 1, \ldots, M.$$

We say that the system is stable if and only if all the queues in the system are stable.

7.3.2 Queue-length relations for the embedded chain

The queue-length relations will be established separately for the case that the servers are at different queues and for the case that the servers are at the same queue. We recursively relate the number of customers present at the departure from the server-location state to the number present upon entering this state, cf. the approach in Sect. 4.4.4. To this end, we mark three type of events, viz., a jump out of the current

server-location state, a customer arrival at a queue with at least one idle server, and a service completion event (at any queue). Notice that the latter two events may coincide in case a customer is routed to another queue upon service completion. Further, we define for $k \geq 1$, $\mathbf{N}_k^l := (N_{k,1}^l, \ldots, N_{k,M}^l)$, where $N_{k,i}^l$, $i = 1, \ldots, M$, denotes the number of customers at Q_i just after the epoch of the k-th marked event at server-location state l and we let $\mathbf{N}_0^l$ refer to the number of customers present upon entering server-location state l. Finally, we denote by the random variable $\Upsilon_l \geq 1$ the number of marked events that occurs during a residence at server-location state l. It follows from our assumptions on the exponential interarrival and service times that the sequence $\{\mathbf{N}_k^l\}_{k=0}^{\infty}$ is a Markov chain.

Similar as for the single-server case, we define the queue-length p.g.f.'s $\phi_k^l(\mathbf{z})$ and $\phi_k^{s,l}(\mathbf{z})$. That is, we let $\phi_k^l(\mathbf{z})$ be the joint p.g.f. of the number of customers at all queues at the k-th marked event epoch at server-location state l and marked event k is *not* the final marked event during the residence (i.e., marked event $k+1$ will occur), while for $\phi_k^{s,l}(\mathbf{z})$ event k is indeed the final marked event (i.e., marked event k is a jump out of the current server-location state, and marked event $k+1$ will not occur). Thus, formally we define for $k \geq 1$,

$$\begin{aligned} \phi_k^l(\mathbf{z}) &:= \mathbb{E}[\mathbf{z}^{\mathbf{N}_k^l} \mathbf{1}_{\{\Upsilon_l > k\}}], \\ \phi_k^{s,l}(\mathbf{z}) &:= \mathbb{E}[\mathbf{z}^{\mathbf{N}_k^l} \mathbf{1}_{\{\Upsilon_l = k\}}]. \end{aligned}$$

Let $N(T)$ be the number of arrivals during a random period T. Recall that I_i denotes the (exponential) interarrival time of customers to Q_i. We define $\phi_0^l(\mathbf{z}) := \alpha^l(\mathbf{z})$, i.e., $\phi_0^l(\mathbf{z})$ is the p.g.f. of the number of customers present when entering the given server-location state $l = (l_1, l_2)$. Finally, we define $\beta^l(\mathbf{z})$ to be the p.g.f. of the number of customers in the system at the instant that server-location state l is left. The expression for $\beta^l(\mathbf{z})$ is stated without proof in the following lemma (see [H2]).

$$\beta^l(\mathbf{z}) = \sum_{k=1}^{\infty} \phi_k^{s,l}(\mathbf{z}). \tag{7.2}$$

The complementary equation which relates the queue length at the entrance of a server-location state to the queue length at the departure instant from the previous state is as follows.

$$\phi_0^l(\mathbf{z}) := \alpha^l(\mathbf{z}) = \sum_{j \in \mathcal{L}_1 \times \mathcal{L}_2} q_{j,l} \beta^j(\mathbf{z}),$$

where $q_{j,l}$, the probability that given that the server-location state is l the preceding server-location state has been j, is given by (see, e.g., [56]):

$$q_{j,l} = \frac{\tau_j \cdot s_{j,l}}{\tau_l}. \tag{7.3}$$

It should be emphasized that this latter equation holds for general service disciplines as long as the server-location process is one-dimensional. When the server-location process is multi-dimensional, it remains valid for the pure exponential time-limited discipline. This is due to the fact that the jumps in the server-location process occur independently of the number of customers at the queues. However, this property is not satisfied for most of the common service disciplines, such as the exponential, gated, and k-limited disciplines, and as a result Eq. (7.3) does not apply anymore. In other words, the adopted pure exponential time-limited service discipline is crucial for the tractability of our multi-server polling model.

We continue with the determination of $\phi_k^{s,l}(\mathbf{z})$ as to solve for $\beta^l(\mathbf{z})$ via Eq. (7.2). This will be done consecutively for the case that the two servers visit the same queue and for the case that the servers visit different queues.

7.3.2.1 Two servers visit the same queue

Next, assume that server S_1 and server S_2 visit the same queue. We note that during a server visit the time until the next marked event depends on the queue length at $Q_{l_1}(= Q_{l_2})$. In particular, it matters whether there are zero, one, or two or more customers at this queue. To this end, we introduce the following random variables:

- X^0: time until a customer arrival to Q_{l_1};
- X^1: time until a customer arrival to Q_{l_1} or a service completion at Q_{l_1}, given $N_{l_1} = 1$;
- X^+: time until a service completion at Q_{l_1}, given $N_{l_1} \geq 2$.

Again, we let the random variable Y account for the time until the server-location process moves to another state, i.e., either one of the servers moves away (individual-server case) or both servers move away (coupled-server case). The distribution of these random variables is as follows: X^0 is $\exp(\lambda_{l_1})$-distributed, X^1 is $\exp(\lambda_{l_1} + \mu_{l_1})$-distributed, X^+ is $\exp(\mu^+)$-distributed, and Y is $\exp(\xi^l)$-distributed. The parameter μ^+ is strictly positive (see also Remark 7.1 below). Then, by analogy with the results of the single-server case in Chapter 4, $\phi_k^l(\mathbf{z})$ and $\phi_k^{s,l}(\mathbf{z})$, $l = (l_1, l_1)$, $k = 1, 2, \ldots$, are

recursively given by:

$$\begin{aligned}
\phi_k^l(\mathbf{z}) = \; & \phi_{k-1}^l(\mathbf{z})\mid_{z_{l_1}=0} \cdot \mathbb{E}[\mathbf{z}^{N(X^0)}\mathbf{1}_{\{Y>X^0\}}] \cdot z_{l_1} \\
& + z_{l_1} \cdot \left(\frac{\phi_{k-1}^l(\mathbf{z}) - \phi_{k-1}^l(\mathbf{z})\mid_{z_{l_1}=0}}{z_{l_1}} \right) \mid_{z_{l_1}=0} \\
& \quad \times \mathbb{E}[\mathbf{z}^{N(X^1)}\mathbf{1}_{\{Y>X^1\}}] \Big(\mathbb{P}(\text{arr. to } Q_{l_1} \mid Y > X^1) \cdot z_{l_1} \\
& \quad + \mathbb{P}(\text{serv. at } Q_{l_1} \mid Y > X^1) \cdot \frac{r^{l_1}(\mathbf{z})}{z_{l_1}} \Big) \\
& + \left(\phi_{k-1}^l(\mathbf{z}) - z_{l_1} \cdot \left(\frac{\phi_{k-1}^l(\mathbf{z}) - \phi_{k-1}^l(\mathbf{z})\mid_{z_{l_1}=0}}{z_{l_1}} \right) \mid_{z_{l_1}=0} \right. \\
& \quad \left. - \phi_{k-1}^l(\mathbf{z})\mid_{z_{l_1}=0} \right) \cdot \mathbb{E}[\mathbf{z}^{N(X^+)}\mathbf{1}_{\{Y>X^+\}}] \cdot \frac{r^{l_1}(\mathbf{z})}{z_{l_1}} \; , \qquad (7.4)
\end{aligned}$$

$$\begin{aligned}
\phi_k^{s,l}(\mathbf{z}) = \; & \phi_{k-1}^l(\mathbf{z})\mid_{z_{l_1}=0} \cdot \mathbb{E}[\mathbf{z}^{N(Y)}\mathbf{1}_{\{Y<X^0\}}] \\
& + z_{l_1} \cdot \left(\frac{\phi_{k-1}^l(\mathbf{z}) - \phi_{k-1}^l(\mathbf{z})\mid_{z_{l_1}=0}}{z_{l_1}} \right) \mid_{z_{l_1}=0} \cdot \mathbb{E}[\mathbf{z}^{N(Y)}\mathbf{1}_{\{Y<X^1\}}] \\
& + \left(\phi_{k-1}^l(\mathbf{z}) - z_{l_1} \cdot \left(\frac{\phi_{k-1}^l(\mathbf{z}) - \phi_{k-1}^l(\mathbf{z})\mid_{z_{l_1}=0}}{z_{l_1}} \right) \mid_{z_{l_1}=0} \right. \\
& \quad \left. - \phi_{k-1}^l(\mathbf{z})\mid_{z_{l_1}=0} \right) \cdot \mathbb{E}[\mathbf{z}^{N(Y)}\mathbf{1}_{\{Y<X^+\}}], \qquad (7.5)
\end{aligned}$$

where for both expressions the three terms on the right-hand side refer to the cases with zero, one and more than one customer present, respectively. Furthermore, we may write:

$$\begin{aligned}
\phi_0^l(\mathbf{z}) &= \sum_{j \in \mathcal{L}_1 \times \mathcal{L}_2} q_{j,l}\beta^j(\mathbf{z}), \\
\mathbb{E}[\mathbf{z}^{N(X^0)}\mathbf{1}_{\{Y>X^0\}}] &= \tilde{X}^0(\xi^l + \textstyle\sum_{j\neq l_1}\lambda_j(1-z_j)), \\
\mathbb{E}[\mathbf{z}^{N(X^1)}\mathbf{1}_{\{Y>X^1\}}] &= \tilde{X}^1(\xi^l + \textstyle\sum_{j\neq l_1}\lambda_j(1-z_j)), \\
\mathbb{E}[\mathbf{z}^{N(X^+)}\mathbf{1}_{\{Y>X^+\}}] &= \tilde{X}^+(\xi^l + \textstyle\sum_{j}\lambda_j(1-z_j)), \\
\mathbb{E}[\mathbf{z}^{N(Y)}\mathbf{1}_{\{Y<X^0\}}] &= \tilde{Y}(\lambda_{l_1} + \textstyle\sum_{j\neq l_1}\lambda_j(1-z_j)), \\
\mathbb{E}[\mathbf{z}^{N(Y)}\mathbf{1}_{\{Y<X^1\}}] &= \tilde{Y}(\lambda_{l_1} + \mu_{l_1} + \textstyle\sum_{j\neq l_1}\lambda_j(1-z_j)), \\
\mathbb{E}[\mathbf{z}^{N(Y)}\mathbf{1}_{\{Y<X^+\}}] &= \tilde{Y}(\mu^+ + \textstyle\sum_{j}\lambda_j(1-z_j)).
\end{aligned}$$

The conditional probabilities above are given by:

$$\begin{aligned}\mathbb{P}(\text{arr. to } Q_{l_1} \mid Y > X^1) &= \lambda_{l_1}/(\lambda_{l_1} + \mu_{l_1}) \\ &= 1 - \mathbb{P}(\text{serv. at } Q_{l_1} \mid Y > X^1),\end{aligned}$$

so that now all terms in Eqs. (7.4) and (7.5) are explicitly characterized.

Remark 7.1. *(Single-server emulation) The introduction of the parameter μ^+ gives us the freedom to modify the service rate when there are two or more customers at Q_{l_1}. Indeed, setting μ^+ equal to μ_{l_1} leads to the special case of having only one server at the queue. In fact, it provides a more complicated alternative to the analysis in Sect. 4.4.4 with exponential service times. Clearly, the perhaps most natural choice of $\mu^+ = 2 \cdot \mu_{l_1}$ leads to the case of having two servers at the queue.*

7.3.2.2 Two servers visit a different queue

Assume that server S_1 and server S_2 visit a different queue. We note that during a server visit the time until the next marked event depends on the queue length at Q_{l_1} and Q_{l_2}. In particular, it matters whether these queues are empty or nonempty. Hence, it is convenient to introduce the following random variables:

- X^{00}: time until a customer arrival to Q_{l_1} or Q_{l_2};
- X^{0+}: time until a customer arrival to Q_{l_1} or a service completion at Q_{l_2}, given $N_{l_2} \geq 1$;
- X^{+0}: time until a service completion at Q_{l_1} or customer arrival at Q_{l_2}, given $N_{l_1} \geq 1$.

Besides, as a generalization for the case that both queues Q_{l_1} and Q_{l_2} are nonempty, we introduce:

- X^{*j}: time until a service completion at Q_{l_j}, given $N_{l_1} \geq 1$ and $N_{l_2} \geq 1$, for $j = 1, 2$.

Further, we let Y account for the time until one of the servers moves away from the queue. Finally, we introduce the parameter $\gamma,\ 0 \leq \gamma \leq 1$ which will be used to prioritize one of the queues, Q_{l_1} or Q_{l_2}, when both of them are nonempty.

The distribution of the above-mentioned random variables is as follows: X^{00} is $\exp(\lambda_{l_1} + \lambda_{l_2})$-distributed, X^{0+} is $\exp(\lambda_{l_1} + \mu_{l_2})$-distributed, X^{+0} is $\exp(\mu_{l_1} + \lambda_{l_2})$-distributed, X^{*j} is $\exp(\mu^{*j})$-distributed, $j = 1, 2$, and Y is $\exp(\xi^l)$-distributed. The parameters $\mu^{*j},\ j = 1, 2$, are strictly positive. Then, by analogy with the single-server results of Sect. 4.4.4, $\phi_k^l(\mathbf{z})$ and $\phi_k^{s,l}(\mathbf{z})$, $l = (l_1, l_2)$, with $l_1 \neq l_2,\ k = 1, 2, \ldots,$

are recursively given by:

$$\begin{aligned}
\phi_k^l(\mathbf{z}) = \; & \phi_{k-1}^l(\mathbf{z})\,|_{(z_{l_1}=0,z_{l_2}=0)} \cdot \mathbb{E}[\mathbf{z}^{N(X^{00})}\mathbf{1}_{\{Y>X^{00}\}}] \\
& \quad \times \left(\mathbb{P}(\text{arr. to } Q_{l_1} \mid Y > X^{00}) \cdot z_{l_1} + \mathbb{P}(\text{arr. to } Q_{l_2} \mid Y > X^{00}) \cdot z_{l_2}\right) \\
& + \phi_{k-1}^l(\mathbf{z})\,|_{(z_{l_1}=0,z_{l_2}>0)} \cdot \mathbb{E}[\mathbf{z}^{N(X^{0+})}\mathbf{1}_{\{Y>X^{0+}\}}] \\
& \quad \times \Bigg(\mathbb{P}(\text{arr. to } Q_{l_1} \mid Y > X^{0+}) \cdot z_{l_1} \\
& \qquad + \mathbb{P}(\text{serv. at } Q_{l_2} \mid Y > X^{0+}) \cdot \frac{r^{l_2}(\mathbf{z})}{z_{l_2}}\Bigg) \\
& + \phi_{k-1}^l(\mathbf{z})\,|_{(z_{l_1}>0,z_{l_2}=0)} \cdot \mathbb{E}[\mathbf{z}^{N(X^{+0})}\mathbf{1}_{\{Y>X^{+0}\}}] \\
& \quad \times \Bigg(\mathbb{P}(\text{serv. at } Q_{l_1} \mid Y > X^{+0}) \cdot \frac{r^{l_1}(\mathbf{z})}{z_{l_1}} \\
& \qquad + \mathbb{P}(\text{arr. to } Q_{l_2} \mid Y > X^{+0}) \cdot z_{l_2}\Bigg) \\
& + \phi_{k-1}^l(\mathbf{z})\,|_{(z_{l_1}>0,z_{l_2}>0)} \cdot \Bigg(\gamma \cdot \mathbb{E}[\mathbf{z}^{N(X^{*1})}\mathbf{1}_{\{Y>X^{*1}\}}] \cdot \frac{r^{l_1}(\mathbf{z})}{z_{l_1}} \\
& \qquad + (1-\gamma) \cdot \mathbb{E}[\mathbf{z}^{N(X^{*2})}\mathbf{1}_{\{Y>X^{*2}\}}] \cdot \frac{r^{l_2}(\mathbf{z})}{z_{l_2}}\Bigg), \qquad (7.6)
\end{aligned}$$

$$\begin{aligned}
\phi_k^{s,l}(\mathbf{z}) \;=\; & \phi_{k-1}^l(\mathbf{z})\,|_{(z_{l_1}=0,z_{l_2}=0)} \cdot \mathbb{E}[\mathbf{z}^{N(Y)}\mathbf{1}_{\{Y<X^{00}\}}] \\
& + \phi_{k-1}^l(\mathbf{z})\,|_{(z_{l_1}=0,z_{l_2}>0)} \cdot \mathbb{E}[\mathbf{z}^{N(Y)}\mathbf{1}_{\{Y<X^{0+}\}}] \\
& + \phi_{k-1}^l(\mathbf{z})\,|_{(z_{l_1}>0,z_{l_2}=0)} \cdot \mathbb{E}[\mathbf{z}^{N(Y)}\mathbf{1}_{\{Y<X^{+0}\}}] \\
& + \phi_{k-1}^l(\mathbf{z})\,|_{(z_{l_1}>0,z_{l_2}>0)} \\
& \quad \times \left(\gamma \cdot \mathbb{E}[\mathbf{z}^{N(Y)}\mathbf{1}_{\{Y<X^{*1}\}}] + (1-\gamma) \cdot \mathbb{E}[\mathbf{z}^{N(Y)}\mathbf{1}_{\{Y<X^{*2}\}}]\right), \qquad (7.7)
\end{aligned}$$

where (via standard generating function manipulation),

$$\begin{aligned}
\phi_{k-1}^l(\mathbf{z})\,|_{(z_{l_1}=0,z_{l_2}>0)} &= \mathbb{E}[\mathbf{z}^{\phi_{k-1}^l}\mathbf{1}_{\{N_{l_1}=0\}}\mathbf{1}_{\{N_{l_2}>0\}}] \\
&= \phi_{k-1}(\mathbf{z})\,|_{z_{l_1}=0} - \phi_{k-1}(\mathbf{z})\,|_{(z_{l_1}=0,z_{l_2}=0)},
\end{aligned}$$

and similarly,

$$\begin{aligned}
\phi_{k-1}^l(\mathbf{z})\,|_{(z_{l_1}>0,z_{l_2}=0)} &= \phi_{k-1}(\mathbf{z})\,|_{z_{l_2}=0} - \phi_{k-1}(\mathbf{z})\,|_{(z_{l_1}=0,z_{l_2}=0)}, \\
\phi_{k-1}^l(\mathbf{z})\,|_{(z_{l_1}>0,z_{l_2}>0)} &= \phi_{k-1}(\mathbf{z}) - \phi_{k-1}(\mathbf{z})\,|_{z_{l_1}=0} - \phi_{k-1}(\mathbf{z})\,|_{z_{l_2}=0} \\
&\quad + \phi_{k-1}(\mathbf{z})\,|_{(z_{l_1}=0,z_{l_2}=0)}\,.
\end{aligned}$$

Further, we have that:

$$\begin{aligned}
\phi_0^l(\mathbf{z}) &= \sum_{j\in\mathcal{L}_1\times\mathcal{L}_2} q_{j,l}\beta^j(\mathbf{z}),\\
\mathbb{E}[\mathbf{z}^{N(X^{00})}\mathbf{1}_{\{Y>X^{00}\}}] &= \tilde{X}^{00}(\xi^l + \textstyle\sum_{j\neq l_1,l_2}\lambda_j(1-z_j)),\\
\mathbb{E}[\mathbf{z}^{N(X^{0+})}\mathbf{1}_{\{Y>X^{0+}\}}] &= \tilde{X}^{0+}(\xi^l + \textstyle\sum_{j\neq l_1}\lambda_j(1-z_j)),\\
\mathbb{E}[\mathbf{z}^{N(X^{+0})}\mathbf{1}_{\{Y>X^{+0}\}}] &= \tilde{X}^{+0}(\xi^l + \textstyle\sum_{j\neq l_2}\lambda_j(1-z_j)),\\
\mathbb{E}[\mathbf{z}^{N(X^{*k})}\mathbf{1}_{\{Y>X^{*k}\}}] &= \tilde{X}^{+l}(\xi^l + \textstyle\sum_{j}\lambda_j(1-z_j)),\ k=1,2,\\
\mathbb{E}[\mathbf{z}^{N(Y)}\mathbf{1}_{\{Y<X^{00}\}}] &= \tilde{Y}(\lambda_{l_1} + \lambda_{l_2} + \textstyle\sum_{j\neq l_1,l_2}\lambda_j(1-z_j)),\\
\mathbb{E}[\mathbf{z}^{N(Y)}\mathbf{1}_{\{Y<X^{0+}\}}] &= \tilde{Y}(\lambda_{l_1} + \mu_{l_2} + \textstyle\sum_{j\neq l_1}\lambda_j(1-z_j)),\\
\mathbb{E}[\mathbf{z}^{N(Y)}\mathbf{1}_{\{Y<X^{+0}\}}] &= \tilde{Y}(\mu_{l_1} + \lambda_{l_2} + \textstyle\sum_{j\neq l_2}\lambda_j(1-z_j)),\\
\mathbb{E}[\mathbf{z}^{N(Y)}\mathbf{1}_{\{Y<X^{*k}\}}] &= \tilde{Y}(\mu^{*k} + \textstyle\sum_{j}\lambda_j(1-z_j)),\ k=1,2,
\end{aligned}$$

where $q_{j,l}$ is the probability that given the current server-location state l the preceding server-location state has been j. The conditional probabilities are given by:

$$\begin{aligned}
\mathbb{P}(\text{arr. to } Q_{l_1} \mid Y > X^{00}) &= \lambda_{l_1}/(\lambda_{l_1}+\lambda_{l_2})\\
&= 1 - \mathbb{P}(\text{arr. to } Q_{l_2} \mid Y > X^{00}),\\
\mathbb{P}(\text{arr. to } Q_{l_1} \mid Y > X^{0+}) &= \lambda_{l_1}/(\lambda_{l_1}+\mu_{l_2})\\
&= 1 - \mathbb{P}(\text{serv. at } Q_{l_2} \mid Y > X^{0+}),\\
\mathbb{P}(\text{serv. at } Q_{l_1} \mid Y > X^{+0}) &= \mu_{l_1}/(\mu_{l_1}+\lambda_{l_2})\\
&= 1 - \mathbb{P}(\text{arr. to } Q_{l_2} \mid Y > X^{+0}).
\end{aligned}$$

Thus, all the terms that appear in Eqs. (7.6) and (7.7) are now explicitly known. We note that the derivation of these expression is essentially a matter of case distinction. However, let us comment on the somewhat strange coefficient of $\phi_{k-1}^l(\mathbf{z})\mid_{(z_{l_1}>0,z_{l_2}>0)}$ in Eq. (7.6), i.e., the term:

$$\gamma\cdot\mathbb{E}[\mathbf{z}^{N(X^{*1})}\mathbf{1}_{\{Y>X^{*1}\}}]\cdot\frac{r^{l_1}(\mathbf{z})}{z_{l_1}} + (1-\gamma)\cdot\mathbb{E}[\mathbf{z}^{N(X^{*2})}\mathbf{1}_{\{Y>X^{*2}\}}]\cdot\frac{r^{l_2}(\mathbf{z})}{z_{l_2}}, \quad (7.8)$$

whereas conform the other coefficients in Eq. (7.6), one might have well expected to see:

$$\begin{aligned}
&\mathbb{E}[\mathbf{z}^{N(X^{++})}\mathbf{1}_{\{Y>X^{++}\}}]\cdot\Bigg(\mathbb{P}(\text{serv. at } Q_{l_1} \mid Y > X^{++})\cdot\frac{r^{l_1}(\mathbf{z})}{z_{l_1}}\\
&+ \mathbb{P}(\text{serv. at } Q_{l_2} \mid Y > X^{++})\cdot\frac{r^{l_2}(\mathbf{z})}{z_{l_2}}\Bigg),
\end{aligned}$$

with X^{++} being $\exp(\mu_{l_1} + \mu_{l_2})$-distributed. Though, it is readily verified that for the parameter choices:

$$\begin{aligned} \gamma &= \mathbb{P}(\text{serv. at } Q_{l_1} \mid Y > X^{++}) = 1 - \mathbb{P}(\text{serv. at } Q_{l_2} \mid Y > X^{++}) \\ &= \frac{\mu_{l_1}}{\mu_{l_1} + \mu_{l_2}}, \\ \mu^{*1} &= \mu^{*2} = \mu_{l_1} + \mu_{l_2}, \end{aligned}$$

these terms indeed match. The reason to adopt the more general form, Eq. (7.8), is that it enables the modelling of resource sharing applications. This will be demonstrated in more detail in Sect. 7.4.2.

7.3.3 Steady-state probabilities

Towards obtaining the steady-state probabilities for the two-server polling system, we will compute the queue-length probabilities at departure instants from a server-location state. These latter probabilities are found using the common iterative approach. However, it should be noted that, depending on the specific service strategy, different relations must be used here. In particular, to analyze the independent-server case the results of Sect. 7.3.2.2 must be used, while to analyze the coupled-server case, the results of Sect. 7.3.2.1 will serve as input for the iterative solution method. As for the individual-server case, the servers may be either at the same queue or at different queues, the specific server-location state implies which results to use.

Next, we can write down an iterative scheme in terms of Discrete Fourier Transforms for the p.g.f. of the joint-queue length distribution as shown in Table 7.2. We emphasize that this scheme is not necessarily the most efficient in terms of computation time. Limited experience has shown that choosing the next state in another fashion (e.g., dependent on the fraction of time a location state is visited) may lead in certain cases to a significant decrease of computation time compared to the solution presented in Table 7.2. However, optimization of such kind is outside the scope of this thesis.

The steady-state joint queue-length probabilities can readily be obtained (cf. the discussion in Chapter 4), since the residence time in any server-location state is exponentially distributed. Specifically, the p.g.f. of these probabilities is given by:

$$P(\mathbf{z}) = \sum_{l \in \mathcal{L}} \beta^l(\mathbf{z}) \cdot \omega_l,$$

where ω_l is provided in Eq. (7.1).

7.4 Examples

In this section, we will discuss two examples to show the applicability of the two-server polling analysis. The first example is a polling system with two servers which

Algorithm 7.2. *Pseudo-code of the iterative scheme for determining* $\check{\beta}^l(\boldsymbol{k}), \forall_l, \forall_{\boldsymbol{k}}$.

$\check{\beta}^\eta(\boldsymbol{k}) = 1,\ \forall_\eta,\ \forall_{\boldsymbol{k}}$; *(start with an empty system)*
FOR $\zeta = 1, \ldots, \lvert\mathcal{L}_1 \times \mathcal{L}_2\rvert$
set $l := \zeta$;
REPEAT
$\bar{\beta}^l(\boldsymbol{k}) = \check{\beta}^l(\boldsymbol{k}),\ \forall_{\boldsymbol{k}}$;
set $j := 0$;
set $\phi_0^l(\boldsymbol{k}) = \sum_j q_{j,l}\check{\beta}^j(\boldsymbol{k})$;
REPEAT
set $j := j + 1$;
IF $l_1 \neq l_2$
compute $\phi_j^l(\boldsymbol{k}),\ \forall_{\boldsymbol{k}}$ *using Eq. (7.6);*
compute $\phi_j^{s,l}(\boldsymbol{k}),\ \forall_{\boldsymbol{k}}$ *using Eq. (7.7);*
ELSEIF $l_1 = l_2$
compute $\phi_j^l(\boldsymbol{k}),\ \forall_{\boldsymbol{k}}$ *using Eq. (7.4);*
compute $\phi_j^{s,l}(\boldsymbol{k}),\ \forall_{\boldsymbol{k}}$ *using Eq. (7.5);*
END IF
compute $\check{\beta}^l(\boldsymbol{k}) = \sum_{i=1}^{j} \phi_i^{s,l}(\boldsymbol{k}),\ \forall_{\boldsymbol{k}}$;
UNTIL $1 - Re(\check{\beta}^l(\boldsymbol{0})) < \delta$
set $l := MOD(l, M) + 1$;
UNTIL $\lvert Re(\check{\beta}^\zeta(\boldsymbol{k})) - Re(\bar{\beta}^\zeta(\boldsymbol{k}))\rvert < \epsilon,\ \forall_{\boldsymbol{k}}$
END

move independently and cyclically through the system. The second example is a three-hop tandem model for multi-hop data communication under intermittent network connectivity. For both examples, we describe in detail how the parameters should be set and how the server-location process should be chosen. The resulting specification can be inserted directly into Algorithm 7.2, so that eventually the steady-state probabilities can be found. We stress that the service discipline in both examples is the pure exponential time-limited discipline.

7.4.1 Cyclic polling model with independent servers

A two-server cyclic polling model (see Fig. 7.1) is an extension of the single-server cyclic polling model. In the standard single-server model, the server serves the queues one by one in a cyclic fashion, i.e., it visits the queues in the order $Q_1, Q_2, \ldots, Q_M, Q_1$, etc. Here, the two servers S_1 and S_2 independently visit the queues in a cyclic order.

Hence, it may also happen that both servers serve the same queue.

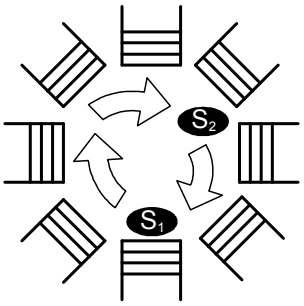

Figure 7.1: Cyclic polling model with two servers.

Customer arrivals to Q_i occur with rate $\lambda_i \geq 0,\ i = 1, \ldots, M$. Upon completion of their service, customers leave the system. Thus, the customer routing probabilities are given by $r_{ij} = 1$, if $j = 0$, and 0 otherwise, for $i = 1, \ldots, M$. In server-location state l, server $S_j,\ j = 1, 2$, remains at Q_{l_j} for an exponentially distributed period of time with parameter ξ_{l_j}. Thus, the total jump rate out of server-location state l is $\xi^l = \xi_{l_1} + \xi_{l_2}$. The state space of the server-location process is given by $\mathcal{L}_1 \times \mathcal{L}_2$ where $\mathcal{L}_1, \mathcal{L}_2 = \{1, \ldots, M\}$ with server routing probabilities for $n \geq 0$ and $l = (l_1, l_2)$:

$$s_{(l_1^n, l_2^n),(l_1^{n+1}, l_2^{n+1})} = \begin{cases} \frac{\xi^l_{l_1}}{\xi^l_{l_1} + \xi^l_{l_2}}, & \text{if } l_1^{n+1} = l_1^n + 1 \text{ and } l_2^{n+1} = l_2^n; \\ \frac{\xi^l_{l_2}}{\xi^l_{l_1} + \xi^l_{l_2}}, & \text{if } l_2^{n+1} = l_2^n + 1 \text{ and } l_1^{n+1} = l_1^n; \\ 0, & \text{otherwise,} \end{cases}$$

where we set $l_k^n + 1 = 1$ if $l_k^n = M,\ k = 1, 2$. The probabilities $q_{j,l}$, required to set the initial state $\phi_0^l(\mathbf{z})$, then follow from solving the stationary distribution of the Markov chain for the server-location process and applying Eq. (7.3).

The servers may either be at the same queue or at different queues in the system. If the servers are at a different queue, then the analysis of Sect. 7.3.2.2 must be applied with an appropriate choice of the parameters γ, μ^{*1}, and μ^{*2}. We note that the parameter γ refers to the probability that a service completion occurs at Q_{l_1} given that indeed a completion occurs. The parameters μ^{*1} and μ^{*2} refer both to the total service rate when both visited queues are nonempty. Hence, it is readily found that this specific model is represented by the following parameter choices:

$$\begin{aligned} \gamma &= \frac{\mu_{l_1}}{\mu_{l_1} + \mu_{l_2}}, \\ \mu^{*1} &= \mu^{*2} = \mu_{l_1} + \mu_{l_2}. \end{aligned}$$

If the servers are at the same queue, then the analysis of Sect. 7.3.2.1 can directly

Mobile queue position, (V_2, V_3) :	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
Communication links:	$\{L_1, L_2\}$	$\{L_1, L_3\}$	$\{L_2\}$	$\{L_3\}$
Server-location state, (l_1, l_2):	$(1,2)$	$(1,3)$	$(2,2)$	$(3,3)$

Table 7.1: Mapping of mobile queue positions (V_2, V_3) to the available links and the server-location state (l_1, l_2).

be applied with the following choice for the parameter μ^+:

$$\mu^+ = \mu_{l_1} + \mu_{l_2}.$$

7.4.2 Multi-hop tandem model for data communication

Consider a three-hop tandem model consisting of four queues (see Fig. 7.2). Data packets are generated at Q_1 and are destined for Q_4 which operates solely as a sink node and will be excluded from the rest of the analysis. Communication links L_1 connect Q_i with $Q_{i+1},\ i = 1, 2, 3$. However, the links are not always available for data transmission as the intermediate queues, Q_2 and Q_3, are mobile and move within a local region. Each of these queues is either in position $V_j = 0,\ j = 2, 3$ (leftmost position) or in position $V_j = 1,\ j = 2, 3$ (rightmost position). It is assumed that mobile queue $Q_j,\ j = 2, 3$, remains at its position for an exponentially distributed period of time with parameter ξ_j (independent of the position being 0 or 1). The mapping of the positions of the mobile queues to the available links is given in Table 7.4.2.

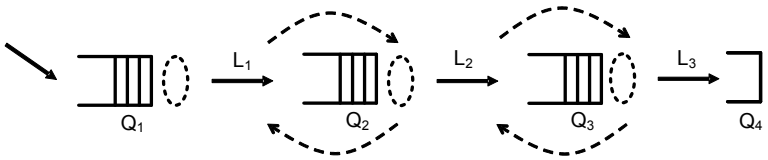

Figure 7.2: Three-hop tandem model.

The exogenous Poisson arrival rates to the queues are as follows: $\lambda_1 \geq 0$, while $\lambda_i = 0,\ i = 2, 3$. Upon completion of their service, customers move to the next queue and thus $r_{ij} = 1$, if $j = i + 1$, and 0 otherwise, for $i = 1, 2$, and $r_{30} = 1$. Next, we should relate the position of the mobile queues to the queues that are being served. We will say that if links L_i and $L_j,\ j \neq i$, are available then Q_i and Q_j are being served. If there is only one link L_i available, then we will say that the server-location state is (i, i) and that Q_i is effectively served by one server only (see Remark 7.1). Using the prior mapping from (V_2, V_3) to the links, the indirect mapping from (V_2, V_3) to the server-location state (l_1, l_2) is then readily obtained and also given in Table 7.4.2. For instance, if Q_2 and Q_3 are both in their rightmost position, i.e., $(V_2, V_3) = (1, 1)$, then as only link L_3 is available both servers are

	(l_1^{n+1}, l_2^{n+1}) :	$(1,2)$	$(1,3)$	$(2,2)$	$(3,3)$
(l_1^n, l_2^n) :					
$(1,2)$		0	θ	$1-\theta$	0
$(1,3)$		$1-\theta$	0	0	θ
$(2,2)$		θ	0	0	$1-\theta$
$(3,3)$		0	$1-\theta$	θ	0

Table 7.2: Server routing probabilities $s_{(l_1^n,l_2^n),(l_1^{n+1},l_2^{n+1})}$.

assigned to Q_3. The one-step server routing probabilities for epoch $n \geq 0$ to $n+1$ with $l = (l_1^n, l_2^n)$ follow directly from the mobility process of Q_2 and Q_3 and are given in Table 7.2 where $\theta = \xi_2/(\xi_2 + \xi_3)$.

The recursive relations for the cases with one available link, $(V_2, V_3) = (1,0)$ and $(V_2, V_3) = (1,1)$, are derived along the analysis for the coupled-server case (see Sect. 7.3.2.1) and then setting:

$$\mu^+ = \mu_{l_1} (= \mu_{l_2}).$$

The other cases $(V_2, V_3) = (0,0)$ and $(V_2, V_3) = (0,1)$, with two links available, can be analyzed using the analysis for the two servers being at a different queue (see Sect. 7.3.2.2). The parameter γ reflects here the fairness of the underlying medium access control (MAC) protocol. According to a fair MAC protocol, each transmission contending for the medium will have an equal probability to start. Hence, we will set $\gamma = 1/2$. Besides, we note that only the "winner" of this contention is allowed to transmit as in a local neighborhood one transmission may occur at a time (to prevent packet collisions to occur). Hence, in case $(V_2, V_3) = (0,0)$, the model is fully specified by the following parameter choices:

$$\begin{aligned} \gamma &= 1/2, \\ \mu^{*j} &= \mu_{l_j},\ j = 1,2. \end{aligned}$$

In case $(V_2, V_3) = (0,1)$, assuming that both transmissions may start and be successful simultaneously, the parameters should be chosen identical as for the cyclic two-server polling model with the servers being at a different queue, i.e.,

$$\begin{aligned} \gamma &= \frac{\mu_{l_1}}{\mu_{l_1} + \mu_{l_2}}, \\ \mu^{*1} &= \mu^{*2} = \mu_{l_1} + \mu_{l_2}. \end{aligned}$$

Otherwise, if the MAC protocol would not allow for simultaneous transmissions (i.e., it protects the transmission over link L_1 from interference caused by the transmission over L_3), then the parameters should be set as for the case $(V_2, V_3) = (0,0)$.

7.5 Discussion

7.5.1 Nonzero switch-over times

According to the coupled-server strategy, the servers always move as one group, so that generally distributed switch-over times can easily be incorporated.

On the contrary, for the individual-server strategy, the servers move away from a queue only one at a time. It may thus happen that a server is switching while the other server continues serving customers at a queue. Also, it may happen that both servers are switching simultaneously. These scenarios appear hard to incorporate in our modelling framework for generally distributed switch-over times. Hence, we will restrict ourselves to exponential switch-over times. The switch-over times are then most adequately incorporated by constructing additional server-location states which account for the switching periods of the servers.

Consider nonzero switch-over times $C_{i,j}$ for a server to move from Q_i to Q_j which follow an exponential distribution with mean $c_{i,j}$. Including the switch-over movements in the server-location process leads to an expansion from the original M to now $M + M^2$ states. We note that these M^2 states account for switches between all M states, thus allowing also for self-loops and for asymmetric switch-over times. If both servers are switching in state l, say server S_1 from i_1 to i_2 and server S_2 from j_1 to j_2, then we may readily write:

$$\beta^l(\mathbf{z}) = \alpha^l(\mathbf{z}) \cdot \tilde{W}\left(\sum_j \lambda_j(1 - z_j)\right),$$

where $\tilde{W}$ is the LST of the random variable X which is $\exp(c_{i_1,i_2}^{-1} + c_{j_1,j_2}^{-1})$-distributed. If only one server is switching in state l, say S_1 is at i and S_2 switches from j_1 to j_2, then the analysis is analogous to the single-server analysis but now with a modified intensity, viz., $\xi^l = \xi_i + c_{j_1,j_2}^{-1}$, to leave the current server-location state. If none of the servers is switching, then the analysis introduced in Sect. 7.3.2 can directly be applied.

7.5.2 Three or more servers

The analysis presented here may readily be extended to the case of three or more servers. The key element is that for each state of the server-location process (which will live on a state space with more than two dimensions) one is able to describe the evolution of the joint queue-length process properly. It may also happen that in some states several queues are visited by multiple servers, while others only by one or zero. For such a scenario, one has to combine the analysis given in Sects. 7.3.2.1 and 7.3.2.2 in an appropriate way. We strongly believe that such only leads to distinguishing between more and more cases as the number of servers increases, but that the general approach remains the same.

7.5.3 A limited number of servers per queue

There may also be a limit on the maximum number of servers allowed at a single queue in such systems with multiple servers (see, e.g., [2, 75]). We note that such a feature can readily be modelled within our framework by appropriately setting the routing probabilities $s_{l,j},\ l,j \in \mathcal{L}_1 \times \mathcal{L}_2$ of the server-location process. Notice that if this limit is taken equal to K, the total number of servers, then the servers move indeed independently (as was considered in the example of the cyclic polling model in Sect. 7.4.1).

7.6 Concluding remarks

We have presented in this chapter a complete framework for the computation of the steady-state joint queue-length distribution for the basic multi-server polling model extended with customer routing. The key relations in this framework, which capture each the queue-length evolution in a specific server-location state, are determined in a recursive manner (see Eq. (7.2)). For all states together, these relations form a system of equations that can be solved in an iterative manner due to the pure time-limited service discipline assumed. The steady-state queue-length distribution readily follows from the exponentiality assumption on the time limits. Clearly, these results allow also for evaluation of the delay measures in such polling systems. The applicability of the results is illustrated along two constructive examples. Finally, we have indicated how nonzero switch-over times, a larger number of servers and per-queue service limits can be incorporated in the analysis.

The most common server strategies for multi-server polling models are the coupled-server and the individual-server strategy. While in the literature typically only one of these cases is considered, our framework allows to study both cases in a unified manner. Moreover, this framework allows to construct the key relations also for other service disciplines, such as exponential, gated, or exhaustive time-limited disciplines, along the recursive manner described. Unfortunately, for the individual-server case, connecting these key relations for the different queues as to obtain the joint queue-length distribution at entrance and departure instants of server-location states appears intractable. Conversely, for the coupled-servers case, which assumes a server-routing process that is identical to the routing process in a single-server polling model, the system of equations can indeed be solved and thus the queue-length distribution at these embedded epochs may be obtained.

CHAPTER

8

Transient analysis for the basic polling system

8.1 Introduction

We have analyzed a polling model with a single server operating under the pure exponential time-limited discipline. Particularly, in Sect. 5.3, we have applied results for the transient behavior of the M/G/1 queue to connect the embedded queue-length processes at the start and the end of a server visit. The main ideas of that analytical approach can also be applied to study polling models with multiple servers operating under the pure time-limited discipline. The difference in the analysis resides in the fact that the location process of the servers becomes multi-dimensional. Hence, we talk of server-location states instead of server visits to a queue. The key relation within the approach in Sect. 5.3 has its multi-server counterpart which describes the relation between the p.g.f.'s of the queue length at the entrance of a specific server-location state, $\alpha^l(\mathbf{z})$, and the departure from the same state, $\beta^l(\mathbf{z})$:

$$\beta^l(\mathbf{z}) = \mathcal{F}(\alpha^l)(\mathbf{z}), \tag{8.1}$$

where l refers to the server-location state and $\mathcal{F}$ is again some operator.

In spite of the small volume of literature on multi-server polling models, this specific relation for the multi-server case has been discussed in a couple of articles.

Borst [14] has considered Eq. (8.1) for the coupled-servers strategy with the servers operating under the exhaustive and gated service discipline by studying an M/M/c model with service interruptions and extending the decomposition ideas of Fuhrmann and Cooper [41]. For the multi-queue system, the (per-queue) key equations are derived and together with the equations that account for the queue-length evolution during the switch-over times a system of equations in terms of p.g.f.'s is formulated. Specifically, the models for which indeed a solution of this system may be found include: a single server with 1-limited service in a one or two-queue system, an infinite number of servers and deterministic service times, and two servers and two queues with exhaustive service. Vlasiou and Yechiali [103] consider the relation for the same strategy for the special case of an infinite number of servers operating according to the pure time-limited discipline. The model is motivated from an application of road-traffic control. By investigating the evolution of the queue-length during a visit, the authors derive a recursive expression for the p.g.f. of the joint queue-length distribution at polling instants. Contrary, research efforts on the key relation for the independent-servers strategy have not been reported in the literature to date.

In this chapter, we will focus exclusively on the key relation, Eq. (8.1), for a two-server polling model for both server strategies with the servers operating under the pure exponential time-limited discipline. We have analyzed this relation between $\alpha^l(\mathbf{z})$ and $\beta^l(\mathbf{z})$ in the previous chapter in an indirect manner by conditioning on specific events during the residence in a server-location state. Conversely, here we will follow a direct approach by studying the transient behavior of the queue length for a specific server-location state. To this end, we follow a similar approach as in Sect. 5.3. In the first scenario, when the servers are at the same queue, we may apply known results for the transient behavior of an M/M/2 queue (see [90]). In the second scenario, when the servers are at different queues, the approach is less straightforward and involves complex analysis techniques. The final results for the key relation may be incorporated in the framework as presented in Chapter 7 to obtain also the steady-state queue-length probabilities.

The remainder of this chapter is organized as follows. First, we give the model description in Sect. 8.2. Next, we present the queue-length analysis in Sect. 8.3. This will be done separately for the case that the servers visit the same queue and for the case that they visit different queues. The chapter will be concluded in Sect. 8.4.

8.2 Model

Let us consider the basic multi-server polling model of $M \geq 2$ queues with $K = 2$ servers as described in Sect. 1.3.5. Thus, customers arrive according to a Poisson process with rate λ_i and require an exponentially distributed amount of service with mean $1/\mu_i$. Customers that have been served depart immediately from the system, i.e., no customer routing. A customer is served by at most one server at a time and we will denote by N_i the number of customers at Q_i. The service discipline at all

queues for both servers is the pure exponential time-limited discipline. Conform the description in Sect. 1.3.4, the server-location process is a two-dimensional process on $\mathcal{L}_1 \times \mathcal{L}_2$ driven by the transition matrix $S = \{s_{l,j}\}_{l,j \in \mathcal{L}_1 \times \mathcal{L}_2}$. We let ξ^l, $l \in \mathcal{L}_1 \times \mathcal{L}_2$, denote the (total) departure intensity of the servers in server-location state $l = (l_1, l_2)$. We neglect the switch-over times here as these do not play a role in the analysis.

Before we get to the analysis, we introduce the following notation:

- x_t: number of customers at time t at Q_{l_1};
- $\mathbf{N}_l^s = (N_{l,1}^s, \ldots, N_{l,M}^s)$: number of customers at all queues at the entrance of server-location state l;
- $\mathbf{N}_l^e = (N_{l,1}^s, \ldots, N_{l,M}^e)$: number of customers at all queues at the departure from server-location state l;
- $N_{l,j}(t)$: number of customers at Q_j at time t during a residence in server-location state l;
- $\alpha^l(\mathbf{z})$: p.g.f. of $\mathbf{N}_l^s$;
- $\beta^l(\mathbf{z})$: p.g.f. of $\mathbf{N}_l^e$.

8.3 Analysis

The stability condition for the polling system is identical to the condition given in Sect. 7.3.1. Notice that since no customer routing is allowed here, $\gamma_i = \lambda_i$, $i = 1, \ldots, M$. In the analysis, we will consider the following two scenarios:

- both servers at Q_{l_1}, i.e., the server-location state reads $l = (l_1, l_1)$;
- one server at Q_{l_1} and one server at Q_{l_2}, $l_1 \neq l_2$, and $l = (l_1, l_2)$.

It is good to notice that these scenarios cover both the coupled-servers and the individual-servers strategy. Consecutively, we study the key relation, Eq. (8.1), for both scenarios. This will be done along the lines of Sect. 5.3.

8.3.1 Two servers visit the same queue

During the residence in this server-location state $l = (l_1, l_1)$, the queue-length process behaves in fact identical to the process of a transient M/M/2 queue. This transient process is well studied in the literature and explicit expressions are known for the transient queue-length probabilities. These results will constitute the core of the final relation between the p.g.f.'s above.

8.3.1.1 The evolution of the queue length

We consider the server-location state (l_1, l_1). The queue-length process at Q_{l_1} is then a birth-and-death process, and more specifically it evolves as the queue length of an M/M/2 queue. The queue-length processes at the other queues in the system are pure-birth processes. Let us next focus on the process at Q_{l_1} and whenever appropriate we will leave out the subscript l_1 to improve the readability. We define the time-dependent queue-length probabilities of interest as follows:

$$p_{hj}(t) := \begin{cases} \mathbb{P}(x_t = j | x_0 = h), & h, j = 0, 1, \ldots, \\ 0, & \text{otherwise.} \end{cases}$$

We consider the function $\pi_h(r, s)$ which is defined in terms of $p_{hj}(t)$ as follows:

$$\pi_h(r, s) := \sum_{j=0}^{\infty} r^j \int_0^{\infty} e^{-st} p_{hj}(t) dt, \ h = 0, 1, \ldots, \tag{8.2}$$

and which is explicitly provided by Saaty [90] as

$$\begin{aligned} \pi_h(r, s) &= \frac{1}{-\lambda_{l_1} r^2 + (2\mu_{l_1} + \lambda_{l_1} + s) \cdot r - 2\mu_{l_1}} \cdot \Bigg\{ r^{h+1} - \\ &\qquad \frac{(1 - r)(2\mu_{l_1} + (s + \lambda_{l_1})r)}{(1 - \hat{\mu}_{l_1}(s))(2\mu_{l_1} + (s + \lambda_{l_1})\hat{\mu}_{l_1}(s))} \cdot \hat{\mu}_{l_1}(s)^{h+1} \Bigg\}, \end{aligned} \tag{8.3}$$

where $\hat{\mu}_{l_1}(s)$ is the root x with the absolute value less than one of the quadratic equation $x = \mu_{l_1}/(\mu_{l_1} + s + \lambda_{l_1}(1 - x))$. This readily gives:

$$\hat{\mu}_{l_1}(s) = \frac{1}{2\lambda_{l_1}} \cdot \left(\lambda_{l_1} + 2\mu_{l_1} + s - \sqrt{(\lambda_{l_1} + 2\mu_{l_1} + s)^2 - 8\mu_{l_1}\lambda_{l_1}} \right).$$

For notational convenience, we introduce the following definitions:

$$\begin{aligned} \xi^{l*} &:= \xi^l + \sum_{j \neq l_1} \lambda_j (1 - z_j), \\ V^l(\mathbf{z}) &:= -\lambda_{l_1} z_{l_1}^2 + (2\mu_{l_1} + \lambda_{l_1} + \xi^{l*}) \cdot z_{l_1} - 2\mu_{l_1}. \end{aligned}$$

Then, we can present the following theorem which establishes the relation between $\beta^l(\mathbf{z})$ and $\alpha^l(\mathbf{z})$. The proof of the theorem will be given in Sect. 8.3.1.2.

Theorem 8.1 (Two servers visit the same queue)**.**

$$\beta^l(\boldsymbol{z}) = d_1(\boldsymbol{z}) \cdot (\alpha^l(\boldsymbol{z}) - \alpha^l(\boldsymbol{z}^*_{l_1})) + d_2(\boldsymbol{z}) \cdot \alpha^l(\boldsymbol{z}^*_{l_1}), \tag{8.4}$$

where

$$\begin{aligned} d_1(\boldsymbol{z}) &= \frac{\xi^l}{V^l(\boldsymbol{z})} \cdot z_{l_1}, \\ d_2(\boldsymbol{z}) &= \frac{\xi^l}{V^l(\boldsymbol{z})} \cdot \frac{(z_{l_1} - \hat{\mu}_{l_1}(\xi^{l*}))(2\mu_{l_1} + (\xi^{l*} + \lambda_{l_1}) z_{l_1} \hat{\mu}_{l_1}(\xi^{l*}))}{(1 - \hat{\mu}_{l_1}(\xi^{l*}))(2\mu_{l_1} + (\xi^{l*} + \lambda_{l_1})\hat{\mu}_{l_1}(\xi^{l*}))}, \end{aligned} \tag{8.5}$$

and $\alpha^l(\boldsymbol{z}^*_{l_1}) := \alpha^l(z_1, \ldots, z_{l_1 - 1}, \hat{\mu}_{l_1}(\xi^{l*}), z_{l_1 + 1}, \ldots, z_M)$.

Remark 8.2 (Single-server model)**.** *The final expression Eq. (8.4) strongly resembles the expression for the single-server model (see Remark 5.4 with $r^{l_1}(\boldsymbol{z}) = 1$). In fact, the shape of the expressions is identical (i.e., neglecting the change from μ_{l_1} to $2\mu_{l_1}$) except for the extra ratio that appears in Eq. (8.5), viz.:*

$$\frac{2\mu_{l_1} + (\xi^{l*} + \lambda_{l_1}) z_{l_1} \hat{\mu}_{l_1}(\xi^{l*})}{2\mu_{l_1} + (\xi^{l*} + \lambda_{l_1}) \hat{\mu}_{l_1}(\xi^{l*})}.$$

8.3.1.2 Proof of Theorem 8.1

We prove the expression for $\beta^l(\mathbf{z})$ as given in Theorem 8.1 along the same lines as we proved the corresponding theorem for the single-server case (see Sect. 5.3). That means that we derive first the conditional p.g.f. $\beta^l_{\mathbf{n}}(\mathbf{z}) := \mathbb{E}[\mathbf{z}^{\mathbf{N}^e_l} | \mathbf{N}^s_l = \mathbf{n}]$ and then uncondition on $\mathbf{N}^s_l$, the number of customers present at the entrance of server-location state l.

The final expression for $\beta^l_{\mathbf{n}}(\mathbf{z})$ is given in the following lemma.

Lemma 8.3.

$$\beta^l_{\boldsymbol{n}}(\boldsymbol{z}) = \xi^l \cdot \pi_{n_{l_1}}(z_{l_1}, \xi^{l*}) \cdot \prod_{j \neq l_1} z_j^{n_j}. \tag{8.6}$$

Proof. Let $A_{l,j}(t)$ denote the number of arrivals to Q_j during a residence in server-location state l. Starting from the definition of the p.g.f., we condition on the exponential timer:

$$\begin{aligned}
\beta^l_{\mathbf{n}}(\mathbf{z}) &= \sum_{m_1=0}^{\infty} \cdots \sum_{m_M=0}^{\infty} z_1^{m_1} \cdots z_M^{m_M} \mathbb{P}(\mathbf{N}^e_l = \mathbf{m} | \mathbf{N}^s_l = \mathbf{n}) \\
&= \int_0^{\infty} \xi^l e^{-\xi^l t} \sum_{m_1=0}^{\infty} \cdots \sum_{m_M=0}^{\infty} z_1^{m_1} \cdots z_M^{m_M} \mathbb{P}(\mathbf{N}_l(t) = \mathbf{m} | \mathbf{N}_l(0) = \mathbf{n}) dt.
\end{aligned}$$

After some simple rearrangements and using that given t the queue-length process at Q_{l_1} is independent of the aggregate arrival process to the other queues, we obtain the following:

$$\begin{aligned}
\int_0^{\infty} & \xi^l e^{-\xi^l t} \sum_{m_1=0}^{\infty} \cdots \sum_{m_M=0}^{\infty} z_1^{m_1 - n_1} \cdots z_M^{m_M - n_M} \\
& \times \mathbb{P}(\{A_{l,j}(t) = m_j - n_j,\ \forall_{j \neq i}\} | \mathbf{N}_l(0) = \mathbf{n}) \\
& \times \sum_{m_{l_1}} z_{l_1}^{m_{l_1}} \mathbb{P}(N_{l,l_1}(t) = m_{l_1} | \mathbf{N}_l(0) = \mathbf{n})\ dt \cdot \prod_{j \neq i} z_j^{n_j}.
\end{aligned}$$

Noting that $N_{l,l_1}(t)$ depends only on $\mathbf{N}_l(0)$ through $N_{l,l_1}(0)$, we retrieve $p_{n_{l_1} m_{l_1}}(t)$

and eventually find that $\beta^l_{\mathbf{n}}(\mathbf{z})$ equals:

$$\int_0^\infty \xi^l e^{-\xi^{l*}t} \sum_{m_{l_1}=0}^{\infty} z_{l_1}^{m_{l_1}} p_{n_{l_1} m_{l_1}}(t)dt \cdot \prod_{j \neq l_1} z_j^{n_j}.$$

Then, using the definition of (8.2), the expression for $\beta^l_{\mathbf{n}}(\mathbf{z})$ is readily obtained. □

Proof of Theorem 8.1. Essentially, the proof follows immediately by unconditioning $\beta^l_{\mathbf{n}}(\mathbf{z})$ on the state $\mathbf{n} = (n_1, \ldots, n_M)$ at the entrance of server-location state l. The result of this operation is shown below. Equation (8.7) follows by substitution of Eq. (8.6) into the definition of $\beta^l(\mathbf{z})$. We note that the final expression, Eq. (8.8), follows from inserting the explicit expression for $\pi_h(z_{l_1}, \xi^{l*}),\ h \geq 0$, which is given in Eq. (8.3), and some simple manipulations.

$$\begin{aligned}
\beta^l(\mathbf{z}) &= \sum_{n_1=0}^{\infty} \cdots \sum_{n_M=0}^{\infty} \beta^l_{\mathbf{n}}(\mathbf{z}) \mathbb{P}(\mathbf{N}^s_l = \mathbf{n}) \\
&= \sum_{n_1=0}^{\infty} \cdots \sum_{n_M=0}^{\infty} \mathbb{P}(\mathbf{N}^s_l = \mathbf{n}) \cdot \prod_{j \neq l_1} z_j^{n_j} \cdot \xi^l \cdot \pi_{n_{l_1}}(z_{l_1}, \xi^{l*}) \qquad (8.7) \\
&= \sum_{n_1=0}^{\infty} \cdots \sum_{n_M=0}^{\infty} \mathbb{P}(\mathbf{N}^s_l = \mathbf{n}) \cdot \prod_{j \neq l_1} z_j^{n_j} \\
&\quad \times \frac{\xi^l}{-\lambda_{l_1} z_{l_1}^2 + (2\mu_{l_1} + \lambda_{l_1} + \xi^{l*}) \cdot z_{l_1} - 2\mu_{l_1}} \cdot \Bigg(z_{l_1}^{n_{l_1}+1} - \\
&\qquad \frac{(1 - z_{l_1})(2\mu_{l_1} + (\xi^{l*} + \lambda_{l_1}) z_{l_1})}{(1 - \hat{\mu}_{l_1}(\xi^{l*}))(2\mu_{l_1} + (\xi^{l*} + \lambda_{l_1})\hat{\mu}_{l_1}(\xi^{l*}))} \cdot \hat{\mu}_{l_1}(\xi^{l*})^{n_{l_1}+1} \Bigg) \\
&= \frac{\xi^l}{-\lambda_{l_1} z_{l_1}^2 + (2\mu_{l_1} + \lambda_{l_1} + \xi^{l*}) \cdot z_{l_1} - 2\mu_{l_1}} \cdot \Bigg(z_{l_1} \cdot \alpha^l(\mathbf{z}) - \\
&\qquad \frac{(1 - z_{l_1})(2\mu_{l_1} + (\xi^{l*} + \lambda_{l_1}) z_{l_1}) \cdot \hat{\mu}_{l_1}(\xi^{l*})}{(1 - \hat{\mu}_{l_1}(\xi^{l*}))(2\mu_{l_1} + (\xi^{l*} + \lambda_{l_1})\hat{\mu}_{l_1}(\xi^{l*}))} \cdot \alpha^l(\mathbf{z}^*_{l_1}) \Bigg), \qquad (8.8)
\end{aligned}$$

where $\alpha^l(\mathbf{z}^*_{l_1}) := \mathbb{E}[z_1^{N^s_{l,1}} \cdots \hat{\mu}_{l_1}(\xi^{l*})^{N^s_{l,l_1}} \cdots z_M^{N^s_{l,M}}]$. □

8.3.2 Two servers visit different queues

Next, we study the expression for the p.g.f. $\beta^l(\mathbf{z})$ for the scenario that the two servers visit a different queue, i.e., $l = (l_1, l_2)$ with $l_1 \neq l_2$. To this end, we will focus first on the conditional p.g.f. $\beta^l_{\mathbf{n}}(\mathbf{z}) := \mathbb{E}[\mathbf{z}^{\mathbf{N}^e_l} | \mathbf{N}^s_l = \mathbf{n}]$. Finally, we discuss how to uncondition on $\mathbf{N}^s_l$, the number of customers present at the entrance of the server-location state l.

8.3.2.1 The conditional p.g.f. $\beta^l_{\mathbf{n}}(\mathbf{z})$

We approach the analysis of the conditional p.g.f. $\beta^l_{\mathbf{n}}(\mathbf{z})$ also similar to the analysis of the single-server polling model (see Sect. 5.3). For convenience, let us first define ξ^{l*} as follows.

$$\xi^{l*} := \xi^l + \sum_{j \neq l_1, l_2} \lambda_j (1 - z_j).$$

Let further $A_{l,j}(t)$ denote the number of arrivals to Q_j during a residence in server-location state l. Starting from the definition of the p.g.f., we condition on the duration of the timer, which is exponentially distributed with rate ξ^l, yielding:

$$\begin{aligned}
\beta^l_{\mathbf{n}}(\mathbf{z}) &= \sum_{m_1=0}^{\infty} \cdots \sum_{m_M=0}^{\infty} z_1^{m_1} \cdots z_M^{m_M} \mathbb{P}(\mathbf{N}^e_l = \mathbf{m} | \mathbf{N}^s_l = \mathbf{n}) \\
&= \int_0^{\infty} \xi^l e^{-\xi^l t} \sum_{m_1=0}^{\infty} \cdots \sum_{m_M=0}^{\infty} z_1^{m_1} \cdots z_M^{m_M} \\
&\qquad \times \ \mathbb{P}(\mathbf{N}_l(t) = \mathbf{m} | \mathbf{N}_l(0) = \mathbf{n}) dt.
\end{aligned}$$

After some rearrangements and using that given t the queue-length processes at Q_{l_1} and Q_{l_2} are independent of the aggregate arrival process to the other queues, we obtain:

$$\begin{aligned}
\beta^l_{\mathbf{n}} = \int_0^{\infty} \xi^l e^{-\xi^l t} &\left(\sum_{m_l=0}^{\infty} z_l^{m_l - n_l} \right)_{l \neq l_1, l_2} \\
&\times \mathbb{P}(\{A_{l,j}(t) = m_j - n_j, \ \forall_{j \neq l_1, l_2}\} | \mathbf{N}_l(0) = \mathbf{n}) \cdot \prod_{j \neq l_1, l_2} z_j^{n_j} \\
&\times \sum_{m_{l_1}=0}^{\infty} z_{l_1}^{m_{l_1}} \mathbb{P}(N_{l,l_1}(t) = m_{l_1} | \mathbf{N}_l(0) = \mathbf{n}) \\
&\times \sum_{m_{l_2}=0}^{\infty} z_{l_2}^{m_{l_2}} \mathbb{P}(N_{l,l_2}(t) = m_{l_2} | \mathbf{N}_l(0) = \mathbf{n}) \ dt.
\end{aligned}$$

Noting that $N_{l,l_1}(t)$ and $N_{l,l_2}(t)$ depend only on $\mathbf{N}_l(0) = (N_{l,1}(0), \ldots, N_{l,M}(0))$ through $N_{l,l_1}(0)$ and $N_{l,l_2}(0)$, respectively, we retrieve $p_{n_{l_1} m_{l_1}}(t)$ and $p_{n_{l_2} m_{l_2}}(t)$ and find for $\beta^l_{\mathbf{n}}(\mathbf{z})$:

$$\beta^l_{\mathbf{n}}(\mathbf{z}) = \xi^l \int_0^{\infty} e^{-\xi^{l*} \cdot t} \sum_{m_{l_1}=0}^{\infty} z_{l_1}^{m_{l_1}} p_{n_{l_1} m_{l_1}}(t) \sum_{m_{l_2}=0}^{\infty} z_{l_2}^{m_{l_2}} p_{n_{l_2} m_{l_2}}(t) dt \cdot \prod_{j \neq l_1, l_2} z_j^{n_j}. \tag{8.9}$$

Unfortunately, we cannot readily simplify this expression much further as we could do for the single-server case or the case of the servers being at the same queue (see Sect. 8.3.1). This is due to the fact that the integrand involves now a product of two functions of t, whereas before it contained only a single function. Thus, we proceed by rewriting the expression to a complex integral using a property of the Laplace Transform (LT) (see, e.g., [62]). This property is in fact the counterpart of the common convolution in the time domain.

Property 8.4. *(Laplace Transform convolution in the frequency domain)*

$$\int_0^\infty e^{-st} f(t)g(t)dt = \frac{1}{2\pi I} \int_{c-I\cdot\infty}^{c+I\cdot\infty} F(p)G(s-p)dp, \tag{8.10}$$

where $F(\cdot)$ is the LT of $f(t)$ and $G(\cdot)$ of $g(t)$, I is the imaginary unit, and c is a real value which depends on the convergence areas of $F(\cdot)$ and $G(\cdot)$.

We note that with regard to Eq. (8.9) the transforms $F(\cdot)$ and $G(\cdot)$ are known; in fact, these were established already in the single-server analysis (see Thm. 5.3). More precisely, we have:

$$\int_0^\infty e^{-st} \sum_{m_{l_1}=0}^{\infty} z_{l_1}^{m_{l_1}} p_{n_{l_1} m_{l_1}}(t)dt = a_l(z_{l_1}, s) \cdot z_{l_1}^{n_{l_1}} + b_l(z_{l_1}, s) \cdot \hat{\mu}_{l_1}(s)^{n_{l_1}},$$

where,

$$\begin{aligned}
a_{l_1}(z_{l_1}, s) &= \frac{z_{l_1}}{z_{l_1} - \tilde{X_{l_1}}(s + \lambda_{l_1}(1 - z_{l_1}))} \cdot \frac{(1 - \tilde{X}_{l_1}(s + \lambda_{l_1}(1 - z_{l_1})))}{\lambda_{l_1}(1 - z_{l_1}) + s} \\
&= \frac{z_{l_1}}{(1 - z_{l_1})(\lambda_{l_1} z_{l_1} - \mu_{l_1}) + s \cdot z_{l_1}}, \\
b_{l_1}(z_{l_1}, s) &= \frac{z_{l_1} - 1}{z_{l_1} - \tilde{X}_{l_1}(s + \lambda_{l_1}(1 - z_{l_1}))} \cdot \frac{\tilde{X}_{l_1}(s + \lambda_{l_1}(1 - z_{l_1}))}{\lambda_{l_1}(1 - \hat{\mu}_{l_1}(s)) + s} \\
&= \frac{(z_{l_1} - 1)\mu_{l_1}}{((1 - z_{l_1})(\lambda_{l_1} z_{l_1} - \mu_{l_1}) + s \cdot z_{l_1})\,(\lambda_{l_1}(1 - \hat{\mu}_{l_1}(s)) + s)},
\end{aligned}$$

where X_{l_1} is $\exp(\mu_{l_1})$-distributed and:

$$\hat{\mu}_{l_1}(s) = \frac{1}{2\lambda_{l_1}} \cdot \left(\lambda_{l_1} + \mu_{l_1} + s - \sqrt{(\lambda_{l_1} + \mu_{l_1} + s)^2 - 4\lambda_{l_1}\mu_{l_1}}\right).$$

In particular, this single-server result implies for the two-server polling model that:

$$\int_0^\infty e^{-st} \sum_{m_{l_1}=0}^{\infty} z_{l_1}^{m_{l_1}} p_{n_{l_1} m_{l_1}}(t) \sum_{m_{l_2}=0}^{\infty} z_{l_2}^{m_{l_2}} p_{n_{l_2} m_{l_2}}(t)dt \tag{8.11}$$

$$= \frac{1}{2\pi I} \int_{c-I\cdot\infty}^{c+I\cdot\infty} \left(a_{l_1}(z_{l_1}, p) \cdot z_{l_1}^{n_{l_1}} + b_{l_1}(z_{l_1}, p) \cdot \hat{\mu}_{l_1}(p)^{n_{l_1}}\right) \times \left(a_{l_2}(z_{l_2}, s-p) \cdot z_{l_2}^{n_{l_2}} + b_{l_2}(z_{l_2}, s-p) \cdot \hat{\mu}_{l_2}(s-p)^{n_{l_2}}\right) dp. \tag{8.12}$$

Here, c is strictly positive and real, while p and s are complex with $Re(s) > c$.

Hence, we have essentially rewritten the original expression for $\beta^l_{\mathbf{n}}(\mathbf{z})$ in terms of a complex integral by exploiting Property 8.4. Next, we will study the roots that appear in these expressions, since the roots typically play a crucial role in the evaluation of complex integrals.

Evaluation of the roots For convenience, let us introduce the following definition:

$$P_{l_1}(z_{l_1}, p) := a_{l_1}(z_{l_1}, p) \cdot z_{l_1}^{n_{l_1}} + b_{l_1}(z_{l_1}, p) \cdot \hat{\mu}_{l_1}(p)^{n_{l_1}}.$$

Noting that $\hat{\mu}_{l_1}(p)$ is a solution of the equation $x = \tilde{X}_{l_1}(s+\lambda_{l_1}(1-x))$ and performing some simple calculations, we may readily rewrite $P_{l_1}(z_{l_1}, p)$ to (cf. [23, p.80]):

$$P_{l_1}(z_{l_1}, p) = \frac{(1 - z_{l_1})\hat{\mu}_{l_1}(p)^{n_{l_1}+1} - (1 - \hat{\mu}_{l_1}(p))z_{l_1}^{n_{l_1}+1}}{\{\lambda_{l_1} z_{l_1}^2 - (\mu_{l_1} + \lambda_{l_1} + p)z_{l_1} + \mu_{l_1}\}(1 - \hat{\mu}_{l_1}(p))}, \tag{8.13}$$

where the root $\hat{\mu}_{l_1}(p)$ satisfies $|\hat{\mu}_{l_1}(p)| < 1$, for $Re(p) > 0$.

It is important to notice that the transformation of Eq. (8.11) now leads to complex integration over p. Hence, we focus here on the roots of $P_{l_1}(z_{l_1}, p)$ as function of p. We note that $P_{l_1}(z_{l_1}, p)$ is an analytic function for $|z_{l_1}| < 1$ and $Re(p) > 0$. This means that a root of the denominator of Eq. (8.13) on this analytic domain should also let the numerator of $P_{l_1}(p, \rho)$ vanish. Let us next consider when the denominator of Eq. (8.13) evaluates to zero. It follows that we must solve for p :

$$\lambda_{l_1} z_{l_1}^2 - (\mu_{l_1} + \lambda_{l_1} + p)z_{l_1} + \mu_{l_1} = 0, \tag{8.14}$$

and

$$1 - \hat{\mu}_{l_1}(p) = 0. \tag{8.15}$$

The first equality, Eq. (8.14), is solved for $p = p_{*,1}$ with

$$p_{*,1} = \frac{(\mu_1 - \lambda_{l_1} \cdot z_{l_1})(1 - z_{l_1})}{z_{l_1}},$$

where it is good to notice that $Re(p_{*,1})$ is not necessarily strictly positive for $|z_{l_1}| < 1$. Next, we check what happens if we insert this solution into the denominator of Eq. (8.13). Inserting $p = p_{*,1}$ into $\hat{\mu}_{l_1}(p)$ gives:

$$\begin{aligned}
\hat{\mu}_{l_1}(p_{*,1}) &= \frac{\mu_{l_1} + \lambda_{l_1} + p_{*,1}}{2\lambda_{l_1}} - \frac{1}{2\lambda_{l_1}}\sqrt{(\mu_{l_1} + \lambda_{l_1} + p_{*,1})^2 - 4\lambda_{l_1}\mu_{l_1}} \\
&= \frac{\mu_{l_1} + \lambda_{l_1} \cdot z_{l_1}^2}{2\lambda_{l_1} \cdot z_{l_1}} - \frac{1}{2\lambda_{l_1}}\sqrt{\left(\frac{\mu_{l_1} + \lambda_{l_1} \cdot z_{l_1}^2}{z_{l_1}}\right)^2 - 4\lambda_{l_1}\mu_{l_1}} \\
&= \frac{\mu_{l_1} + \lambda_{l_1} \cdot z_{l_1}^2}{2\lambda_{l_1} \cdot z_{l_1}} - \frac{1}{2\lambda_{l_1} z_{l_1}}\sqrt{(\mu_{l_1} - \lambda_{l_1} \cdot z_{l_1}^2)^2}.
\end{aligned} \tag{8.16}$$

The square root that appears in Eq. (8.16) should be chosen such that $Re(p_{*,1}) > 0$ and $|z_{l_1}| < 1$ are satisfied, i.e.,

$$\sqrt{(\mu_{l_1} - \lambda_{l_1} \cdot z_{l_1}^2)^2} = \mu_{l_1} - \lambda_{l_1} \cdot z_{l_1}^2.$$

Substituting this result into Eq. (8.16) gives that:

$$\hat{\mu}_{l_1}(p_{*,1}) = z_{l_1}.$$

It is then readily seen that for $\hat{\mu}_{l_1}(p_{*,1}) = z_{l_1}$ the numerator of Eq. (8.13) indeed evaluates to zero. The second equality, Eq. (8.15), is solved for $p = p_{*,2}$ with

$$\hat{\mu}_{l_1}(p_{*,2}) = 1.$$

It follows that for $p_{*,2} = 0$ (which forces $\hat{\mu}_{l_1}(p_{*,2}) = 1$) the numerator of Eq. (8.13) does not evaluate to zero. In other words, $p = 0$ is the only irremovable pole of $P_{l_1}(z_{l_1}, p)$. Analogously, we may find for $P_{l_2}(z_{l_2}, s - p)$ that $p = s$ is the only irremovable pole in the complex plane $\mathbb{C}$.

Evaluation of the complete integral We have considered the roots of the integrand of Eq. (8.12). However, this complex integrand contains also a square root which depends on the integration parameter p. This square root typically yields branch points which are the points ρ^* where the square root in $\hat{\mu}_{l_1}(p)$ evaluates to zero, i.e.,

$$\sqrt{(\mu_{l_1} + \lambda_{l_1} + \rho^*)^2 - 4\lambda_{l_1}\mu_{l_1}} = 0.$$

Solving the equation $(\mu_{l_1} + \lambda_{l_1} + x)^2 - 4\lambda_{l_1}\mu_{l_1} = 0$ for x readily gives:

$$x = -(\lambda_{l_1} + \mu_{l_1}) \pm 2\sqrt{\lambda_{l_1}\mu_{l_1}}.$$

This shows that the branch points related to $P_{l_1}(p, \rho)$, which we denote by $b_{l_1,1}$ and $b_{l_1,2}$, are real and negative, i.e.,

$$\begin{aligned} b_{l_1,1} &= -(\lambda_{l_1} + \mu_{l_1}) - 2\sqrt{\lambda_{l_1}\mu_{l_1}}, \\ b_{l_1,2} &= -(\lambda_{l_1} + \mu_{l_1}) + 2\sqrt{\lambda_{l_1}\mu_{l_1}}, \end{aligned}$$

whereas the branch points related to $P_{l_2}(p, s-p)$, which we denote by $b_{l_2,1}$ and $b_{l_2,2}$, are given by:

$$\begin{aligned} b_{l_2,1} &= s - \left(-(\lambda_{l_2} + \mu_{l_2}) + 2\sqrt{\lambda_{l_2}\mu_{l_2}}\right) = s + (\lambda_{l_2} + \mu_{l_2}) - 2\sqrt{\lambda_{l_2}\mu_{l_2}}, \\ b_{l_2,2} &= s - \left(-(\lambda_{l_2} + \mu_{l_2}) - 2\sqrt{\lambda_{l_2}\mu_{l_2}}\right) = s + (\lambda_{l_2} + \mu_{l_2}) + 2\sqrt{\lambda_{l_2}\mu_{l_2}}. \end{aligned}$$

Since s is complex with $Re(s) > c > 0$, the branch points $b_{l_2,1}$ and $b_{l_2,2}$ lie in the right half-plane (with respect to $p = c$) and might well be strictly complex.

Computation of the complete integral The problem of finding $\beta^l_{\mathbf{n}}(\mathbf{z})$ is reduced to calculating the complex integral over the line C_0 from $c - I \cdot W$ to $c + I \cdot W$, for $W \to \infty$. To facilitate this computation, we will construct a closed contour and evaluate the integral over this contour. We construct this contour by adding a curve C_1 in the left half-plane (w.r.t. $p = c$) which connects the endpoints of C_0 (see Fig. 8.1). For clearness of presentation, let us define $f(p) := P_{l_1}(z_{l_1}, p) \cdot P_{l_2}(z_{l_2}, s-p)$.

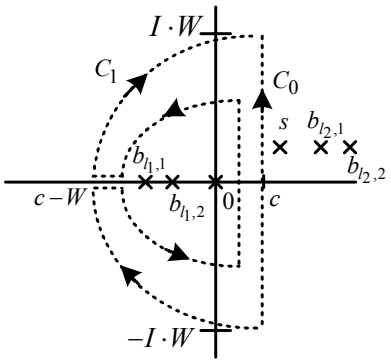

Figure 8.1: Sketch of the contour and extreme points in the complex plane.

Notice that $f(p)$ is analytic in the interior of the closed contour formed by C_0 and C_1. Thus, it follows from Cauchy's Integral Theorem that:

$$\int_{C_0} f(p)dp = \int_{C_1} f(p)dp.$$

The contour C_1 can be split into several sub-contours as shown in Fig. 8.2. Accordingly, we can write for the integral over C_1:

$$\begin{aligned}\int_{C_1} f(p)dp &= \int_{C_{1a}} f(p)dp + \int_{C_{1b}} f(p)dp + \int_{C_{1c}} f(p)dp \\ &\quad + \int_{C_{1d}} f(p)dp + \int_{C_{1e}} f(p)dp.\end{aligned}$$

It can readily be observed that the integrals over C_{1b} and C_{1d} cancel out, since these lie at the same Riemann surface. Hence, we redefine the contours as shown in Fig. 8.3 with C_3 being a closed contour of finite length which encloses the branch points $b_{l_1,1}$ and $b_{l_1,2}$, and the pole at $p = 0$. This leads to the equation:

$$\int_{C_0} f(p)dp = \int_{C_2} f(p)dp + \int_{C_3} f(p)dp.$$

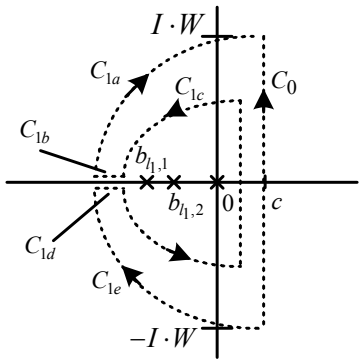

Figure 8.2: Sketch of the sub-contours of C_1.

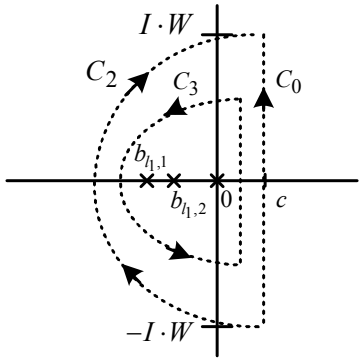

Figure 8.3: Sketch of the adjusted sub-contours of C_1.

The integral over C_2 will vanish as W goes to infinity. To see this, let us rewrite $\hat{\mu}_{l_1}(p)$ as follows:

$$\begin{aligned}
\hat{\mu}_{l_1}(p) &= \frac{1}{2\lambda_{l_1}} \cdot \left(\lambda_{l_1} + \mu_{l_1} + p - \sqrt{(\lambda_{l_1} + \mu_{l_1} + p)^2 - 4\lambda_{l_1}\mu_{l_1}}\right) \\
&= \frac{1}{2\lambda_{l_1}} \cdot \left(\lambda_{l_1} + \mu_{l_1} + p - \sqrt{(\lambda_{l_1} + \mu_{l_1} + p)^2 - 4\lambda_{l_1}\mu_{l_1}}\right) \\
&\quad \times \frac{\lambda_{l_1} + \mu_{l_1} + p + \sqrt{(\lambda_{l_1} + \mu_{l_1} + p)^2 - 4\lambda_{l_1}\mu_{l_1}}}{\lambda_{l_1} + \mu_{l_1} + p + \sqrt{(\lambda_{l_1} + \mu_{l_1} + p)^2 - 4\lambda_{l_1}\mu_{l_1}}} \\
&= \frac{1}{2\lambda_{l_1}} \cdot \frac{4\lambda_{l_1}\mu_{l_1}}{\lambda_{l_1} + \mu_{l_1} + p + \sqrt{(\lambda_{l_1} + \mu_{l_1} + p)^2 - 4\lambda_{l_1}\mu_{l_1}}}.
\end{aligned}$$

This demonstrates that for $|p| \to \infty$, $\hat{\mu}_{l_1}(p) = O(|p|^{-1})$. It can be observed (see Eq. (8.13)) that also both $P_{l_1}(z_{l_1}, p)$ and $P_{l_2}(z_{l_2}, s - p) = O(|p|^{-1})$. Thus, it follows

that:

$$\int_{C_2} f(p)dp \leq \left| \int_{C_2} f(p)dp \right| \leq \pi W \cdot \frac{c}{W^2} = O(W^{-1}) \to 0, \text{ for } W \to \infty,$$

for some constant $c > 0$.

Hence, we eventually find for the integral over C_0:

$$\int_{C_0} f(p)dp = \int_{C_3} f(p)dp. \tag{8.17}$$

Combining Eqs. (8.9), (8.10) and (8.17) provides us with the final expression for $\beta^l_{\mathbf{n}}(\mathbf{z})$:

$$\beta^l_{\mathbf{n}}(\mathbf{z}) = \xi^{*l} \cdot \left(\frac{1}{2\pi I} \cdot \int_{C_3} P_{l_1}(z_{l_1}, p) \cdot P_{l_2}(z_{l_2}, \xi^{*l} - p)dp \right) \cdot \prod_{j \neq l_1, l_2} z_j^{n_j},$$

where $P_k(z_k, r)$ is given in Eq. (8.13).

8.3.2.2 The unconditional p.g.f. $\beta^l(\mathbf{z})$

The unconditional p.g.f. of the joint queue-length at departures epochs from a server-location state is given by:

$$\begin{aligned}
\beta^l(\mathbf{z}) &= \sum_{n_1=0}^{\infty} \cdots \sum_{n_M=0}^{\infty} \beta^l_{\mathbf{n}}(\mathbf{z}) \mathbb{P}(\mathbf{N}^s_l = \mathbf{n}) \\
&= \sum_{n_1=0}^{\infty} \cdots \sum_{n_M=0}^{\infty} \xi^{*l} \cdot \left(\frac{1}{2\pi I} \cdot \int_{C_3} P_{l_1}(z_{l_1}, p) \cdot P_{l_2}(z_{l_2}, \xi^{*l} - p)dp \right) \\
&\quad \times \prod_{j \neq l_1, l_2} z_j^{n_j} \mathbb{P}(\mathbf{N}^s_l = \mathbf{n}).
\end{aligned} \tag{8.18}$$

We note that the contour C_3 is of finite length and it can be constructed such that the function $P_{l_1}(z_{l_1}, p) \cdot P_{l_2}(z_{l_2}, \xi^{*l} - p)$ is finite everywhere on C_3. More specifically, by choosing C_3 sufficiently distant from the extreme points, so that $|\hat{\mu}_{l_1}(p)| < 1$ on C_3. This is feasible since $\hat{\mu}_{l_1}(p) = O(|p|^{-1})$. In particular, this allows to interchange summation and integration in Eq. (8.18), so that we may write:

$$\begin{aligned}
\beta^l(\mathbf{z}) &= \frac{\xi^{*l}}{2\pi I} \cdot \Bigg(\int_{C_3} a(z_{l_1}, p) a_{l_2}(z_{l_2}, \xi^{*l} - p) \cdot \alpha^l(\mathbf{z})dp \\
&\quad + \int_{C_3} a(z_{l_1}, p) \cdot b_{l_2}(z_{l_2}, \xi^{*l} - p) \cdot \alpha^l(\mathbf{z})|_{z_{l_2} = \hat{\mu}_{l_2}(\xi^{*l} - p)} dp \\
&\quad + \int_{C_3} b(z_{l_1}, p) \cdot a_{l_2}(z_{l_2}, \xi^{*l} - p) \cdot \alpha^l(\mathbf{z})|_{z_{l_1} = \hat{\mu}_{l_1}(p)} dp \\
&\quad + \int_{C_3} b(z_{l_1}, p) \cdot b_{l_2}(z_{l_2}, \xi^{*l} - p) \cdot \alpha^l(\mathbf{z})|_{z_{l_1} = \hat{\mu}_{l_1}(p), z_{l_2} = \hat{\mu}_{l_2}(\xi^{*l} - p)} dp \Bigg).
\end{aligned} \tag{8.19}$$

The complex integrals in Eq. (8.19) can be computed along standard numerical evaluation methods (see, e.g., [100]).

8.4 Concluding remarks

We have presented a transient analysis for a two-server polling model operating under the pure exponential time-limited discipline. Particularly, we have focussed on the relation that describes the evolution of the joint queue length between the entrance and departure epochs of a specific server-location state. Two scenarios have been studied requiring both a different analysis. The first scenario, referring to two servers being at the same queue, is analyzed by exploiting transient results of the M/M/2 queue. The second scenario, with the servers being at different queues, is numerically resolved via a subtle transformation to a complex integration problem. As a result, for this scenario, the exact form of the key relation, Eq. (8.1), is not as explicit as for the single-server model as one has to resort to numerical evaluation techniques.

The extension of the presented approach to a larger number of servers is not straightforward for the scenario with the servers being at different queues. More concretely, consider a polling system with three servers and all of them being at different queues. In such case, we can no longer appeal to Proposition 8.4 to transform our problem into a complex integration problem. Nevertheless, the other, coupled-servers, scenario appears still amenable for a tractable analysis, since one may appeal to known transient results for the M/M/c queue, $c \geq 3$ (see, e.g., [57, 90]). In particular, this implies that our analytical approach is applicable to polling systems with a finite number of coupled servers which operate under the pure exponential time-limited discipline.

SELF-REFERENCES

[H1] A. Al Hanbali, R. de Haan, R. J. Boucherie, and J. C. W. van Ommeren. Delay in a tandem queueing model with mobile queues: Two analytical approximations, Working paper, 2009.

[H2] A. Al Hanbali, R. de Haan, R. J. Boucherie, and J.-K. van Ommeren. A tandem queueing model for delay analysis in disconnected ad hoc networks. In *Proc. of ASMTA*, Nicosia, Cyprus, 2008.

[H3] A. Al Hanbali, R. de Haan, R. J. Boucherie, and J.-K. van Ommeren. Time-limited and k-limited polling systems: A matrix geometric solution. In *Proc. of SMCTools*, Athens, Greece, 2008.

[H4] M. de Graaf et al. Advances in emergency networking. In *Proc. of IEEE Conference on Wireless Rural and Emergency Communications*, Rome, Italy, 2007.

[H5] M. de Graaf et al. Easy wireless: broadband ad-hoc networking for emergency services. In *Proc. of MedHoc*, Corfu, Greece, 2007.

[H6] R. de Haan, R. J. Boucherie, and J.-K. van Ommeren. The impact of interference on optimal multi-path routing in ad hoc networks. In *Proc. of ITC*, Ottawa, Canada, 2007.

[H7] R. de Haan, R. J. Boucherie, and J.-K. van Ommeren. A polling model with an autonomous server, Research Memorandum 1845, University of Twente, 2007.

REFERENCES

[1] M. Abolhasan, T. Wysocki, and E. Dutkiewicz. A review of routing protocols for mobile ad hoc networks. *Ad Hoc Networks*, 2:1–22, 2004.

[2] M. Ajmone Marsan, L. F. de Moraes, S. Donatelli, and F. Neri. Analysis of symmetric nonexhaustive polling with multiple servers. In *Proc. of Infocom*, San Francisco, CA, United States, 1990.

[3] G. Anastasi, M. Conti, M. D. Francesco, and A. Passarella. Energy conservation in wireless sensor networks: A survey. *Ad Hoc Networks*, 7(3):537–568, 2009.

[4] M. Balazinska and P. Castro. Characterizing mobility and network usage in a corporate wireless local-area network. In *Proc. of MobiSys*, San Francisco, CA, United States, 2003.

[5] N. Bansal and Z. Liu. Capacity, delay and mobility in wireless ad-hoc networks. In *Proc. of IEEE Infocom*, San Francisco, United States, Apr. 2003.

[6] L. N. Bhuyan, D. Ghosal, and Q. Yang. Approximation analysis of single and multiple ring networks. *IEEE Trans. on Computers*, 38(7):1027–1040, 1989.

[7] G. Bianchi. Performance analysis of the IEEE 802.11 distributed coordination function. *IEEE Journal on Selected Areas in Communications*, 18(3):535–547, 2000.

[8] N. Bisnik and A. A. Abouzeid. Queueing network models for delay analysis of multihop wireless ad hoc networks. *Ad Hoc Networks*, 7(1):79–97, 2009.

[9] J. Blanc. A numerical approach to cyclic-service queueing models. *Queueing Systems*, 6(1):173–188, 1990.

[10] J. Blum, A. Eskandarian, and L. J. Hoffman. Performance characteristics of inter-vehicle ad hoc networks. In *Proc. of IEEE Intl. Conf. on Intelligent Transportation Systems*, Shanghai, China, 2003.

[11] C. Boldrini, M. Conti, and A. Passarella. Users mobility models for opportunistic networks: the role of physical locations. In *Proc. of WRECOM*, Rome, Italy, 2007.

[12] S. Borst, O. Boxma, and M. Combe. Collection of customers: a correlated M/G/1 queue. In *Proc. of ACM Sigmetrics/Performance*, Newport, RI, United States, 1992.

[13] S. C. Borst. A polling system with a dormant server. Report BS-R9313, Centrum voor Wiskunde en Informatica, Amsterdam, The Netherlands, 1993.

[14] S. C. Borst. Polling systems with multiple coupled servers. *Queueing Systems*, 20(3-4):369–393, 1995.

[15] S. C. Borst and R. D. van der Mei. Waiting time approximations for multiple-server polling systems. *Performance Evaluation*, 31:163–182, 1998.

[16] O. J. Boxma, S. Schlegel, and U. Yechiali. A note on an M/G/1 queue with a waiting server, timer and vacations. *Analytic Methods in Applied Probability: In Memory of Fridrikh Karpelevich, American Mathematical Society Translations*, 2(207):25–35, 2002.

[17] O. J. Boxma, S. Schlegel, and U. Yechiali. Two-queue polling models with a patient server. *Annals of Operations Research*, 112:101–121, 2002.

[18] O. J. Boxma and J. A. Weststrate. Waiting times in polling systems with Markovian routing. In *Messung, Modellierung und Bewertung von Rechensystemen und Netzen, eds. G. Stiege and J.S. Lie (Springer Verlag)*, pages 89–104, Berlin, 1989.

[19] S. Browne and O. Kella. Parallel service with vacations. *Operations Research*, 43(5):870–878, 1995.

[20] S. Browne and G. Weiss. Dynamic priority rules when polling with multiple parallel servers. *Operations Research Letters*, 12:129–137, 1992.

[21] I. Chlamtac, M. Conti, and J. J.-N. Liu. Mobile ad hoc networking: imperatives and challenges. *Ad Hoc Networks*, 1(1):13–64, 2003.

[22] E. G. Coffman, G. Fayolle, and I. Mitrani. Two queues with alternating service periods. In *Performance '87: Proc. of the 12th IFIP WG 7.3 International Symposium on Computer Performance Modelling, Measurement and Evaluation*, pages 227–239, 1988.

[23] J. W. Cohen. *The Single Server Queue.* Elsevier Science Publishers, second edition, 1992.

[24] M. Conti and S. Giordano. Multihop ad hoc networking: The theory, IEEE Communications Magazine, April, 2007.

[25] R. Cooper and G. Murray. Queues served in cyclic order. *The Bell System Technical Journal*, 48(3):675–689, 1969.

[26] Delay Tolerant Networking Research Group. Website: http://www.dtnrg.org.

[27] M. Dow. Explicit inverses of Toeplitz and associated matrices. *ANZIAM J.*, 44:185–215, 2003.

[28] M. Eisenberg. Two queues with changeover times. *Operations Research*, 19(2):386–401, 1971.

[29] M. Eisenberg. Queues with periodic service and changeover times. *Operations Research*, 20(2):440–451, 1972.

[30] F. Ekman, A. Keränen, J. Karvo, and J. Ott. Working day movement model. In *Proc. of MobilityModels*, Hong Kong, China, 2008.

[31] I. Eliazar and U. Yechiali. Polling under the randomly timed gated regime. *Stochastic Models*, 14(1-2):79–93, 1998.

[32] I. Eliazar and U. Yechiali. Randomly timed gated queueing systems. *SIAM Journal of Applied Mathematics*, 59(2):423–441, 1998.

[33] W. Enkelmann. Fleetnet - applications for inter-vehicle communication. In *Proc. of IEEE Intelligent Vehicles Symposium*, Columbus, OH, United States, 2003.

[34] G. Fayolle and R. Iasnogorodski. Two coupled processors: The reduction to a Riemann-Hilbert problem. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47:325–351, 1979.

[35] W. Feller. *An Introduction to Probability Theory and Its Applications, Volume II.* John Wiley & Sons, 1966.

[36] M. J. Ferguson and Y. J. Aminetzah. Exact results for nonsymmetric token ring systems. *IEEE Trans. on Communications*, 33(3):223–231, 1985.

[37] C. Fricker and M. R. Jaïbi. Monotonicity and stability of periodic polling systems. *Queueing Systems*, 15(1-4):211–238, 1994.

[38] C. Fricker and M. R. Jaïbi. Stability of a polling model with a Markovian scheme. INRIA report 2278, 1994.

[39] I. Frigui and A.-S. Alfa. Analysis of a time-limited polling system. *Computer Communications*, 21(6):558–571, 1998.

[40] S. W. Fuhrmann. A decomposition result for a class of polling models. *Queueing Systems*, 11(1-2):109–120, 1992.

[41] S. W. Fuhrmann and R. B. Cooper. Stochastic decompositions in the M/G/1 queue with generalized vacations. *Operations Research*, 33(5):1117–1129, 1985.

[42] D. P. Gaver. A waiting line with interrupted service, including priorities. *Journal of Royal Statistical Society*, 24(1):73–90, 1962.

[43] D. P. Gaver, P. A. Jacobs, and G. Latouche Finite Birth-and-Death Models in Randomly Changing Environments. *Advances in Applied Probability*, 16:715–731, 1984.

[44] R. Groenevelt, P. Nain, and G. Koole. Message delay in mobile ad hoc networks. In *Proc. of Performance*, Juan-les-Pins, France, 2005.

[45] M. Grossglauser and D. Tse. Mobility increases the capacity of ad-hoc wireless networks. In *Proc. of IEEE Infocom*, Anchorage, AK, United States, 2001.

[46] P. Gupta and P. R. Kumar. The capacity of wireless networks. *IEEE Trans. on Information Theory*, 46(2):388–404, 2000.

[47] R. Gupta, J. Musacchio, and J. Walrand. Sufficient rate constraints for QoS flows in ad-hoc networks, UCB/ERL Technical Memorandum M04/42, 2004.

[48] T. Henderson, D. Kotz, and I. Abyzov. The changing usage of a mature campus-wide wireless network. In *Proc. of MobiCom*, Philadelphia, PA, United States, 2004.

[49] P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, and C. Diot. Pocket switched networks and human mobility in conference environments. In *Proc. of ACM SIGCOMM first workshop on delay tolerant networking and related topics*, Philadelphia, PA, United States, 2005.

[50] J. R. Jackson. Networks of waiting lines. *Operations Research*, 5(4):518–521, 1957.

[51] K. Jain, J. Padhye, V. Padmanabhan, and L. Qiu. Impact of interference on multi-hop wireless network performance. In *Proc. of Mobicom*, San Diego, CA, United States, 2003.

[52] C. E. Jones, K. M. Sivalingam, P. Agrawal, and J. C. Chen. A survey of energy efficient network protocols for wireless networks. *Wireless Networks*, 7:343–358, 2001.

[53] P. Juang, H. Oki, Y. Wang, M. Martonosi, L.-S. Peh, and D. Rubinstein. Energy-efficient computing for wildlife tracking: Design tradeoff and early experiences with Zebranet. In *Proc. of ASPLOS*, San Jose, CA, United States, 2002.

[54] T. Katayama. Waiting time analysis for a queueing system with time-limited service and exponential timer. *Naval Research Logistics*, 48:638–651, 2001.

[55] J. Keilson and L. D. Servi. A distributional form of Little's law. *Operations Research Letters*, 7(5):223–227, 1988.

[56] F. Kelly. *Reversibility and Stochastic Networks*. John Wiley and Sons Ltd, first edition, 1979.

[57] W. D. Kelton and A. M. Law. The transient behavior of the M/M/s queue, with implications for steady-state simulation. *Operations Research*, 33(2):378–396, 1985.

[58] M. Kodialam and T. Nandagopal. Characterizing achievable rates in multi-hop wireless networks: The joint routing and scheduling problem. In *Proc. of MobiCom*, San Diego, CA, United States, 2003.

[59] A. G. Konheim, H. Levy, and M. M. Srinivasan. Descendant set: An efficient approach for the analysis of polling systems. *IEEE Trans. on Communications*, 42(2/3/4):1245–1253, 1994.

[60] S. Kumar, V. S. Raghavan, and J. Deng. Medium Access Control protocols for ad hoc wireless networks: A survey. *Ad Hoc Networks*, 4(3):326–358, 2006.

[61] LAN MAN Standards Committee of the IEEE Computer Society. Information technology - telecommunications and information exchange between systems - local and metropolitan area networks - specific requirements - part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications, 1999.

[62] B. Lathi. *Linear Systems and Signals*. Oxford University Press, second edition, 2005.

[63] S.-J. Lee and M. Gerla. AODV-BR: Backup routing in ad hoc networks. In *Proc. of WCNC*, Chicago, IL, United States, 2000.

[64] S.-J. Lee and M. Gerla. Split multipath routing with maximally disjoint paths in ad hoc networks. In *Proc. of IEEE ICC*, Helsinki, Finland, 2001.

[65] T. Lee. Analysis of infinite servers polling systems with correlated input and state dependent vacations. *European Journal of Operations Research*, 115:392–412, 1998.

[66] T. Lee. Analysis of random polling system with an infinite number of coupled servers and correlated input process. *Computers & Operations Research*, 30:2003–2020, 2003.

[67] K. K. Leung. Cyclic-service systems with probabilistically-limited service. *IEEE Journal on Selected Areas in Communications*, 9(2):185–193, 1991.

[68] K. K. Leung. Cyclic-service systems with non-preemptive time-limited service. *IEEE Trans. on Communications*, 42(8):2521–2524, 1994.

[69] H. Levy and M. Sidi. Polling systems: Applications, modeling, and optimization. *IEEE Trans. on Communications*, 38(10):1750–1760, 1990.

[70] R. Litjens, H. van den Berg, R. J. Boucherie, F. Roijers, and M. Fleuren. Performance analysis of wireless LANs: an integrated packet/flow level approach. In *Proc. of ITC-18*, Berlin, Germany, 2003.

[71] J. D. C. Little. A proof for the queuing formula: L = λ W. *Operations Research*, 9(3):383–387, 1961.

[72] M. M. Ibrahim, A. Al Hanbali, and P. Nain. Delay and resource analysis in MANETs in presence of throwboxes. *Performance Evaluation*, 64:933–947, 2007.

[73] R. D. van der Mei and S. C. Borst. Analysis of multiple-server polling systems by means of the power-series algorithm. *Stochastic Models*, 13(2):339–369, 1997.

[74] R. de Moraes, H. Sadjadpour, and J. J. G. Luna-Aceves. Mobility-capacity-delay trade-off in wireless ad hoc networks. *Ad Hoc Networks*, 4(5):607–620, 2006.

[75] R. J. T. Morris and Y. T. Wang. Some results for multi-queue systems with multiple cyclic servers. In *Performance of Computer Communication Systems, H. Rudin and W. Bux (eds.)*, pages 245–258, 1984.

[76] S. Mueller, R. P. Tsang, and D. Ghosal. Multipath routing in mobile ad hoc networks: Issues and challenges. *Lecture Notes in Computer Science*, 2004.

[77] M. Musolesi and C. Mascolo. A community based mobility model for ad hoc network research. In *Proc. of RealMAN*, Florence, Italy, 2006.

[78] A. Nasipuri, R. Castañeda, and S. R. Das. Peformance of multipath routing for on-demand protocols in mobile ad hoc networks. *ACM/Baltzer Mobile Networks and Applications (MONET) Journal*, 6:339–349, 2001.

[79] G. L. Nemhauser and L. A. Wolsey. *Integer and Combinatorial Optimization*. John Wiley & Sons, 1988.

[80] M. F. Neuts. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. Dover Publications, Inc., 1994.

[81] J. C. W. van Ommeren. The discrete-time single-server queueing model. *Queueing Systems*, 8(1):279–294, 1991.

[82] J. Padhye, S. Agarwal, V. Padmanabhan, L. Qiu, A. Rao, and B. Zill. Estimation of link interference in static multi-hop wireless networks. In *Proc. of IMC*, Berkeley, CA, United States, 2005.

[83] K. Papagiannaki, M. Yarvis, and W. S. Conner. Experimental characterization of home wireless networks and design implications. In *Proc. of IEEE Infocom*, Barcelona, Spain, 2006.

[84] M. R. Pearlman, Z. J. Haas, P. Sholander, and S. S. Tabrizi. On the impact of alternate path routing for load balancing in mobile ad hoc networks. In *Proc. of MobiHoc*, Boston, MA, United States, 2000.

[85] L. Pelusi, A. Passarella, and M. Conti. Opportunistic networking: data forwarding in disconnected mobile ad hoc networks, IEEE Communications Magazine, Nov., 2006.

[86] P. P. Pham and S. Perreau. Increasing the network performance using multipath routing mechanism with load balance. *Ad Hoc Networks*, 2:433–459, 2004.

[87] W. H. Press, W. T. Vetterling, S. A. Teukolsky, and B. P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, 1992.

[88] J. Resing. Polling systems and multitype branching processes. *Queueing Systems*, 13(4):409–426, 1993.

[89] S. M. Ross. *Stochastic Processes*. John Wiley & Sons, 1996.

[90] T. L. Saaty. Time-dependent solution of the many server Poisson queue. *Operations Research*, 8(6):755–772, 1960.

[91] A. Schrijver. *Combinatorial Optimization: Polyhedra and Efficiency*. Springer, 2006.

[92] G. Sharma, R. R. Mazumdar, and N. B. Shroff. Delay and capacity trade-offs in mobile ad hoc networks: A global perspective. In *Proc. of IEEE Infocom*, Barcelona, Spain, 2006.

[93] T. Small and Z. Haas. The shared wireless infostation model - a new ad hoc networking paradigm (or where there is a whale, there is a way). In *Proc. of MobiHoc*, Annapolis, MD, United States, 2003.

[94] T. Small and Z. Haas. Resource and performance tradeoffs in delay-tolerant wireless networks. In *Proc. of ACM SIGCOMM first workshop on delay tolerant networking and related topics*, Philadelphia, PA, United States, 2005.

[95] E. de Souza e Silva, H. R. Gail, and R. R. Muntz. Polling systems with server timeouts and their application to token passing networks. *IEEE Trans. on Networking*, 3(5):560–575, 1995.

[96] T. Spyropoulos, K. Psounis, and C. S. Raghavendra. Efficient routing in intermittently connected mobile networks: The single-copy case. *IEEE/ACM Trans. on Networking*, 16(1):63–76, 2008.

[97] H. Takagi. Queueing analysis of polling systems: An update. *Chapter of Stochastic Analysis of Computer and Communication Systems*, pages 267–318, 1990.

[98] H. Takagi. Queueing analysis of polling models: progress in 1990-1994. In *Frontiers in Queueing: Models, Methods and Problems, J.H. Dshalalow (ed.)*, pages 119–146. CRC Press, Boca Raton, 1997.

[99] M. Tangemann and K. Sauer. Performance analysis of the timed token protocol of FDDI and FDDI-II. *IEEE Journal on Selected Areas in Communications*, 9(2):271–278, 1991.

[100] E. C. Titchmarsh. *The Theory of Functions*. Oxford University Press, second edition, 1993.

[101] A. Tsirigos and Z. J. Haas. Multipath routing in the presence of frequent topological changes. *IEEE Communications Magazine*, 39(11):132–138, 2001.

[102] V. M. Vishnevskii and O. V. Semenova. Mathematical methods to study the polling systems. *Automation and Remote Control*, 67(2):173–220, 2006.

[103] M. Vlasiou and U. Yechiali. M/G/∞ polling systems with random visit times. *Probability in the Engineering and Informational Sciences*, 22:81–106, 2008.

[104] C.-L. Wang and R. Wolff. Work-conserving tandem queues. *Queueing Systems*, 49(3-4):283–296, 2005.

[105] E. Winands, I. Adan, and G. van Houtum. Mean value analysis for polling systems. *Queueing Systems*, 54(1):35–44, 2006.

[106] R. Wolff. Poisson arrivals see time averages. *Operations Research*, 30(2):223–231, 1982.

[107] K. Wu and J. Harms. Multipath routing for mobile ad hoc networks. *Journal of Communication and Networks*, 4(1), 2002.

[108] J. Xie, M. J. Fischer, and C. M. Harris. Workload and waiting time in a fixed-time loop system. *Computers Operations Research*, 24(8):789–803, 1997.

[109] M. Zazanis. A Palm calculus approach to functional versions of Little's law. *Stochastic Processes and their Applications*, 74(2):195–201, 1998.

[110] E. Zhang, G. Neglia, J. Kurose, and D. Towsley. Performance modeling of epidemic routing. *Computer Networks*, 51(10):2867–2891, 2007.

Summary

This thesis presents models for the performance analysis of a recent communication paradigm: *mobile ad hoc networking*. The objective of mobile ad hoc networking is to provide wireless connectivity between stations in a highly dynamic environment. These dynamics are driven by the mobility of stations and by breakdowns of stations, and may lead to temporary disconnectivity between parts of the network. As wireless communication tends to become ubiquitous in everyday's life, robust and accurate performance models are necessary to predict traffic delays, detect bottlenecks, and to assess other relevant quality of service measures in networks driven by this paradigm. Applications of this novel paradigm can be found in telecommunication services, but also in manufacturing systems, road-traffic control, animal monitoring and emergency networking.

A queueing model that arises quite naturally as a performance model for mobile ad hoc networking is a specific *polling model*. Polling models are queueing systems consisting of multiple queues served by one or more servers. In this thesis, we consider time-limited polling models as these capture the uncontrollable characteristic of link availability in mobile ad hoc networks in a straightforward fashion. Particularly, we introduce and analyze a novel, so-called pure exponential time-limited, service discipline in the context of polling systems. This discipline does not qualify as a "simple" and readily tractable, branching-type, discipline, but belongs to a class for which closed-form expressions for traditional performance measures are unlikely to exist. Consequently, a significant part of this thesis deals with the analysis of these specific time-limited polling systems.

Our main focus is on the network performance in terms of stability, buffer levels at the stations and transfer delays of the data packets in a mobile ad hoc network. In Chapter 1, we discuss the motivation for the thesis and state the performance issues for mobile ad hoc networking. Also, we present the concept of polling systems both for systems with one and for systems with multiple servers and review their analysis. We introduce formally our *basic* single-server and multi-server polling systems that will be considered in this thesis.

In Part I, we turn to network capacity and stability issues. Particularly, Chapter 2 presents an analytical framework to assess the performance trade-off between the level of signal interference and multi-path routing. Using multiple paths for peer-to-peer data transmission at first sight improves performance, but as it involves many stations it also causes additional interference with respect to single-path routing. This interference leads to restrictions in the number of transmission opportunities or to unsuccessful packet exchanges, so that capacity is wasted anyhow. Numerical experiments for the network capacity in arbitrary, finite-sized networks provide insight in whether or not to employ multiple paths for data transmission. In Chapter 3, the stability of exponential time-limited polling models is the subject of study. Stability conditions prescribe limits on the amount of traffic that can be sustained by the system. Exceeding these limits leads to instable behavior of the system, so that the establishment of stability conditions is indeed a fundamental keystone in the performance analysis of polling models. In the remainder of the thesis, we assume a stable environment and concentrate on the buffer levels at the stations in the mobile ad hoc network and the delays experienced by the data packets. This is done via the analysis of specific time-limited polling systems.

Part II treats the analysis of single-server polling systems. These polling systems effectively represent performance models for networks in which a single communication link can be active at a time. This occurs for instance for very small networks or networks comprising stations that are fully connected, i.e., all stations in the network are able to sense each other's transmissions. In Chapter 4, we provide an analysis of our basic single-server polling system which operates under the pure time-limited discipline. In particular, we present a framework to compute the joint queue-length distribution at specific time instants and also under stationarity. The key relation within this framework, which captures the evolution of the queue lengths during a visit of the server, is obtained in a recursive fashion. Moreover, we point out how this analysis can be extended to account for more general system configurations. These enhancements greatly expand the applicability of the presented techniques with regard to the development of accurate performance models for mobile ad hoc networking. An alternative analysis of the basic single-server polling system is presented in Chapter 5. This is done by applying results of the transient analysis of the well-studied M/G/1 queue to our polling system. Specifically, this provides us with a direct, non-recursive, relation for the queue-length evolution during a visit which is a more elegant and faster to evaluate counterpart of the recursive procedure in Chapter 4. Moreover, using similar techniques, we obtain such a relation for the

exhaustive time-limited service discipline. The underlying analytical framework of Chapters 4 and 5 appears most suitable for networks with a small number of queues and a light to moderate load. To analyze other network configurations, in Chapter 6, we resort to a different analytical tool, namely approximations. This is a valuable tool when exact methods become intractable or computationally infeasible. We consider a product-form approximation for the joint queue-length distribution of the basic single-server polling model. This approximation is a fruitful alternative from a computational viewpoint for the queue-length distribution of the basic polling model. In addition, we consider a sojourn-time approximation for a two-hop tandem queueing model which can be seen as a first-step model for delay analysis in opportunistic or delay-tolerant networks. This second approximation is analytically more involved and provides an accurate, closed-form expression for the mean sojourn time at the relay queue.

Finally, Part III concerns the analysis of polling systems with multiple servers operating under the pure time-limited discipline. Polling systems with multiple servers are typically much harder to analyze than their single-server counterparts. This is due to the fact that the process describing the location of the servers becomes multi-dimensional, which also leads to different server strategies. Typically, servers move either coupled as one group or individually through the system. The systems with individual servers reflect communication networks with multiple active links, while the coupled-servers strategy is more likely to appear in a manufacturing environment. In Chapter 7, we set up a framework to compute the joint queue-length probabilities in a multi-server polling system operating under this pure exponential time-limited discipline. The specific service discipline enables us to analyze such systems effectively. The analysis builds on the recursive analysis introduced in Chapter 4. The final chapter, Chapter 8, studies the basic two-server polling system by using a transient analysis conform Chapter 5. For the individual-servers strategy, this approach comes down to solving a complex analysis problem, which is rather difficult to extend to more than two servers. Conversely, the coupled-servers strategy is amenable for a tractable analysis that readily allows to progress to systems with three servers and beyond.

Samenvatting

Dit proefschrift beschouwt wiskundige modellen voor de prestatieanalyse van een recent communicatie-paradigma: *mobiele ad hoc netwerken*. Mobiele ad hoc netwerken worden gekenmerkt door het ontbreken van centrale coördinatie, draadloze communicatie tussen de aanwezige stations en een dynamische netwerkstructuur. Deze dynamiek wordt grotendeels veroorzaakt door de mobiliteit van de stations en het uitvallen van stations op onverwachte momenten. Dit heeft tot gevolg dat delen van het netwerk tijdelijk niet verbonden zijn. Daar draadloze communicatie steeds prominenter aanwezig is in ons alledaagse leven zijn robuuste en nauwkeurige prestatiemodellen benodigd om vertragingen van datapakketten te schatten, knelpunten op te sporen of andere relevante kwaliteitsmaten te bepalen voor zulke netwerken. Toepassingen van dit nieuwe communicatie-paradigma kunnen uiteraard worden gevonden in telecommunicatiediensten, maar ook in productiesystemen, regelsystemen voor wegverkeer en calamiteitnetwerken.

Een bijzondere plaats in het proefschrift wordt ingenomen door wiskundige modellen die bekend staan als *wachtrijmodellen*. Een klassiek wachtrijmodel (ook wel kortweg wachtrij genaamd) bestaat uit de volgende componenten: een proces dat de aankomsten van klanten beschrijft, een proces dat de hoeveelheid werk beschrijft die een klant meebrengt en een bediende die deze klanten bedient. De genoemde processen zijn meestal stochastisch van aard wat wil zeggen dat zij gebaseerd zijn op een onderliggende kansverdeling. De standaardprocedure is dat een klant bij de wachtrij arriveert, daar wacht op zijn beurt, wordt geholpen door de bediende en ten slotte de wachtrij weer verlaat. Er kunnen allerlei variaties van wachtrijmodellen

geconstrueerd worden door bijvoorbeeld het aantal toegestane klanten in de wachtrij te beperken, het aantal bedienden te variëren, de volgorde van bediening van de klanten aan te passen of door prioriteiten aan bepaalde klanten toe te wijzen.

Een specifiek wachtrijmodel dat op natuurlijke wijze naar boven komt als prestatiemodel voor mobiele ad hoc netwerken is een zogenaamd *pollingmodel*. Pollingmodellen zijn wachtrijsystemen die bestaan uit meerdere wachtrijen welke zijn gekoppeld doordat de klanten van de verschillende wachtrijen de aanwezige bedienden moeten delen. In dit proefschrift beschouwen we pollingmodellen waarbij de tijd dat een bediende klanten bij een wachtrij bedient begrensd is. Deze zogeheten *bedieningsdiscipline* benadert het stochastische proces van de beschikbaarheid van communicatielinks in mobiele ad hoc netwerken. In het bijzonder introduceren en analyseren we een nieuwe bedieningsdiscipline voor pollingmodellen waaronder de bediende altijd precies een stochastisch verdeelde tijd bij een wachtrij blijft alvorens naar een volgende wachtrij te gaan. Deze discipline valt niet binnen de bekende klasse van relatief eenvoudig te analyseren bedieningsdisciplines, maar behoort tot de klasse waarvoor het onwaarschijnlijk is dat gesloten uitdrukkingen voor eenvoudige prestatiematen (zoals de gemiddelde rijlengte) bestaan. Dit maakt deze discipline naast praktisch waardevol ook zeker theoretisch interessant en aldus is een belangrijk deel van dit proefschrift aan de analyse van deze specifieke bedieningsdiscipline gewijd.

Onze aandacht in dit proefschrift is voornamelijk gericht op de netwerkprestaties in termen van stabiliteit, rijlengtes bij de stations en de verblijftijden van de datapakketten in een mobiel ad hoc netwerk. In hoofdstuk 1 motiveren we het onderzoek en benoemen we de specifieke aspecten van mobiele ad hoc netwerken die van invloed zijn op de netwerkprestatie. Daarnaast presenteren we het concept van pollingsystemen zowel met één als met meerdere bedienden en geven we een overzicht van de analyse zoals bekend uit de literatuur. Tevens introduceren we onze *standaard* pollingmodellen met één bediende en met meerdere bedienden welke een centrale rol zullen spelen in de rest van het proefschrift.

In deel één van dit proefschrift concentreren we ons op de capaciteit en de stabiliteit van het netwerk. Specifiek presenteren we in hoofdstuk 2 een analytisch kader om de relatie tussen interferentie van radiosignalen en het routeren over meerdere paden met betrekking tot de capaciteit te bestuderen. Op het eerste gezicht lijkt het gebruik van meerdere paden voor datacommunicatie positief voor de netwerkprestatie, maar aangezien op deze manier vele stations betrokken zijn bij het verzenden van datapakketten neemt de interferentie ten opzichte van routeren over één enkel pad sterk toe. Deze interferentie zorgt ervoor dat stations minder vaak data kunnen versturen of dat data verloren gaat, zodat in beide gevallen de capaciteit van het netwerk weer afneemt. Aan de hand van numerieke experimenten voor kleine en middelgrote netwerken hebben we inzichten verkregen op de vraag wanneer wel en wanneer niet dataverkeer over meerdere paden te verzenden. In hoofdstuk 3 bewijzen we de stabiliteitscondities voor een klasse van pollingsystemen. Dit zijn pollingsystemen met één bediende waarbij de tijd dat de bediende doorbrengt bij een station begrensd is

door een exponentieel verdeelde stochast. Stabiliteitscondities schrijven voor hoeveel verkeer er maximaal door het netwerk kan worden verwerkt. Het overschrijden van deze limieten leidt tot instabiel gedrag en daarom vormen zulke condities eigenlijk het fundament voor de prestatieanalyse van pollingmodellen. In het vervolg van het proefschrift zullen we aannemen dat aan deze condities is voldaan en zullen we ons volledig concentreren op de rijlengten bij de stations en de verblijftijden van de datapakketten in het netwerk.

Het tweede deel van dit proefschrift behandelt de analyse van pollingmodellen met één bediende wiens bedieningstijd bij een wachtrij begrensd is. Deze modellen kunnen gebruikt worden als prestatiemodel voor netwerken waarin op ieder tijdstip slechts één communicatielink tegelijk actief kan zijn. Dit geeft een realistisch model voor kleine netwerken en ook voor netwerken waarbij alle stations zendactiviteit van andere stations direct kunnen observeren. In hoofdstuk 4 analyseren we in detail ons standaard pollingmodel met één server. We presenteren een aanpak om de gezamenlijke rijlengteverdeling numeriek te bepalen op specifieke tijdstippen en tevens voor het systeem in evenwicht. De belangrijkste relatie binnen deze aanpak beschrijft het verloop van de rijlengteverdeling tijdens een bezoek van de bediende aan de wachtrij en deze relatie wordt recursief bepaald. Daarnaast presenteren we een aantal interessante modeluitbreidingen. Deze uitbreidingen vergroten de toepasbaarheid van de gepresenteerde technieken aanzienlijk met betrekking tot de ontwikkeling van nauwkeurige prestatiemodellen voor mobiele ad hoc netwerken. Een alternatieve analyse van de bovengenoemde relatie wordt gegeven in hoofdstuk 5. Deze afleiding is uitgevoerd door resultaten van het tijdsafhankelijke gedrag van het bekende M/G/1 wachtrijmodel toe te passen op ons specifieke pollingmodel. Op deze manier leiden we een directe, niet-recursieve, relatie af voor de verandering van de rijlengteverdeling gedurende een bezoek van de bediende aan een wachtrij. Deze relatie is niet alleen eleganter, maar ook sneller numeriek te evalueren dan de recursieve relatie uit hoofdstuk 4. Bovendien kunnen we aan de hand van de gebruikte methoden eenzelfde soort relatie afleiden voor soortgelijke bedieningsdisciplines waarbij de bedieningstijd van een bediende door een exponentiële stochast begrensd is. De onderliggende aanpak van de hoofdstukken 4 en 5 is bij uitstek geschikt voor het analyseren van netwerken met een klein aantal stations en een laag tot gemiddeld aanbod van dataverkeer. Om ook andere netwerkconfiguraties te kunnen analyseren gebruiken we in hoofdstuk 6 een alternatieve wiskundige methode namelijk benaderingen. Dit is een waardevolle methode vooral wanneer exacte methoden niet langer mogelijk zijn of teveel rekentijd vereisen. We bestuderen eerst een zogenaamde productvormbenadering voor de gezamenlijke rijlengteverdeling voor het standaard pollingmodel met één server. Deze benadering vermindert de benodigde rekenkracht aanzienlijk, omdat nu slechts de rijlengte voor elk station in isolatie hoeft te worden berekend. De benadering is nauwkeurig voor uiteenlopende waarden van de modelparameters. Daarnaast beschouwen we een benadering voor de verblijftijd van een datapakket in een specifiek tandem wachtrijmodel bestaande uit drie stations. Dit model kan worden gezien als een eerste stap in de analyse van netwerken waarin er tijdelijk geen route bestaat

tussen verzender en ontvanger. De analyse achter deze benadering is relatief ingewikkeld, maar uiteindelijk verkrijgen we een eenvoudige, gesloten uitdrukking voor de gemiddelde verblijftijd in het centrale station die uiterst nauwkeurig is.

Ten slotte, in het derde en laatste gedeelte van dit proefschrift analyseren we pollingmodellen met meerdere bedienden die allen voor een exponentieel verdeelde tijd bij een wachtrij werken. De analyse voor modellen met meerdere bedienden is veel lastiger dan voor dezelfde modellen met slechts één bediende. Dit komt onder andere doordat het dynamische proces dat de posities van de bedienden beschrijft nu meerdimensionaal wordt. Tevens leidt het tot diverse strategieën voor de bedienden die zich nu ofwel gekoppeld als één groep ofwel individueel door het systeem verplaatsen. De systemen met individuele bedienden representeren in feite communicatienetwerken met meerdere actieve links, terwijl toepassingen van de gekoppelde bedienden eerder in een productieomgeving terug te vinden zijn. In hoofdstuk 7 beschrijven we een aanpak om de gezamenlijke rijlengteverdeling te bepalen voor ons standaard pollingmodel met meerdere bedienden. Het is goed om te benadrukken dat voor de toepasbaarheid van deze aanpak de specifieke bedieningsdiscipline een cruciale rol speelt. De gebruikte analyse borduurt voort op de recursieve analyse uit hoofdstuk 4. In het laatste hoofdstuk, hoofdstuk 8, bestuderen we het standaard pollingsysteem met twee bedienden aan de hand van de tijdsafhankelijke analyse van hoofdstuk 5. Voor de strategie met individuele bedienden komt deze aanpak neer op het toepassen van technieken uit de complexe functietheorie. Het gebruik van deze methode wordt echter bijzonder lastig in het geval van meer dan twee bedienden. De strategie met gekoppelde bedienden daarentegen leent zich beter voor een gestructureerde wiskundige analyse en kan derhalve wel worden uitgebreid worden naar systemen met drie of meer bedienden.

About the author

Roland de Haan was born in Geldermalsen (The Netherlands) on October 25, 1980. He obtained his VWO diploma at the Koningin Wilhelmina College in Culemborg in 1998. Early 2004, he received his master's degree in Industrial and Applied Mathematics from the Eindhoven University of Technology. His final master's project on performance modelling of TCP was carried out at TNO-ICT (Delft). Directly afterwards, he successfully applied for a one-year scholarship which enabled him to work as a research assistant at the Instituto de Telecomunicações in Lisbon (Portugal) on analytical models for multiple-access protocols. In 2005, Roland returned to the Netherlands and started working towards his PhD degree in the Stochastics Operations Research group at the University of Twente. He will defend his thesis on June 4, 2009.