Processor-Sharing Queues and Resource Sharing in Wireless LANs

ISBN 978-90-9021792-5

Beta
Research School for Operations
Management and Logistics

This thesis is number D-97 of the thesis series of the Beta Research School for Operations Management and Logistics. The Beta Research School is a joint effort of the departments of Technology Management, and Mathematics and Computer Science at the Technische Universiteit Eindhoven and the Centre for Telematics and Information Technology at the University of Twente. Beta is the largest research centre in the Netherlands in the field of operations management in technology-intensive environments. The mission of Beta is to carry out fundamental and applied research on the analysis, design and control of operational processes.

PROCESSOR-SHARING QUEUES AND RESOURCE SHARING IN WIRELESS LANS

PROEFSCHRIFT

ter verkrijging van

de graad van doctor aan de Universiteit Twente,

op gezag van de rector magnificus,

prof.dr. W.H.M. Zijm,

volgens besluit van het College voor Promoties

in het openbaar te verdedigen

op vrijdag 1 juni 2007 om 15.00 uur

door

Sing Kwong Cheung

geboren op 2 juli 1979

te Amersfoort

Dit proefschrift is goedgekeurd door de promotoren:

prof.dr. R.J. Boucherie
prof.dr. J.L. van den Berg

Acknowledgments

This monograph is the result of my research carried out at the Stochastic Operations Research Group, University of Twente. At this point, I would like to express my gratitude to those who helped and encouraged me throughout the years in accomplishing this thesis.

First of all, I would like to thank my promoter and supervisor Richard Boucherie for his excellent and no-nonsense guidance in doing research. I greatly benefited from his creative ideas and his expertise in stochastics, as well as his critical attitude in many aspects of our joint work. Next, I am grateful to my second promoter Hans van den Berg for his excellent advices, his inspiring enthusiasm and his help in numerous occasions.

In the past few years, the possibility to have joint meetings with Richard and Hans – at a non-mobile place – was quite rare. Occasionally, we managed to have meetings in the first-class compartment of the *international train* from Hengelo to Amersfoort. These enjoyable meetings in the train were very efficient, without losing too much time in our busy schedules. Their professional guidance has made me grow as a researcher and has also developed me in a broader sense as a person.

I am further grateful to all other members of my promotion committee: prof.dr.ir. Sem Borst, prof.dr. Chris Blondia, prof.dr. Henk Tijms, prof.dr. Wim Albers and dr.ir. Geert Heijenk, for their thorough examination of the manuscript and for their comments.

I also take great pleasure in thanking Sindo Núñez-Queija from CWI and TNO Information and Communication Technology (TNO ICT) for his contribution in this thesis. In addition, his good advices for my presentation at the PERFORMANCE conference of Juan-les-Pins in 2005 were highly appreciated, as well as his never ending enthusiasm. I also wish to thank Remco Litjes and Frank Roijers, both from TNO ICT, for their contribution in this thesis and also for our trip to the most adventurous part of the Great Wall of China during the ITC conference of Beijing in 2005.

Next, I would like to acknowledge the Korea Science and Engineering Foundation (KOSEF) and the Netherlands Organisation for Scientific Research (NWO) for providing funds for my research visit to South-Korea in 2006. I am grateful to Bara Kim from Korea University and Jeongsim Kim from Chungbuk National University for their contribution in this thesis and their kindness during my stay in Korea. I also want to express my

gratitude to professor Bong Dae Choi, director of the Telecommunications Mathematics Research Centre (TMRC) of Korea University, and professor Gang Uk Hwang from the Korean Advanced Institute of Science and Technology (KAIST). I sincerely thank them for their hospitality and kindness, and I also want to thank their graduate students (my office-mates) for the offered and pleasant help in many aspects during my visit. Thanks to Yunzi, Youngkyo, Seonmi, Sangkyu, Jerim and Jaesun from TMRC of Korea University and thanks to Yunju, Yoora and Bongjoo from KAIST. *Gamsahamnida!* Also, thanks to all other students who helped me, but it's impossible to name them all.

I owe a particular word of gratitude to Rein Nobel from the Vrije Universiteit for encouraging me to pursue a PhD after my Econometrics study, as well as his long-lasting contacts with professor Bong Dae Choi. After all, the latter fact has led to joint conferences between Korea and The Netherlands on Queueing Theory and its Applications to Telecommunication Systems, and therefore, it has also led to my research visit to Korea and the (partial) accomplishment of my thesis.

At the University of Twente, I would like to thank my former room- or corridor-mates Nicky, Irwan and Bas for enduring my presence and for helping me out in numerous occasions (corridor-mates due to the fire disaster of the TWRC building). Furthermore, I am also thankful to the other members of the Stochastic Operations Research Group for providing a pleasant working environment. Thyra, Werner, Jan-Kees, Nelly, Judith, Maurits, Tom, Roland, Denis and Yana – Thank you!

At a personal level, I greatly appreciated the interest expressed and support offered throughout the years by family and friends. I thank Steven and Lai Ting for their willingness to assist me during the thesis defense. Michaela, I thank you for your unconditional love and support. You are the greatest joy in my life, always giving me power and energy. Without your support, I would not have succeeded!

Sing-Kong Cheung
Amersfoort, April 2007

Contents

Processor-Sharing Queues and Resource Sharing in Wireless LANs

Chapter 1

Preliminaries and scope of the thesis

In many aspects of daily life queueing phenomena may be observed when service facilities (counters, buses, telephone lines, Internet services) cannot facilitate the users of these services immediately. The typical example where queues arise is at counters of supermarkets, where customers usually have to wait for some time before they receive service. Since waiting is considered as an unpleasant experience, we would like to reduce waiting times and meet a certain Quality-of-Service level for the customers. On the other hand, it makes economic sense to have queues, since resources are often scarce and/or expensive. In order to design a service system optimally from an economical perspective, it is desirable to study the characteristics and effects of congestion phenomena which influence the performance of the system. These characteristics and effects may be adequately studied with mathematical methods from queueing theory.

Queueing systems are mainly characterized by the random nature in which customers arrive at the system and by the variable amount of work that the customers require from the service facility. Another important characteristic is called the service discipline, which defines how the resources are allocated to the customers. The complex interaction between these characteristics has a significant impact on the performance of the system and on the individual customers.

The first queueing-theoretic models were developed in the early 20-th century for the dimensioning of telephony systems. Later, queueing theory was successfully applied in operations research and management science, in particular for production planning. Nowadays, queueing theory plays a prominent role in the performance analysis of a wide range of systems in computer communications, logistics, and manufacturing.

1.1 Queueing theory and communication networks

Queueing theory plays an important role in the design and performance analysis of (tele)communication systems. The Erlang loss model [37] is one of the most traditional and basic types of queueing models, originally developed for the performance analysis of circuit-switched telephony systems. In these systems, newly arriving calls are either admitted into the system or rejected (i.e., blocked and lost), depending on the availability of a free circuit (channel). The key performance measure is the fraction of blocked calls. The Erlang loss model is still much applied, e.g., for the planning of mobile communication networks (GSM) and optical networks.

Another important type of queueing model arising in the performance analysis of telecommunication systems is the so-called processor-sharing (PS) model. PS models were originally developed for the analysis of time-sharing in computer communication systems [62, 63] in the 1960s. The key property of the PS service discipline is that the common resource is fairly shared among all jobs present in the system. Typically, PS models are applied to the performance analysis of data transfers in (packet-switched) communication networks. Important performance measures in these models are the sojourn times of the jobs and their throughputs. Over the past decade PS models have attracted new attention for the performance analysis of fair bandwidth sharing mechanisms such as the Transmission Control Protocol (TCP) [50]. TCP deploys a flow-control protocol that reacts dynamically to network congestion, and can be considered as one of the most important enablers of the huge expansion of the Internet since the 1980s. The PS discipline – as a convenient modeling abstraction – continues to play an instrumental role in the design and operations of communication systems as motivated by various technological trends, particularly also for wireless networks [21, 68].

This thesis is primarily devoted to PS queueing models as basic models for resource sharing in communication networks. In particular, we perform extensive mathematical analyses in order to quantitatively characterize the performance of these models. In addition, applying the results of our mathematical analyses, we present a performance study of data transfers in Wireless Local Area Networks (WLANs).

Basic concepts of classical queueing models

Figure 1.1 depicts the classical queueing model, which is conventionally denoted as the G/G/1 queue. The notation ‘G/G/1’ was proposed by Kendall [52], and the first symbol G reflects that the time between two consecutive customer arrivals (interarrival time) is a positive random variable with a general probability distribution. The second G reflects the general probability distribution for the service requirements of the customers, and the 1 refers to the single server. Customers who have received their full service leave the system. The interarrival times are assumed to be independent and identically distributed

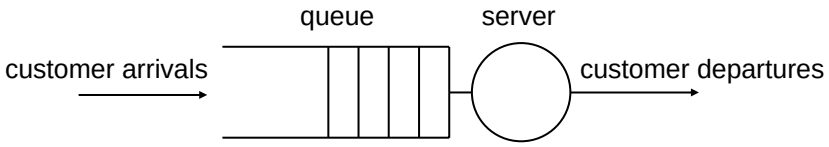

Figure 1.1: *The G/G/1 single-server queueing model.*

(i.i.d.). The service requirements also form a sequence of i.i.d. random variables and are independent of the interarrival times. The notation GI is sometimes used when the independence assumption needs to be emphasized. When the arrival process is assumed to be a Poisson process then we have the M/G/1 model, where the M stands for the Markovian (or Memoryless) nature of the Poisson process. The symbol M is also used in the second position of Kendall's notation, if the service requirements are exponentially distributed. Other common service requirement distributions are the deterministic distribution (denoted with D) and phase-type distributions (usually denoted with PH).

To describe the G/G/1 queueing system we also need to define how the server capacity is allocated to the customers in the system. The most natural service discipline is the First Come First Served (FCFS) discipline, where the customers are served in order of arrival. In certain systems other service disciplines may be more appropriate, such as the Last Come First Served (LCFS) discipline or a priority discipline. As mentioned above, in this thesis the focus is on processor-sharing (PS) service disciplines, where all customers in the system are served in parallel.

Other variations on the basic G/G/1 queueing model are, for example, queues with fluctuating service capacity, server vacations, or customer impatience, which all may have a significant influence on the performance. The performance measures of interest strongly depend on system-specific objectives and model-specific assumptions. The most common performance measures are queue lengths, waiting times (or: delays), and sojourn times (or: transfer times). For a thorough introduction on queueing theory we refer to Kleinrock [64], Cohen [33], Takagi [96, 97, 98], and Tijms [99].

In the next section we consider several processor-sharing queueing models in some more detail, in particular the *egalitarian* processor-sharing (EPS) model and the *discriminatory* processor-sharing (DPS) model with multiple customer classes; the EPS and DPS queues play a central role in this thesis. In Section 1.3, (PS) queueing models with time-fluctuating service capacity are introduced, which are studied in this thesis as well.

1.2 Processor-sharing queueing models

The processor-sharing (PS) discipline has gained a prominent role in queueing theory over the past few decades. Kleinrock [62, 63] introduced the simplest and best known egalitarian processor-sharing (EPS) discipline, in which a single server assigns each customer a fraction $1/n$ of the service capacity when $n > 0$ customers are in the system; the total service rate is equally shared among all customers present.

A drawback of the EPS discipline is the inability to model heterogeneous time-sharing systems in which users from different classes get different shares of the capacity, i.e., EPS cannot differentiate the Quality-of-Service among users. Two main generalizations are proposed as multi-class extensions of the single-class EPS model: the generalized processor-sharing (GPS) and the discriminatory processor-sharing (DPS) model.

In the GPS model each customer class is guaranteed a minimum service rate by assigning class j a non-negative weight ϕ_j, $j = 1, ..., K$ (as a minimum share of the capacity c), where K is the number of classes. When no customers are present in a class, its share of the capacity is distributed among the other active classes. The service rate is distributed across (non-empty) classes, irrespective of the actual number of customers present. For more details on GPS we refer to [83, 101] and references therein.

The other generalization of EPS, the discriminatory processor-sharing (DPS) discipline, was originally introduced under the name Priority Processor Sharing by Kleinrock [63]. The range of applications for DPS is extremely large; see e.g. [5, 19, 29, 48, 54, 75] and also Chapter 8. In DPS, a customer of type k receives a service rate $\alpha_k / \sum_{j=1}^{K} \alpha_j n_j$, according to the set of weights $\{\alpha_j : j = 1, ..., K\}$, when n_j customers of type j, $j = 1, ..., K$, are present in the system. If all weights α_j are equal, then we have the ordinary EPS queue. Exact analysis of DPS turns out to be difficult compared to ordinary PS; results for DPS are scarce in the queueing literature.

We also mention that Cohen [32] generalized the standard (E)PS model to a PS model in which each customer receives a service rate according to an arbitrary positive function $\phi(n)$. This model is also called *generalized processor-sharing*, which should not be confused with the GPS model as in [83, 101]. Cohen's GPS model is in fact an EPS model with queue-dependent service capacity. By appropriate choice of the queue-dependent service capacity $\phi(n)$, this model includes a very wide class of service disciplines, and significantly enhances the modeling capabilities of the standard EPS model. In many applications it models the main factors determining the performance, while on the other hand, it is simple enough to be analytically tractable, see e.g. [13, 68] and Chapter 8.

A particularly relevant performance measure for PS models is the sojourn time of a job in the system. For the egalitarian PS model, several expressions for the sojourn time distribution have been obtained in terms of Laplace-Stieltjes transforms, see Section 2.3. However, these transform expressions are not particularly insightful or readily applicable for computational purposes. Transform expressions for discriminatory PS seem not to

exist in manageable form, see Section 2.4. The above motivates the derivation of bounds and approximations for EPS and DPS, which also give new insights in the system behavior of various PS models.

1.3 Queues with time-fluctuating service capacity

PS queueing models with time-fluctuating service capacity form another important class of models of practical interest. For example, document transmissions in the Internet and file downloads from web servers may experience high variation in transmission rates, due to the interaction with other traffic streams [46]. In particular, for TCP-driven traffic flows, which are responsive to temporary network congestion, the effectively available transmission capacity (and hence the flow throughput) is highly affected by the presence of traffic generated by other applications (e.g., voice or video) that rely on unresponsive transport protocols such as UDP. In the queueing theory community, it has long been recognized that fluctuating service capacity has a decisive effect on the perceived performance, as is witnessed in the extensive literature on queueing models describing congestion phenomena subject to varying service capacity [34, 39, 45, 46, 79, 96].

There is a rich variety of models in which the available service capacity alternates between a positive value and complete absence of service, including unreliable servers, server vacations and service failures [39, 96]. These models allow for closed-form solutions for many performance measures. The situation changes completely when the service capacity can vary between several positive values. In the latter case, there are no general results available that describe the relation between performance measures and system parameters, such as the traffic arrival process and the service process. For the specific class of Markovian queueing models with a G/M/1 structure, there are efficient numerical solutions to determine performance measures using matrix-geometric techniques [66], in particular if the structure is further specialized to fall within the framework of Quasi Birth and Death (QBD) processes. However, the dependence of the performance on the system parameters still remains largely hidden in the solution of linear equations. For understanding the system behavior, an explicit relation between system parameters and performance measures is of the utmost importance.

1.4 Overview of the thesis

In this first chapter we have set the scope of this thesis. We have briefly outlined the role of queueing theory in application areas such as communication networks, and we gave an introduction to several relevant processor-sharing queueing models, in particular egalitarian and discriminatory PS, which play a central role in this thesis. In addition, we briefly discussed queues with time-fluctuating service capacity. The remainder of the thesis is presented as follows.

In Chapter 2 we introduce Wireless Local Area Networks (WLANs), paying particular attention to the resource sharing mechanism deployed in these networks. Next, we discuss the mapping of these resource sharing mechanisms to PS queueing models, for analyzing the network performance at flow-level. In addition, we give an overview of the literature on egalitarian and discriminatory PS models.

In Chapter 3, the (steady-state) queue length distribution in egalitarian PS models is investigated. In particular, we obtain a decomposition result of the marginal queue length distribution for multi-class egalitarian processor-sharing models. We show that the marginal queue length distribution for each class equals the queue length distribution of an equivalent single-class PS model with a random number of permanent customers. Permanent customers are customers who never leave the system, and the decomposition result implies linear relations between the marginal queue length probabilities, which also hold for other EPS models such as Cohen's *generalized* PS model. The decomposition result plays a crucial role in the derivation of bounds for the moments of the sojourn time distribution in the classical EPS queue (see Chapter 4), and motivates the approximation method for DPS models presented in Chapter 5. Chapter 3 is based on the results from Cheung, Van den Berg, and Boucherie [27].

In Chapter 4 we study the moments of the (conditional) sojourn time $T(x)$ of a customer given its initial service requirement $x > 0$, for the classical M/G/1 EPS queue. In particular, we derive explicit lower and upper bounds for all moments of $T(x)$, and these bounds are insensitive to the service requirement distribution apart from its mean. The lower bounds are based on Jensen's inequality, while the upper bounds are derived from the so-called 'instantaneous' sojourn time $\widehat{T}(x)$. The instantaneous sojourn time will be defined as the sojourn time of a customer with an infinitesimally small service requirement ($x \approx 0$). In addition to the bounds, stochastic ordering and moment ordering results for the sojourn time distribution are obtained. These results provide further support for the observation that egalitarian PS is a fair resource discipline. Chapter 4 builds upon the analysis of Cheung, Van den Berg, and Boucherie [28], and the proof for the upper bound is related to the PS model with random number of permanent customers from Chapter 3.

In Chapter 5, an approximation method for evaluating the mean sojourn time in *general* discriminatory processor-sharing (GDPS) models is proposed and investigated. The method is based on an approximate decomposition motivated by the exact decomposition result for EPS queues obtained in Chapter 3. The numerically efficient method is also applicable for DPS models with queue-dependent service capacity and queue-dependent service weights. We numerically show that the approximation is accurate for small to moderate differences in service weights. Chapter 5 is based upon Cheung, Van den Berg, and Boucherie [27].

In Chapter 6 we analyze the M/M/1 queue with discriminatory PS service discipline. Building on the work of Kim and Kim [56], we obtain the first and second moments of the so-called slowdown for this queue, which is defined as the ratio of the sojourn time to its job size, i.e., $T(x)/x$ for a job of size $x > 0$. The slowdown is a measure for queueing fairness: jobs in the standard EPS queue have a constant mean slowdown, i.e., the mean slowdown is independent of $x > 0$, which reflects the fairness of the EPS service discipline. For DPS, we discuss that a job of a certain size may sometimes be treated better or worse (in terms of slowdown) compared to a similar queueing model with equal weights, depending on the job size, arrival- and service rates, and the weights of all classes. This is illustrated with numerical examples. Chapter 6 presents the analysis of Cheung, Kim, and Kim [26].

In Chapter 7 we analyze a queueing system with time-fluctuating service capacity. The service capacity fluctuations are assumed to be driven by an independent Markov process. We allow the queue to be overloaded in some of the server states. In all but a few special cases, either exact analysis is not tractable, or the dependence of system performance in terms of input parameters (such as the traffic load) is hidden in complex or implicit characterizations. Various asymptotic regimes have been considered to develop insightful approximations. In particular, the so-called 'quasi-stationary' approximation has proven extremely useful under the assumption of uniform stability. In this chapter, we refine the quasi-stationary analysis to allow for temporary instability. We study the 'effective system load' and introduce a notion of 'adjusted stability'. For this we rely on a detailed analysis of the time needed to recover from the excess load after a low service rate period. The recovery time can be seen to be associated to the workload process in a so-called fluid queue driven by the same Markov process as the original queue. Chapter 7 presents the analysis of Cheung, Boucherie, and Núñez-Queija [25].

Finally, in Chapter 8, we present a flow-level performance analysis of Wireless Local Area Networks (WLANs) with Quality-of-Service (QoS) support. In particular, we propose a modeling and analysis approach based on the mapping of resource sharing

mechanisms in QoS enabled WLANs to GDPS queueing models. The priority weights in the GDPS flow-level model depend on the WLAN QoS differentiation parameters and on the number of active users in the system. Our analytical modeling approach is validated by detailed system simulations of QoS enabled WLANs. Chapter 8 builds upon the work of Cheung, Van den Berg, Boucherie, Litjens, and Roijers [29] and uses the GDPS approximation method developed in Chapter 5.

Chapter 2

Resource sharing in WLANs and processor-sharing queues

2.1 Wireless Local Area Networks (WLANs)

Wireless communication has become an integral part of our modern daily life as illustrated by the regular use of mobile cellular phones (GSM) allowing us to communicate almost anywhere and anytime. Another example that illustrates the relevance of wireless communications is the increasing deployment of Wireless Local Area Networks (wireless LANs or WLANs).

Wireless LAN was originally developed to be used for the connection (via a so-called Access Point) of mobile computing devices, such as laptops, to wired LANs. However, WLANs are now increasingly used for additional purposes, including Internet access and Voice over IP (VoIP) telephony. The popularity of wireless LANs is primarily due to the convenience of location freedom, cost efficiency, and ease of integration with other networks and network components. The majority of computers sold to consumers today come pre-equipped with all necessary WLAN technology. Wireless LANs also fulfill the need for an additional public wireless access solution in hot spots (e.g. train stations, airports, coffee shops, malls, etc.), besides the access provided by mobile cellular networks such as GSM/GPRS and UMTS. Wireless LANs offer low-cost capacity and higher bandwidths to end-users without sacrificing the inherently scarce and expensive capacity of cellular networks.

IEEE 802.11 standards

The development of WLAN technology is primarily driven by the data communications industry. After the first products appeared around 1990, the WLAN market grew substantially, and it has led to the release of the international IEEE 802.11 standard by the Institute of Electrical and Electronics Engineers (IEEE) in 1997. This first IEEE 802.11 legacy comprised of three different physical-layer technologies (including one based on infrared transmission) and it specifies two raw data rates of 1 and 2 megabits per second (Mbit/s). A weakness of this original specification was that it offered so many choices to vendors and customers that interoperability could hardly be realized. To ensure interoperability the next generation WiFi (Wireless Fidelity) or the IEEE 802.11B (supplemented) standard [2] was ratified in 1999.

The IEEE 802.11B standard supports data rates of 1, 2, 5.5, and 11 Mbit/s. Another standard (IEEE 802.11A) was ratified in 1999 for deployment in the 5 gigahertz (5 GHz) radio spectrum, achieving up to 54 Mbit/s. The third standard (IEEE 802.11G) was ratified in 2003, which combines the long range of 802.11B with the high throughput of 802.11A. The 802.11G standard works at a maximum data rate of 54 Mbit/s and is also backward-compatible with 802.11B products. Both 802.11B and 802.11G operate in the overly used 2.4 GHz frequency band.

One limitation of the previous standards is that they only support so-called 'best-effort' services; there is no notion of high or low priority traffic in the network and hence no Quality-of-Service (QoS) differentiation can be achieved. In 2001 the IEEE announced a new taskgroup to develop a QoS enhanced version of the original standard. The IEEE 802.11E protocol as of late 2005 has been approved as a standard that defines a set of QoS enhancements. The IEEE 802.11E standard is considered of critical importance for delay-sensitive applications, such as VoIP and Streaming Multimedia.

More standards are under development, for example the IEEE 802.11N and the IEEE 802.11P standards. The envisioned 802.11N standard is estimated to reach a theoretical 540 Mbit/s of real data throughput, by using so-called MIMO technology (multiple-input multiple-output), and is expected to be formalized in October 2008. The IEEE 802.11P standard will be used in Intelligent Transportation Systems (ITS) applications, for example, to increase safety for cars on highways, and is expected to be formalized in April 2009. For an up-to-date overview of all IEEE 802.11 standards available and under development we refer to [1].

2.2 Resource sharing in WLANs

In wireless LANs, data packets are sent over-the-air using a common radio channel. Therefore, 'packet collisions' may occur when different stations (users) are transmitting at the same time and the signal interference is too high. In order to regulate the access of stations to the wireless channel and to achieve appropriate sharing of the radio transmission resources, a Medium Access Control (MAC) protocol is used. In contrast to wired Ethernet, which uses Carrier Sense Multiple Access with Collision Detection (CSMA/CD), wireless LANs mostly use a Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) MAC protocol; collision detection is not possible in WLANs due to the nature of the wireless channel. The CSMA/CA based MAC protocol basically determines the resource sharing in IEEE 802.11 WLANs.

2.2.1 The CSMA/CA protocol

The CSMA/CA mechanism deploys a random access scheme to avoid packet collisions. If all stations would be 'greedy' in the sense they always attempt to access the channel immediately when the channel is sensed idle, then collisions would occur and the actually achieved throughput would be zero (unless there is only one station connected to an Access Point in the network). CSMA/CA in WLANs tries to improve the performance by attempting to be 'less greedy' on the channel.

The principle of the CSMA/CA MAC protocol is as follows. A station has to wait a random number of time slots before it is permitted to access the medium, which is regulated by a so-called back-off counting procedure. Every station has its own back-off counter, which is a non-negative integer value and sampled from an initial interval ('back-off window'). For every time slot the station senses the network idle, the back-off counter is decremented by one. When the back-off counter of a particular station reaches zero, the station transmits a packet. Hence, when the back-off counters of multiple stations reach zero at the same time slot, multiple packets are sent and a collision occurs. In that case, the stations will make another attempt to send their packets by restarting the back-off procedure with a new back-off counter, sampled from a larger back-off window. The latter has the effect of reducing the probability of collisions at the next attempts to send the packets (Collision Avoidance, or congestion control). For more technical details of the CSMA/CA mechanism used in the IEEE 802.11B and 802.11E standards we refer to Chapter 8.

Figure 2.1 gives a schematic view of how a CSMA/CA based MAC protocol distributes the available channel capacity among the different users (stations) in the wireless network on the packet-level. In this figure, we assume that there are three stations in the WLAN network all having multiple packets waiting for transmission. Note, that the channel remains unused at certain time slots which is due to the overhead created by the

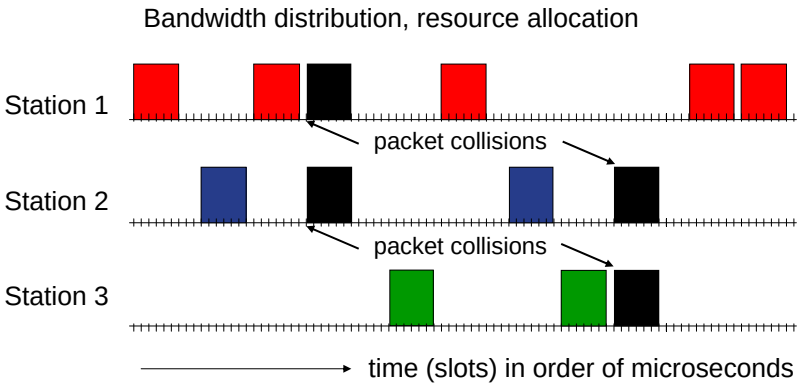

Figure 2.1: *A simplified abstraction of the resulting random access mechanism in a wireless channel with three stations at the packet-level.*

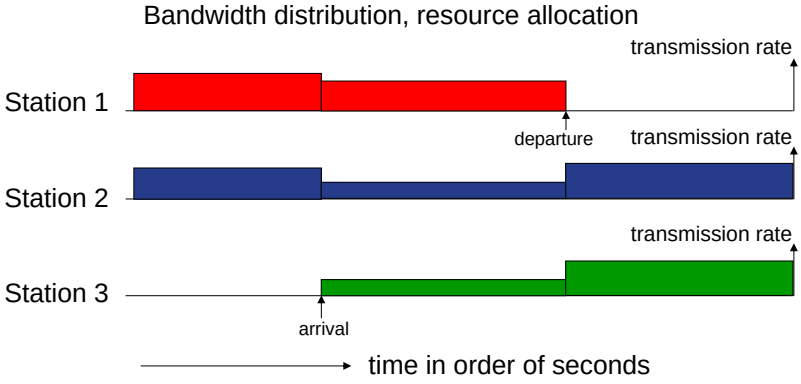

Figure 2.2: *A simplified abstraction of the bandwidth sharing mechanism in a wireless channel at the flow-level with dynamic user behavior.*

back-off counting mechanism. Channel capacity is in particular also wasted when collisions occur – illustrated by black boxes in Figure 2.1. As a result, the channel capacity is distributed among the users in a certain random manner, and actually depends on the network parameter settings and the network state. For example, when the number of active users in the network increases, there will be generally more collisions, and hence the aggregate system throughput will be lower.

By treating different types of traffic differently at the MAC level, Quality-of-Service differentiation (QoS) can be achieved. This may be desired since communication sessions usually also differ in QoS requirements. For instance, streaming video and Voice over Wireless IP conversation require a more or less constant throughput to ensure image and speech quality. On the other hand, data applications ('elastic traffic') are usually able to adapt their transmission rates to fluctuations in the available bandwidth. As mentioned in Section 2.1, the IEEE 802.11E standard enables QoS differentiation in WLANs. This is achieved, among others, by using different back-off window sizes for different traffic types.

2.2.2 Resource sharing in WLANs at the flow-level time-scale

From the user's perspective, for many data applications, the flow-level performance is more important than packet-level performance; a user is mostly only interested in the total transfer time of the file, and not in the transfer times of the individual packets.

Figure 2.2 shows what happens if the time-scale of Figure 2.1 is "zoomed out", from packet-level to a larger time-scale, i.e., the flow-level. At the origin of the time axis in Figure 2.2 we assume that there are two active stations in the network, where Station 1 has a higher average transmission rate than Station 2 – indicating that the traffic of Station 1 is treated with some higher priority than that of Station 2 (e.g., by using QoS differentiation at the MAC layer). On the flow-level, the capacity of the wireless network seems not to be shared in a random way anymore. Instead, on this larger time-scale, the packet-level randomness smoothes out over time and the stations experience a more or less fixed throughput, as long as the network state does not change.

Changes in the network state occur, for example, when additional stations become active initiating new data transfers or when stations finish their transmission. So, in practice, the actual throughput of a communication session will vary randomly over time – depending on the number of active stations – which in turn influences the time instants at which connections are terminated. In Figure 2.2, at the time instant when the new Station 3 joins the network and initiates a file transfer, the capacity allocation changes. Stations 2 and 3 are allocated the same capacity when these stations are active, which indicates that we do not differentiate the QoS between Stations 2 and 3. At the time instant when the "high-priority" Station 1 has finished its file transfer, Station 1 becomes inactive and the transmission rates for the other active stations will increase.

2.2.3 Processor-sharing modeling

In the field of performance evaluation of computer and communication systems, the PS discipline has been widely adopted as a convenient paradigm for modeling bandwidth sharing (e.g., see [14, 34]). The motivation for PS models in the WLAN setting can also be seen from Figure 2.2, where the actual capacity allocation is depicted from the flow-level point-of-view. The resource sharing mechanism for 'best-effort' WLANs may be modeled with an egalitarian processor-sharing discipline, whereas the QoS enabled WLANs may be modeled with a discriminatory processor-sharing discipline.

In these types of systems, a particularly relevant performance measure from a user's perspective is the file transfer time $T(x)$ for a file of a given size $x > 0$. Characterizing the distribution of the transfer time $T(x)$ is an important problem, where $T(x)$ could be class-specific for systems with multiple traffic types (e.g., DPS modeling of QoS enabled WLAN). In the following sections we give a more detailed overview of the basic performance results available in the literature for the egalitarian and discriminatory processor-sharing queues.

2.3 Egalitarian processor-sharing queueing literature

The *egalitarian* processor-sharing (EPS) queue has several appealing properties. From a practical point-of-view, one of the nice properties is that a small job can not get stuck behind large jobs, since fair sharing policies prevent large jobs from hogging the server, which is in sharp contrast to policies such as the First Come First Served (FCFS) discipline. FCFS policies have a significant negative impact on the performance of the system, when the service requirements are highly variable.

The stationary queue length distribution in the M/G/1 EPS queue is the geometric distribution [87, 88, 32]:

$$\pi_n = (1-\rho)\rho^n, \quad \text{for } n = 0, 1, ...,$$

which only depends on the traffic load $\rho = \lambda \mathbb{E}X < 1$, and where λ is the Poisson arrival rate and X is the random variable denoting the service requirement. The queue length distribution is said to be insensitive to the service requirement distribution $\mathbb{P}(X \leq x)$; it only depends on the service requirement distribution through its mean. By Little's law [69], the mean sojourn time is insensitive to the service requirement distribution as well. However, the sojourn time distribution does not have a nice closed-form characterization, which is in contrast to the simple geometric queue length distribution. Determining the sojourn time distribution in PS queues turned out to be a rather challenging problem, even for exponential service requirements.

Many studies in the literature have focused on the analysis of the sojourn time $T(x)$ conditioned on the initial service requirement $x > 0$. For the M/M/1 EPS queue, Klein-

rock [62, 63] showed that the conditional mean sojourn time is given by:

$$\mathbb{E}T(x) = x/(1-\rho),$$

which is proportional to the initial service requirement $x > 0$ of a customer. This proportionality result reflects a certain fairness of the EPS discipline, and was extended to the M/G/c case with general distributed service requirements and multiple servers in Sakata, Noguchi, and Oizumi [87, 88]; also see Kleinrock [65]. The proportionality and insensitivity results remain valid for Cohen's *generalized* PS model [32]. In the work he also obtained generalizations of known results for classical networks such as product-form and insensitivity property of the joint queue length distribution in general closed and open networks with multiple customer types.

For the M/M/1 EPS queue, Coffman, Muntz, and Trotter [31] first derived a closed-form expression for the Laplace-Stieltjes transform (LST) of the sojourn time $T_n(x)$ conditioned on the service requirement $x > 0$ and the number of customers $n \geq 0$ seen upon arrival. They showed that the LST of the "delay" $T_n(x) - x$ is given by:

$$\mathbb{E}\left[e^{-s(T_n(x)-x)}\right] = \frac{(1-\rho r^2)e^{-\lambda x(1-r)}}{(1-\rho r)+\rho r(1-r)e^{-\mu x(1-\rho r^2)/r}}\gamma^n, \quad \text{for } \mathrm{Re}(s) \geq 0,$$

where $\mu^{-1} = \mathbb{E}X$ is the mean service requirement, with

$$\gamma = \frac{r(1-\rho r)+(1-r)e^{-\mu x(1-\rho r^2)/r}}{(1-\rho r)+\rho r(1-r)e^{-\mu x(1-\rho r^2)/r}}$$

and r is the smaller root the the quadratic equation

$$\lambda r^2 - r(\lambda+\mu+s)+\mu = 0.$$

Based on this work, Morrison [76] established an expression for the distribution function $\mathbb{P}(T > t)$ of the (unconditional) sojourn time T, which can be represented as, for $t > 0$,

$$\mathbb{P}(T>t) = 2\int_0^\pi \left(\frac{\exp\{-\theta\left[2\sqrt{\rho}-(1+\rho)\cos\theta\right]\}}{(1-\rho)\sin\theta} - \frac{(1-\rho)^2 t}{1+\rho-2\sqrt{\rho}\cos\theta}\right)\times$$
$$\times\,(1-\rho)^{-1}\left(1+\exp\left\{\frac{-\pi[2\sqrt{\rho}-(1+\rho)\cos\theta]}{(1-\rho)\sin\theta}\right\}\right)^{-1}\sin\theta d\theta.$$

This result is useful for numerical evaluation of $\mathbb{P}(T > t)$ for different values of the traffic intensity $\rho < 1$, and also for investigation of heavy-traffic behavior when ρ is close to 1. An alternative expression for the LST of the sojourn time distribution conditioned only on the number of customers seen upon arrival was found by Sengupta and Jagerman [91].

For the M/G/1 EPS queue, the exact determination of the distribution of the conditional sojourn time $T(x)$ given its initial service requirement $x > 0$ was an open problem for a long time. Several analytic solutions have been obtained. Yashkov [106] first found an expression for the conditional sojourn time in terms of double Laplace transforms by writing the sojourn time as a functional on a branching process; the LST $v(s,x) = \mathbb{E}\left[e^{-sT(x)}\right]$ of $T(x)$, for $\mathrm{Re}(s) \geq 0$, is expressed as:

$$v(s,x) = \frac{(1-\rho)e^{-(s+\lambda)x}}{\psi(s,x) - \lambda\int_0^x e^{-(s+\lambda)y}\psi(s,x-y)\overline{B}(y)dy - \lambda e^{-(s+\lambda)x}\int_x^\infty \overline{B}(y)dy}$$

with

$$\psi(s,x) = \frac{1}{2\pi i}\int_{-i\cdot\infty}^{+i\cdot\infty} \frac{q+s+\lambda\beta(q+s+\lambda)}{(q+s+\lambda)[q+\lambda\beta(q+s+\lambda)]}e^{qx}dq$$

and the LST of the service requirement distribution $\beta(s) = \mathbb{E}e^{-sX}$ and its complementary distribution function $\overline{B}(y) = \mathbb{P}(X > y)$.

Via different approaches, similar results for $v(s,x)$ were obtained by Ott [82], Schassberger [89], and Van den Berg and Boxma [16]. However, although several transform expressions have been obtained, these expressions are fairly complex, and not particularly insightful or readily applicable for computational purposes. The expression for $v(s,x)$ obtained by Zwart and Boxma [112] which avoids the contour integrals in the expressions of [106, 82, 89], is the most explicit expression that is obtained in the literature. They showed that $v(s,x)$ can be written as

$$v(s,x) = \left(\sum_{n=0}^{\infty}\frac{s^n}{n!}\alpha_n(x)\right)^{-1}, \tag{2.1}$$

where the coefficients $\alpha_n(x)$ are related to the waiting-time distribution in an equivalent M/G/1 queue with FCFS service discipline: $\alpha_0(x) := 1$, and for $n \geq 1$,

$$\alpha_n(x) = \frac{n}{(1-\rho)^n}\int_{y=0}^{x}(x-y)^{n-1}R^{(n-1)*}(y)dy, \tag{2.2}$$

with $R^{n*}(y)$ denoting the n-fold convolution of the waiting-time distribution $R(y)$ in the M/G/1 FCFS queue. It is shown that for $n \geq 1$ (cf. [112]):

$$R^{n*}(y) = (1-\rho)^n\sum_{m=0}^{\infty}\binom{m+n-1}{n-1}\rho^m\widetilde{B}^{m*}(y), \quad \text{and } R^{0*}(y) := 1, \tag{2.3}$$

where $\widetilde{B}^{m*}(y)$ is the m-fold convolution of the integrated tail- or excess service requirement distribution $\widetilde{B}(y) = \frac{1}{\mathbb{E}X}\int_0^y(1-B(u))du$.

As a consequence of the form of the LST in (2.1), it is shown in [112] that the moments $v_k(x) = \mathbb{E}T(x)^k$ can be calculated recursively, as $v_0(x) := 1$ and for $k \geq 1$,

$$v_k(x) = -\sum_{j=1}^{k} \binom{k}{j} v_{k-j}(x)\alpha_j(x)(-1)^j. \tag{2.4}$$

In particular, the first three moments are given by, in terms of the coefficients $\alpha_j(x)$,

$$\begin{aligned} v_1(x) &= \alpha_1(x) = \frac{x}{1-\rho}, \\ v_2(x) &= 2\frac{x^2}{(1-\rho)^2} - \alpha_2(x), \\ v_3(x) &= 6\frac{x^3}{(1-\rho)^3} - 6\frac{x}{1-\rho}\alpha_2(x) + \alpha_3(x). \end{aligned}$$

The known expressions for $v(s,x)$ in the M/G/1 EPS queue lead to, at best, complicated recursive formulas for the moments which have mainly been examined only asymptotically, see e.g. [112]. In Chapter 4 we will derive insensitive upper and lower bounds for all moments of the conditional sojourn time distribution, which provide further support for the fairness principle of EPS.

For additional and related work on EPS queues, readers may refer to Asare and Foster [7], Yashkov [107, 108, 109], Grishechkin [43], Kitayev [59], Whitt [104], Ward and Whitt [102], Núñez-Queija [78, 79], Guillemin and Boyer [44], Bansal [11], Egorova, Zwart, and Boxma [36], Mandjes and Zwart [71], Brandt and Brandt [23], Hampshire, Harchol-Balter, and Massey [46], Kim and Kim [57], and references therein. For a recent survey on sojourn time asymptotics we refer to Borst, Núñez-Queija, and Zwart [22].

2.4 Discriminatory processor-sharing queueing literature

The analysis for EPS models does not easily extend to *discriminatory* processor-sharing (DPS) models. Therefore, results for DPS are scarce in the literature. Most notably, the simple geometric queue length distribution for the ordinary EPS queue does not have any counterpart for DPS, and tractable transform results for the sojourn time distribution seem not to exist, not even for exponential service requirements.

After Kleinrock's introduction of the Priority Processor Sharing model [63] in 1967, Fayolle, Mitrani, and Iasnogorodski [38] made the most important progress in the DPS analysis in 1980. They showed that the conditional mean sojourn times $\mathbb{E}T_i(x)$, for class $i = 1, ..., K$, satisfy a system of integro-differential equations for the M/G/1 DPS queue. Unfortunately, the system of equations that was first obtained by O'Donovan [81]

contained an error. The corrected form of the integro-differential equations are given by [38]:

$$\mathbb{E}T_i'(x) = 1 + \sum_{j=1}^{K} \int_0^{\infty} \lambda_j \frac{\alpha_j}{\alpha_i} \mathbb{E}T_j'(y) \left[1 - B_j\left(y + \frac{\alpha_j}{\alpha_i}x\right)\right] dy \qquad (2.5)$$
$$+ \int_0^{x} \mathbb{E}T_i'(y) \sum_{j=1}^{K} \lambda_j \frac{\alpha_j}{\alpha_i} \left[1 - B_j\left(\frac{\alpha_j}{\alpha_k}(x-y)\right)\right] dy,$$

with $\mathbb{E}T_i'(x) := \frac{d}{dx}\mathbb{E}T_i'(x)$ and initial condition $\mathbb{E}T_i(0) = 0$ for all classes $i = 1, ..., K$, and service requirement distribution $B_j(x)$ for class $j = 1, ..., K$. Under the assumption that the second moments of the service requirements distributions are finite, it is shown in [38] that the system of integro-differential equations (2.5) has a unique solution:

$$\mathbb{E}T_i(x) = \alpha_i \int_0^{x/\alpha_i} a(t)dt + \int_0^{x/\alpha_i} b(t)dt, \quad \text{for } i = 1, ..., K, \qquad (2.6)$$

where $a(x)$ is the unique solution of the defective renewal equation

$$a(x) = 1 + \int_0^{x} a(y)\Psi(x-y)dy$$

with $\Psi(x) = \sum_{j=1}^{K} \lambda_j \alpha_j \left(1 - B_j(\alpha_j x)\right)$, and $b(x)$ satisfies

$$b(x) = \sum_{j=1}^{K} \lambda_j \alpha_j^2 \int_0^{\infty} a(y) \left[1 - B_j(\alpha_j(x+y))\right] dy$$
$$+ \int_0^{\infty} b(y)\Psi(x+y)dy + \int_0^{x} b(y)\Psi(x-y)dy.$$

In the case of exponentially distributed service requirements, [38] derived closed-form expressions for the conditional mean sojourn times and obtained the unconditional mean sojourn times from a system of linear equations.

Further progress in the DPS analysis was made in 1996 by Rege and Sengupta [85]. They obtained the moments of the queue length distributions as the solutions to linear equations for the case of exponential service requirements, and they also proved a heavy-traffic limit theorem for the joint queue length distribution. More recently, these results were extended to phase-type distributions by Van Kessel, Núñez-Queija, and Borst [53]. Other more recent results were obtained by Kim and Kim [56], who found the moments of the sojourn time in the M/M/1 DPS queue as a solution of linear simultaneous equations (also see Chapter 6 where the slowdown in the M/M/1 DPS queue is studied).

Avrachenkov, Ayesta, Brown, and Núñez-Queija [10] showed that the mean queue lengths of all classes are finite under the usual stability condition $\rho < 1$, regardless of the

higher moments of the service requirements. In particular, they showed that the integro-differential equations (2.5) has the unique solution (2.6) without the assumption of finite second moments of the service requirements, leaving the usual stability condition as a sufficient condition. They also showed that the conditional sojourn times of the different customer classes are stochastically ordered according to the DPS weights. For a recent survey on DPS we refer to Altman, Avrachenkov, and Ayesta [4].

Chapter 3

Decomposition of EPS models into permanent customer queues

In this chapter we obtain an exact and analytically tractable decomposition for multi-class egalitarian PS models. More specifically, for a two-class egalitarian PS queue with Poisson arrivals at rates λ_1, λ_2, when (N_1, N_2) is the joint steady-state queue length, we show that the marginal queue length N_i is in distribution equal to $\widetilde{N}_i(N_j)$, with $i \neq j$. The latter random variable denoted as $\widetilde{N}_i(N_j)$ can be interpreted as the steady-state queue length of a *single*-class M/G/1 EPS queue with arrival rate λ_i and with a random number of *permanent* customers distributed as N_j (i.e., the marginal queue length of the *other* type in the original two-class EPS queue). Permanent customers are customers who never leave the system, or equivalently, customers who have an infinitely large service requirement.

The queue length decomposition result implies that the marginal queue length distribution for class 1 factorizes over the number of class 2 customers, where the factorizing coefficients are equal to the queue length probabilities of an isolated EPS queue for class 1, given that class 2 customers are permanent in the system. This decomposition result can be generalized for an arbitrary number of classes K, and similar results hold for other EPS models, such as EPS networks with feedback customers (see Section 3.4), and (E)PS models with queue-dependent but *balanced* class capacities (see Section 3.5). Based on the queue length decomposition result we propose an approximation method for discriminatory PS models in Chapter 5.

3.1 Multi-class PS models

In this section we introduce a general single-server processor-sharing model with K customer classes and we introduce the notation used. Customers arrive at a single server according to individual and independent Poisson processes with rate $\lambda_i > 0$ for customer class i, $i = 1, ..., K$. The service requirements of class i customers are i.i.d. random variables with a general distribution $B_i(x) = \mathbb{P}(X_i \leq x)$ with mean $\mathbb{E}X_i$. Denote the load of class i by $\rho_i = \lambda_i \mathbb{E}X_i$, and the total offered load by $\rho := \sum_{j=1}^{K} \rho_j$. The server shares its capacity among all customers present in the system. Denote by $\mathbf{n} = (n_1, ..., n_K)$ the system state with n_j customers of type j. The server capacity may be dependent on the system state. When the system state is $\mathbf{n}$, the total rate class i receives is $\phi_i(\mathbf{n})$. All customers within class i share the capacity $\phi_i(\mathbf{n})$ in an egalitarian manner, i.e., each customer in class i receives rate $\phi_i(\mathbf{n})/n_i$. The total server capacity is denoted by $\phi(\mathbf{n}) := \sum_{i=1}^{K} \phi_i(\mathbf{n})$. We make the following assumption.

Assumption 3.1. $\phi_i(\mathbf{n}) = 0$ *if and only if* $n_i = 0$.

This general model describes a very wide class of service disciplines. In particular, it includes the following special cases of processor-sharing models.

a. EPS (with fixed capacity): $\phi_i(\mathbf{n}) = \frac{n_i}{\sum_{j=1}^{K} n_j}$.

b. EPS with queue-dependent service capacity: $\phi_i(\mathbf{n}) = \frac{n_i \phi(\mathbf{n})}{\sum_{j=1}^{K} n_j}$.

c. DPS (with fixed capacity): $\phi_i(\mathbf{n}) = \frac{\alpha_i n_i}{\sum_{j=1}^{K} \alpha_j n_j}$.

This model framework also covers GPS [83, 101] models, and DPS models with queue-dependent service capacity $\phi(\mathbf{n})$ and queue-dependent service weights $\alpha_i(\mathbf{n})$. The DPS models with both queue-dependent capacity and queue-dependent weights are called *generalized* discriminatory processor-sharing (GDPS) models.

The egalitarian PS models **a.** and **b.** are analytically tractable (when $\phi(\mathbf{n})$ only depends on $\mathbf{n}$ through its sum $n_1 + ... + n_K$). Note that in the *generalized* PS model studied by Cohen [32], $\phi(\mathbf{n})$ only depends on $\mathbf{n}$ through its sum $n_1 + ... + n_K$. In particular, analytical expressions are available for the equilibrium distributions of the numbers of customers simultaneously present in the system (and marginal distributions), mean number of customers $\mathbb{E}N_i$ of class i, mean sojourn time $\mathbb{E}T_i$ and conditional mean sojourn time $\mathbb{E}T_i(x)$ of a class i customer given its initial service requirement $x > 0$. For GDPS models, these expressions have not yet been obtained in tractable form.

3.2 Queue length decomposition

In this section, we first establish queue length decomposition results for the ordinary egalitarian PS model. Results for more general egalitarian PS models are treated in Sections 3.4 and 3.5.

Consider an egalitarian processor-sharing model with two types of customers (indexed by $l = 1, 2$), class capacity functions $\phi_l(\mathbf{n}) := \phi_l(n_1, n_2) = \frac{n_l}{n_1+n_2}$, and where the second class of customers is possibly an aggregate of several other classes. Let the random vector (N_1, N_2) denote the joint steady-state queue length of this processor-sharing model; its distribution has the product-form (cf. [32, 51])

$$\mathbb{P}(N_1 = i; N_2 = j) = (1-\rho)\binom{i+j}{j}\rho_1^i\rho_2^j, \tag{3.1}$$

when the stability condition is satisfied, i.e., $\rho := \rho_1 + \rho_2 < 1$, and is insensitive to the service requirement distributions apart from their means; see e.g. [20]. By appropriate summation of (3.1) the marginal queue length probabilities are given by, for $i, j \in \mathbb{Z}_+$,

$$\mathbb{P}(N_1 = i) = \frac{1-\rho}{1-\rho_2}\left(\frac{\rho_1}{1-\rho_2}\right)^i, \tag{3.2}$$

$$\mathbb{P}(N_2 = j) = \frac{1-\rho}{1-\rho_1}\left(\frac{\rho_2}{1-\rho_1}\right)^j. \tag{3.3}$$

Theorem 3.2 shows how the marginal steady-state queue length probabilities of the two-class EPS queue can be related through the negative binomial probabilities $a(i, j)$ and $b(j, i)$, defined as

$$a(i,j) := \mathbb{P}\left(\widetilde{N}_1(j) = i\right) = (1-\rho_1)^{j+1}\binom{i+j}{i}\rho_1^i, \tag{3.4}$$

$$b(j,i) := \mathbb{P}\left(\widetilde{N}_2(i) = j\right) = (1-\rho_2)^{i+1}\binom{i+j}{j}\rho_2^j, \tag{3.5}$$

with

$$\sum_{i=0}^{\infty} a(i,j) = 1 \text{ for all } j \in \mathbb{Z}_+, \text{ and } \sum_{j=0}^{\infty} b(j,i) = 1 \text{ for all } i \in \mathbb{Z}_+.$$

The random variable denoted as $\widetilde{N}_k(m)$ has a distribution which is the $(m+1)$-fold convolution of the distribution of $\widetilde{N}_k$, and $\widetilde{N}_k$ denotes the steady-state queue length of an *isolated* M/G/1 EPS queue with arrival rate λ_k and general service requirement distribution $B_k(x)$, i.e.,

$$\mathbb{P}\left(\widetilde{N}_k = n\right) = (1-\rho_k)\rho_k^n.$$

Note that $\widetilde{N}_k(0) \equiv \widetilde{N}_k$, and assume that $\widetilde{N}_i$ is independent of N_j, for $i \neq j$.

Theorem 3.2. *For $i, j = 1, 2$ and $i \neq j$, the marginal queue length N_i is in distribution equal to the random variable $\widetilde{N}_i(N_j)$, i.e.,*

$$N_i \stackrel{d}{=} \widetilde{N}_i(N_j),$$

where $\widetilde{N}_i(N_j) := \sum_{l=0}^{N_j} \widetilde{N}_{i,l}$, with $\left\{\widetilde{N}_{i,l}\right\}_{l\geq 0}$ i.i.d. and distributed as $\widetilde{N}_i$.

Proof. First observe that the following equality holds by combining the equations (3.1)-(3.5):

$$\mathbb{P}(N_1 = i; N_2 = j) = a(i,j)\mathbb{P}(N_2 = j) = b(j,i)\mathbb{P}(N_1 = i).$$

Hence, by conditioning on $\{N_2 = j\}$ and independence of $\widetilde{N}_1$ and N_2, we have for all $i \in \mathbb{Z}_+$

$$\mathbb{P}\left(\widetilde{N}_1(N_2) = i\right) = \sum_{j=0}^{\infty} a(i,j)\mathbb{P}(N_2 = j) = \mathbb{P}(N_1 = i),$$

i.e., $\widetilde{N}_1(N_2) \stackrel{d}{=} N_1$. Analogously, we have that $\widetilde{N}_2(N_1) \stackrel{d}{=} N_2$. ∎

Corollary 3.3. *From Theorem 3.2 we obtain the following set of linear equations:*

$$\mathbb{P}(N_1 = i) = \sum_{j=0}^{\infty} a(i,j)\mathbb{P}(N_2 = j),$$
$$\mathbb{P}(N_2 = j) = \sum_{i=0}^{\infty} b(j,i)\mathbb{P}(N_1 = i).$$

The above Corollary forms the basic approximation assumption for DPS models in Chapter 5. The decomposition theorem can be generalized to an arbitrary number of classes K, and also to multi-class egalitarian PS models with a queue-dependent service capacity that only depends on the total number of customers in the system; see Section 3.5.

Theorem 3.2 can be interpreted as follows. In the two-class EPS model, the marginal queue length N_1 is in distribution equal to a queue length from a related M/G/1 queue with permanent customers. The latter M/G/1 queue has j additional permanent customers with probability $\mathbb{P}(N_2 = j)$. To this end, note that the queue length distribution in an ordinary M/G/1 EPS queue with j permanent customers, equals the $(j+1)$-fold convolution of the queue length distribution of the same model without permanent customers; see Van den Berg [15]. The remarkable fact is that $a(i,j) = \mathbb{P}\left(\widetilde{N}_1 = i \mid j \text{ permanent customers}\right)$ and

$$\mathbb{P}\left(N_1 = i \mid N_2 = j\right) = \frac{\mathbb{P}\left(N_1 = i; N_2 = j\right)}{\mathbb{P}\left(N_2 = j\right)}$$

are identical and independent of ρ_2.

From the class 1 point-of-view in the original two-class processor-sharing model, it seems as if class 1 behaves according to an ordinary single-class M/G/1 EPS queue with arrival rate λ_1 and with j additional permanent customers in the system, if j customers of type 2 are present in the system. Furthermore, if there is a customer arrival (resp. departure) for type 2 in the system, then it seems as if class 1 *instantaneously* 'jumps' to the same M/G/1 model with arrival rate λ_1, but now with $j+1$ (resp. $j-1$) permanent customers, and as if the new equilibrium (steady-state behavior) is instantaneously attained at the jump epoch.

3.3 Sojourn time decomposition

After establishing the queue length decomposition result, a natural question is whether or not a similar decomposition result holds for the sojourn time distribution. It can be shown that a similar decomposition holds for the first moment of the conditional sojourn time distribution (see Theorem 3.4), where $T_i(x)$ is the sojourn time for customer type i with initial service requirement $x > 0$ in the original two-class PS model, and $\widetilde{T}_i(x; N_j)$ is the conditional sojourn time of the isolated M/G/1 EPS queue with arrival rate λ_i and a random number of permanent customers N_j. For a fixed $m \geq 0$, the distribution of $\widetilde{T}_i(x; m)$ is given by the $(m+1)$-fold convolution of the distribution of $\widetilde{T}_i(x; 0) = \widetilde{T}_i(x)$, where $\widetilde{T}_i(x)$ is the conditional sojourn time in the single-class M/G/1 EPS queue with traffic load ρ_i and service requirement distribution $B_i(x)$ (e.g. see [15, 104]).

Theorem 3.4. *For all $x \geq 0$, the conditional mean sojourn times can be decomposed into*

$$\mathbb{E}T_1(x) = \mathbb{E}\left(\widetilde{T}_1(x; N_2)\right) \equiv \sum_{j=0}^{\infty} \frac{(j+1)x}{1-\rho_1} \mathbb{P}(N_2 = j), \tag{3.6}$$

$$\mathbb{E}T_2(x) = \mathbb{E}\left(\widetilde{T}_2(x; N_1)\right) \equiv \sum_{i=0}^{\infty} \frac{(i+1)x}{1-\rho_2} \mathbb{P}(N_1 = i), \tag{3.7}$$

where $(m+1)x/(1-\rho_k)$ is the mean conditional sojourn time of an isolated M/G/1 PS queue with arrival rate λ_k, service requirement distribution $B_k(x)$ and $m \geq 0$ permanent customers.

Proof. From the equations (3.2)-(3.3), and (3.6)-(3.7) it is readily verified that

$$\mathbb{E}\left(\widetilde{T}_1(x; N_2)\right) = \sum_{j=0}^{\infty} \mathbb{E}\left(\widetilde{T}_1(x) \mid j \text{ permanent customers}\right) \cdot \mathbb{P}(N_2 = j),$$

and

$$\mathbb{E}\left(\widetilde{T}_1(x;N_2)\right) = \sum_{j=0}^{\infty} \frac{(j+1)x}{1-\rho_1} \cdot \frac{1-\rho_1-\rho_2}{1-\rho_1} \left(\frac{\rho_2}{1-\rho_1}\right)^j$$
$$= \frac{1-\rho_1-\rho_2}{(1-\rho_1)^2} \left(\frac{1-\rho_1}{1-\rho_1-\rho_2}\right)^2 x = \frac{x}{1-\rho_1-\rho_2}.$$

Analogously, $\mathbb{E}\left(\widetilde{T}_2(x;N_1)\right) = x/(1-\rho_1-\rho_2)$ which is the same as the well-known result: $\mathbb{E}T_1(x) = \mathbb{E}T_2(x) = x/(1-\rho)$. ∎

Example 3.5. *The similar decomposition for the sojourn time distribution conditional on the initial service requirement $x > 0$, i.e.,*

$$T_1(x) \stackrel{d}{=} \widetilde{T}_1(x;N_2), \tag{3.8}$$

$$T_2(x) \stackrel{d}{=} \widetilde{T}_2(x;N_1), \tag{3.9}$$

does not hold in general. To this end, take for example $\lambda_1 > 0$ and $\lambda_2 = 0$, then it is not difficult to see that

$$\widetilde{T}_1(x;N_2) \stackrel{d}{=} \widetilde{T}_1(x), \tag{3.10}$$

$$\widetilde{T}_2(x;N_1) \stackrel{d}{=} (N_1+1)x, \tag{3.11}$$

since $N_2 = 0$. The random variable $(N_1+1)x$ is insensitive to the service requirement distributions, while $\widetilde{T}_1(x)$ is not insensitive to the service requirement distribution. However, since the sojourn times are conditional on a fixed service requirement (in the original two-class EPS model), it must hold that $T_1(x) \stackrel{d}{=} T_2(x)$ and hence (3.8) and (3.9) can not hold.

Obviously, the result for unconditional mean sojourn times is similar, since it also follows directly from the exact decomposition result for queue length distributions and Little's law. For higher moments of the (conditional) sojourn distribution, it can be easily seen that a similar decomposition does not hold in general; see Example 3.5. However, if both customer classes have the same service requirement distribution, then a certain stochastic ordering result can be proven (see Chapter 4). More specifically, when the arrival rates satisfy $\lambda_1 \geq \lambda_2$, then the moment ordering

$$\mathbb{E}\left(\widetilde{T}_1(x;N_2)\right)^k \leq \mathbb{E}\left(\widetilde{T}_2(x;N_1)\right)^k$$

holds for all $x \geq 0$ and $k \geq 2, k \in \mathbb{N}$. From this moment ordering result, insensitive upper bounds for *all* moments of the sojourn time distribution (conditioned on the initial service requirement) for an ordinary M/G/1 EPS queue will be derived in Chapter 4. In addition, these upper bounds have a special structure with so-called Eulerian numbers in the expressions.

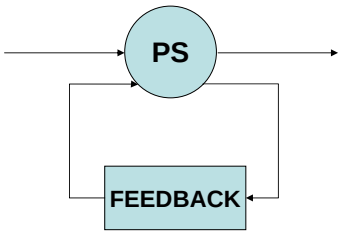

Figure 3.1: *A network with one EPS node and one feedback node.*

3.4 An EPS network with feedback node

Now we consider a processor-sharing network with an egalitarian PS node and a node used by a single feedback customer, see Figure 3.1. Exogenous customer arrivals at the PS node form a Poisson process with rate $\lambda > 0$, and these customers are served at the PS node with i.i.d. service requirements; generally distributed with mean $\mathbb{E}X$. The service requirement for the feedback customer at the PS node is generally distributed and denoted by the random variable Z. After service completion of the feedback customer at the PS node, the feedback customer is routed to the feedback node (with probability 1) where he spends a generally distributed time Y. After this random time Y at the feedback node, the feedback customer joins the PS node for a service requirement Z.

If we denote by $\mathbb{P}(N^{PS} = n)$ the steady-state distribution of the number of (non-feedback) customers at the PS node, then it is readily verified that the following decomposition holds:

$$\mathbb{P}(N^{PS} = n) = \xi \cdot \pi_0(n) + (1 - \xi) \cdot \pi_1(n),$$

where $\pi_m(\cdot)$ is the steady-state queue length distribution of the isolated M/G/1 PS queue with arrival rate λ and with m permanent customers, and ξ is the steady-state probability (fraction of time) that the feedback customer is at the feedback node in the network, i.e.,

$$\pi_0(n) = (1 - \rho)\rho^n,$$
$$\pi_1(n) = (1 - \rho)^2(n + 1)\rho^n,$$

with $\xi = \frac{\mathbb{E}Y}{\mathbb{E}Y + \mathbb{E}Z/(1-\rho)}$ and $\rho := \lambda\mathbb{E}X$.

Take for example, X exponentially distributed with mean μ^{-1}, Y exponentially distributed with mean ν^{-1}, Z exponentially distributed with mean θ^{-1}, and define the joint distribution

$$\pi\,(n, i) := \mathbb{P}(N^{PS} = n; N_{FB}^{PS} = i), \quad \text{for } n \geq 0 \text{ and } i \in \{0, 1\},$$

where N^{PS} is the number of non-feedback customers, and N_{FB}^{PS} is the number of feedback customers, both at the PS node in steady-state. In this case, the global balance

equations are given by, for $n = 0$:

$$(\lambda + \nu)\pi(0,0) = \theta \cdot \pi(0,1) + \mu \cdot \pi(1,0),$$
$$(\lambda + \theta)\pi(0,1) = \nu \cdot \pi(0,0) + \frac{1}{2}\mu \cdot \pi(1,1),$$

and, for $n \geq 1$:

$$(\lambda + \mu + \nu)\pi(n,0) = \lambda \cdot \pi(n-1,0) + \frac{\theta}{n+1}\pi(n,1) + \mu \cdot \pi(n+1,0),$$
$$(\lambda + \frac{\theta}{n+1} + \frac{n\mu}{n+1})\pi(n,1) = \lambda \cdot \pi(n-1,1) + \nu \cdot \pi(n,0) + \frac{n+1}{n+2}\mu \cdot \pi(n+1,1).$$

Then, it can be verified that the solution is given by, for $n \geq 0$,

$$\pi(n,0) = (1-\rho)\rho^n \cdot \xi,$$
$$\pi(n,1) = (1-\rho)^2(n+1)\rho^n \cdot (1-\xi),$$

where

$$\xi = \frac{\theta(1-\rho)}{\theta(1-\rho)+\nu} = \frac{1/\nu}{1/\nu + 1/(\theta(1-\rho))}.$$

In fact, the solution also satisfies the *detailed balance* equations (see e.g. [51]), and the result can be extended to multiple feedback customers where the feedback node is a so-called BCMP [12] node. The feedback node may also be replaced by a BCMP network.

3.5 Multi-class EPS with queue-dependent capacity

Now we consider the egalitarian processor-sharing queue with K customer classes, with the total service capacity dependent on the system state $\mathbf{n}$ through its sum $n_1 + ... + n_K$, cf. [32]. More precisely, $\phi(\mathbf{n}) = \varphi(\mathbf{n} \cdot \mathbf{e})$, for all $\mathbf{n} \neq \mathbf{0}$ (null vector), where $\mathbf{e}$ is the vector with 1-entries of appropriate length, $\mathbf{n} \cdot \mathbf{e}$ denotes the inner product and where $\varphi(\cdot) : \mathbb{N} \to \mathbb{R}_+$ is an arbitrary positive function. Serving the customers in egalitarian manner is equivalent to

$$\frac{\phi_i(\mathbf{n})}{n_i} = \frac{\varphi(\mathbf{n} \cdot \mathbf{e})}{\mathbf{n} \cdot \mathbf{e}}, \text{ for all } i = 1, ..., K. \tag{3.12}$$

Remark 3.6. ***(balance property)*** *The class capacities $\phi_i(\mathbf{n})$ are uniquely characterized and balanced by*

$$\phi_i(\mathbf{n}) = \frac{\Phi(\mathbf{n} - \mathbf{e}_i)}{\Phi(\mathbf{n})},$$

where $\Phi(\mathbf{n})$ is the so-called balance function, and $\mathbf{e}_i$ is the i-th unity vector of appropriate length; see [51, 20]. It is said that the class capacities $\phi_i(\mathbf{n})$ are balanced if a

function $\Phi(\mathbf{n})$ exists such that $\phi_i(\mathbf{n}) = \Phi(\mathbf{n} - \mathbf{e}_i)/\Phi(\mathbf{n})$ is satisfied, and equivalently, the class capacities $\phi_i(\mathbf{n})$ are balanced if

$$\frac{\phi_i(\mathbf{n} - \mathbf{e}_j)}{\phi_i(\mathbf{n})} = \frac{\phi_j(\mathbf{n} - \mathbf{e}_i)}{\phi_j(\mathbf{n})}, \quad \textit{for all } i, j, \textit{ and } n_i > 0, n_j > 0.$$

From Eq. (3.12) and the balance property, we get

$$\Phi(\mathbf{n}) = \frac{(\mathbf{n} \cdot \mathbf{e})!}{\prod_{i=1}^{K} n_i!} \left(\prod_{j=1}^{\mathbf{n} \cdot \mathbf{e}} \varphi(j) \right)^{-1},$$

and without restriction $\varphi(0) \equiv 1$. The joint steady-state queue length distribution $\pi(\mathbf{n}) := \mathbb{P}(N_1 = n_1; ...; N_K = n_K)$ is given by the product-form

$$\pi(\mathbf{n}) = (\mathbf{n} \cdot \mathbf{e})! \left(G \prod_{j=1}^{\mathbf{n} \cdot \mathbf{e}} \varphi(j) \right)^{-1} \prod_{i=1}^{K} \rho_i^{n_i}/n_i!, \text{ for } \mathbf{n} \neq \mathbf{0},$$

with $\rho_i = \lambda_i \mathbb{E}X_i$ and a normalizing constant G, see [20]. We will show that the marginal distributions of $\pi(\mathbf{n})$ can be decomposed into queue length distributions of (isolated) queues with permanent customers.

Theorem 3.7. *For multi-class egalitarian processor-sharing models, with balanced class capacities*

$$\phi_k(\mathbf{n}) = \frac{\varphi(\mathbf{n} \cdot \mathbf{e})}{\mathbf{n} \cdot \mathbf{e}} n_k,$$

the marginal steady-state queue length distribution can be decomposed into

$$N_k \stackrel{d}{=} \widetilde{N}_k(N_{-k}),$$

for all $k = 1, ..., K$, and where N_{-k} is defined by $N_{-k} := \sum_{i=1, i \neq k}^{K} N_i$.

Proof. The decomposition for class k follows from the observation that

$$\prod_{j=1}^{\mathbf{n} \cdot \mathbf{e}} \varphi(j)^{-1} \equiv \left(\prod_{l=1}^{n_k} \varphi(l + (\mathbf{n} \cdot \mathbf{e} - n_k)) \right)^{-1} \left(\prod_{j=1}^{\mathbf{n} \cdot \mathbf{e} - n_k} \varphi(j) \right)^{-1}, \text{ for } n_k \geq 1, \quad (3.13)$$

with $\varphi(0) \equiv 1$. Hence, with (3.13) and by appropriate summation of $\pi(\mathbf{n})$, the marginal queue length distribution for class k equals

$$\mathbb{P}(N_k = n_k) \sim \sum_{\substack{n_1, ..., n_{k-1} \\ n_{k+1}, ..., n_K}} (\mathbf{n} \cdot \mathbf{e})! \left(\prod_{j=1}^{\mathbf{n} \cdot \mathbf{e}} \varphi(j) \right)^{-1} \prod_{i=1}^{K} \rho_i^{n_i}/n_i! =$$

$$= \sum \frac{(\mathbf{n}\cdot\mathbf{e})!}{n_k!} \left\{ \prod_{l=1}^{n_k} \varphi(l + (\mathbf{n}\cdot\mathbf{e} - n_k))^{-1} \rho_k^{n_k} \right\} \left(\prod_{j=1}^{\mathbf{n}\cdot\mathbf{e}-n_k} \varphi(j)^{-1} \prod_{i\neq k}^{K} \frac{\rho_i^{n_i}}{n_i!} \right), \quad (3.14)$$

where the symbol $\sim$ denotes equality up to a multiplicative constant. The proof is readily completed, by observing that the expression between parentheses in equation (3.14) is equivalent to the queue length distribution for type k in isolation and with the number of $(\mathbf{n}\cdot\mathbf{e} - n_k)$ permanent customers of the other classes, i.e.,

$$\mathbb{P}(\widetilde{N}_k = n_k \mid \mathbf{n}\cdot\mathbf{e} - n_k \text{ permanent customers}).$$

The expression after the parentheses in (3.14) is equivalent to the marginal steady-state probability $\mathbb{P}(N_{-k} = \mathbf{n}\cdot\mathbf{e} - n_k)$, after appropriate summation. ∎

3.6 Conclusions

In this chapter, we obtained a decomposition result for the queue length distributions in egalitarian processor-sharing models. In particular, for a multi-class egalitarian processor-sharing model, the marginal steady-state queue length distribution N_k for class k, satisfies $N_k \stackrel{d}{=} \widetilde{N}_k(N_{-k})$. The latter random variable $\widetilde{N}_k(N_{-k})$ can be interpreted as a random variable denoting the queue length of an isolated processor-sharing queue for class k, where the other customer types are permanent customers in the system and N_{-k} represents the random variable of the total number of permanent customers. This result remains valid for egalitarian processor-sharing models with a queue-dependent system capacity that only depends on the total number of customers, or equivalently, for balanced PS networks.

For DPS models, these queue length decomposition results cannot hold in general, since DPS models are not balanced and not insensitive to the service requirement distributions. However, we will investigate a similar decomposition of the marginal queue length distribution into queue length distributions of PS models with permanent customers, as an approximation in Chapter 5.

Chapter 4

Stochastic orderings for the sojourn time in the M/G/1 EPS queue

In the present chapter we derive results for the moments of the conditional sojourn time $T(x)$ in the classical M/G/1 EPS queue, and we also study the sojourn time when the initial service requirement is arbitrarily small. We define $\widehat{T}(x)$ as the *instantaneous* sojourn time, i.e., the sojourn time of a customer with infinitesimally small service requirement. We show that the instantaneous sojourn time for arbitrarily small $x > 0$ leads to a moment ordering result between $\widehat{T}(x)$ and $T(x)$ for arbitrary $x > 0$. More specifically, the main result of this chapter is that the moments of the instantaneous sojourn time provide upper bounds for all moments of the conditional sojourn time, which generalizes the upper bound for the second moment in Van den Berg [15]. Additionally, stochastic ordering results for the M/G/1 EPS queue and also for EPS models with a random number of permanent customers are obtained.

The upper bounds have the valuable characteristic of insensitivity requiring only knowledge of the traffic intensity ρ, and not of higher moments of the service requirement distribution. The upper bounds are also tight in a few appropriate senses, namely for all jobs with a small service requirement ($x \to 0$), and for all jobs in systems with heavy-traffic ($\rho \to 1$) or light-traffic ($\rho \to 0$). The latter valuable property follows from the fact that for $\rho \to 0$, the upper bounds coincide with the insensitive lower bounds given by the Jensen's inequality.

This chapter uses some results of the M/G/1 EPS queue from Chapter 2. The remainder of this chapter is organized as follows. In Section 4.1 we establish the existence of insensitive upper (and lower) bounds for *all* conditional moments of the sojourn time,

with a particular polynomial structure in ρ. In Section 4.2 we derive a Laplace transform ordering result and show that $T(x)$ belongs to the so-called $\mathcal{L}$-class of life-time distributions. Then, in Section 4.3, the instantaneous sojourn time $\widehat{T}(x)$ is introduced and we give readily applicable expressions using the so-called Eulerian numbers. Via stochastic comparison of the permanent customer model in Section 4.4, we prove our main result that the moments of $\widehat{T}(x)$ provide upper bounds for the moments of $T(x)$ in Section 4.5. The instantaneous sojourn time can also be seen as the conditional sojourn time in a so-called quasi-stationary regime, see Section 4.6. The stochastic and moment ordering results in this chapter provide simple characterizations of $T(x)$ under EPS, which provide further support for the observation that the EPS service discipline is 'fair' from a tagged customer's perspective.

4.1 Moment bounds for the conditional sojourn time

In this section, we establish the existence of insensitive bounds for all moments of the conditional sojourn time distribution, which have the form:

$$1 \leq (1-\rho)^k v_k(x)/x^k \leq \psi_{k-1}(\rho),$$

where $\psi_{k-1}(\rho)$ is a polynomial in ρ of (at most) degree $k-1$ and with non-negative coefficients. For the second moment of the conditional sojourn time in the M/G/1 EPS queue, Van den Berg [15] obtained the following simple bounds:

$$\frac{1}{(1-\rho)^2}x^2 \leq v_2(x) \leq \frac{1+\rho}{(1-\rho)^2}x^2, \tag{4.1}$$

simply by using the fact that $R(0) = 1-\rho > 0$, see Eq. (2.3). We note that the upper bound for the second moment is $100\rho\%$ larger than the lower bound, and these bounds only depend on the mean service requirement and *not* on the second and higher moments. From (4.1) it is also interesting to note that $T(x)$ has a coefficient of variation between 0 and $\sqrt{\rho}$.

By using the recursive formula (2.4) for $v_k(x)$ and 'ignoring' the alternating term $(-1)^j$, the following crude upper bound for all moments can be given:

$$v_k(x) \leq k! \left(\frac{(e-1)x}{1-\rho}\right)^k,$$

see also Zwart [111]. As a consequence of this bound, the sojourn time $T(x)$ is always light-tailed conditional upon its service requirement.

The crude bound for the second moment is always worse than the upper bound given in (4.1), since $1+\rho < 2 < 2!(e-1)^2$. Furthermore, for $\rho \to 0$, we have the attractive property that the upper and lower bounds in (4.1) coincide. Now we will generalize (4.1)

for *all* moments, by using the recursive formula (2.4) and using the simple observation as for the second moment in [15], to obtain 'tight' bounds with a similar structure as (4.1).

Theorem 4.1. *For all $k \geq 2$, there exist non-negative constants $c_i^k \geq 0$, such that $v_k(x)$ is bounded by*

$$\frac{1}{(1-\rho)^k}x^k \leq v_k(x) \leq \frac{\psi_{k-1}(\rho)}{(1-\rho)^k}x^k, \tag{4.2}$$

where $\psi_{k-1}(\rho) = \sum_{i=0}^{k-1} c_i^k \rho^i$ is a polynomial in ρ of degree $k-1$ (if k even) or $k-2$ (if k odd) and $c_0^k = 1$.

Proof. The lower bound in (4.2) is straightforward by applying Jensen's inequality. For the upper bound, we note that $1-\rho \leq R(y) \leq 1$ and hence $(1-\rho)^n \leq R^{n*}(y) \leq 1$. Therefore, by using (2.2) we obtain upper and lower bounds for $\alpha_n(x)$:

$$\frac{x^n}{1-\rho} \leq \alpha_n(x) \leq \frac{x^n}{(1-\rho)^n}, \quad \text{for } n \geq 1, \text{ and } \alpha_0(x) := 1. \tag{4.3}$$

Existence of the upper bound as in (4.2) is obtained by induction. We rewrite the recursive formula (2.4) as

$$\overline{v}_k(x) = -\sum_{j=1}^{k} \binom{k}{j} \overline{v}_{k-j}(x)\overline{\alpha}_j(x)(-1)^j,$$

where $\overline{v}_k(x) := (1-\rho)^k v_k(x)/x^k$ and $\overline{\alpha}_j(x) := (1-\rho)^j \alpha_j(x)/x^j$. From (4.3), bounds for $\overline{\alpha}_j(x)$ are given by:

$$(1-\rho)^{j-1} \leq \overline{\alpha}_j(x) \leq 1, \quad \text{for } j \geq 1.$$

Assume (induction hypothesis) that the following bounds hold for $\overline{v}_{k-1}(x), \overline{v}_{k-2}(x), ...$:

$$1 \leq \overline{v}_{k-j}(x) \leq \psi_{k-j-1}(\rho),$$

and where the bounds for $\overline{v}_0(x)$ and $\overline{v}_1(x)$ are satisfied by definition, since $\overline{v}_0(x) := 1$, and $\overline{v}_1(x) := 1$. Then, we have bounds for the product $\overline{v}_{k-j}(x)\overline{\alpha}_j(x)$, for $j = 1, ..., k$:

$$1+\sum_{i=1}^{j-1}(-\rho)^i \binom{j-1}{i} = (1-\rho)^{j-1} \leq \overline{v}_{k-j}(x)\overline{\alpha}_j(x) \leq \psi_{k-j-1}(\rho) = 1+ \sum_{i=1}^{k-j-1} c_i^{k-j}\rho^i.$$

Now, apply induction and take into account the alternating term $(-1)^j$ (hence we need both upper and lower bounds for $\overline{v}_{k-j}(x)\overline{\alpha}_j(x)$) to obtain the upper bound for $\overline{v}_k(x)$. Hence, straightforward term-by-term bounding (and 'splitting the positive and negative

terms' in the recursive formula) gives:

$$\begin{aligned}\overline{v}_k(x) &= \sum_{\substack{j=1\\ j:\text{ odd}}}^{k}\binom{k}{j}\overline{v}_{k-j}(x)\overline{\alpha}_j(x) - \sum_{\substack{j=2\\ j:\text{ even}}}^{k}\binom{k}{j}\overline{v}_{k-j}(x)\overline{\alpha}_j(x)\\ &\le \sum_{\substack{j=1\\ j:\text{ odd}}}^{k}\binom{k}{j}\left\{1+\sum_{i=1}^{k-j-1}c_i^{k-j}\rho^i\right\} - \sum_{\substack{j=2\\ j:\text{ even}}}^{k}\binom{k}{j}\left\{1+\sum_{i=1}^{j-1}(-\rho)^i\binom{j-1}{i}\right\}\\ &\equiv \sum_{i=0}^{k-1}c_i^k\rho^i = \psi_{k-1}(\rho). \end{aligned} \tag{4.4}$$

By definition of the coefficients c_i^k in (4.4) and by comparing the terms, it is not difficult to see that: $c_0^k = 1$ for all k, $c_{k-1}^k = 1$ if k is even, and $c_{k-1}^k = 0$ if k is odd. Furthermore, it can be shown that $c_i^k \ge 0$, where the coefficients are constructed as in (4.4). However, for the existence of an upper bound of the described structure, it is not necessary to show that $c_i^k \ge 0$, since c_i^k can always be chosen sufficiently large (and finite). ∎

Remark 4.2. *In principle, we can apply the 'alternating' procedure to obtain a lower bound as well. However, the resulting lower bound is always worse than the Jensen's lower bound. The Jensen's lower bound is obtained via the recursive formula if the coefficient $\alpha_n(x)$ is replaced by the upper bound $x^n/(1-\rho)^n$ for all $n \ge 1$. Hence, the procedure of term-by-term bounding as in the proof of Theorem 4.1 for obtaining a lower bound (as well as for an upper bound) for $v_k(x)$, is too conservative. The latter fact can also be argued from the dependency of the coefficients $\{\alpha_n(x), \alpha_{n+1}(x), ...\}$, $n \ge 2$. For example, if $\alpha_n(x)$ is 'close to its lower bound' $x^n/(1-\rho)$, then $\alpha_{n+1}(x)$ is generally 'not close to its upper bound' $x^{n+1}/(1-\rho)^{n+1}$. In fact, for a fixed $x > 0$, if it holds that $\alpha_n(x) = x^n/(1-\rho)$ for some $n \ge 2$, then necessarily $\alpha_n(x) = x^n/(1-\rho)$ for all $n \ge 2$. The latter observation will be important (see the similar observation in Lemma 4.14); the sequence $\alpha_n(x) = x^n/(1-\rho)$ for $n \ge 1$, also uniquely defines the so-called instantaneous sojourn time, which will be introduced in Section 4.3.*

For the second moment, we obtain $c_1^2 = 1$ since $k = 2$ is even, and the upper bound is the same as in (4.1). As a direct consequence of Theorem 4.1 we have the following Corollary 4.3, which states that all conditional moments are finite in the stable M/G/1 EPS system. This result is in sharp contrast with the stable M/G/1 FCFS queue, which provides further support for the observation that EPS is a 'fair' service discipline.

Corollary 4.3. *If $\rho < 1$, then $v_k(x) < \infty$ for all $k \ge 1$.*

For the moments of the sojourn time in the stable M/G/1 FCFS queue, it is known that the k-th moment exists if and only if $\mathbb{E}X^{k+1}$ is finite. For the PS case, the k-th moment of the (unconditional) sojourn time exists if and only if $\mathbb{E}X^k$ is finite (see [112]).

4.2 Laplace transform ordering and $\mathcal{L}$-class

In this section we obtain some stochastic ordering results for the distribution of $T(x)$. For stochastic ordering theory we refer to Stoyan [95], and Shaked and Shanthikumar [92]. We first establish a Laplace transform ordering for $T(x)$. Then, a characterization that $T(x)$ belongs to the so-called $\mathcal{L}$-class of life time distributions will be derived, which is related to the Laplace transform ordering.

The stochastic ordering in Laplace transforms denoted by $Y \geq_{Lt} X$, for any non-negative random variables X and Y, i.e., $v(s) = \mathbb{E}e^{-sY} \leq \mathbb{E}e^{-sX} = w(s)$, $\mathrm{Re}(s) \geq 0$, is generally a weak ordering; it only implies $\mathbb{E}Y \geq \mathbb{E}X$. If in addition $\mathbb{E}Y = \mathbb{E}X$ is known besides the ordering $v(s) \leq w(s)$, then it can be easily shown that $\mathbb{E}Y^2 \leq \mathbb{E}X^2$, see Proposition 4.4. Implications for higher moments cannot be made in general. For $T(x)$ in the M/G/1 EPS case we have a stronger Laplace transform ordering result; see Theorem 4.6.

Proposition 4.4. *For any non-negative random variables X and Y, with $\mathbb{E}Y = \mathbb{E}X$ and the LST ordering $v(s) = \mathbb{E}e^{-sY} \leq \mathbb{E}e^{-sX} = w(s)$, $\mathrm{Re}(s) \geq 0$, it holds that: $\mathbb{E}Y^2 \leq \mathbb{E}X^2$.*

Proof. By $\mathbb{E}Y = \mathbb{E}X$, the tangent line of $v(s)$ at $s = 0$ is equal to the tangent line of $w(s)$ at $s = 0$. Then, by convexity and analyticity of LSTs, and the ordering $v(s) \leq w(s)$, it is readily seen that $\frac{d^2}{ds^2}v(s) \leq \frac{d^2}{ds^2}w(s)$ for s in a neighborhood of 0. Hence, $\mathbb{E}Y^2 \leq \mathbb{E}X^2$. ∎

Definition 4.5. ***(Klefsjö [61])*** *It is said that $T(x)$ belongs to the $\mathcal{L}$-class of life time distributions if the LST ordering $v(s,x) \leq z(s,x)$ holds, $\mathrm{Re}(s) \geq 0$, where $z(s,x)$ is the LST of an exponential distribution with mean $x/(1-\rho)$.*

Theorem 4.6. *For the stable M/G/1 EPS queue, the LST $v(s,x)$ of $T(x)$ is bounded by*

$$\begin{aligned} e^{-sx/(1-\rho)} &\leq v(s,x) \\ &\leq \widehat{v}(s,x) = \frac{1-\rho}{e^{sx}-\rho} \\ &\leq z(s;x) = \frac{1}{1+sx/(1-\rho)}, \quad \mathrm{Re}(s) \geq 0, \end{aligned}$$

where $\widehat{v}(s,x) := \mathbb{E}e^{-s(N+1)x}$ is the LST of $(N+1)x$; and $z(s;x)$ is the LST of an exponential random variable with mean $x/(1-\rho)$. In addition, $T(x) \in \mathcal{L}$, i.e., the conditional sojourn time belongs to the $\mathcal{L}$-class of life time distributions.

Proof. The bounds

$$e^{-sx/(1-\rho)} \leq v(s,x) \leq \frac{1-\rho}{e^{sx}-\rho}$$

follow straightforwardly from (2.1) with the bounds (4.3), and it is also straightforwardly shown that $\frac{1-\rho}{e^{sx}-\rho}$ coincides with $\widehat{v}(s,x) := \mathbb{E}e^{-s(N+1)x}$, if $\rho < 1$. The inequality $\widehat{v}(s,x) \leq z(s;x)$ follows from:

$$\frac{1-\rho}{e^{sx}-\rho} \leq \frac{1-\rho}{1+sx-\rho} =: z(s,x),$$

where $z(s,x)$ is clearly the LST of an exponential distribution with mean $x/(1-\rho)$. Hence, $T(x) \in \mathcal{L}$. ∎

Distributions belonging to the $\mathcal{L}$-class of life time distributions always have a finite second moment, and the coefficient of variation is not greater than one; see e.g. [17, 67]. More interestingly, although the $\mathcal{L}$-class is a wide class of distributions, Klar [60] obtained explicit and sharp 'reliability bounds' for any $\mathcal{L}$-class distribution. As an application of these reliability bounds (Theorem 4.1 from [60]), for the conditional sojourn time distribution in the M/G/1 EPS queue, we obtain the next corollary.

Corollary 4.7. *For $y \leq x/(1-\rho)$ we have the insensitive lower bound*

$$\mathbb{P}(T(x) > y) \geq 1 - \frac{1}{(y(1-\rho)/x)^2 - 2y(1-\rho)/x + 2}.$$

For $y > x/(1-\rho)$ we have the insensitive upper bound

$$\mathbb{P}(T(x) > y) \leq \frac{1}{(y(1-\rho)/x)^2 - 2y(1-\rho)/x + 2}.$$

Remark 4.8. *Stronger results for the reliability bounds exist for life time distributions belonging to subclasses of the $\mathcal{L}$-class.*

4.3 The instantaneous sojourn time

In this section we introduce the *instantaneous* sojourn time $\widehat{T}(x)$, defined as the sojourn time of an infinitesimally small job. The key idea is as follows. A customer with a (very small) initial service requirement $x > 0$ arrives at the system in steady state, say at time t_0. By the PASTA (Poisson Arrivals See Time Averages) property, the tagged customer sees n other customers upon arrival with probability $\pi_n = (1-\rho)\rho^n$. If we denote the remaining service requirements of the n other customers at time t_0 by x_i, $i = 1, ..., n$, then we may assume that $x << \min_{i=1,...,n} x_i$. Furthermore, we assume that x is small enough such that no other customers arrive during the time interval $[t_0, t_0 + (n+1)x)$. Under these assumptions, it is as if the tagged customer arrived at a system with n *permanent* customers with probability π_n and with no other arriving customers.

Hence, the sojourn time of the tagged customer is $(n+1)x$ with probability π_n. We define the *instantaneous* sojourn time as

$$\widehat{T}(x) = (N+1)x,$$

where N is distributed as $\mathbb{P}(N=n) = \pi_n$. The k-th moment of the true sojourn time $T(x)$ can be approximated with the k-th moment of the instantaneous sojourn time $\widehat{v}_k(x)$, as $x \to 0$,

$$v_k(x) \approx \widehat{v}_k(x) := \mathbb{E}\widehat{T}(x)^k = \sum_{n=0}^{\infty} \pi_n \left\{(n+1)x\right\}^k, \quad k \in \mathbb{N}.$$

Clearly, it holds that $\widehat{v}_1(x) = \mathbb{E}(N+1)x = x/(1-\rho)$, and thus the instantaneous sojourn time $\widehat{T}(x)$ as an approximation for $T(x)$, is exact for the first moment $v_1(x)$ and even for arbitrary $x > 0$. This will be an important fact for the rest of the chapter. It can also be shown that $T(x)/x \xrightarrow{d} N+1$, as $x \to 0$; the convergence in distribution (denoted by $\xrightarrow{d}$) follows from:

$$\lim_{x\to 0} \frac{\alpha_n(x)}{x^n} = \frac{1}{1-\rho}, \quad \text{for } n \geq 1,$$

cf. (2.2) and (2.3), and hence

$$v(s/x;x)^{-1} = 1 + \sum_{n=1}^{\infty} \frac{s^n}{n!} \frac{\alpha_n(x)}{x^n} \to \frac{e^s - \rho}{1-\rho} \quad \text{as } x \to 0,$$

where $\frac{e^s-\rho}{1-\rho}$ is the reciprocal of the LST of $(N+1)$, see also the last comment in Remark 4.2 and Theorem 4.6.

Interestingly, the moments $\widehat{v}_k(x)$ can be written explicitly by using the so-called Eulerian numbers. Eulerian numbers appear in many contexts in various fields of mathematics (number theory, combinatorics) and in various special functions (sinc functions, polylogarithms, etc.); we refer to [24, 42, 93, 94, 103] and references therein. An interpretation of the Eulerian number $\left\langle {k \atop j} \right\rangle$ is that it counts the total number of permutations of the ordered set $\{1, ..., k\}$ that have j 'permutation ascents'. The first five rows of the Euler's number triangle are given by

$$\begin{array}{ccccccccc}
 & & & & \left\langle {1 \atop 0} \right\rangle & & & & \\
 & & & \left\langle {2 \atop 0} \right\rangle & & \left\langle {2 \atop 1} \right\rangle & & & \\
 & & \left\langle {3 \atop 0} \right\rangle & & \left\langle {3 \atop 1} \right\rangle & & \left\langle {3 \atop 2} \right\rangle & & \\
 & \left\langle {4 \atop 0} \right\rangle & & \left\langle {4 \atop 1} \right\rangle & & \left\langle {4 \atop 2} \right\rangle & & \left\langle {4 \atop 3} \right\rangle & \\
\left\langle {5 \atop 0} \right\rangle & & \left\langle {5 \atop 1} \right\rangle & & \left\langle {5 \atop 2} \right\rangle & & \left\langle {5 \atop 3} \right\rangle & & \left\langle {5 \atop 4} \right\rangle
\end{array}$$

with the corresponding values:

$$\begin{array}{ccccccccccc}
 & & & & & 1 & & & & & \\
 & & & & 1 & & 1 & & & & \\
 & & & 1 & & 4 & & 1 & & & \\
 & & 1 & & 11 & & 11 & & 1 & & \\
 & 1 & & 26 & & 66 & & 26 & & 1 &
\end{array}$$

Shenton and Bowman [93, 94] studied geometric distributions, and obtained an unusual recurrence relation for its cumulants, with 'cumulant components' that involve Eulerian numbers. To the best of our knowledge, the raw moments of a *shifted* geometric distribution on $\{1, 2, ...\}$, are not explicitly stated in the existing literature, in the form of Theorem 4.9.

Theorem 4.9. *The k-th moment of the instantaneous sojourn time is given by*

$$\widehat{v}_k(x) = \frac{x^k}{(1-\rho)^k} \sum_{j=0}^{k-1} \left\langle {k \atop j} \right\rangle \rho^j, \quad \textit{for } k \in \mathbb{N}.$$

Proof. Use the identity (e.g., see [103]):

$$\sum_{n=1}^{\infty} n^k r^n \equiv \frac{1}{(1-r)^{k+1}} \sum_{i=0}^{k} \left\langle {k \atop i} \right\rangle r^{k-i} = \frac{r}{(1-r)^{k+1}} \sum_{i=0}^{k-1} \left\langle {k \atop i} \right\rangle r^{k-i-1},$$

where the last equality sign relies on the fact that $\left\langle {k \atop k} \right\rangle = 0$. Then, we readily derive

$$\frac{\widehat{v}_k(x)}{x^k} = \sum_{n=0}^{\infty} \pi_n (n+1)^k = \frac{1-\rho}{\rho} \sum_{n=1}^{\infty} n^k \rho^n = \frac{\sum_{i=0}^{k-1} \left\langle {k \atop i} \right\rangle \rho^{k-i-1}}{(1-\rho)^k} = \frac{\sum_{j=0}^{k-1} \left\langle {k \atop j} \right\rangle \rho^j}{(1-\rho)^k},$$

where $\left\langle {k \atop i} \right\rangle = \left\langle {k \atop k-i-1} \right\rangle$ is used in the last equality sign (by symmetry of Euler's number triangle). ∎

In Figure 4.1, as an illustration of the instantaneous sojourn time, we have depicted $v_k(x)$ and $\widehat{v}_k(x)$ for the M/M/1 EPS queue (together with the Jensen's lower bound $v_1(x)^k$), for $k = 2, 3, 4$, on a small and large scale for x, respectively. Figure 4.2 depicts the moments, all properly scaled with $(1-\rho)^k/x^k$ for the large scale of x. As expected, $\widehat{v}_k(x)$ is a good approximation for $v_k(x)$ when x is small (and even for x up to the mean $\mathbb{E}X$). The approximation is loose for large x, since $T(x)/x \xrightarrow{\mathbb{P}} 1/(1-\rho)$ as $x \to \infty$, where $\xrightarrow{\mathbb{P}}$ denotes convergence in probability. In fact, for $k \geq 2$, we have an asymptotic estimate whenever $\mathbb{E}X^2 < \infty$ (cf. [112]):

$$v_k(x) = \left(\frac{x}{1-\rho}\right)^k + \frac{\lambda k(k-1)\mathbb{E}X^2}{2(1-\rho)^{k+1}} x^{k-1} + o(x^{k-1}), \quad x \to \infty.$$

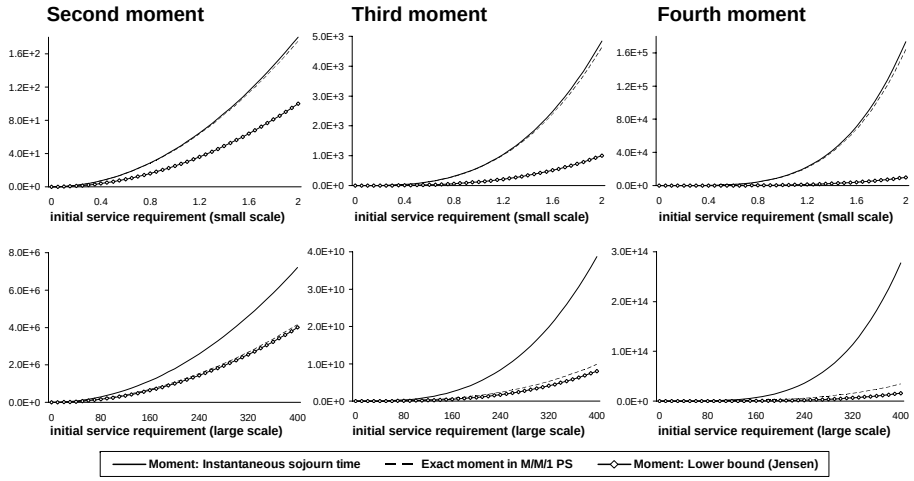

Figure 4.1: *The moments $v_k(x)$ in the M/M/1 EPS queue with $\lambda = 0.4$, $\mathbb{E}X = 2, \rho = 0.8$, and the instantaneous sojourn time moments $\widehat{v}_k(x)$, and the Jensen's lower bound $v_1(x)^k$, for $k = 2, 3, 4$. Upper graphs: $x \in (0, \mathbb{E}X]$ (small scale for x); and lower graphs: $x \in (0, 200\mathbb{E}X]$ (large scale).*

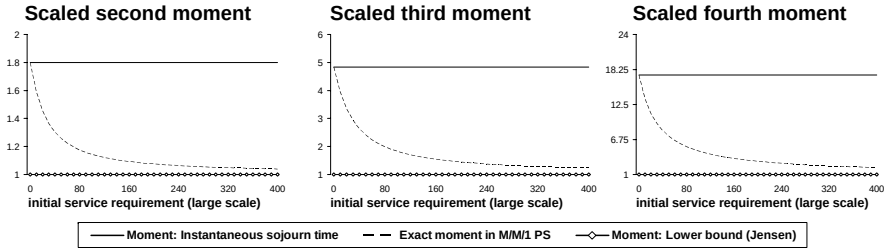

Figure 4.2: *The scaled moments $(1-\rho)^k v_k(x)/x^k$ in the M/M/1 EPS queue ($\lambda = 0.4$, $\mathbb{E}X = 2$), with $(1-\rho)^k \widehat{v}_k(x)/x^k$ as the scaled instantaneous sojourn time moments and the lower bound of 1.*

Note that $\widehat{v}_k(x)$ does not use knowledge of the higher moments of the service requirement distribution; $\widehat{v}_k(x)$ is also properly defined when $\mathbb{E}X^2 = \infty$.

Interestingly, the approximation for the second moment:

$$v_2(x) \approx \widehat{v}_2(x) = \frac{1+\rho}{(1-\rho)^2}x^2$$

for small $x > 0$, yields in fact $v_2(x) \leq \widehat{v}_2(x)$ for arbitrary $x > 0$, see (4.1). This might suggest that the moments of the instantaneous sojourn time $\widehat{v}_k(x)$ are upper bounds for $v_k(x)$, for all $k \geq 2$. In Section 4.5 we will prove our main result that $v_k(x) \leq \widehat{v}_k(x)$ for all $x \geq 0$ and $k \in \mathbb{N}$; see Theorem 4.17. The upper bounds hold under general service requirement distributions. An intuitive explanation is given in the next remark.

Remark 4.10. ***(intuition for the upper bound)*** *In the instantaneous sojourn time analysis ($x \to 0$) we assumed that during a time interval of length $(n+1)x$, there is no other activity in the system. When n is large upon arrival, then this is not very likely: $\widehat{T}(x)$ overestimates the true sojourn time $T(x)$ when n is large upon arrival of the tagged customer, and underestimates $T(x)$ when n is small upon arrival. Apparently, for the first moment: over- and under estimation outweigh each other (weighted with probability π_n). For higher moments: overestimation is weighted more heavily than underestimation, since $\rho < 1$ and thus the queue length process shows a negative drift for a large initial value of the number of customers present in the system.*

We note that $T(x)$ and $\widehat{T}(x)$ have a similar heavy-traffic behavior (when $\rho \to 1$). From the identity $\sum_{j=0}^{k-1} \left\langle {k \atop j} \right\rangle = k!$ (i.e., the row sums of the Euler's number triangle), it follows that:

$$\lim_{\rho \to 1}(1-\rho)^k \widehat{v}_k(x) = k! x^k.$$

For $T(x)$ it is known that: $\lim_{\rho\to 1}(1-\rho)^k v_k(x) = k!x^k$ (cf. [112]), or in fact,

$$\mathbb{P}\left((1-\rho)T(x)/x \leq y\right) \to 1 - e^{-y}, \quad \text{as } \rho \to 1,\ y \geq 0,$$

i.e., $(1-\rho)T(x)/x$ converges in distribution to an exponential random variable with mean 1, when $\rho \to 1$ (cf. [109]). We also note that $\widehat{v}_k(x) = v_k(x) = x^k$ (deterministic) for $\rho \to 0$.

The above observations and Remark 4.10 suggest that $\widehat{v}_k(x)$ are (insensitive) upper bounds and tight in an appropriate sense. Furthermore, since $\{k!x^k/(1-\rho)^k\}_{k\geq 1}$ is the moment sequence of an exponentially distributed random variable $Z(x)$ with mean $x/(1-\rho)$, it seems that $T(x)$ is 'less variable' than $Z(x)$, in the *convex stochastic order* sense. In the next Section 4.4 and 4.5 we obtain more precise stochastic ordering results together with the formal proof that the instantaneous sojourn time moments are upper bounds for $v_k(x)$, for all $x \geq 0$ and $k \in \mathbb{N}$, with Eulerian numbers as coefficients for the polynomials in ρ.

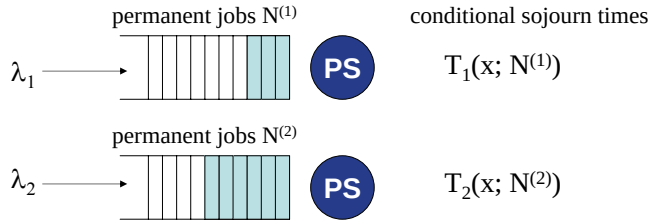

Figure 4.3: *Comparison of two related single-server M/G/1 EPS queues with random number of permanent customers.*

4.4 Model with random number of permanent customers

In this section, we will prove our main result (Theorem 4.17):

$$v_k(x) \leq (1 + \sum_{i=1}^{k-1} \left\langle {k \atop i} \right\rangle \rho^i) / \left[(1-\rho)/x\right]^k .$$

This moment ordering result follows from a more general moment ordering result between two PS queues with a random number of *permanent* customers, see Theorem 4.15 and Figure 4.3. In Section 4.3, the instantaneous sojourn time $\widehat{T}(x)$ is defined as the sojourn time of an infinitesimally small job. Alternatively, $\widehat{T}(x)$ can also be viewed as the sojourn time of a customer (with an arbitrary service requirement x) that enters a PS system with no other arriving customers but with a random number of permanent customers that is distributed as π_n. The latter viewpoint turns out to be convenient for proving our main result in Section 4.5.

We proceed with constructing two related single-server EPS queues as illustrated in Figure 4.3; both M/G/1 queues have the same service requirement distribution $B(x)$ with mean $\mathbb{E}X$; they only have different Poisson arrival rates, λ_1 and λ_2 respectively. We let $T_i(x)$ denote the conditional sojourn time in the M/G/1 EPS queue with arrival rate λ_i, $i = 1, 2$. Next, we define the random variable $T_i(x; n)$ as the conditional sojourn time in the M/G/1 EPS queue with arrival rate λ_i, but now modified with n permanent customers in the system. The distribution of $T_i(x; n)$ is given by the $(n+1)$-fold convolution of the distribution of $T_i(x)$, also see Section 3.3 in Chapter 3. Note that $T_i(x) \equiv T_i(x; 0)$.

Remark 4.11. *Note that we use the notation $T_i(x; n)$ in Chapter 4, while we use $\widetilde{T}_i(x; n)$ in Chapter 3; the tilde only indicates (and emphasizes) that the sojourn time refers to an isolated single-class queue, which can be omitted in this chapter concerning the classical single-class M/G/1 EPS queue.*

We let $N^{(i)}$, $i = 1, 2$, be geometrically distributed with probability density function

$$\mathbb{P}(N^{(i)} = n) = \frac{1-\rho}{1-\rho_i}\left(\frac{\rho-\rho_i}{1-\rho_i}\right)^n, \quad n \in \mathbb{N} \cup \{0\}. \tag{4.5}$$

Hence the random variable $T_i(x; N^{(i)})$ can be interpreted as the conditional sojourn time in the M/G/1 EPS queue with arrival rate λ_i and with a random number of permanent customers distributed as $N^{(i)}$, where $\rho_i = \lambda_i \mathbb{E}X$, and assume $\rho = \rho_1 + \rho_2 < 1$.

It is not difficult to show that $\mathbb{E}T_i(x; N^{(i)}) = x/(1-\rho)$, and $T_i(x; N^{(i)}) \in \mathcal{L}$, $i = 1, 2$. In general, we will prove that the following moment ordering holds:

$$\mathbb{E}T_1(x; N^{(1)})^k \leq \mathbb{E}T_2(x; N^{(2)})^k,$$

if the values for the arrival rates satisfy $\lambda_1 \geq \lambda_2$; see Theorem 4.15. First we derive the LST of $T_i(x; N^{(i)})$, for $i = 1, 2$.

Lemma 4.12. *For $i = 1, 2$, the LST defined by $\widehat{v}^{(i)}(s; x) = \mathbb{E}e^{-sT_i(x;N^{(i)})}$, $\mathrm{Re}(s) \geq 0$, for the random variable $T_i(x; N^{(i)})$ is expressed by*

$$\widehat{v}^{(i)}(s; x) = \left(\sum_{n=0}^{\infty} \frac{s^n}{n!}\widehat{\alpha}_n(x, \rho_i)\right)^{-1},$$

where $\widehat{\alpha}_n(x, \rho_i)$ is defined by: $\widehat{\alpha}_0(x, \rho_i) := 1, \widehat{\alpha}_1(x, \rho_i) := x/(1-\rho)$ and for $n \geq 2$:

$$\widehat{\alpha}_n(x, \rho_i) = \frac{n}{1-\rho}\sum_{m=0}^{\infty} \rho_i^m \binom{m+n-2}{n-2} \int_{y=0}^{x} (x-y)^{n-1} \widetilde{B}^{m*}(y)dy. \tag{4.6}$$

Proof. It holds that $\widehat{v}^{(i)}(s; x) = \sum_{n=0}^{\infty} \left(v^{(i)}(s; x)\right)^{n+1} \mathbb{P}(N^{(i)} = n)$ by definition of $T_i(x; N^{(i)})$ and conditioning on the event $\{N^{(i)} = n\}$, where

$$v^{(i)}(s; x) = \left(\sum_{n=0}^{\infty} \frac{s^n}{n!}\alpha_n^{(i)}(x)\right)^{-1}$$

is the LST of $T_i(x; 0)$. Straightforward calculations give

$$\begin{aligned}
\widehat{v}^{(i)}(s; x) &= \frac{(1-\rho)v^{(i)}(s; x)}{1-\rho+(\rho-\rho_i)\left(1-v^{(i)}(s; x)\right)} \\
&= \left(\frac{1}{v^{(i)}(s; x)} + \frac{\rho-\rho_i}{1-\rho}\left(\frac{1}{v^{(i)}(s; x)} - 1\right)\right)^{-1} \\
&= \left(\sum_{n=0}^{\infty} \frac{s^n\alpha_n^{(i)}(x)}{n!} + \frac{\rho-\rho_i}{1-\rho}\sum_{n=1}^{\infty} \frac{s^n\alpha_n^{(i)}(x)}{n!}\right)^{-1} \\
&=: \left(\sum_{n=0}^{\infty} \frac{s^n}{n!}\widehat{\alpha}_n(x, \rho_i)\right)^{-1},
\end{aligned}$$

where $\widehat{\alpha}_n(x,\rho_i)$ is defined by: $\widehat{\alpha}_0(x,\rho_i) = \alpha_0^{(i)}(x) = 1$, and for $n \geq 1$:

$$\widehat{\alpha}_n(x,\rho_i) = \alpha_n^{(i)}(x)\left(1 + \frac{\rho - \rho_i}{1-\rho}\right) = \frac{1-\rho_i}{1-\rho}\alpha_n^{(i)}(x),$$

which leads to $\widehat{\alpha}_1(x,\rho_i) = (1-\rho_i)\alpha_1^{(i)}(x)/(1-\rho) = x/(1-\rho)$ and for $n \geq 2$, it is given by the expression (4.6); cf. (2.2) and (2.3) for the ordinary M/G/1 EPS queue with workload ρ_i. ∎

As a direct consequence of Lemma 4.12, the moments of the random variables $T_i(x;N^{(i)})$, $i = 1, 2$, satisfy a similar recursion as for an ordinary M/G/1 EPS queue.

Corollary 4.13. *For all $x \geq 0$, the moments defined by $\widehat{v}_k^{(i)}(x) = \mathbb{E}\left\{T_i(x;N^{(i)})\right\}^k$, are recursively given by $\widehat{v}_0^{(i)}(x) = 1$, and for $k \geq 1$:*

$$\widehat{v}_k^{(i)}(x) = -\sum_{j=1}^{k}\binom{k}{j}\widehat{v}_{k-j}^{(i)}(x)\widehat{\alpha}_j(x,\rho_i)(-1)^j.$$

In order to prove the general moment ordering: $\widehat{v}_k^{(1)}(x) \leq \widehat{v}_k^{(2)}(x)$, for $\lambda_1 \geq \lambda_2$ (Theorem 4.15), we first need the following Lemma 4.14, which implies that if we have that $\widehat{v}_k^{(1)}(x) = \widehat{v}_k^{(2)}(x)$ for some $k \geq 2$ and for some $x > 0$, then $T_1(x;N^{(1)})$ and $T_2(x;N^{(2)})$ are equally distributed. The converse statement is clearly true.

Lemma 4.14. *For any $x > 0$, and for any $k \geq 2$ (both fixed), the following equivalence holds (provided $\rho_1 + \rho_2 < 1$):*

$$\widehat{v}_k^{(1)}(x) = \widehat{v}_k^{(2)}(x) \text{ if and only if } \rho_1 = \rho_2.$$

For $k = 1$, we have $\widehat{v}_1^{(1)}(x) = \widehat{v}_1^{(2)}(x) = x/(1-\rho)$, irrespective of the ordering between ρ_1 and ρ_2.

Proof. For $x > 0$ fixed, we will first prove the following equivalent statements:
(i) $\rho_1 = \rho_2$
(ii) For all $n \geq 2$: $\widehat{\alpha}_n(x,\rho_1) = \widehat{\alpha}_n(x,\rho_2)$
(iii) For some $n \geq 2$: $\widehat{\alpha}_n(x,\rho_1) = \widehat{\alpha}_n(x,\rho_2)$

Clearly, (i) implies (ii), which in turn implies (iii). Now we will show the non-trivial implication (iii) ⇒ (i). For $x > 0$, suppose that for some $n \geq 2$: $\widehat{\alpha}_n(x,\rho_1) = \widehat{\alpha}_n(x,\rho_2)$. Then, it follows from the structure of (4.6) that $\rho_1 = \rho_2$. To this end, note that $\widehat{\alpha}_n(x,\rho_i)$ is a real-valued polynomial in ρ_i with non-negative coefficients, and with positive coefficients if $x > 0$ (also observe that $\widetilde{B}(y)$ is a proper distribution function). Hence for $x > 0$, $\widehat{\alpha}_n(x,\rho_1)$ and $\widehat{\alpha}_n(x,\rho_2)$ are the same strictly increasing continuous functions (on $\mathbb{R}_+$), only evaluated at a different point, at ρ_1 and ρ_2 respectively. Hence, if

$\widehat{\alpha}_n(x,\rho_1) = \widehat{\alpha}_n(x,\rho_2)$, then necessarily $\rho_1 = \rho_2$.

For any $k \geq 2$ and for any $x > 0$ (both fixed), we now proceed with proving the implication:

$$\widehat{v}_k^{(1)}(x) = \widehat{v}_k^{(2)}(x) \Rightarrow \rho_1 = \rho_2. \tag{4.7}$$

The above implication (4.7) must hold for all $k \geq 2$ and all $x > 0$ (and it is *not* the same as the statement: $\{$For all $k \geq 2$ and all $x > 0$: $\widehat{v}_k^{(1)}(x) = \widehat{v}_k^{(2)}(x)\} \Rightarrow \{\rho_1 = \rho_2\}$).

The implication (4.7) is straightforward for the second moment, since $\widehat{v}_2^{(1)}(x) = \widehat{v}_2^{(2)}(x)$ is equivalent to $\widehat{\alpha}_2(x,\rho_1) = \widehat{\alpha}_2(x,\rho_2)$, cf. the recursive formula in Corollary 4.13 with equal first moments, and thus $\rho_1 = \rho_2$ by the equivalent statements (i)-(ii)-(iii). For $k > 2$, the implication (4.7) is not trivial, mainly due to the presence of the alternating term $(-1)^j$ in the recursive formula. However, the statements (i)-(ii)-(iii) are equivalent in a strong sense. For example, if $\widehat{\alpha}_k(x,\rho_1) \neq \widehat{\alpha}_k(x,\rho_2)$ for *some* $k \geq 2$, then $\rho_1 \neq \rho_2$ and also $\widehat{\alpha}_k(x,\rho_1) \neq \widehat{\alpha}_k(x,\rho_2)$ for *all* $k \geq 2$; see also the similar observation in Remark 4.2.

Hence, if $\widehat{v}_k^{(1)}(x) = \widehat{v}_k^{(2)}(x)$, then we have two mutually exclusive possibilities:
(a) $\widehat{\alpha}_k(x,\rho_1) = \widehat{\alpha}_k(x,\rho_2)$
(b) $\widehat{\alpha}_k(x,\rho_1) \neq \widehat{\alpha}_k(x,\rho_2)$
The fact that $\widehat{v}_k^{(1)}(x) = \widehat{v}_k^{(2)}(x)$ implies either (a) or (b). By contradiction, we will now show that possibility (b) cannot occur. So, suppose $\widehat{v}_k^{(1)}(x) = \widehat{v}_k^{(2)}(x) \Rightarrow \widehat{\alpha}_k(x,\rho_1) \neq \widehat{\alpha}_k(x,\rho_2)$, but $\widehat{\alpha}_k(x,\rho_1) \neq \widehat{\alpha}_k(x,\rho_2)$ is equivalent to $\rho_1 \neq \rho_2$. Hence, assuming (b) true is the same as

$$\widehat{v}_k^{(1)}(x) = \widehat{v}_k^{(2)}(x) \Rightarrow \rho_1 \neq \rho_2, \tag{4.8}$$

which is obviously false, since the negation of (4.8), i.e., $\rho_1 = \rho_2 \Rightarrow \widehat{v}_k^{(1)}(x) \neq \widehat{v}_k^{(2)}(x)$ is clearly false. Thus, the assumption that (b) holds is false, and hence $\widehat{v}_k^{(1)}(x) = \widehat{v}_k^{(2)}(x)$ implies possibility (a) which in turn is equivalent to $\rho_1 = \rho_2$ by the strong equivalences (i)-(ii)-(iii). ∎

Theorem 4.15. *For $x > 0$ and $\rho = \rho_1 + \rho_2 < 1$, if $\rho_1 \geq \rho_2$, then we have the moment ordering*

$$\widehat{v}_k^{(1)}(x) \leq \widehat{v}_k^{(2)}(x), \quad \textit{for all } k \in \mathbb{N}.$$

Proof. For the first moment we have $\widehat{v}_1^{(1)}(x) = \widehat{v}_1^{(2)}(x)$, irrespective of the ordering between ρ_1 and ρ_2. By Lemma 4.14, if $\rho_1 \neq \rho_2$, then it holds that $\widehat{v}_k^{(1)}(x) \neq \widehat{v}_k^{(2)}(x)$ for all $k \geq 2$ and all $x > 0$. Now consider the strict ordering $\rho_1 > \rho_2$. Lemma 4.14 guarantees for $\rho_1 > \rho_2$, that $\widehat{v}_k^{(1)}(x)$ and $\widehat{v}_k^{(2)}(x)$ cannot coincide for any $x > 0$ and $k \geq 2$. Then, continuity of $\widehat{v}_k^{(i)}(x)$ in x implies for $\rho_1 > \rho_2$, that $\widehat{v}_k^{(1)}(x)$ and $\widehat{v}_k^{(2)}(x)$ cannot cross each

other for any $k \geq 2$, as a function of $x > 0$. Hence, either (a) or (b) holds:
(a) $\{\widehat{v}_k^{(1)}(x) < \widehat{v}_k^{(2)}(x) \text{ for all } x > 0\}$,
(b) $\{\widehat{v}_k^{(1)}(x) > \widehat{v}_k^{(2)}(x) \text{ for all } x > 0\}$.

The proof is completed, if we can find a $x^* > 0$ such that for all $k \geq 2$, if $\rho_1 > \rho_2$, then:

$$\widehat{v}_k^{(1)}(x^*) < \widehat{v}_k^{(2)}(x^*).$$

This can be done by choosing x^* large enough, since $\frac{T_i(x)}{x} \xrightarrow{\mathbb{P}} \frac{1}{1-\rho_i}$, as $x \to \infty$, and

$$\frac{T_i(x; N^{(i)})}{x} \xrightarrow{d} \frac{N^{(i)}+1}{1-\rho_i}, \text{ as } x \to \infty.$$

It is readily verified that (cf. the proof of Theorem 4.9 and the geometric distribution (4.5)):

$$\mathbb{E}\left(\frac{N^{(i)}+1}{1-\rho_i}\right)^k = \frac{1 + \sum_{j=1}^{k-1} \left\langle {k \atop j} \right\rangle \left(\frac{\rho-\rho_i}{1-\rho_i}\right)^j}{(1-\rho)^k},$$

hence if $\rho_1 > \rho_2$, then $\frac{\rho_2}{1-\rho_1} = \frac{\rho-\rho_1}{1-\rho_1} < \frac{\rho-\rho_2}{1-\rho_2} = \frac{\rho_1}{1-\rho_2}$, and

$$\lim_{x\to\infty} \frac{\widehat{v}_k^{(1)}(x)}{x^k} = \mathbb{E}\left(\frac{N^{(1)}+1}{1-\rho_1}\right)^k \tag{4.9}$$

$$< \mathbb{E}\left(\frac{N^{(2)}+1}{1-\rho_2}\right)^k = \lim_{x\to\infty} \frac{\widehat{v}_k^{(2)}(x)}{x^k}, \text{ for all } k \geq 2, \tag{4.10}$$

and with equality sign in (4.10) if and only if $\rho_1 = \rho_2$ (and for $k = 1$: $\widehat{v}_1^{(1)}(x) = \widehat{v}_1^{(2)}(x) = \frac{x}{1-\rho}$). Hence if $\rho_1 \geq \rho_2$, then $\widehat{v}_k^{(1)}(x) \leq \widehat{v}_k^{(2)}(x)$ for all $k \geq 1$ and all $x > 0$. ∎

Theorem 4.15 can be interpreted as follows. For a fixed $x > 0$, if the sojourn time $T_2(x; N^{(2)})$ is very large, then this is more likely due to the presence of many permanent customers in the system (large λ_1 implies that $N^{(2)}$ is 'stochastically' large) rather than a large arrival rate of non-permanent customers (large λ_2). By construction, the (random) number of permanent customers in system i is $N^{(i)}$ (label by "system i" the model that corresponds to $T_i(x; N^{(i)})$, $i = 1, 2$). Interestingly, the number of non-permanent customers in system i is in distribution equal to $N^{(j)}$, $i \neq j$, for $i, j \in \{1, 2\}$; cf. Theorem 3.2 (queue length decomposition result) in Chapter 3. Hence, if $\lambda_1 > \lambda_2$, then there are 'on average' more non-permanent and less permanent customers in system 1 compared to system 2. However, both systems have 'on average' an equal total number of customers (permanent plus non-permanent) regardless of the ordering between λ_1 and λ_2, which also explains the equality of the first moments: $\mathbb{E}T_1(x; N^{(1)}) = \mathbb{E}T_2(x; N^{(2)})$.

Remark 4.16. *We conjecture that $T_1(x;N^{(1)}) \leq_{cx} T_2(x;N^{(2)})$ holds if $\lambda_1 \geq \lambda_2$, i.e., the random variables are ordered in the convex stochastic order sense (see [95, 92]). Then, it is said that the random variable $T_2(x;N^{(2)})$ is more variable (more likely to take extreme values) than the variable $T_1(x;N^{(1)})$. The first moments are necessarily equal. A sufficient condition for convex stochastic ordering is the so-called Karlin & Novikoff cut-criterion, cf. [95], which states that two random variables X and Y are convex stochastic ordered if the means are equal and the corresponding distribution functions cross each other once and exactly once. The difficulty to verify the cut-criterion is that we do not have the distribution functions explicitly. We note that the cut-criterion and the intuition for the conjecture given in the instantaneous sojourn time analysis, are similar (see Remark 4.10).*

4.5 Insensitive and tight bounds

The main result of this chapter that the moments of the instantaneous sojourn time are upper bounds for the moments of the conditional sojourn time in the M/G/1 EPS queue, is given in the next theorem.

Theorem 4.17. *In the stable M/G/1 EPS queue, we have the insensitive lower and upper bounds for all moments of the conditional sojourn time $T(x)$, for $x \geq 0$ and $k \in \mathbb{N}$:*

$$\frac{1}{(1-\rho)^k}x^k \leq v_k(x) \leq \frac{1+\sum_{j=1}^{k-1}\left\langle {k \atop j} \right\rangle \rho^j}{(1-\rho)^k}x^k.$$

Proof. The result is trivial for the lower bound and for $x = 0$. For $x > 0$, we consider the special case of $\rho_2 = 0$ in Theorem 4.15. Then, for all $\rho \equiv \rho_1 \geq \rho_2 = 0$ and $\rho < 1$ it holds that

$$v_k(x) \equiv \widehat{v}_k^{(1)}(x) \leq \widehat{v}_k^{(2)}(x) = x^k \mathbb{E}(N+1)^k,$$

since 'with probability 1' we have: $N^{(1)} = 0$, $T_2(x;0) \equiv x$, and

$$T_1(x;N^{(1)}) \equiv T(x),$$
$$T_2(x;N^{(2)}) \equiv T_2(x;N) \stackrel{d}{=} x(N+1) \equiv \widehat{T}(x),$$

where $\mathbb{P}(N^{(2)} = n) = (1-\rho)\rho^n$, and $\widehat{T}(x)$ is the instantaneous sojourn time as defined in Section 4.3; also see Example 3.5 in Chapter 3. ∎

The special choice of $\rho_2 = 0$ in Theorem 4.15 is essentially the same as the assumptions made in the instantaneous sojourn time analysis, as in Section 4.3. For $\rho_2 \to 0$:

$$T_2(x;N^{(2)}) \stackrel{d}{\to} \widehat{T}(x) = (N^{(2)}+1)x,$$

as if the tagged customer arrived at a system with n permanent customers with probability $\mathbb{P}(N^{(2)} = n)$ and with no other arriving customers ($\rho_2 = 0$).

Remark 4.18. *With the moments of the instantaneous sojourn time as upper bounds, i.e., $v_k(x) \leq \widehat{v}_k(x)$, and the Chebyshev-Markov inequalities $\mathbb{P}(T(x) > y) \leq \frac{1}{y^k} v_k(x)$ for all $k \geq 1$, an insensitive upper bound for the tail probability $\mathbb{P}(T(x) > y)$ can be given*

$$\mathbb{P}(T(x) > y) \leq \min_{k \geq 1} \frac{1 + \sum_{i=1}^{k-1} \left\langle {k \atop i} \right\rangle \rho^i}{\left(y(1-\rho)/x\right)^k}, \quad \text{for } y \geq x > 0,\ \rho < 1. \tag{4.11}$$

The improvement upon Corollary 4.7 is considerable, particularly for large y. For $y \leq x/(1-\rho)$ the bound is not useful. However, if $y(1-\rho)/x > 1$, then both the numerator and denominator on the right-hand-side of (4.11) increase in k, but the denominator will be dominant for a certain k^*, defined by $k^* \equiv k^*(y;x) = \arg\min_{k \geq 1} \frac{\widehat{v}_k(x)}{y^k}$. We omit the details of the latter statement.

4.6 Quasi-stationary and fluid regime

The main result of this chapter (Theorem 4.17) can be related to the so-called 'fluid' and 'quasi-stationary' regimes, and to a result obtained by Delcoigne, Proutière, and Régnié [34]. They obtained the following (increasing convex) stochastic ordering:

$$W^{fl} \leq_{icx} W \leq_{icx} W^{qs},$$

for the stationary workload W in the M/G/1 EPS queue with *time-varying* service capacity. Their bounds correspond to the workload in the 'fluid' and 'quasi-stationary' regimes. As noted in [34], it proves much more difficult to derive similar results for the mean sojourn time.

In this chapter, $\widehat{T}(x)$ can also be viewed as the sojourn time $T^{qs}(x)$ in a quasi-stationary regime, which can be obtained by considering a (modified) M/G/1 EPS queue with fixed capacity, arrival rate λs and service requirement X/s. The (perturbation) parameter $s > 0$ represents the 'speed' of the queue length process, and does not influence the queue length distribution. In the limit $s \to 0$, the queue length process freezes in some initial state, yielding the quasi-stationary regime, and it can be shown that $T^{qs}(x) \stackrel{d}{=} \widehat{T}(x)$. For the sojourn time $T^{fl}(x)$ in the fluid regime (i.e., for the limit $s \to \infty$), it can be shown that $T^{fl}(x)$ is constant and equal to $x/(1-\rho)$. Analogous to the insensitive bounds in [34], we conjecture that it holds that:

$$x/(1-\rho) \leq_{icx} T(x) \leq_{icx} \widehat{T}(x),$$

where the $\leq_{icx}$-ordering could be replaced by the $\leq_{cx}$-ordering (since the means are equal). The above is also summarized (and generalized) by the following moment ordering result.

Theorem 4.19. *For $s > 0$, denote by $T^{(s)}(x)$ the sojourn time conditional on the initial service requirement $x > 0$ in the M/G/1 EPS queue with arrival rate λs, service requirement X/s, and with traffic load $\rho = \lambda \mathbb{E}X < 1$ (independent of $s > 0$); denote by $T(x) = T^{(1)}(x)$ the conditional sojourn time in the unscaled system. Then, for all $x > 0$, and $k \in \mathbb{N}$, we have the moment orderings:*

$$\begin{aligned} \mathbb{E}T(x)^k &\leq \mathbb{E}T^{(s)}(x)^k, \quad \text{if } s \leq 1, \\ \mathbb{E}T^{(s)}(x)^k &\leq \mathbb{E}T(x)^k, \qquad \text{if } s \geq 1. \end{aligned}$$

In fact, the first moments are identical: $\mathbb{E}T(x) = \mathbb{E}T^{(s)}(x) = x/(1-\rho)$ for all $s > 0$. Furthermore, for all $k \geq 2$, and $x > 0$ (both fixed), we have the equivalence

$$\mathbb{E}T^{(s)}(x)^k = \mathbb{E}T(x)^k \text{ if and only if } s = 1.$$

Proof. Analogous to Lemma 4.14 and Theorem 4.15 from Section 4.4. ∎

From the above moment ordering result an alternative proof for the main result (Theorem 4.17) can be given, analogous to Sections 4.4 and 4.5. The case $s \to \infty$ corresponds to Jensen's lower bound for the moments of the conditional sojourn time $T(x)$ in the M/G/1 EPS queue.

4.7 Conclusions and extensions

In this chapter, we have investigated the sojourn time $T(x)$ conditional on the initial service requirement $x > 0$ in the classical M/G/1 queue with egalitarian PS. In particular, we have studied all moments of $T(x)$ and we have obtained upper and lower bounds. The main result (Theorems 4.17 and 4.15) is that there exist upper bounds, given by $\left(1 + \sum_{i=1}^{k-1} \left\langle {k \atop i} \right\rangle \rho^i\right) x^k/(1-\rho)^k$, where $\left\langle {k \atop i} \right\rangle$ are Eulerian numbers, and they only depend on x and the traffic intensity $\rho < 1$. A lower bound follows easily from Jensen's inequality. The main result has been proved via stochastic comparisons of two related EPS models with a random number of permanent customers.

An attractive feature of the upper bound of the above structure is that it is independent of second and higher moments of the service requirement distribution. Another attractive feature is that the upper bound coincides with Jensen's lower bound when $\rho \to 0$. Moreover, the k-th moment of $T(x)$ and the above upper bound, converge to the same expression, after proper scaling when $\rho \to 1$. The upper bounds of the above structure with Eulerian numbers are in fact the moments of the so-called *instantaneous* sojourn time $\widehat{T}(x)$, i.e., the sojourn time of a customer with an infinitesimally small initial service requirement ($x \to 0$). If the initial service requirement $x > 0$ is arbitrary (and not necessarily small), the instantaneous sojourn time also corresponds to the sojourn time of

a tagged customer entering a PS system with no other arrivals but with a random number of permanent customers. The instantaneous sojourn time is also the sojourn time in a so-called quasi-stationary regime.

By studying the higher moments and providing insensitive upper bounds, we provide further support for the observation that EPS is a 'fair' service discipline. In the stable M/G/1 EPS system, excessive behavior of other customers in the system always has a limited influence on the sojourn time of the tagged customer. Intuitively, from a tagged customer point-of-view, the influence of the service requirements of other customers on the sojourn time of the tagged customer, is nearly insensitive. Even when there is a customer with infinite service requirement, the influence of this *permanent* customer on non-permanent customers is limited.

The influence of other customers is even more limited for jobs with a small initial service requirement; the sojourn time of a small job may be reasonably approximated by the instantaneous sojourn time. Moreover, it provides tight upper bounds for the higher moments of $T(x)$. However, for very large x, these upper bounds are 'quite loose'. This can be seen as the price that must be paid for obtaining *insensitive* upper bounds. Nevertheless, the moments of $T(x)$ are always bounded from above by the moments of $\widehat{T}(x)$, which in turn are bounded from above by the moments of an exponential random variable with mean $x/(1-\rho)$, regardless of the service requirement distribution (even when the service requirement has an infinite second moment).

We conclude this chapter with the remark that considerable attention has been paid in the literature to the exact analysis of the sojourn time in the M/G/1 EPS queue. Relatively little work has been done on the investigation of the practical implications of the results. The discovery of simple bounds for *all* moments of $T(x)$ stimulates the investigation of simple but nevertheless good approximations for the distribution of $T(x)$, the moments and the tail probabilities. In addition, a logical next step is to investigate if similar results also hold for other PS queues. For extensions of PS service disciplines, such as the *discriminatory* PS, it is to be expected that the nice properties (regarding the instantaneous sojourn time) are lost. For the M/G/1 queue with the egalitarian PS discipline and queue-dependent service capacity [32] it may be worthwhile to investigate if the structures remain valid.

Chapter 5

An approximation for DPS models

In this chapter we apply the queue length decomposition result for EPS models from Chapter 3 to *general* discriminatory processor-sharing (GDPS) models to obtain an efficient and analytically tractable approximation of the queue length distribution and mean sojourn times. The basic approximation assumption is that the linear equations in Corollary 3.3 from Chapter 3 also hold for (G)DPS models. More specifically, if one type of customers is treated as permanent in a general two-class DPS model, then the model is analytically tractable for the non-permanent class, with reduced service capacity that is exogenously given. The approximations are obtained as solutions of linear systems of equations.

We investigate the approximation error if this assumption is made for DPS models. By the exact queue length decomposition results for egalitarian PS models (with queue-dependent and balanced class capacities), our method provides exact results if applied to these PS models.

5.1 Approximation method

In this section we propose an approximation method for (unconditional) mean sojourn times in GDPS models. The basic assumption is that an isolated customer class in DPS is considered to behave like an egalitarian PS model with reduced capacity and a *random environment* that is exogenously determined. In the exact (G)DPS model this is obviously not the case, since the *random environments* for the different isolated queues in (G)DPS are interrelated and not independent.

For sake of notational convenience, we first consider a two-class GDPS model where

the class capacities $\phi_l(\mathbf{n}) = \phi_l(n_1, n_2)$ are arbitrary non-negative functions, for $l = 1, 2$. In addition, we assume a finite number of service positions for both customer types separately ($N_1 \leq m$ and $N_2 \leq n$); customers of type i ($i = 1, 2$) finding its own queue full upon arrival are blocked and lost, which is not a crucial assumption.

5.1.1 System of equations for two classes

If one customer type is treated as permanent in the system, then the model is analytically tractable for the non-permanent type. More precisely, the probabilities $a(i, j)$ and $b(j, i)$ are easily computed in closed form by (see [32, 68])

$$a(i,j) = \mathbb{P}\left(\widetilde{N}_1 = i \mid j \text{ permanent of type-2}\right) = \frac{\rho_1^i \varphi_{1,i}(j)}{\sum_{k=0}^{m} \rho_1^k \varphi_{1,k}(j)}, \tag{5.1}$$

$$b(j,i) = \mathbb{P}\left(\widetilde{N}_2 = j \mid i \text{ permanent of type-1}\right) = \frac{\rho_2^j \varphi_{2,j}(i)}{\sum_{k=0}^{n} \rho_2^k \varphi_{2,k}(i)}, \tag{5.2}$$

where

$$\varphi_{1,i}(j) = \left(\prod_{k=1}^{i} \phi_1(k,j)\right)^{-1}, \varphi_{1,0}(j) \equiv 1 \text{ for all } j = 0, 1, ..., n,$$

$$\varphi_{2,j}(i) = \left(\prod_{k=1}^{j} \phi_2(i,k)\right)^{-1}, \varphi_{2,0}(i) \equiv 1 \text{ for all } i = 0, 1, ..., m.$$

Note that $a(i, j)$ is identical to the conditional steady-state queue length probability $\mathbb{P}(N_1 = i \mid N_2 = j)$ for balanced (egalitarian) PS models; see Chapter 3. Our basic approximation assumption for (G)DPS models is that the linear equations given in Corollary 3.3 is applicable. Under this assumption, we approximate the marginal queue length probabilities $\eta_i = \mathbb{P}(N_1 = i)$ and $\xi_j = \mathbb{P}(N_2 = j)$ by solving the following set of linear equations:

$$\eta = \mathbf{A}\xi, \text{ and } \xi = \mathbf{B}\eta, \tag{5.3}$$

where $\eta = (\eta_0, \eta_1, ..., \eta_m)^T$, $\xi = (\xi_0, \xi_1, ..., \xi_n)^T$, and the matrices are given by

$$\mathbf{A} = \begin{pmatrix} a(0,0) & a(0,1) & \cdots & a(0,n) \\ a(1,0) & a(1,1) & \cdots & a(1,n) \\ \vdots & \vdots & \ddots & \vdots \\ a(m,0) & a(m,1) & \cdots & a(m,n) \end{pmatrix},$$

$$\mathbf{B} = \begin{pmatrix} b(0,0) & b(0,1) & \cdots & b(0,m) \\ b(1,0) & b(1,1) & \cdots & b(1,m) \\ \vdots & \vdots & \ddots & \vdots \\ b(n,0) & b(n,1) & \cdots & b(n,m) \end{pmatrix}.$$

It is not difficult to give conditions such that the (approximated) probability vectors η and ξ are uniquely determined after normalization. The system of equations (5.3) is also equivalent to $\eta = (\mathbf{AB})\,\eta$, or $\xi = (\mathbf{BA})\,\xi$, which can be interpreted as 'solving the equation $\pi = \pi\mathcal{P}$', where $\mathcal{P}$ is a transition matrix of a discrete-time Markov chain. In many practical (G)DPS models, it is easily verified that the product matrices $(\mathbf{AB})^T$ and $(\mathbf{BA})^T$, have row sums equal to one and do not have negative entries (irreducible, regular stochastic matrices). It is sufficient to have (Assumption 3.1 from Chapter 3): $\phi_j(\mathbf{n}) > 0$ for all j, and for all vectors $\mathbf{n}$ with $n_j > 0$, to guarantee uniqueness of η and ξ, up to a multiplicative constant.

The approximated (unconditional) mean sojourn time for each class follows from Little's law, and in our case with finite capacity (blocking) we have the approximation:

$$\mathbb{E}\widehat{T}_1^{approx} = \frac{1}{\lambda_1(1-\eta_m)} \sum_{i=0}^{m} i \cdot \eta_i, \text{ and } \mathbb{E}\widehat{T}_2^{approx} = \frac{1}{\lambda_2(1-\xi_n)} \sum_{j=0}^{n} j \cdot \xi_j, \qquad (5.4)$$

The proposed approximation method is exact for egalitarian PS models with *balanced* class capacities. The steady-state queue length distribution is insensitive to the service requirement distributions, if and only if the class capacities are *balanced* (see [20]). Hence, the approximation (5.1)-(5.4) can not be exact for PS models with *unbalanced* class capacities, since the approximation is insensitive to the service requirement distributions.

5.1.2 System of equations for three classes

In principle, our approximation can be applied for a general number of customer classes K. The method seems very efficient, since only linear systems need to be solved. However, significantly more computational effort is needed for increasing K. To illustrate the complexity, let us consider the case of $K = 3$ classes. Suppose that the class capacities $\phi_l(\mathbf{n}) = \phi_l(n_1, n_2, n_3)$, $l = 1, 2, 3$, are given in a three-class GDPS model with system states $(N_1, N_2, N_3) = (i, j, k)$. The approximated marginal steady state queue length probabilities denoted by $\eta_i = \mathbb{P}(N_1 = i)$, $\xi_j = \mathbb{P}(N_2 = j)$, $\zeta_k = \mathbb{P}(N_3 = k)$ are uniquely obtained from the linear equations (5.5)-(5.7), up to a multiplicative constant:

$$\begin{cases} \eta_i &= \sum_j (\sum_k \alpha(i \mid j,k)\pi_{3,2}(k \mid j))\xi_j &=: \sum_j a_{i,j}\xi_j \\ \eta_i &= \sum_k (\sum_j \alpha(i \mid j,k)\pi_{2,3}(j \mid k))\zeta_k &=: \sum_k b_{i,k}\zeta_k \end{cases}, \qquad (5.5)$$

$$\begin{cases} \xi_j &= \sum_i (\sum_k \beta(j \mid i,k)\pi_{3,1}(k \mid i))\eta_i &=: \sum_i c_{j,i}\eta_i \\ \xi_j &= \sum_k (\sum_i \beta(j \mid i,k)\pi_{1,3}(i \mid k))\zeta_k &=: \sum_k d_{j,k}\zeta_k \end{cases}, \qquad (5.6)$$

$$\begin{cases} \zeta_k &= \sum_i (\sum_j \gamma(k \mid i,j)\pi_{2,1}(j \mid i))\eta_i &=: \sum_i e_{k,i}\eta_i \\ \zeta_k &= \sum_j (\sum_i \gamma(k \mid i,j)\pi_{1,2}(i \mid j))\xi_j &=: \sum_j f_{k,j}\xi_j \end{cases}. \qquad (5.7)$$

The coefficients $\alpha(i \mid j,k)$ are given in closed-form formulas, similar to Eq. (5.1), since $\alpha(\cdot \mid j,k)$ is the steady-state queue length distribution for the isolated type 1 queue given that type 2 and 3 customers are permanently in the system. Analogously, the coefficients $\beta(j \mid i,k)$ and $\gamma(k \mid i,j)$ are also easily computed. The pairs of coefficients $\{\pi_{2,1}(j \mid i), \pi_{3,1}(k \mid i)\}$, $\{\pi_{1,2}(i \mid j), \pi_{3,2}(k \mid j)\}$, and $\{\pi_{2,3}(j \mid k), \pi_{1,3}(i \mid k)\}$ are obtained as unique solutions from the linear systems (5.8)-(5.10), similar to the approximation method in case of $K = 2$, up to multiplicative constant:

$$\begin{cases} \pi_{2,1}(j \mid i) &= \sum_k \beta(j \mid i,k)\pi_{3,1}(k \mid i) \\ \pi_{3,1}(k \mid i) &= \sum_j \gamma(k \mid i,j)\pi_{2,1}(j \mid i) \end{cases}, \text{ for all } i, \tag{5.8}$$

$$\begin{cases} \pi_{1,2}(i \mid j) &= \sum_k \alpha(i \mid j,k)\pi_{3,2}(k \mid j) \\ \pi_{3,2}(k \mid j) &= \sum_i \gamma(k \mid i,j)\pi_{1,2}(i \mid j) \end{cases}, \text{ for all } j, \tag{5.9}$$

$$\begin{cases} \pi_{1,3}(i \mid k) &= \sum_j \alpha(i \mid j,k)\pi_{2,3}(j \mid k) \\ \pi_{2,3}(j \mid k) &= \sum_i \beta(j \mid i,k)\pi_{1,3}(i \mid k) \end{cases}, \text{ for all } k. \tag{5.10}$$

The systems (5.5)-(5.7) written in matrix form: $\eta = \mathbf{A}\xi$, $\eta = \mathbf{B}\zeta$, $\xi = \mathbf{C}\eta$, $\xi = \mathbf{D}\zeta$, $\zeta = \mathbf{E}\eta$, and $\zeta = \mathbf{F}\xi$, are efficiently solved by e.g. the following two systems

$$\eta = (\mathbf{ACBFDE})\,\eta,$$
$$\xi = (\mathbf{CADEBF})\,\xi,$$

with normalization $\eta \cdot \mathbf{e} = 1$ and $\xi \cdot \mathbf{e} = 1$, and where the system for determining ζ is automatically satisfied and normalized if we have the unique solution for η. For increasing K, it seems that convenient notation may overcome the increase in complexity.

5.2 Numerical examples

In this section, we numerically investigate our approximation method with exact results in case of exponential service requirements, for the two- and three-class DPS models with fixed weights and fixed capacity. For numerical experiments and an application of the GDPS model with queue-dependent service weights and queue-dependent service capacity, we refer to Chapter 8.

Conservation law

In this subsection, we obtain a conservation law for the unconditional mean sojourn times in a DPS queue, which turns out to be useful in improving the approximations for the lowest priority class. The practical use of a conservation law is that if we are able to obtain accurate approximations of $\mathbb{E}T_k$ for customer classes $k = 1, ..., K-1$, then an accurate approximation for class K follows automatically.

Theorem 5.1. *For a K-class DPS queue with fixed capacity, fixed weights α_i, Poisson input λ_i, and exponential service requirements with mean $\mathbb{E}X_i$, $i = 1, ..., K$, the following conservation law for the unconditional mean sojourn times holds:*

$$\sum_{j=1}^{K} \rho_j \mathbb{E}T_j = \sum_{i=1}^{K} \frac{\rho_i}{1-\rho} \mathbb{E}X_i, \text{ independently of } (\alpha_1, ..., \alpha_K). \tag{5.11}$$

Proof. The result follows from the conservation law [10] for DPS models and from the fact that $(1 - B_i(x))\, dx = \frac{1}{\mu_i} dB_i(x)$ and $\mathbb{E}X_i^2 = 2\,(\mathbb{E}X_i)^2 = 2/\mu_i^2$, in case of exponential service requirements X_i with distribution function $B_i(x) = 1 - e^{-\mu_i x}$, for $x > 0$. ∎

5.2.1 Two-class DPS queue

In the two-class DPS model we refer to type 1 customers as the *high* priority customers and to type 2 customers as the *low* priority customers ($\alpha_1 > \alpha_2$). In case of exponential service requirements, exact closed-form expressions are given by (cf. [38]):

$$\mathbb{E}T_1 = \frac{1}{\mu_1(1-\rho_1-\rho_2)} \left(1 + \frac{\mu_1\rho_2(\alpha_2-\alpha_1)}{\mu_1\alpha_1(1-\rho_1) + \mu_2\alpha_2(1-\rho_2)}\right),$$
$$\mathbb{E}T_2 = \frac{1}{\mu_2(1-\rho_1-\rho_2)} \left(1 + \frac{\mu_2\rho_1(\alpha_1-\alpha_2)}{\mu_1\alpha_1(1-\rho_1) + \mu_2\alpha_2(1-\rho_2)}\right).$$

The approximated mean sojourn times $\mathbb{E}\widehat{T}_1^{approx}$ and $\mathbb{E}\widehat{T}_2^{approx}$ are calculated from the equations (5.1)-(5.4) and with infinite buffer capacity ($m = n = \infty$). The direct approximation $\mathbb{E}\widehat{T}_2^{approx}$ (based on decomposition) for the low priority class can be improved. The improved approximation, denoted by $\mathbb{E}\widehat{T}_2^{imp}$, is based on the conservation law (5.11) and the direct approximation $\mathbb{E}\widehat{T}_1^{approx}$ for the high priority class.

Figure 5.1 provides graphs for the exact and approximated mean sojourn times for both classes with $\mathbb{E}X_1 = \mathbb{E}X_2 = 1$, and for different values of α_1/α_2. For class 2, in addition, the improved approximation $\mathbb{E}\widehat{T}_2^{imp}$ is included. Figure 5.1 gives results as function of $\rho = \rho_1 + \rho_2$, with $\rho_1 = \rho_2 = \rho/2$. As can be seen from these graphs, the approximation for $\mathbb{E}T_1$ is reasonable up to a traffic load $\rho = 0.9$ for weight ratios $1 \leq \alpha_1/\alpha_2 \leq 5$. The approximation for $\mathbb{E}T_2$ breaks down with increasing difference in weights. However, the approximation $\mathbb{E}\widehat{T}_2^{imp}$ that uses $\mathbb{E}\widehat{T}_1^{approx}$ to approximate $\mathbb{E}T_2$ is accurate for all weight ratios. For a discussion of the quality of the approximation, we refer to Section 5.3.

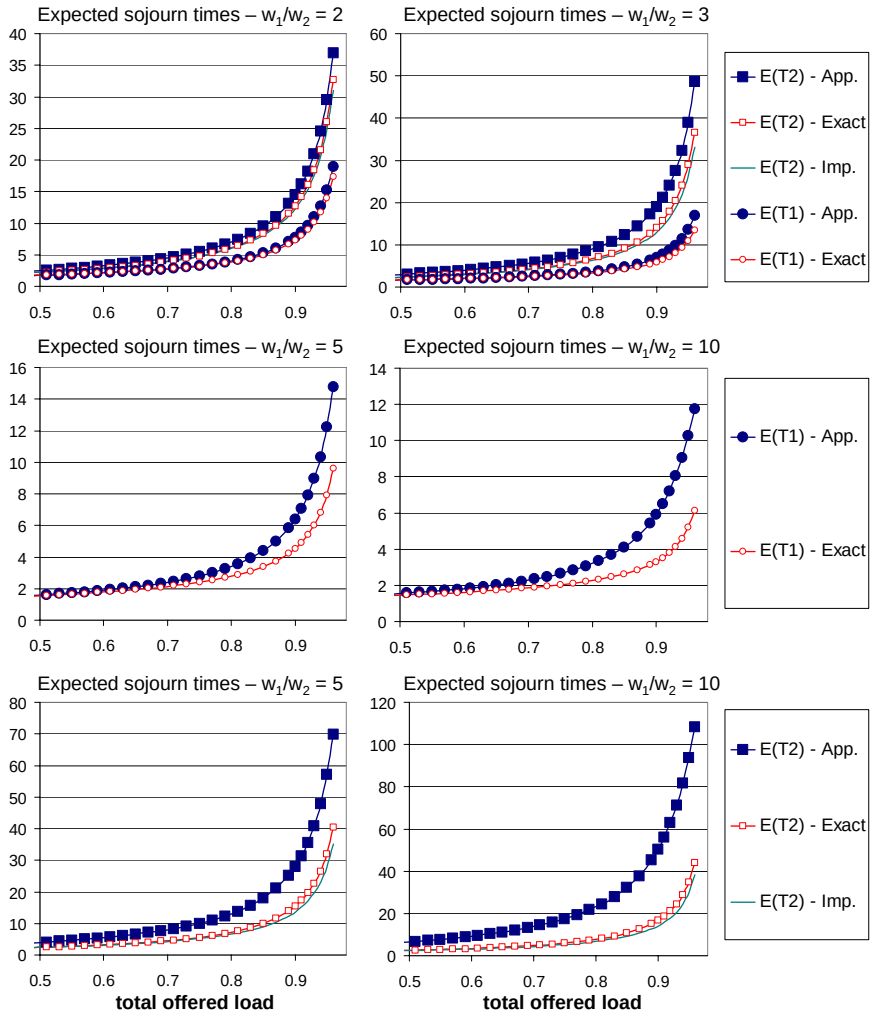

Figure 5.1: *Exact and approximated mean sojourn times for the 2-class M/M/1 DPS queue, with* $\mathbb{E}X_1 = \mathbb{E}X_2 = 1$*, and for weight ratios* $\frac{\alpha_1}{\alpha_2} \in \{2, 3, 5, 10\}$.

5.2.2 Three-class DPS queue

For the three-class DPS model, we consider the following numerical examples with mean service requirements $\mathbb{E}X_1 = 2$, and $\mathbb{E}X_2 = \mathbb{E}X_3 = 1$. The exact values for $\mathbb{E}T_j$, are obtained from [38] as solution of the linear system, for all $j = 1, 2, 3$:

$$\left(1 - \sum_{i=1}^{K} \frac{\lambda_i \alpha_i}{\mu_i \alpha_i + \mu_j \alpha_j}\right) \mathbb{E}T_j - \sum_{i=1}^{K} \frac{\lambda_i \alpha_i}{\mu_i \alpha_i + \mu_j \alpha_j} \mathbb{E}T_i = \frac{1}{\mu_j}.$$

Figure 5.2 provides graphs for the exact and approximated mean sojourn times for the three classes and for two sets of weights $\alpha = (\alpha_1, \alpha_2, \alpha_3)$, respectively for $\alpha = (2, 2, 1)$ and $\alpha = (3, 2, 1)$. Figure 5.3 provides approximated and exact mean sojourn times for $\alpha = (5, 3, 1)$ and $\alpha = (10, 3, 1)$. The approximated mean sojourn times $\mathbb{E}\widehat{T}_j^{approx}, j = 1, 2, 3$, are calculated according to the system of equations (5.5)-(5.10) and by applying Little's law. The figures are provided as function of the total load $\rho := \rho_1 + \rho_2 + \rho_3$, with $\rho_1 = \rho_2 = \rho_3 = \rho/3$. In addition, in Figure 5.3, an improved approximation $\mathbb{E}\widehat{T}_3^{imp}$ is included, based on the conservation law (5.11) and based on the direct approximations $\mathbb{E}\widehat{T}_1^{approx}$ and $\mathbb{E}\widehat{T}_2^{approx}$ of the other types.

As can be seen from the graphs (Figures 5.2 and 5.3), the approximations for $\mathbb{E}\widehat{T}_j^{approx}$, are accurate as long as the set of weights is 'more or less balanced'. It seems that our approximation improves for $K = 3$ customer classes. This can be explained by the fact that an additional customer class can increase the *balance* between the classes, see the next subsection.

5.3 Discussion

In this section we discuss the quality of our approximation $\mathbb{E}\widehat{T}_j^{approx}$ for $\mathbb{E}T_j$. In particular, in the case of $K = 2$ customer classes, numerical examples indicate that the approximation for the lower priority class $\mathbb{E}\widehat{T}_2^{approx}$ is poor when the ratio of weight α_1/α_2 is extremely large (unbalanced), whereas the improved approximation $\mathbb{E}\widehat{T}_2^{imp}$ is accurate.

Our basic approximation assumption is that the various customer classes in DPS models are treated as single-class egalitarian PS queues with queue-dependent (and reduced) service capacity. Supported by the queue length decomposition result for balanced egalitarian PS models, the isolated single-class PS queues in a multi-class egalitarian PS queue are exactly related to the other isolated customer queues.

When the ratio of weights α_1/α_2 is large, then from a class 2 point-of-view, the queue behaves as an ON-OFF processor-sharing queue [79]. As an illustration, Figure 5.4 shows the typical behavior of the queue length processes $N_i(t)$ for a two-class DPS queue under heavy load and large ratio α_1/α_2. From a class 2 point-of-view, it seems

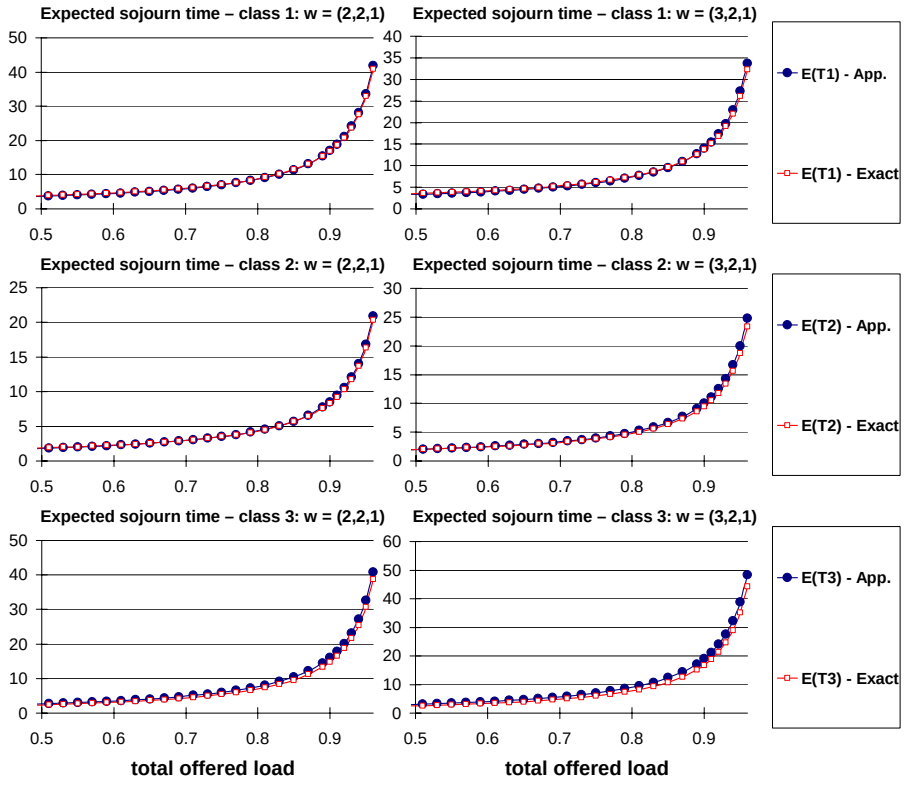

Figure 5.2: *Exact and approximated mean sojourn times for the 3-class M/M/1 DPS queue (with* $\mathbb{E}X_1 = 2, \mathbb{E}X_2 = \mathbb{E}X_3 = 1$*), for weights* $\alpha = (2, 2, 1)$ *and* $\alpha = (3, 2, 1)$.

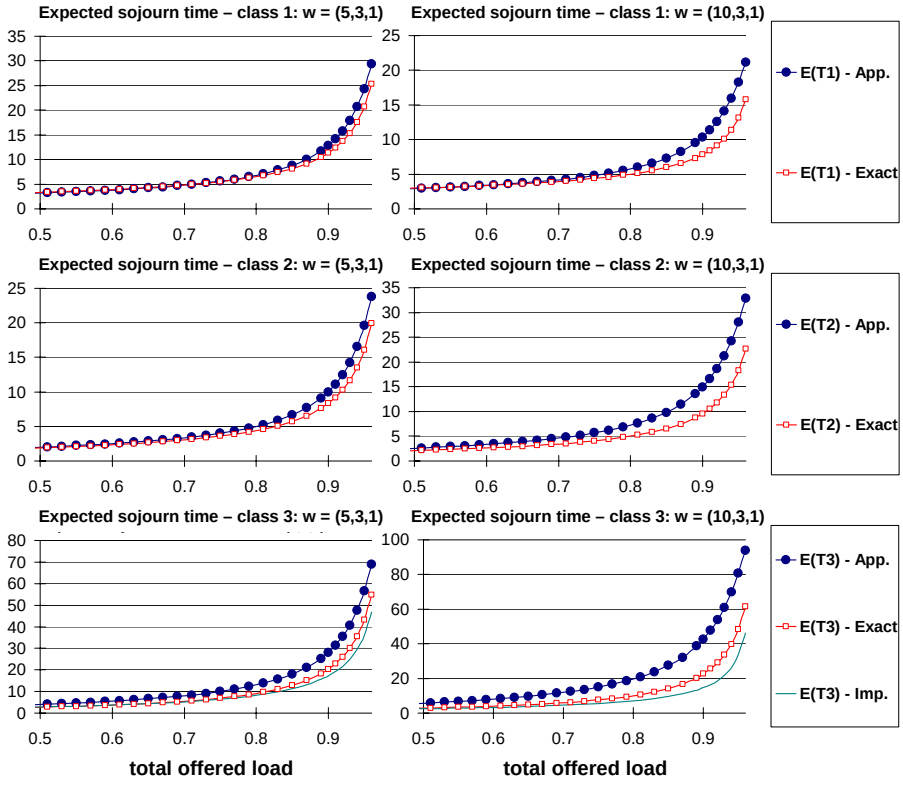

Figure 5.3: *Exact and approximated mean sojourn times for the 3-class M/M/1 DPS queue (with* $\mathbb{E}X_1 = 2, \mathbb{E}X_2 = \mathbb{E}X_3 = 1$*), for weights* $\alpha = (5, 3, 1)$ *and* $\alpha = (10, 3, 1)$.

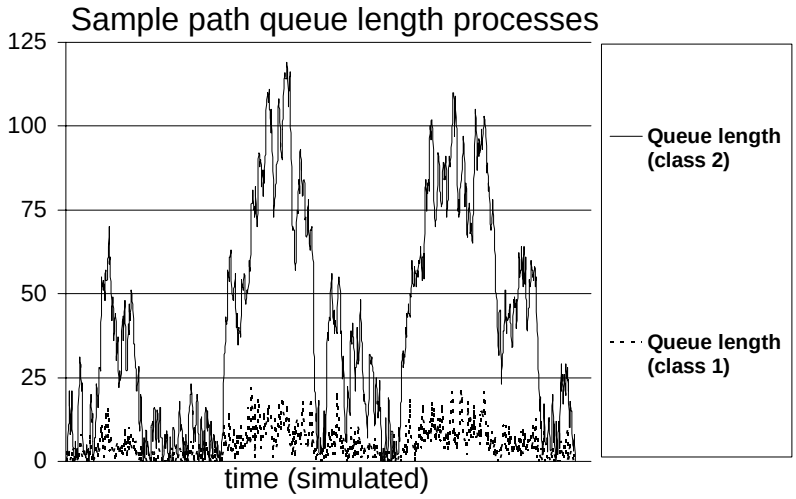

Figure 5.4: *A sample path of the queue length process $N_1(t)$ and $N_2(t)$, for a 2-class M/M/1 DPS model with $\alpha_1/\alpha_2 = 10$, $\lambda_1 = \lambda_2 = 0.49$, and $\mathbb{E}X_1 = \mathbb{E}X_2 = 1$.*

as if a *burst of permanent customers* (of size α_1/α_2) arrives, when a single customer of type 1 arrives in the original two-class DPS model. Therefore, when the number of class 1 customers gets large enough, the service process for class 2 may seem frozen (OFF period), and the queue length process for class 2 increases rapidly. However, since the high priority customers (class 1) reside in the system for a relatively short time period (class 1 gets a large share of the capacity), the queue length for the high priority class will decrease rapidly. When there is no high priority customer in the system, the low priority class receives all the available service capacity despite the large ratio of weights α_1/α_2 (ON period), and the queue length for the low priority class decreases significantly.

In the original two-class DPS model, the isolated customer class 2 has a *random environment* that is severely influenced by the 'burstiness' of class 1 (seen from class 2 point-of-view), while from an isolated class 1 point-of-view, it seems as if class 1 behaves according to its own single-class and isolated (egalitarian) PS queue, with a random environment that is less fluctuating over time, compared to the isolated class 2 point-of-view.

For the case of $K \geq 3$ customer classes, similar behavior is present in the DPS model. The queue length process of the highest priority class has a significant influence on the

queue length process of the lowest priority class, and not the other way round. However, in the case that more classes are present in the system, with service weights that are in between the highest and lowest priority class, the influence of the highest priority class on the random environment of lowest priority class may be less than in the case of $K = 2$.

5.4 Conclusions and extensions

In this chapter we have proposed an approximation for mean sojourn times in *general* DPS models, which is motivated by the queue length decomposition result for EPS models. The numerically efficient method is also applicable for GDPS models with queue-dependent service capacity and queue-dependent service weights. Numerical results have indicated that our approximation is accurate for a wide range of the weight ratios and for moderate loads. The approximation error is small for all loads if the DPS queue has 'nearly balanced' class capacities, which is in agreement with the exact queue length decomposition results. In heavy traffic and for extreme asymmetric weights, the approximation breaks down. However, the insights provided in this chapter suggest other approximations for these regimes, e.g., exploit PS models with ON-OFF periods. Further theoretical study and improvements of the approximation scheme are also desired.

Chapter 6

Slowdown for the M/M/1 DPS queue

In this chapter we investigate the so-called slowdown measures in the M/M/1 DPS queue. The slowdown is a way to measure how fairly jobs are treated by a service discipline (e.g. see [9, 47, 105]), and the mean slowdown is often used as a measure of system performance as opposed to the more traditional mean sojourn time. In general, it is desirable that a job's sojourn time should be correlated with its size, that is, we would like small jobs to have small sojourn times. The slowdown $S(x)$ of a job of size $x > 0$ is defined as $S(x) := T(x)/x$, which eliminates the dominating effect of large jobs in the sojourn time measure and introduces a bias towards small jobs. The (conditional) mean slowdown in the M/G/1 EPS queue is $1/(1-\rho)$, which only depends on the offered load $\rho < 1$ (insensitivity) and is also independent of the job size. Therefore, EPS is often considered as a "fair" service discipline; also see Chapter 4.

The mean slowdown in the M/G/1 DPS queue depends on the job size, the job (class) type, and furthermore service requirement distributions. In the case of exponential service requirements, we obtain the first and second moments of the slowdown in this chapter. In addition, we provide numerical examples and give some insights in the behavior of the slowdown measures for different traffic classes.

It is a priori not clear how the 'unfairness' depends on the job size and the weights. The jobs with the smallest DPS weight, call it the lowest priority class, are obviously treated 'unfairly' under DPS compared to EPS. In particular, short jobs of the lowest priority class are treated the 'most unfairly' (in terms of mean slowdown). It is also intuitively clear that the jobs with the largest DPS weight, are treated better under DPS than under EPS, in terms of the slowdown measure. More interestingly, when the DPS model has three or more job classes, it is not immediately clear how the jobs of the 'mid-

dle classes' (classes with weights in between the largest and lowest weights) are treated. Depending on the parameters, it is possible that the middle class jobs are always treated better or worse than the EPS service discipline. However, in some specific scenario settings, some of the middle class jobs are treated better than, and some of the middle class jobs are treated worse than under the EPS service discipline (depending on the job size).

The remainder of this chapter is organized as follows. In Section 6.1, we give a short review of the M/M/1 DPS results which are used in the current chapter. In Section 6.2, we obtain the first and second moments of the slowdown. In Section 6.3, we provide numerical examples and give some insights in the behavior of the slowdown measures. In Section 6.4 we provide our concluding remarks.

6.1 Preliminaries

In this subsection we give a short review of the M/M/1 DPS results which are used in the current chapter. For the proofs we refer to Rege and Sengupta [85] and Kim and Kim [56]. In the DPS service discipline, all jobs present in the system are simultaneously served according to the set of weights $\{\alpha_i > 0,\, i = 1, ..., K\}$. If there are n_i class i jobs present in the system, then each class i job receives a fraction $\alpha_i / \sum_{j=1}^{K} \alpha_j n_j$ of the fixed service capacity. The weights are denoted in vector form $\alpha = [\alpha_1 \cdots \alpha_K]^T$. Analogously we denote by $\lambda = [\lambda_1 \cdots \lambda_K]^T$ the vector of arrival rates. The exponential service rates are denoted by μ_i, for $i = 1, ..., K$.

6.1.1 Moments of the number of jobs

Let N_i, $i = 1, ..., K$, be the number of class i jobs in the system in steady state, and define $Q(z_1, ..., z_K)$ as the joint probability generating function (pgf) of the number of each class job in the system in steady state: $Q(z_1, ..., z_K) \equiv \mathbb{E}\left(z_1^{N_1} \cdots z_K^{N_K}\right)$. From this pgf, define the following moments for $j, k = 1, ..., K$:

$$L_j^1 \equiv \frac{\partial}{\partial z_j} Q(z_1, ..., z_K)\Big|_{z_1 = \cdots = z_K = 1},$$
$$L_{jk}^2 \equiv \frac{\partial^2}{\partial z_j \partial z_k} Q(z_1, ..., z_K)\Big|_{z_1 = \cdots = z_K = 1}.$$

Note that L_j^1 is the mean number of class j jobs in steady state, i.e., $L_j^1 = \mathbb{E}N_j$.

By Equation (16) of Rege and Sengupta [85], we have the system of linear equations for L_l^1, for $l = 1, ..., K$,

$$L_l^1 - \sum_{j=1}^{K} \alpha_j \frac{\lambda_j L_l^1 + \lambda_l L_j^1}{\alpha_j \mu_j + \alpha_l \mu_l} = \frac{\lambda_l}{\mu_l}. \tag{6.1}$$

Solving for K linear equations (6.1) with K unknowns yields L_l^1, for $l = 1, ..., K$. Furthermore, we have a system of $K(K+1)/2$ equations for L_{jk}^2, $1 \le j \le k \le K$, by Equation (17) of Rege and Sengupta [85] and the fact that $L_{jk}^2 = L_{kj}^2$. The linear simultaneous equations for L_{jk}^2 are given by, for $1 \le j \le k \le K$:

$$L_{jk}^2 - \sum_{i=1}^{K} \alpha_i \frac{\lambda_j L_{ki}^2 + \lambda_k L_{ij}^2 + \lambda_i L_{jk}^2}{\alpha_j \mu_j + \alpha_k \mu_k + \alpha_i \mu_i} = (\alpha_j + \alpha_k) \frac{\lambda_j L_k^1 + \lambda_k L_j^1}{\alpha_j \mu_j + \alpha_k \mu_k},$$

where L_i^1 on the right hand side is obtained by Eq. (6.1).

6.1.2 Moments of the sojourn time

The sojourn time $T_i(x)$ of a class i job, given its initial job size $x > 0$ can also be interpreted as the time necessary for a class i job whose required service requirement is *greater* than x to *attain* service amount x. Let us tag a class i job with service requirement greater than x. When the tagged job attains service x, let $N_{ij}(x)$ denote the number of class j jobs in the system, $j = 1, ..., K$ (*excluding* the tagged job). We introduce the following joint transform

$$R_{ix}(s; z_1, ..., z_K) \equiv \mathbb{E}\left(e^{-sT_i(x)} \; z_1^{N_{i1}(x)} \cdots z_K^{N_{iK}(x)} \right),$$

which is defined for $|z_j| \le 1$, for $j = 1, ..., K$, and $\mathrm{Re}(s) \ge 0$. To find the first and second moments of the sojourn time of class i jobs with service requirement $x > 0$, we define the following moments:

$$M_{ix}^0 \equiv \left. \frac{\partial}{\partial s} R_{ix}(s; z_1, ..., z_K) \right|_{s=0,\; z_1 = \cdots = z_K = 1},$$
$$M_{ix}^j \equiv \left. \frac{\partial}{\partial z_j} R_{ix}(s; z_1, ..., z_K) \right|_{s=0,\; z_1 = \cdots = z_K = 1},$$

and

$$M_{ix}^{00} \equiv \left. \frac{\partial^2}{\partial s^2} R_{ix}(s; z_1, ..., z_K) \right|_{s=0,\; z_1 = \cdots = z_K = 1},$$
$$M_{ix}^{0j} \equiv \left. \frac{\partial^2}{\partial s \partial z_j} R_{ix}(s; z_1, ..., z_K) \right|_{s=0,\; z_1 = \cdots = z_K = 1},$$
$$M_{ix}^{jk} \equiv \left. \frac{\partial^2}{\partial z_j \partial z_k} R_{ix}(s; z_1, ..., z_K) \right|_{s=0,\; z_1 = \cdots = z_K = 1},$$

where $i, j, k = 1, ..., K$. We note that $-M_{ix}^0$ and M_{ix}^{00} are the first and second moments of the sojourn time of class i jobs with service requirement $x > 0$, respectively, i.e.,

$$\mathbb{E}T_i(x) = -M_{ix}^0, \text{ and } \mathbb{E}T_i^2(x) = M_{ix}^{00}.$$

Kim and Kim [56] derived the following system of first-order linear differential equations for the first moment (see Equation (20) of [56]):

$$\frac{d}{dx}\mathbb{E}T_i(x) = \frac{1}{\alpha_i}\alpha^T m_i(x) + 1, \tag{6.2}$$

where the vector function $m_i(x) = [M_{ix}^1\ M_{ix}^2\ \cdots\ M_{ix}^K]^T$ satisfies (see Equation (21) of [56]):

$$\frac{d}{dx}m_i(x) = \frac{1}{\alpha_i}Bm_i(x) + \lambda, \tag{6.3}$$

with the matrix B defined as

$$B = \lambda\alpha^T - diag(\alpha_1\mu_1, ..., \alpha_K\mu_K).$$

Furthermore, we have $\mathbb{E}T_i(0) = 0$ for all $i = 1, ..., K$, and $m_i(0) = [L_1^1 \cdots\ L_K^1]^T \equiv L^1$ by the PASTA property.

Kim and Kim [56] also derived the following system of $(K+1)(K+2)/2$ first-order linear differential equations for the second moment (see Eqs. (24)-(29) of [56]):

$$\frac{d}{dx}\mathbb{E}T_i^2(x) = \frac{2}{\alpha_i}\alpha^T y_i(x) + 2\mathbb{E}T_i(x), \tag{6.4}$$

with $y_i(x) = -[\ M_{ix}^{01}\ \cdots\ M_{ix}^{0K}]^T$ and $y_i(0) = [0 \cdots\ 0]^T$. In addition, we have

$$\frac{d}{dx}y_i(x) = \frac{1}{\alpha_i}Z_i(x)\alpha + \frac{1}{\alpha_i}By_i(x) + \mathbb{E}T_i(x)\lambda + [I + \frac{1}{\alpha_i}diag(\alpha_1, ..., \alpha_K)]m_i(x), \tag{6.5}$$

and

$$\begin{aligned}\frac{d}{dx}Z_i(x) = &\frac{1}{\alpha_i}BZ_i(x) + \frac{1}{\alpha_i}Z_i(x)B^T + [I + \frac{1}{\alpha_i}diag(\alpha_1, ..., \alpha_K)]m_i(x)\lambda^T \\ &+ \lambda m_i(x)^T[I + \frac{1}{\alpha_i}diag(\alpha_1, ..., \alpha_K)],\end{aligned} \tag{6.6}$$

where the matrix $Z_i(x)$ is given by

$$Z_i(x) = \begin{bmatrix} M_{ix}^{11} & \cdots & M_{ix}^{1K} \\ \vdots & \ldots & \vdots \\ M_{ix}^{K1} & \cdots & M_{ix}^{KK} \end{bmatrix}, \text{ and } Z_i(0) = \begin{bmatrix} L_{11}^2 & \cdots & L_{1K}^2 \\ \vdots & \ldots & \vdots \\ L_{K1}^2 & \cdots & L_{KK}^2 \end{bmatrix} \equiv L^2. \tag{6.7}$$

6.2 First and second moment of the slowdown

The unconditional slowdown S_i of a class i job is defined as its sojourn time divided by the job size, i.e., $S_i = T_i/X_i$. The conditional slowdown of class i job whose size

is $x > 0$ is denoted by $S_i(x) = T_i(x)/x$. In this section we obtain the mean of the conditional and unconditional slowdown. Then, we express the second moment of the conditional slowdown in terms of Laplace transforms, and obtain the second moment of the unconditional slowdown. In the remainder of this chapter, when we speak of the unconditional slowdown moments, we will omit the adjective 'unconditional' if no confusion arises. First we need the following lemma.

Lemma 6.1. *We have the following identities:*

(a) The matrix

$$B = \lambda\alpha^T - diag(\alpha_1\mu_1, ..., \alpha_K\mu_K)$$

is diagonalizable and the eigenvalues of B, say κ_j, $j = 1, ..., K$, are all negative. In addition, B can be written as

$$B = [v_1 \ \cdots \ v_K] diag\,(\kappa_1, ..., \kappa_K) \begin{bmatrix} u_1 \\ \vdots \\ u_K \end{bmatrix}, \tag{6.8}$$

with real right eigenvectors v_j and real left eigenvectors u_j satisfying $u_j v_k = \delta_{jk}$.

(b) $\alpha^T(-B)^{-1} = \frac{1}{1-\rho}[\mu_1^{-1} \cdots \mu_K^{-1}]$

(c) $B^{-1}\lambda = \frac{-1}{1-\rho}[\frac{\rho_1}{\alpha_1} \cdots \frac{\rho_K}{\alpha_K}]^T$

(d) $\alpha^T B^{-1}\lambda = \frac{-\rho}{1-\rho}$

Proof. Letting $D = diag(d_1, ..., d_K)$ with $d_i = \frac{\sqrt{\alpha_i}}{\sqrt{\lambda_i}}, i = 1, \ldots, K$, yields

$$DBD^{-1} = (D\lambda)(\alpha^T D^{-1}) - diag(\alpha_1\mu_1, ..., \alpha_K\mu_K).$$

Since $D\lambda = \left[\sqrt{\alpha_1\lambda_1} \ \cdots \ \sqrt{\alpha_1\lambda_1}\right]^T$ and $\alpha^T D^{-1} = \left[\sqrt{\alpha_1\lambda_1} \ \cdots \ \sqrt{\alpha_1\lambda_1}\right]$, we have

$$DBD^{-1} = (D\lambda)(D\lambda)^T - diag(\alpha_1\mu_1, ..., \alpha_K\mu_K),$$

which is a symmetric matrix. This implies that B is diagonalizable and can be written as (6.8). Furthermore, since DBD^{-1} is a real symmetric matrix, the eigenvalues κ_j, $j = 1, ..., K$, of B, are all real, and the eigenvectors v_j and u_j can always be taken to be real. Note that

$$[\mu_1^{-1} \cdots \mu_K^{-1}]B = -(1-\rho)\alpha^T < \mathbf{0}, \tag{6.9}$$

where $\mathbf{0}$ denotes a K-dimensional null-vector, and the inequality between two vectors is interpreted componentwise. Therefore, the eigenvalues κ_j, $j = 1, \ldots, K$, are all negative, and the proof of (a) is completed. Part (b) follows from (6.9), and (c) follows from the identity

$$B[\frac{\rho_1}{\alpha_1} \cdots \frac{\rho_K}{\alpha_K}]^T = (\rho - 1)\lambda.$$

Finally, (d) follows immediately from (b) or (c). ∎

6.2.1 First moment

In this subsection we derive the mean of the conditional slowdown, i.e., $\mathbb{E}S_i(x) = \mathbb{E}T_i(x)/x$ for a class i job with size $x > 0$.

Theorem 6.2. *For the M/M/1 DPS queue, the mean conditional sojourn time and the mean conditional slowdown for class i, $i = 1, ..., K$, are given by*

$$\mathbb{E}T_i(x) = \frac{1}{1-\rho}x + a - \alpha_i b + \sum_{j=1}^{K}(\alpha_i\xi_j - \eta_j)e^{\frac{\kappa_j}{\alpha_i}x}, \tag{6.10}$$

$$\mathbb{E}S_i(x) = \frac{1}{1-\rho} + \frac{a-\alpha_i b}{x} + \sum_{j=1}^{K}(\alpha_i\xi_j - \eta_j)\frac{e^{\frac{\kappa_j}{\alpha_i}x}}{x}, \tag{6.11}$$

where κ_j, $j = 1, \ldots, K$, are eigenvalues of B, and

$$\eta_j = \frac{1}{1-\rho}\left([\mu_1^{-1}\cdots\mu_K^{-1}]v_j\right)(u_j L^1), \tag{6.12}$$

$$\xi_j = \frac{1}{(1-\rho)^2}\left([\mu_1^{-1}\cdots\mu_K^{-1}]v_j\right)\left(u_j\,[\frac{\rho_1}{\alpha_1}\cdots\frac{\rho_K}{\alpha_K}]^T\right), \tag{6.13}$$

with the eigenvectors v_j and u_j given by Lemma 6.1, and a and b are given by

$$a = \frac{1}{1-\rho}\,[\mu_1^{-1}\cdots\mu_K^{-1}]L^1,$$
$$b = \frac{1}{(1-\rho)^2}\,[\mu_1^{-1}\cdots\mu_K^{-1}]\,[\frac{\rho_1}{\alpha_1}\cdots\frac{\rho_K}{\alpha_K}]^T.$$

Proof. Integrating (6.3) and using $m_i(0) = [L_1^1\ \cdots\ L_K^1]^T \equiv L^1$, we obtain

$$\begin{aligned} m_i(x) &= e^{\frac{1}{\alpha_i}Bx}L^1 + e^{\frac{1}{\alpha_i}Bx}\int_0^x e^{-\frac{1}{\alpha_i}Bw}dw\ \lambda \\ &= e^{\frac{1}{\alpha_i}Bx}L^1 + \alpha_i B^{-1}\left(e^{\frac{1}{\alpha_i}Bx} - I\right)\lambda. \end{aligned} \tag{6.14}$$

Similarly, by (6.2), (6.14), together with $\mathbb{E}T_i(0) = 0$ for all $i = 1, ..., K$, we have

$$\begin{aligned} \mathbb{E}T_i(x) &= \frac{1}{\alpha_i}\alpha^T\int_0^x e^{\frac{1}{\alpha_i}Bw}dw\ L^1 + \alpha^T B^{-1}\int_0^x\left(e^{\frac{1}{\alpha_i}Bw} - I\right)dw\ \lambda + x \\ &= \left(1 - \alpha^T B^{-1}\lambda\right)x + \alpha^T B^{-1}(e^{\frac{1}{\alpha_i}Bx} - I)L^1 + \alpha_i\alpha^T B^{-1}(e^{\frac{1}{\alpha_i}Bx} - I)B^{-1}\lambda. \end{aligned}$$

Then, by using Lemma 6.1,

$$\begin{aligned}\mathbb{E}T_i(x) &= \frac{1}{1-\rho}x - \frac{1}{1-\rho}[\mu_1^{-1}\cdots\mu_K^{-1}]\left(e^{\frac{1}{\alpha_i}Bx} - I\right)L^1 \\ &+ \frac{\alpha_i}{(1-\rho)^2}[\mu_1^{-1}\cdots\mu_K^{-1}]\left(e^{\frac{1}{\alpha_i}Bx} - I\right)[\frac{\rho_1}{\alpha_1}\cdots\frac{\rho_K}{\alpha_K}]^T \\ &= \frac{1}{1-\rho}x + \frac{1}{1-\rho}[\mu_1^{-1}\cdots\mu_K^{-1}]L^1 - \frac{\alpha_i}{(1-\rho)^2}[\mu_1^{-1}\cdots\mu_K^{-1}]\,[\frac{\rho_1}{\alpha_1}\cdots\frac{\rho_K}{\alpha_K}]^T \\ &+ \sum_{j=1}^{K}(-\eta_j + \alpha_i\xi_j)e^{\frac{\kappa_j}{\alpha_i}x},\end{aligned}$$

where η_j and ξ_j are given by (6.12) and (6.13). Hence, (6.10) is obtained, and (6.11) follows immediately. ∎

Remark 6.3. *We can rewrite the conditional mean slowdown (6.11) as*

$$\mathbb{E}S_i(x) = \frac{1}{1-\rho} + \sum_{j=1}^{K}(\eta_j - \alpha_i\xi_j)\frac{1-e^{\frac{\kappa_j}{\alpha_i}x}}{x}. \tag{6.15}$$

Since $\kappa_j < 0$ for all $j = 1, ..., K$, the expression (6.11) is suitable for investigation when $x \to \infty$, whereas (6.15) is convenient for investigation when $x \to 0$.

Next we derive an expression for the mean slowdown $\mathbb{E}S_i$ for class i. Note that

$$\mathbb{E}S_i = \int_0^\infty \mathbb{E}S_i(x)\mu_i e^{-\mu_i x}dx.$$

Theorem 6.4. *For the M/M/1 DPS queue, the mean slowdown $\mathbb{E}S_i$ for class i, is given by*

$$\mathbb{E}S_i = \frac{1}{1-\rho} + \mu_i\sum_{j=1}^{K}(\eta_j - \alpha_i\xi_j)\log\left(1 - \frac{\kappa_j}{\alpha_i\mu_i}\right),$$

where η_j and ξ_j are given by (6.12) and (6.13) in Theorem 6.2.

Proof. For $\mathrm{Re}(s) > 0$, define $\widetilde{S}_i(s) \equiv \int_0^\infty e^{-sx}\mathbb{E}S_i(x)dx$ as the Laplace transform (LT) of $\mathbb{E}S_i(x)$. First, it not difficult to verify that the following identities hold:

$$\frac{d}{ds}\int_0^\infty e^{-sx}\frac{1-e^{\frac{\kappa_j}{\alpha_i}x}}{x}dx = \int_0^\infty e^{-sx}\left(e^{\frac{\kappa_j}{\alpha_i}x} - 1\right)dx = \frac{1}{s - \frac{\kappa_j}{\alpha_i}} - \frac{1}{s}, \text{ and}$$

$$\int_0^\infty e^{-sx}\frac{1-e^{\frac{\kappa_j}{\alpha_i}x}}{x}dx = \log\left(\frac{s - \frac{\kappa_j}{\alpha_i}}{s}\right) = \log\left(1 - \frac{\kappa_j}{\alpha_i s}\right).$$

Hence, taking LTs in (6.15) yields

$$\widetilde{S}_i(s) = \frac{1}{1-\rho}\frac{1}{s} + \sum_{j=1}^{K}(\eta_j - \alpha_i\xi_j)\int_0^\infty e^{-sx}\frac{1-e^{\frac{\kappa_j}{\alpha_i}x}}{x}dx$$
$$= \frac{1}{1-\rho}\frac{1}{s} + \sum_{j=1}^{K}(\eta_j - \alpha_i\xi_j)\log\left(1-\frac{\kappa_j}{\alpha_i s}\right).$$

The proof is completed by noting that $\mathbb{E}S_i = \mu_i\widetilde{S}_i(\mu_i)$. ∎

6.2.2 Second moment

In order to derive the second moment of the slowdown $\mathbb{E}S_i^2$, we similarly define the LT of $\mathbb{E}S_i^2(x)$ by $\widetilde{G}_i(s) \equiv \int_0^\infty e^{-sx}\mathbb{E}[S_i^2(x)]dx$, and hence it holds that

$$\frac{d^2}{ds^2}\widetilde{G}_i(s) = \int_0^\infty e^{-sx}\mathbb{E}[T_i^2(x)]dx.$$

First, we give an expression for the above equation as follows.

Lemma 6.5. *The equation for $\frac{d^2}{ds^2}\widetilde{G}_i(s)$ satisfies:*

$$\frac{d^2}{ds^2}\widetilde{G}_i(s) = \frac{2}{\alpha_i^2}\frac{1}{s}\sum_{j=1}^{K}c_j\left\{\alpha^T\sum_{k=1}^{K}\frac{v_k u_k L^2 u_j^T}{(s-\frac{\kappa_k}{\alpha_i})(s-\frac{\kappa_j+\kappa_k}{\alpha_i})}\right.$$
$$+\alpha^T\sum_{k=1}^{K}\sum_{m=1}^{K}\frac{v_k u_k\left(I+\frac{1}{\alpha_i}diag(\alpha_1,.....,\alpha_K)\right)v_m u_m(\lambda+sL^1)(\lambda^T u_j^T)}{s(s-\frac{\kappa_k}{\alpha_i})(s-\frac{\kappa_j+\kappa_k}{\alpha_i})(s-\frac{\kappa_m}{\alpha_i})}$$
$$\left.+\alpha^T\sum_{k=1}^{K}\sum_{m=1}^{K}\frac{v_k u_k\lambda(\lambda+sL^1)^T u_m^T v_m^T\left(I+\frac{1}{\alpha_i}diag(\alpha_1,...,\alpha_K)\right)u_j^T}{s(s-\frac{\kappa_k}{\alpha_i})(s-\frac{\kappa_j+\kappa_k}{\alpha_i})(s-\frac{\kappa_m}{\alpha_i})}\right\}$$
$$+\frac{2}{s^3}\left(1+\frac{1}{\alpha_i}\sum_{j=1}^{K}\frac{\alpha^T v_j u_j(\lambda+sL^1)}{s-\frac{\kappa_j}{\alpha_i}}\right)\left(\frac{1}{\alpha_i}\sum_{k=1}^{K}\frac{\alpha^T v_k u_k\lambda}{s-\frac{\kappa_k}{\alpha_i}}+1\right)$$
$$+\frac{2}{\alpha_i}\sum_{j=1}^{K}\sum_{k=1}^{K}\frac{\alpha^T v_j u_j\left(I+\frac{1}{\alpha_i}diag(\alpha_1,...,\alpha_K)\right)v_k u_k(\lambda+sL^1)}{s^2(s-\frac{\kappa_j}{\alpha_i})(s-\frac{\kappa_k}{\alpha_i})}, \quad (6.16)$$

where the coefficients c_j, $j=1,...,K$ are defined by

$$[c_1\ \cdots\ c_K] = \alpha^T[v_1\ \cdots\ v_K].$$

Proof. For the proof see Section 6.5. ∎

Second moment for the M/M/1 EPS queue

Note that if $\alpha_1 = \alpha_2 = \cdots = \alpha_K$ and $\mu_1 = \mu_2 = \cdots = \mu_K$, then the M/M/1 DPS system is equivalent to a single-class M/M/1 EPS queue. In this case, we can reproduce the known formula for the second moment of the conditional slowdown (cf. [107, 15]) as follows. We may assume that $\alpha_j = \frac{1}{\mu}$ with $\mu = \mu_i$ for all $i, j \in \{1, ..., K\}$. Then, the eigenvalues of $B = \lambda\alpha^T - diag(\alpha_1\mu_1, ..., \alpha_K\mu_K) = \lambda\alpha^T - I$ are given by

$$\begin{aligned} \kappa_1 &= -(1-\rho), \text{ and} \\ \kappa_j &= -1, \text{ for } 2 \le j \le K. \end{aligned} \tag{6.17}$$

We may choose

$$v_1 = \rho^{-1}[\rho_1 \ \cdots \ \rho_K]^T, \text{ and } u_1 = 1^T, \tag{6.18}$$

where 1 denotes a K-dimensional column vector with all components equal to one. Note that

$$\begin{aligned} L^1 &= \frac{1}{1-\rho}[\rho_1 \ \cdots \ \rho_K]^T, \\ L^2 &= \frac{2}{(1-\rho)^2}[\rho_1 \ \cdots \ \rho_K]^T[\rho_1 \ \cdots \ \rho_K], \end{aligned} \tag{6.19}$$

and furthermore,

$$\begin{aligned} \alpha^T v_k &= \mu^{-1} u_1 v_k = \mu^{-1}\delta_{k1}, \\ u_1 L^2 u_j^T &= \frac{2\rho^2}{(1-\rho)^2}\delta_{j1}, \\ \lambda^T u_j^T &= \mu\rho v_1^T u_j^T = \mu\rho\delta_{j1}, \\ \left(\lambda + sL^1\right)^T u_m^T &= (\mu\rho + \frac{s\rho}{1-\rho})v_1^T u_m^T = (\mu\rho + \frac{s\rho}{1-\rho})\delta_{m1}, \\ c_1 &= \frac{1}{\mu}. \end{aligned}$$

Substituting (6.17)-(6.19) into (6.16) leads to

$$\begin{aligned} \frac{d^2}{ds^2}\widetilde{G}_i(s) = &\frac{4\rho^2}{(1-\rho)^2}\frac{1}{s(s+\mu(1-\rho))(s+2\mu(1-\rho))} \\ &+ \frac{4\mu\rho^2}{1-\rho}\frac{1}{s^2(s+\mu(1-\rho))(s+2\mu(1-\rho))} \\ &+ \frac{4\mu\rho^2}{1-\rho}\frac{1}{s^2(s+\mu(1-\rho))(s+2\mu(1-\rho))} \\ &+ \frac{2}{1-\rho}\frac{1}{s^3}\left(1 + \frac{\mu\rho}{s+\mu(1-\rho)}\right) \\ &+ \frac{4\rho}{1-\rho}\frac{1}{s^2(s+\mu(1-\rho))}, \end{aligned}$$

which is simplified to

$$\frac{d^2}{ds^2}\widetilde{G}_i(s) = \frac{2}{(1-\rho)^2}\frac{1}{s^3} + \frac{2\rho}{(1-\rho)^2}\frac{1}{s^2(s+\mu(1-\rho))}.$$

Decomposing the above into partial fractions yields

$$\begin{aligned}\frac{d^2}{ds^2}\widetilde{G}_i(s) = &-\frac{2\rho}{\mu^2(1-\rho)^4}\frac{1}{s} + \frac{2\rho}{\mu(1-\rho)^3}\frac{1}{s^2} + \frac{2}{(1-\rho)^2}\frac{1}{s^3} \\ &+ \frac{2\rho}{\mu^2(1-\rho)^4}\frac{1}{s+\mu(1-\rho)}.\end{aligned} \tag{6.20}$$

By inversion of Laplace transforms, we obtain

$$\mathbb{E}T_i^2(x) = \frac{x^2}{(1-\rho)^2} + \frac{2\rho x}{\mu(1-\rho)^3} - \frac{2\rho}{\mu^2(1-\rho)^4}(1-e^{-\mu(1-\rho)x}), \tag{6.21}$$

$$\mathbb{E}S_i^2(x) = \frac{1}{(1-\rho)^2} + \frac{2\rho}{\mu^2(1-\rho)^4}\frac{\mu(1-\rho)x - 1 + e^{-\mu(1-\rho)x}}{x^2}, \tag{6.22}$$

where (6.21) is the same result as in [107, 15].

Second moment for the M/M/1 DPS queue

Now we express the second moment of the unconditional slowdown for class i jobs in the M/M/1 DPS queue. The second moment $\mathbb{E}S_i^2$ of the unconditional slowdown for class i job is given by

$$\mathbb{E}S_i^2 = \int_0^\infty \mathbb{E}[S_i^2(x)]\mu_i e^{-\mu_i x}dx = \mu_i \widetilde{G}_i(\mu_i).$$

Let us decompose (6.16) into partial fractions

$$\begin{aligned}\frac{d^2}{ds^2}\widetilde{G}_i(s) = &\frac{\epsilon_{i1}^1}{s} + \frac{\epsilon_{i2}^1}{s^2} + \frac{\epsilon_{i3}^1}{s^3} + \sum_{j=1}^{K}\left(\frac{\epsilon_{ij}^2}{s-\frac{\kappa_j}{\alpha_i}} + \frac{\epsilon_{ij}^3}{(s-\frac{\kappa_j}{\alpha_i})^2}\right) \\ &+ \sum_{1\le j\le k\le K}\frac{\epsilon_{ijk}^4}{s-\frac{\kappa_j+\kappa_k}{\alpha_i}},\end{aligned} \tag{6.23}$$

for some constants $\epsilon_{i1}^1, \epsilon_{i2}^1, \epsilon_{i3}^1, \epsilon_{ij}^2, \epsilon_{ij}^3, j = 1, \cdots, K$, and $\epsilon_{ijk}^4, 1 \le j \le k \le K$.

Theorem 6.6. *The second moment of the slowdown for class i jobs is expressed by*

$$\begin{aligned}\mathbb{E}S_i^2 = \frac{1}{(1-\rho)^2} &+ \mu_i \sum_{j=1}^{K} \epsilon_{ij}^2 \left\{ \left(\mu_i - \frac{\kappa_j}{\alpha_i}\right) \log\left(1 - \frac{\kappa_j}{\alpha_i \mu_i}\right) + \frac{\kappa_j}{\alpha_i} \right\} \\ &- \mu_i \sum_{j=1}^{K} \epsilon_{ij}^3 \log\left(1 - \frac{\kappa_j}{\alpha_i \mu_i}\right) \\ &+ \mu_i \sum_{1 \le j \le k \le K} \epsilon_{ijk}^4 \left\{ \left(\mu_i - \frac{\kappa_j + \kappa_k}{\alpha_i}\right) \log\left(1 - \frac{\kappa_j + \kappa_k}{\alpha_i \mu_i}\right) + \frac{\kappa_j + \kappa_k}{\alpha_i} \right\}.\end{aligned}$$

Proof. Integrating (6.23) twice, we get

$$\begin{aligned}\widetilde{G}_i(s) = \epsilon_{i1}^1 (s \log s - s) - \epsilon_{i2}^1 \log s + \frac{\epsilon_{i3}^1}{2} \frac{1}{s} \\ + \sum_{j=1}^{K} \epsilon_{ij}^2 \left\{ \left(s - \frac{\kappa_j}{\alpha_i}\right) \log\left(s - \frac{\kappa_j}{\alpha_i}\right) - \left(s - \frac{\kappa_j}{\alpha_i}\right) \right\} - \sum_{j=1}^{K} \epsilon_{ij}^3 \log\left(s - \frac{\kappa_j}{\alpha_i}\right) \\ + \sum_{1 \le j \le k \le K} \epsilon_{ijk}^4 \left\{ \left(s - \frac{\kappa_j + \kappa_k}{\alpha_i}\right) \log\left(s - \frac{\kappa_j + \kappa_k}{\alpha_i}\right) - \left(s - \frac{\kappa_j + \kappa_k}{\alpha_i}\right) \right\} \\ + C_1 s + C_2, \qquad (6.24)\end{aligned}$$

for some constants C_1 and C_2. We note that for all $a \in \mathbb{R}$:

$$\log(s + a) = \log s + \frac{a}{s} + o(\frac{1}{s}), \quad \text{as } s \to \infty, \qquad (6.25)$$

Substituting (6.25) into (6.24), after some arithmetic, we can rewrite (6.24) as

$$\begin{aligned}\widetilde{G}_i(s) = &\left\{ \epsilon_{i1}^1 + \sum_{j=1}^{K} \epsilon_{ij}^2 + \sum_{1 \le j \le k \le K} \epsilon_{ijk}^4 \right\} s \log s \\ &+ \left\{ C_1 - \epsilon_{i1}^1 - \sum_{j=1}^{K} \epsilon_{ij}^2 - \sum_{1 \le j \le k \le K} \epsilon_{ijk}^4 \right\} s \\ &- \left\{ \epsilon_{i2}^1 + \sum_{j=1}^{K} \epsilon_{ij}^2 \frac{\kappa_j}{\alpha_i} + \sum_{j=1}^{K} \epsilon_{ij}^3 + \sum_{1 \le j \le k \le K} \epsilon_{ijk}^4 \frac{\kappa_j + \kappa_k}{\alpha_i} \right\} \log s + C_2 \\ &+ \left\{ \frac{\epsilon_{i3}^1}{2} + \sum_{j=1}^{K} \epsilon_{ij}^2 \left(\frac{\kappa_j}{\alpha_i}\right)^2 + \sum_{j=1}^{K} \epsilon_{ij}^3 \frac{\kappa_j}{\alpha_i} + \sum_{1 \le j \le k \le K} \epsilon_{ijk}^4 \left(\frac{\kappa_j + \kappa_k}{\alpha_i}\right)^2 \right\} \frac{1}{s} \\ &+ o(\frac{1}{s}), \quad \text{as } s \to \infty.\end{aligned}$$

Since $\lim_{s\to\infty}\widetilde{G}_i(s)=0$, the following conditions should hold:

$$\begin{aligned}
\epsilon^1_{i1} &= -\sum_{j=1}^{K}\epsilon^2_{ij} - \sum_{1\le j\le k\le K}\epsilon^4_{ijk},\\
\epsilon^1_{i2} &= -\sum_{j=1}^{K}\epsilon^2_{ij}\frac{\kappa_j}{\alpha_i} - \sum_{j=1}^{K}\epsilon^3_{ij} - \sum_{1\le j\le k\le K}\epsilon^4_{ijk}\frac{\kappa_j+\kappa_k}{\alpha_i},\\
C_1 &= C_2 = 0.
\end{aligned}$$

Substituting the conditions above into (6.24), after some arithmetic, we obtain

$$\begin{aligned}
\widetilde{G}_i(s) = \frac{\epsilon^1_{i3}}{2}\frac{1}{s} &+ \sum_{j=1}^{K}\epsilon^2_{ij}\left\{\left(s-\frac{\kappa_j}{\alpha_i}\right)\log\left(1-\frac{\kappa_j}{\alpha_i s}\right)+\frac{\kappa_j}{\alpha_i}\right\}\\
&- \sum_{j=1}^{K}\epsilon^3_{ij}\log\left(1-\frac{\kappa_j}{\alpha_i s}\right)\\
&+ \sum_{1\le j\le k\le K}\epsilon^4_{ijk}\left\{\left(s-\frac{\kappa_j+\kappa_k}{\alpha_i}\right)\log\left(1-\frac{\kappa_j+\kappa_k}{\alpha_i s}\right)+\frac{\kappa_j+\kappa_k}{\alpha_i}\right\}. \quad (6.26)
\end{aligned}$$

By (6.23), $\epsilon^1_{i3} = s^3\left(\frac{d^2}{ds^2}\widetilde{G}_i(s)\right)\Big|_{s=0}$. From the latter, together with (6.16) and Lemma 6.1 (d), it follows that

$$\epsilon^1_{i3} = 2\left(1+\sum_{j=1}^{K}\frac{\alpha^T v_j u_j \lambda}{-\kappa_j}\right)^2 = 2\left(1-\alpha^T B^{-1}\lambda\right)^2 = \frac{2}{(1-\rho)^2}. \quad (6.27)$$

Finally, substituting (6.27) into (6.26) and noticing that $\mathbb{E}S_i^2 = \mu_i\widetilde{G}_i(\mu_i)$, we finish the proof. ∎

As a special case of Theorem 6.6 we have the following corollary for the M/M/1 EPS queue.

Corollary 6.7. *For the single-class M/M/1 EPS queue, the second moment of the slowdown S is given by*

$$\mathbb{E}S^2 = \frac{1}{(1-\rho)^2} + \frac{2\rho}{(1-\rho)^4}\left((2-\rho)\log(2-\rho)-(1-\rho)\right).$$

Proof. Use (6.20) and Theorem 6.6 with equal weights α_i and rates μ_i. ∎

This result can also be obtained by integrating the known expression (6.22) by evaluating so-called exponential integrals, i.e., evaluate $\mathbb{E}S^2 = \int_0^\infty \mathbb{E}S^2(x)\mu e^{-\mu x}dx$.

6.3 Numerical examples

In this section we provide some numerical examples to discuss aspects of the slowdown in the DPS queue. For convenience, we refer to the "highest priority class" as the class with the largest weight; the "lowest priority class" as the class with the smallest weight. The classes with weights in between the largest and smallest weights, are labeled as the "middle classes". From the figures in this section, we will observe that

- The conditional mean slowdown of the highest priority class increases as the job size increases;
- The conditional mean slowdown of the lowest priority class decreases as the job size increases;
- It could happen that the conditional mean slowdown of the middle classes is neither increasing nor decreasing. See Figures 6.4-6.6 and 6.11-6.12. This phenomenon was also observed in [55].

It is known that for the M/G/1 DPS queue, the conditional mean slowdown of each class tends to $1/(1-\rho)$ as the job size increases to infinity (see Remark 2 in [38]), which is the same as the conditional mean slowdown of the EPS model. If the conditional mean slowdown of a job with size x is larger (resp. smaller) than $1/(1-\rho)$, then we say that this job (of size x) is treated worse (resp. better) under DPS than under EPS.

6.3.1 Mean slowdown for two-class DPS model

We consider the case of $K = 2$ job classes with weights α_1 and α_2. We assume equal loads of $\rho_1 = \rho_2 = 0.3$, and hence $1/(1-\rho) = 2.5$. We consider the following two DPS models.

Example 1. ***[See Figures 6.1-6.3]*** *We assume $\mu_1 = 2$ and $\mu_2 = 1$; hence $\lambda_1 = 0.6$ and $\lambda_2 = 0.3$, and consider the following weight settings:*
(a.) *$\alpha_1 = 3$ and $\alpha_2 = 1$.*
(b.) *$\alpha_1 = 1$ and $\alpha_2 = 3$.*

Note that in both examples, class 1 has a smaller mean job size than class 2. It is shown in [10, 55] that if $\alpha_1 \geq \alpha_2$, then DPS outperforms EPS from the viewpoint of the mean number of jobs and the mean sojourn time in steady state.

In Figures 6.1 and 6.2, we depict the conditional mean slowdown of each class for Examples 1a and 1b, respectively, varying the job size x. As expected, the overall conditional mean slowdown is better in Example 1a than in Example 1b. It is also intuitively clear that the highest priority class is always treated better than under EPS and the lowest priority class is always treated worse than under EPS. In addition, the conditional mean

slowdown curves of the two classes do not intersect, which also follows from the stochastic ordering result for conditional sojourn times; see Theorem 2 in Avrachenkov, Ayesta, Brown, and Núñez-Queija [10]. As illustrated in Figures 6.1 and 6.2, the conditional mean slowdown for the lowest priority class is much larger for small job sizes x. Short jobs of the lowest priority class are treated relatively the most unfairly, which can be explained by the so-called "ON-OFF" effect: If the ratio of weights α_1/α_2 is large, then from a class 2 point-of-view, the queue behaves as an ON-OFF processor-sharing queue (see Section 5.3 of Chapter 5). When the number of class 1 jobs gets large, the service process for class 2 may seem frozen (OFF period). When there are no high priority class jobs in the system, the low priority class gets full service capacity (ON period). Short low priority jobs experience the ON-OFF effect relatively more than long low priority jobs, since long jobs reside in the system for a longer time, and hence experience more or less the average system with reduced capacity.

In Figure 6.3, we plot the unconditional mean slowdown of each class, varying the weight ratio α_1/α_2. As expected, the mean slowdown of class 1 jobs equals that of class 2 jobs in the case of $\alpha_1/\alpha_2 = 1$. Furthermore, the mean slowdown of class 2 jobs (resp. class 1 jobs) increases (resp. decreases) as the weight ratio α_1/α_2 increases.

6.3.2 Mean slowdown for three-class DPS model

Now we consider the more interesting case of a DPS model with $K = 3$ job classes, with the presence of a 'middle class'. We assume equal loads of $\rho_1 = \rho_2 = \rho_3 = 0.2$, and hence $1/(1-\rho) = 2.5$.

Example 2a. [See Figures 6.4-6.7]

We assume $\mu_1 = 10, \mu_2 = 5$, $\mu_3 = 1$. Take $\alpha_1 = 8$, $\alpha_3 = 1$ and choose α_2 such that $\alpha_3 \leq \alpha_2 \leq \alpha_1$. In this case, class 1 is the highest priority class, class 2 is the middle class and class 3 is the lowest priority class.

Figures 6.4-6.6 show the conditional mean slowdown of each class for different values of α_2, varying the job size x. Figure 6.7 shows the conditional mean slowdown for only the middle class for different values of α_2. We observe that if α_2 is small, then the conditional mean slowdown curve of the middle class is above the curve $1/(1-\rho)$, i.e., middle class jobs are always treated worse under DPS compared to EPS. If the weight α_2 of the middle class is moderate (e.g., $\alpha_2 = 1.5$ or $\alpha_2 = 2.0$), then the slowdown curve of the middle class crosses the curve $1/(1-\rho)$. Sometimes the middle class job is treated worse and sometimes better under DPS compared to EPS, depending on the job size x of the middle class job. If α_2 gets larger, then the conditional mean slowdown curve of the middle class will be always below the curve $1/(1-\rho)$.

Example 2b. [See Figures 6.8-6.14]

We assume $\mu_1 = 10, \mu_2 = 5$, $\mu_3 = 1$. Take $\alpha_1 = 1$, $\alpha_3 = 8$ and choose α_2 such that $\alpha_1 \leq \alpha_2 \leq \alpha_3$. In this case, class 3 is the highest priority class, class 2 is the middle class and class 1 is the lowest priority class.

Figures 6.8 and 6.9 show the conditional mean slowdown of each class for different values of $\alpha_2 = 2.0$ and $\alpha_2 = 3.0$, respectively. Figures 6.10-6.14 show the conditional mean slowdown of only the middle class for different values of α_2.

As illustrated in Figures 6.8-6.11, if α_2 is small, then the conditional mean slowdown curve of the middle class is always above the curve $1/(1-\rho)$; however, the shape of the curve changes. From Figures 6.12 and 6.14, we see that if α_2 is moderate, then the short middle class jobs have a smaller conditional mean slowdown under DPS than under EPS, however, long middle class jobs are still treated unfairly under DPS in this situation. Figures 6.13 and 6.14 illustrate that the conditional mean slowdown curve of the middle class jobs will be always below the curve $1/(1-\rho)$, indicating that if α_2 is large, then the middle class gets served better under DPS for all job sizes. Figure 6.14 also indicates that the ON-OFF effect experienced by class 2 jobs becomes larger when α_2 becomes smaller.

From Figures 6.11 and 6.12, we also mention an a priori counterintuitive fact: neither the shortest job nor longest job is treated the 'most unfairly', but jobs of a certain 'medium size' are treated the 'most unfairly' among all middle class jobs. Similar phenomena are also present in the so-called foreground-background processor-sharing (FBPS) queue, see [80]. An exact and tractable characterization when middle class jobs are treated better or worse in DPS than under EPS seems not straightforward.

6.3.3 Variance of the slowdown for three-class DPS model

In this subsection, we investigate the conditional and unconditional variance of the slowdown for a DPS model with $K = 3$ job classes. We assume equal loads of $\rho_1 = \rho_2 = \rho_3 = 0.2$ with $1/(1-\rho) = 2.5$.

Example 3. [See Figures 6.15-6.17]

Like in Example 2 we assume $\mu_1 = 10, \mu_2 = 5$, and $\mu_3 = 1$. In Figure 6.15, we plot the variance of the conditional slowdown of each class, varying the job size x, for the case when $\alpha_1 = 8, \alpha_2 = 2$ and $\alpha_3 = 1$. The variance of the conditional slowdown of each class decreases as the job size x increases. Furthermore, we observe that the smaller the weight becomes, the larger the variance of the conditional slowdown becomes.

In Figures 6.16 and 6.17, we plot the variance and coefficient of variation for the slowdown of each class, respectively, varying the weight ratio α_1/α_2 $(= \alpha_2/\alpha_3)$. We

observe that the variance and coefficient of variation for the slowdown of a class 3 job (resp. a class 1 job) increases (resp. decreases) as the weight ratio α_1/α_2 increases.

6.4 Conclusion

In this chapter we obtained the first and second moments of the slowdown in the M/M/1 queue with the discriminatory processor-sharing (DPS) service discipline. The slowdown is a queueing fairness measure, which measures how fair jobs are treated by a service discipline. In an M/G/1 queue with egalitarian processor-sharing (EPS) service discipline, all jobs experience the same mean slowdown. Hence EPS is often considered as a fair service discipline for all jobs. In contrast, DPS is aimed at differentiating the Quality-of-Service among the different types of jobs.

How fair jobs are treated under DPS depends on several parameters; in particular, it depends on the set of DPS weights $(\alpha_1, ..., \alpha_K)$ in combination with the mean job sizes $(1/\mu_1, ..., 1/\mu_K)$, and the job size $x > 0$ of a particular class i, $i = 1, ..., K$. The highest priority class with the largest DPS weight is always treated better under DPS than under EPS at the expense of other classes. The lowest priority class is always treated worse under DPS than under EPS. However, the unfairness also depends on the job size; short lowest priority jobs are generally treated the 'most unfairly'. Short highest priority jobs are generally treated the best; these jobs benefit the most from the preemptive priority effect that the highest priority jobs observe.

When there are middle classes, i.e., classes with weights in between the largest and smallest weights, the characterization of the middle class is not straightforward. In the numerical examples we have observed and explained the following possible cases for the middle class:

- All jobs are always treated worse under DPS than under EPS;
- All jobs are always treated better under DPS than under EPS;
- Sometimes short jobs are treated worse and long jobs are treated better under DPS than under EPS;
- Sometimes short jobs are treated better and long jobs are treated worse under DPS than under EPS.

We also observed that the slowdown curve for a middle class job is generally not monotone in the job size $x > 0$, unless the weight of the middle class is sufficiently close to the weight of the highest or lowest priority class. The slowdown curves of the highest and lowest priority classes are increasing and decreasing, respectively. To give an exact and tractable characterization is a challenging task.

The variance of the unconditional slowdown is one of the unfairness measures suggested by Avi-Itzhak, Levy, and Brosh (see Section 3.1 in [9]). They mentioned that the variance of the unconditional slowdown is very hard to compute, and they left the feasibility of the unfairness metric, the variance of the unconditional slowdown, as an open research subject. The analysis for the variance of the unconditional slowdown of the M/M/1 DPS queue in this chapter can be a step towards this subject.

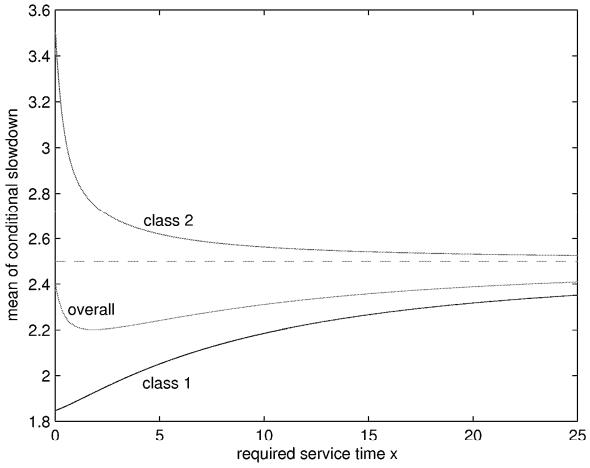

Figure 6.1: *Conditional mean slowdown for Example 1a.*

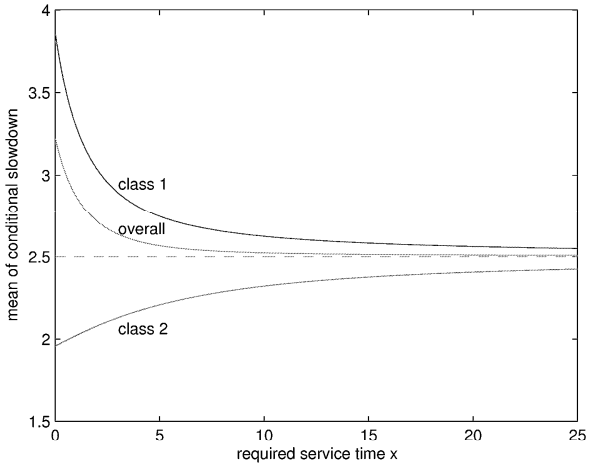

Figure 6.2: *Conditional mean slowdown for Example 1b.*

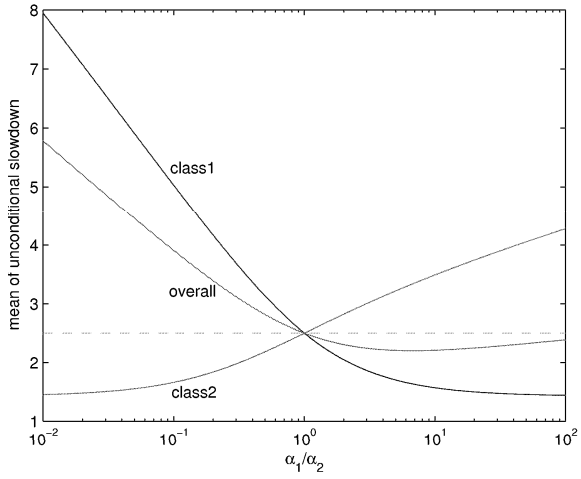

Figure 6.3: *Mean slowdown when* $\mu_1 = 2$ *and* $\mu_2 = 1$.

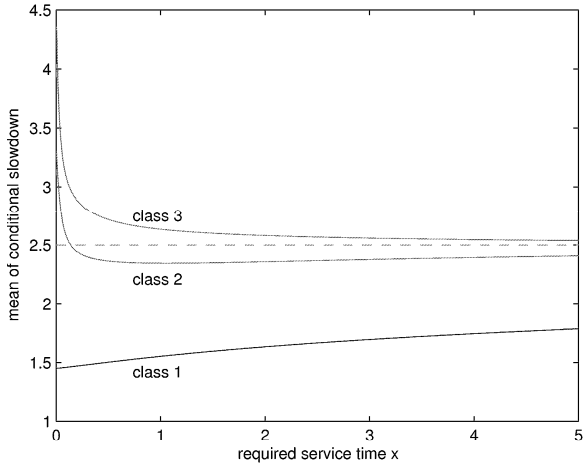

Figure 6.4: *Conditional mean slowdown for Example 2a with* $\alpha_2 = 1.5$.

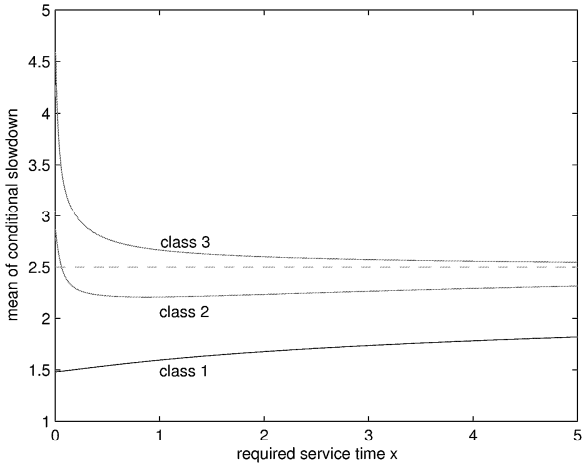

Figure 6.5: *Conditional mean slowdown for Example 2a with* $\alpha_2 = 2.0$.

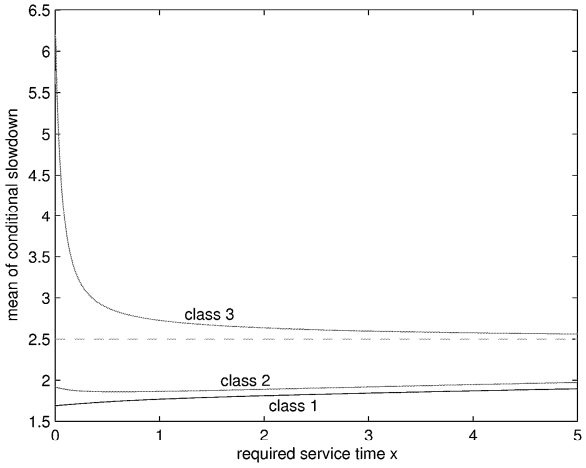

Figure 6.6: *Conditional mean slowdown for Example 2a with* $\alpha_2 = 6.0$.

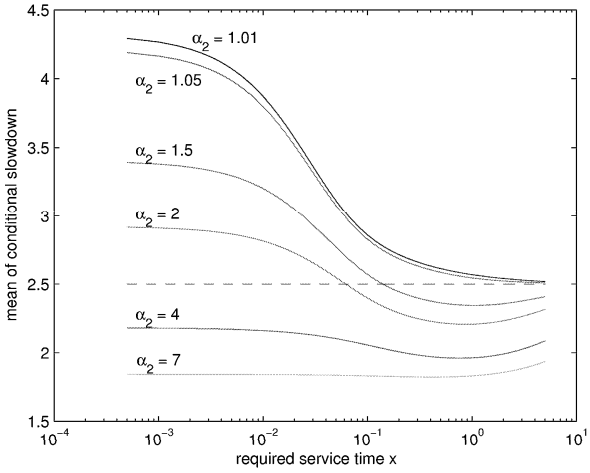

Figure 6.7: *Conditional mean slowdown for middle class in Example 2a.*

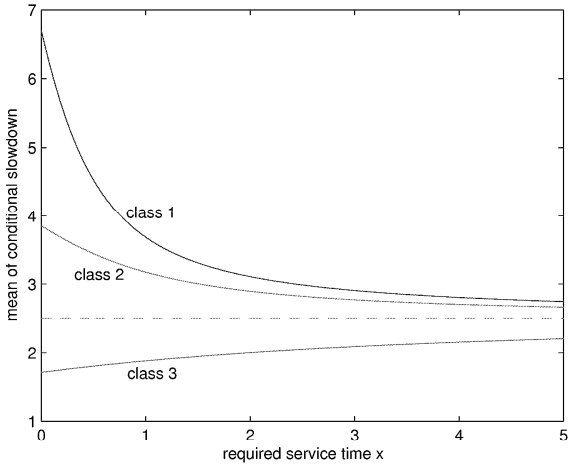

Figure 6.8: *Conditional mean slowdown for Example 2b with* $\alpha_2 = 2.0$.

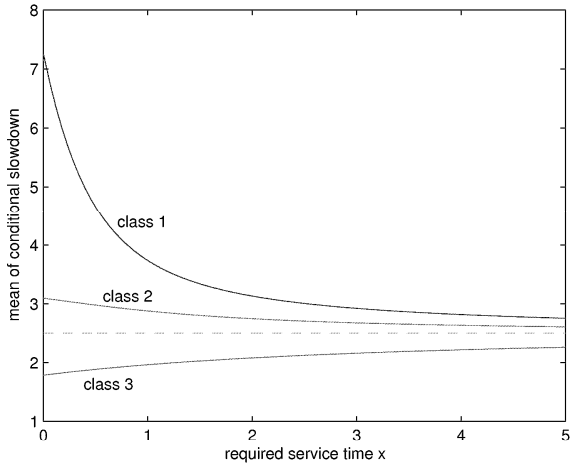

Figure 6.9: *Conditional mean slowdown for Example 2b with* $\alpha_2 = 3.0$.

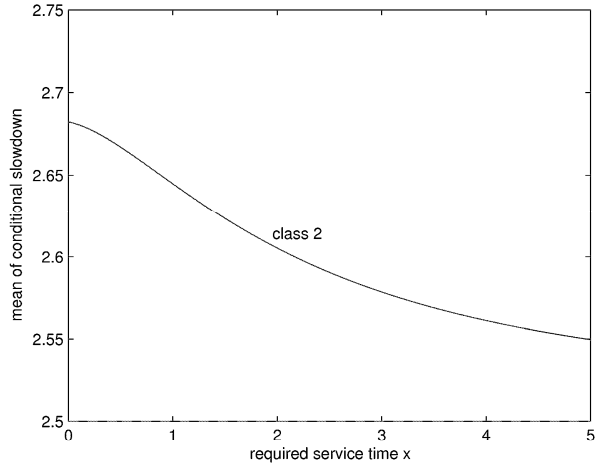

Figure 6.10: *Conditional mean slowdown for middle class in Example 2b with* $\alpha_2 = 4.0$.

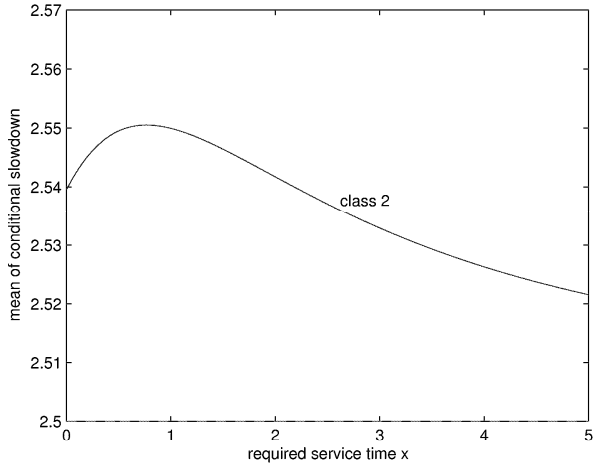

Figure 6.11: *Conditional mean slowdown for middle class in Example 2b with* $\alpha_2 = 4.5$.

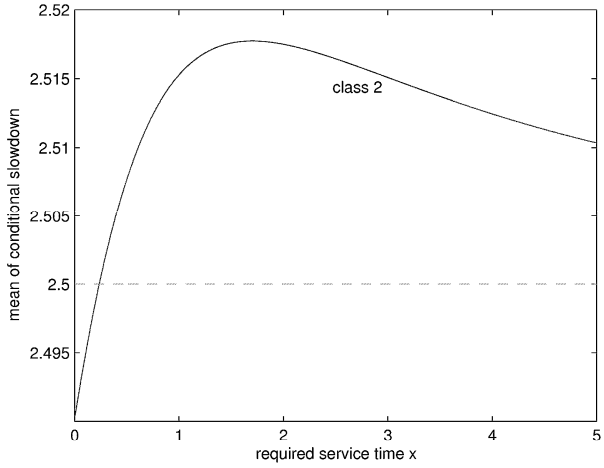

Figure 6.12: *Conditional mean slowdown for middle class in Example 2b with* $\alpha_2 = 4.7$.

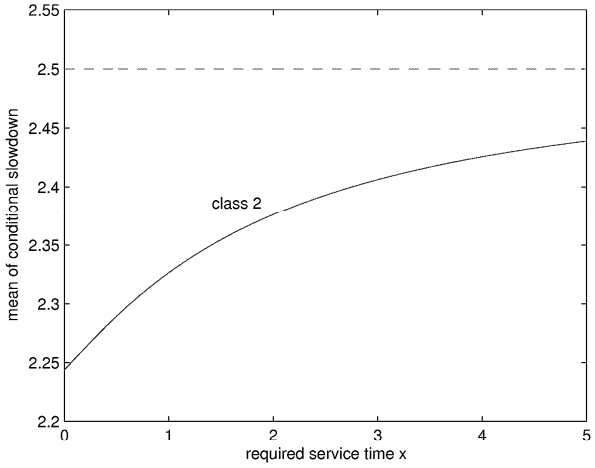

Figure 6.13: *Conditional mean slowdown for middle class in Example 2b with* $\alpha_2 = 6.0$.

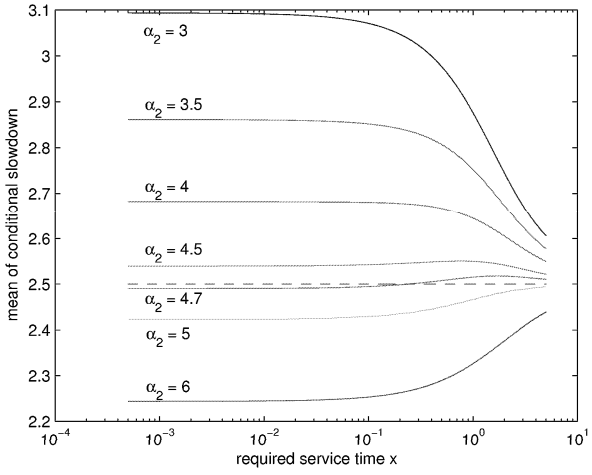

Figure 6.14: *Conditional mean slowdown for middle class in Example 2b.*

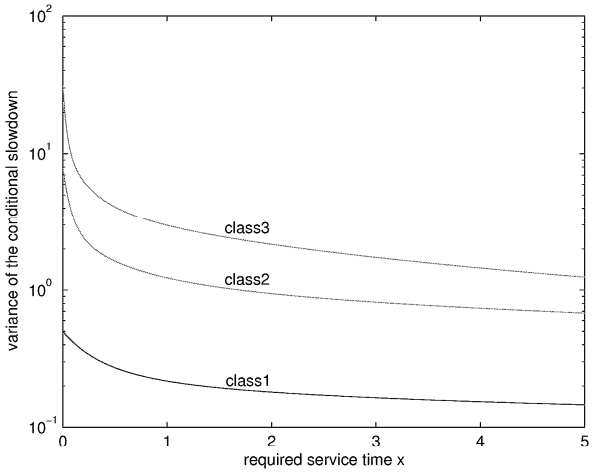

Figure 6.15: *Variance of the conditional slowdown in Example 3.*

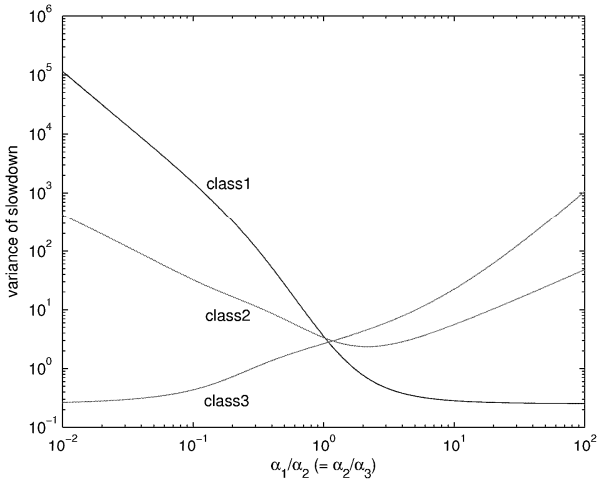

Figure 6.16: *Variance of slowdown in Example 3.*

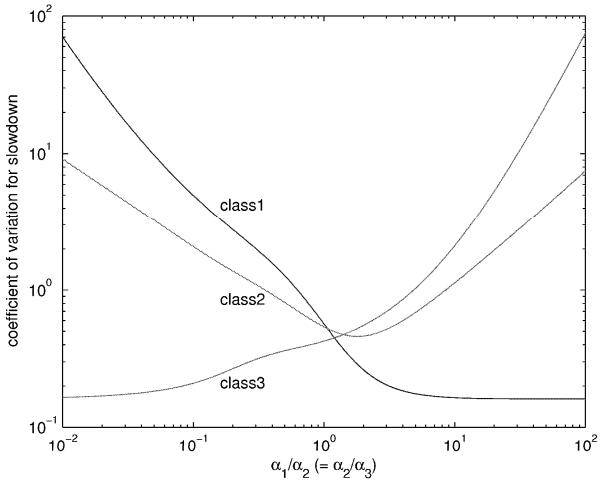

Figure 6.17: *Coefficient of variation for slowdown in Example 3.*

6.5 Proof of Lemma 6.5

Proof. For $s > 0$, let $\widetilde{T}_i(s) \equiv \int_0^\infty e^{-sx}\mathbb{E}T_i(x)dx$, be the Laplace Transform (LT) of $\mathbb{E}T_i(x)$. Taking LTs in (6.2), we readily obtain

$$\widetilde{T}_i(s) = \frac{1}{s}\left(\frac{1}{\alpha_i}\alpha^T\widetilde{m}_i(s) + \frac{1}{s}\right), \tag{6.28}$$

where $\widetilde{m}_i(s)$ is the LT of $m_i(x)$. Similarly, from (6.3), we have $s\widetilde{m}_i(s) - m_i(0) = \frac{1}{\alpha_i}B\widetilde{m}_i(s) + \frac{1}{s}\lambda$, and together with $m_i(0) = L^1$, it can be rewritten as

$$\widetilde{m}_i(s) = \alpha_i(\alpha_i sI - B)^{-1}\left(\frac{1}{s}\lambda + L^1\right). \tag{6.29}$$

Then substitution of (6.29) into (6.28) yields

$$\widetilde{T}_i(s) = \frac{1}{s^2} + \frac{1}{s^2}\alpha^T(\alpha_i sI - B)^{-1}\left(\lambda + sL^1\right). \tag{6.30}$$

Then, taking LTs in (6.4) and using (6.30) leads to

$$\frac{d^2}{ds^2}\widetilde{G}_i(s) = \frac{1}{s}\frac{2}{\alpha_i}\alpha^T\widetilde{y}_i(s) + 2\left(\frac{1}{s^3} + \frac{1}{s^3}\alpha^T(\alpha_i sI - B)^{-1}(\lambda + sL^1)\right), \tag{6.31}$$

where $\widetilde{y}_i(s)$ is the LT of $y_i(x)$. Taking LTs in (6.5) and using (6.29), (6.30), and the fact that $y_i(0) = [0\ \cdots\ 0]^T$, we obtain

$$\begin{aligned}\frac{1}{\alpha_i}(\alpha_i sI - B)\widetilde{y}_i(s) = {} & \frac{1}{\alpha_i}\widetilde{Z}_i(s)\alpha \\ & + \left(\frac{1}{s^2} + \frac{1}{s^2}\alpha^T(\alpha_i sI - B)^{-1}(\lambda + sL^1)\right)\lambda \\ & + \left(I + \frac{1}{\alpha_i}diag(\alpha_1, ..., \alpha_K)\right)\alpha_i(\alpha_i sI - B)^{-1}\left(\frac{1}{s}\lambda + L^1\right),\end{aligned}$$

or, equivalently,

$$\begin{aligned}\widetilde{y}_i(s) = {} & (\alpha_i sI - B)^{-1}\widetilde{Z}_i(s)\alpha \\ & + \alpha_i\left(\frac{1}{s^2} + \frac{1}{s^2}\alpha^T(\alpha_i sI - B)^{-1}(\lambda + sL^1)\right)(\alpha_i sI - B)^{-1}\lambda \\ & + \alpha_i(\alpha_i sI - B)^{-1}\left(I + \frac{1}{\alpha_i}diag(\alpha_1, ..., \alpha_K)\right)\alpha_i(\alpha_i sI - B)^{-1}\left(\frac{1}{s}\lambda + L^1\right),\end{aligned} \tag{6.32}$$

where $\widetilde{Z}_i(s)$ is the LT of $Z_i(x)$. Substitution of (6.32) into (6.31) yields

$$\begin{aligned}
&\frac{d^2}{ds^2}\widetilde{G}_i(s) = \frac{2}{\alpha_i}\frac{1}{s}\alpha^T(\alpha_i sI - B)^{-1}\widetilde{Z}_i(s)\alpha \\
&+ 2\left(\frac{1}{s^3} + \frac{1}{s^3}\alpha^T(\alpha_i sI - B)^{-1}(\lambda + sL^1)\right)\left(\alpha^T(\alpha_i sI - B)^{-1}\lambda + 1\right) \\
&+ \frac{2}{s^2}\alpha^T(\alpha_i sI - B)^{-1}\left(I + \frac{1}{\alpha_i}diag(\alpha_1, ..., \alpha_K)\right)\alpha_i(\alpha_i sI - B)^{-1}(\lambda + sL^1).
\end{aligned} \tag{6.33}$$

Now we need to investigate the expression $\alpha^T(\alpha_i sI - B)^{-1}\widetilde{Z}_i(s)\alpha$. Taking LTs in (6.6) and using (6.29) leads to

$$\begin{aligned}
s\widetilde{Z}_i(s) - Z_i(0) = &\frac{1}{\alpha_i}B\widetilde{Z}_i(s) + \frac{1}{\alpha_i}\widetilde{Z}_i(s)B^T \\
&+ \left(I + \frac{1}{\alpha_i}diag(\alpha_1, ..., \alpha_K)\right)\alpha_i(\alpha_i sI - B)^{-1}\left(\frac{1}{s}\lambda + L^1\right)\lambda^T \\
&+ \lambda\left(\frac{1}{s}\lambda + L^1\right)^T\alpha_i(\alpha_i sI - B^T)^{-1}\left(I + \frac{1}{\alpha_i}diag(\alpha_1, ..., \alpha_K)\right),
\end{aligned}$$

where $Z_i(0) = L^2$, see (6.7). Next, after postmultiplying the above by u_j^T yields

$$\begin{aligned}
&\frac{1}{\alpha_i}((\alpha_i s - \kappa_j)I - B)\widetilde{Z}_i(s)u_j^T \\
&= L^2u_j^T + \left(I + \frac{1}{\alpha_i}diag(\alpha_1, ..., \alpha_K)\right)\alpha_i(\alpha_i sI - B)^{-1}\left(\frac{1}{s}\lambda + L^1\right)(\lambda^T u_j^T) \\
&+ \lambda\left(\frac{1}{s}\lambda + L^1\right)^T\alpha_i(\alpha_i sI - B^T)^{-1}\left(I + \frac{1}{\alpha_i}diag(\alpha_1, ..., \alpha_K)\right)u_j^T,
\end{aligned}$$

and hence

$$\begin{aligned}
&\alpha^T(\alpha_i sI - B)^{-1}\widetilde{Z}_i(s)u_j^T \\
&= \alpha_i\alpha^T(\alpha_i sI - B)^{-1}\left((\alpha_i s - \kappa_j)I - B\right)^{-1}L^2u_j^T \\
&+ \alpha_i^2\alpha^T(\alpha_i sI - B)^{-1}((\alpha_i s - \kappa_j)I - B)^{-1}\left(I + \frac{1}{\alpha_i}diag(\alpha_1, ..., \alpha_K)\right) \\
&\times (\alpha_i sI - B)^{-1}\left(\frac{1}{s}\lambda + L^1\right)(\lambda^T u_j^T) + \alpha_i^2\alpha^T(\alpha_i sI - B)^{-1}((\alpha_i s - \kappa_j)I - B)^{-1} \\
&\times \lambda\left(\frac{1}{s}\lambda + L^1\right)^T(\alpha_i sI - B^T)^{-1}\left(I + \frac{1}{\alpha_i}diag(\alpha_1, ..., \alpha_K)\right)u_j^T.
\end{aligned} \tag{6.34}$$

If we let $[c_1 \ \cdots \ c_K] = \alpha^T [v_1 \ \cdots \ v_K]$, then $\alpha^T = [c_1 \ \cdots \ c_K] \begin{bmatrix} u_1 \\ \vdots \\ u_K \end{bmatrix}$, and equation (6.34) leads to

$$
\begin{aligned}
&\alpha^T(\alpha_i sI - B)^{-1}\widetilde{Z}_i(s)\alpha \\
&= \sum_{j=1}^{K} c_j \alpha^T(\alpha_i sI - B)^{-1}\widetilde{Z}_i(s)u_j^T \\
&= \sum_{j=1}^{K} c_j \left\{ \alpha_i \alpha^T(\alpha_i sI - B)^{-1}((\alpha_i s - \kappa_j)I - B)^{-1}L^2 u_j^T \right. \\
&+\alpha_i^2 \alpha^T(\alpha_i sI - B)^{-1}((\alpha_i s - \kappa_j)I - B)^{-1}\left(I + \frac{1}{\alpha_i} diag(\alpha_1, ..., \alpha_K)\right) \\
&\times(\alpha_i sI - B)^{-1}\left(\frac{1}{s}\lambda + L^1\right)(\lambda^T u_j^T) + \alpha_i^2 \alpha^T(\alpha_i sI - B)^{-1}((\alpha_i s - \kappa_j)I - B)^{-1} \\
&\left.\times\lambda\left(\frac{1}{s}\lambda + L^1\right)^T(\alpha_i sI - B^T)^{-1}\left(I + \frac{1}{\alpha_i} diag(\alpha_1, ..., \alpha_K)\right)u_j^T \right\}. \qquad (6.35)
\end{aligned}
$$

Finally, substitution of (6.35) into (6.33) yields the result (6.16) in Lemma 6.5. ∎

Chapter 7

Queueing models with time-fluctuating service capacity

7.1 Background and introduction

In this chapter we consider queues with time-fluctuating service capacity; also see Section 1.3 for a brief introduction.

For these models, various approaches have been developed to capture the essential dependence of the system performance in terms of parameters such as arrival rates, service rates, etc. One very successful line of research has been the analysis through 'time-scale decomposition'. In short, this approach consists of studying the system performance in two limiting regimes: one, coined the 'fluid regime' [34], in which the dynamics of the modulating environment is sped up to infinity, which in case of independent modulating processes, is equivalent to replacing the server by one working at constant speed equal to the original average speed. This approach in general tends to be much too optimistic and the obtained performance may not be approached even by far in the system with stochastic variations. The other extreme, the 'quasi-stationary' regime, is obtained by assuming that capacity fluctuations are infinitely slow compared to traffic dynamics. Although this approach is usually conservative, it tends to be much too pessimistic and does not serve as a useful approximation in general either. A further complication is that the quasi-stationary limit has no sensible meaning if the service rates can at times be smaller than the arrival rate. In general, this need not lead to instability, but in the quasi-stationary limit the system may be unstable during infinitely long periods which implies that the predicted system performance has no sensible meaning, as buffer contents and transmission delay grow to infinity. This led to the notion of 'uniform stability' where one requires that service rates are never smaller than arrival rates [34].

A recent paper by Hampshire, Harchol-Balter, and Massey [46] points out that in practice uniform stability is not realistic. As a consequence, the authors focus on the transient or time-dependent performance of the system. More precisely, they assume a given realization (in time) of the service rate process and aim at approximating the state of the system at time t. Then they perform a time-acceleration technique similar to the time-decomposition mentioned above, to estimate the transient performance of the system if the arrival and service rates are scaled linearly with a common parameter. This technique is referred to as Uniform Acceleration [70, 73, 74]. For time epochs satisfying a stability condition which is more strict than 'instantaneous stability', the system performance is essentially estimated by that of a constant-rate system. Their ultimate goal in that paper is to estimate the queue length and delay in the system with the EPS service discipline.

The quasi-stationary approximation can usually be interpreted as follows: Given the stationary probabilities of the service rate process, one assumes that the service rate at time 0 is determined according to this distribution, and remains fixed at that value indefinitely. This gives a sensible approximation when the queueing process is stable for all possible service rates. As pointed out above, this approach fails if the queueing process is unstable for some set of service rates that occur with positive probability. One way to circumvent this shortcoming is to examine the process at the time-scale of service rate fluctuations, instead of the time-scale of the queueing dynamics. In other words, instead of taking the service rate variations to be infinitely slow, one assumes the in- and output of the queue to be a piecewise linear function with the slope of this function being determined by the drift of the queue. Effectively, this can be seen as the 'fluid' limit *from the perspective of the service rate process*. Although this terminology can give rise to confusion, the models that result from this approach are commonly referred to as fluid queues, see [6, 90]. When the service rates are governed by a Markov process, the distribution of the buffer content may be exactly determined through spectral analysis. We will discuss this technique in more detail below.

In this chapter, we study the buffer content in a queue with a fluctuating service rate that depends on the state of an exogenous Markov process. This process can, for instance, model the number of transfers of unresponsive flows. The arrival rate of the queue may be larger than the service rate for some states of the Markovian environment. To this end, we refine the quasi-stationary analysis to allow for unstable service states, as is the case with the fluid queue approach, but aim at including the queueing dynamics as well. In particular, we study the evolution in time of the "effective load", which measures the aggregate past input and output rates, cf. [46]. The instantaneous effective load can be described from the stationary measure of the corresponding fluid queue. We then extend this analysis to capture the randomness of the input and output processes of the original queue (in the fluid queue these were essentially replaced by their average values). This gives rise to a notion of adjusted stability which captures the effect of accumulated work during periods in which the queue is unstable. For this we rely on a detailed analysis of

the recovery time, i.e., the time needed to recover from the excess load after a low rate period which may encompass multiple stable periods. The recovery time is associated with the workload process in a fluid queue driven by the same Markov process as the original queue. This allows us to determine, in the limiting regime, the conditional queue length distribution given that the service rate is high (this distribution is defective in the quasi-stationary limit).

The research in this chapter is strongly motivated by the work in [46] that focuses on the transient behavior of a single-server EPS queue with a service rate fluctuating according to a deterministic process. Our model also includes the high and low service rates model of Gupta, Harchol-Balter, Wolf, and Yechiali [45], where a single-server queue with exponential high and low load periods is studied. These authors focus on the impact of the rate of fluctuation between the high and low load periods on the queue length and response times. On the one hand, for the special case of constant service rate their results confirm Ross's conjecture [86] stating that increasing variability in a Poisson arrival process increases the mean customer delay. On the other hand, the results of [45] show that the mean queue length in the quasi-stationary regime can be either higher or lower than in the fluid regime. However, if the service rates vary, then the mean queue length in the quasi-stationary regime is higher than in the fluid regime. In this chapter, we take the analysis one step further and explicitly investigate the queue length during high and low service rate periods. Additional generalizations with respect to [45] include a generalization to multiple service rate levels, which also allows us to consider non-exponential high and low rate periods, and our explicit analysis of the quasi-stationary regime, including the distribution of the number of high service rate periods required for recovering to stability.

This chapter is organized as follows. Section 7.2 presents the single-server queue with a service rate fluctuating according to a Markov process. Section 7.3 introduces effective load, and characterizes its distribution via a fluid queue driven by the same Markov process. Section 7.4 considers the on-off model to provide intuition for our main result presented in Section 7.5, where the qualitatively different behavior between recovery periods and stable periods during high service rate periods is studied, refining the quasi-stationary regime to allow for temporary instability. Our concluding remarks are given in Section 7.6.

7.2 Preliminaries

We consider a queue with Poisson arrivals at rate λ and service requirements are assumed to be *exponentially* distributed with mean 1. The stochastic service rate process $\{\mu(t), t > 0\}$ fluctuates over time and is assumed to be independent of all inter-arrival times and service requirements of the customers before time t. For all realizations of the

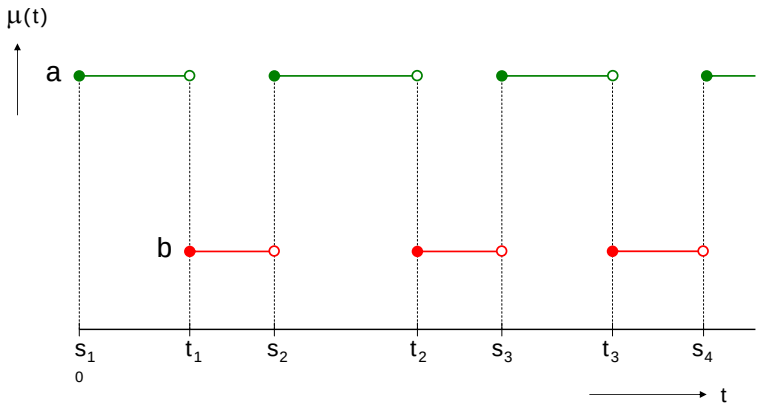

Figure 7.1: *Service rates in high-low model*

service rate process, it is assumed that the sample paths are continuous and differentiable almost everywhere, except on a countable set of isolated points with measure 0. The latter technical condition is satisfied for the particular processes that we will consider.

We will be specifically interested in the case where $\mu(t)$ fluctuates between a high and a low service rate. Under the low service rate the queue is unstable, that is the instantaneous load $\lambda/\mu(t)$ exceeds 1, but we do assume that the high service rate is such that our system is stable (in the long run). Since high and low service rate periods are random variables, high rate periods may be insufficient to recover from the excess load of the low rate period. In the following two sections we develop some intuition for the queue dynamics for two special cases. First for the case of an alternating service rate between a high and a low value. Then we consider a more versatile setting where the service rate may take on more than two values, while the modulating environment is governed by a Markovian birth-death process.

7.2.1 High-low model

Suppose the service rate process alternates between a high and a low value, see Figure 7.1. For a given realization of the service rate process, we let $\{s_i, t_i\}_{i\in\mathbb{N}}$ be the sequence of time points where the service rate switches from the low service rate to the high service rate for the i^{th} time (time s_i) and the first epoch thereafter that it switches back (time t_i). We assume that $0 = s_1 < t_1 < s_2 < t_2 < s_3 < t_3 < \cdots$, and let the time-dependent service rate be given by

$$\mu(t) = \begin{cases} a, & t \in [s_i, t_i) \\ b, & t \in [t_i, s_{i+1}) \end{cases}$$

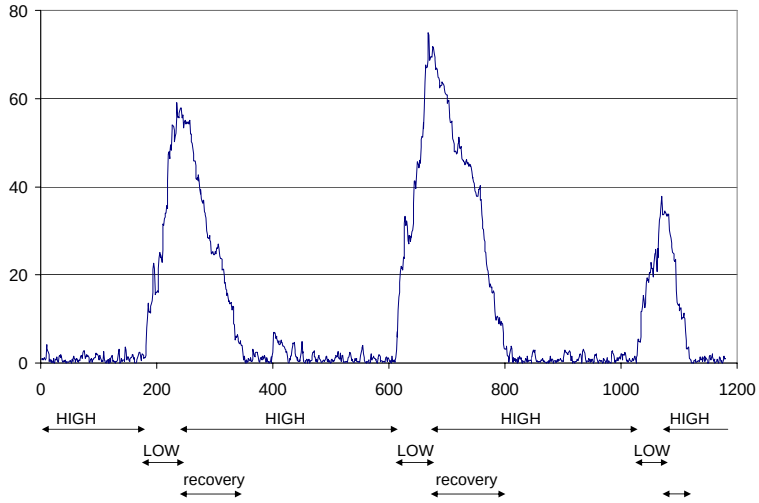

Figure 7.2: *A typical sample path of the queue length process in the alternating high-low model*

for $i \in \mathbb{N}$ and assume that $a > \lambda > b \geq 0$. During the time period $[s_i, t_i)$ with length $A_i = t_i - s_i$, the server works at the *higher* rate a; and during $[t_i, s_{i+1})$ with length $B_i = s_{i+1} - t_i$, the server works at the *lower* rate b. We will refer to this as the *high-low* model. In the special case of $b = 0$ we will speak of an *on-off* model. For now we do not need to make any assumptions on the distributions of the high rate periods A_i and the low rate periods B_i, but in the sequel these will typically be assumed to form two i.i.d. sequences, independent of each other.

We will be particularly interested in the case where λ, a and b are relatively much larger than the typical durations of high rate and low rate periods. We will later formalize this by replacing these parameters with $\eta\lambda$, ηa and ηb and passing the parameter $\eta > 0$ to infinity. In Figure 7.2 we depict a typical realization of the queue length process. The service rate starts off in the higher value a and the process shows stationary behavior. As soon as the service rate switches to the lower value b the queue starts building up. The instantaneous load $\rho(t) = \lambda/\mu(t)$ then exceeds the value 1, i.e., the queue is temporarily unstable. The major trend is characterized by the linear drift $\lambda - b$, but due to the randomness in the arrival and service processes (both Poissonian) there are fluctuations around the linear trend. The top of the curve corresponds to a switching time back to the high rate. Although the linear trend is then negative ($\lambda - a < 0$) it takes a while for the process to reach the level of the typical stationary behavior under the high service rate. Roughly speaking, this *recovery* period lasts until the linear trend hits the horizontal axis;

also see Figures 7.5-7.7 on page 113, for a graphical representation of the behavior when the parameter $\eta > 0$ grows to infinity.

An unconditional (not conditioning on the state of the service process) stationary regime exists under the usual (long-term) stability condition

$$\lambda < \frac{\mathbb{E}A}{\mathbb{E}A + \mathbb{E}B} a + \frac{\mathbb{E}B}{\mathbb{E}A + \mathbb{E}B} b. \tag{7.1}$$

We will assume this condition to hold. The main message from these preliminaries is that there are three types of periods during which the queueing dynamics are intrinsically different: (i) instability periods (when the service rate is low), (ii) recovery periods (when the service rate is high, but the queue has not yet recovered from an instability period), and (iii) quasi-stationarity (the queue behaves as if the service rate is always high). It is crucial to note that some high rate periods may be too short to recover from instability, i.e., a recovery period may be interrupted by one or more instability periods.

The above gives rise to a notion of adjusted stability: the queue has similar dynamics as a stable queue with a constant service rate a with a probability *smaller* than $\frac{\mathbb{E}A}{\mathbb{E}A+\mathbb{E}B}$. We will make this statement more precise later.

7.2.2 Markov modulated service rates

In this chapter we do not aim at full generality of the modulating service process, the arrival process into the queue, nor the service requirement distribution. Nevertheless, there is a particular choice for the service rate process that leads to convenient expressions, while still offering sufficient flexibility to cover, for example, high or low rate periods with non-exponential lengths. We discuss such a setting in this section. We assume that the service rate process $\mu(t)$ is modulated by an independent Markovian background process $M(t)$ with state space $\mathcal{M} = \{0, 1, ..., m\}$, for some integer value m, with possibly $m = \infty$. If the background process is in state i (i.e., if $M(t) = i$), then the service rate at time t is given by $\mu(t) = \mu_i$, for any $i \in \mathcal{M}$. The states $i \in \mathcal{M}$ for which $r_i := \lambda - \mu_i > 0$ are called the low service rate states in which the instantaneous load exceeds 1. The high service rate states are the states $i \in \mathcal{M}$ with $r_i < 0$. It is convenient and not very restrictive to assume $r_i \neq 0$.

We assume that the background process $M(t)$ is a birth-death process on its state space $\mathcal{M}$. If $M(t) = i$, then the birth rate is $\alpha_i > 0$ and the death rate is $\beta_i > 0$. The usual stability condition *for the queue* is given by

$$r_0 + \sum_{i=1}^{m} \left(\prod_{j=0}^{i-1} \frac{\alpha_j}{\beta_{j+1}} \right) r_i < 0, \tag{7.2}$$

which can be interpreted as the mean drift of the queue being negative (e.g., see [90]). If $m = \infty$ we additionally need to assume that the modulating birth-death process itself is

stable which amounts to:

$$\sum_{i=1}^{\infty}\left(\prod_{j=0}^{i-1}\frac{\alpha_j}{\beta_{j+1}}\right)<\infty. \tag{7.3}$$

Note that, for the special case of the high-low service rate model with exponential low and high periods, (7.2) is the same as the usual stability condition (7.1).

Another particularly convenient special case is obtained when there is some $k \geq 0$ so that $r_i \equiv a$ for all $i \leq k$ and $r_i \equiv b$ for all $i > k$ (or with the role of a and b interchanged). This setting will allow for elegant closed-form results.

7.3 Effective load

Following [46] we define the "effective load" at time t as

$$\rho^*(t) \equiv \sup_{0\leq s<t}\frac{\int_s^t \lambda(r)dr}{\int_s^t \mu(r)dr} = \sup_{0\leq s<t}\frac{(t-s)\cdot\lambda}{\int_s^t \mu(r)dr}, \tag{7.4}$$

which measures the aggregate past input and output. This entity will be the basis to a formalization of the concept of adjusted stability mentioned above. Note that, since the service rates $\mu(t)$ constitute a random process, the effective load itself is a random process. As we will see later, the distribution of $\rho^*(t)$ can be obtained from that of the workload in the associated Markov modulated fluid queue with constant fluid arrival rate λ and drain rate $\mu(t)$.

Before turning our attention to the random process ρ^* we first discuss some of its properties for a *given realization* of the service rate process $\mu(t)$, cf. the setting in [46].

7.3.1 Effective load for high-low model

We focus on the high-low model and suppose that the sequence $\{s_i, t_i\}_{i\in\mathbb{N}}$ which determines the high-rate and low-rate periods, is given.

Lemma 7.1. *During the i^{th} high-rate period, that is for $t \in [s_i, t_i)$, $i \geq 2$, we have*

$$\rho^*(t) = \sup_{1\leq j\leq i-1}\frac{(t-t_j)\lambda}{\int_{t_j}^t \mu(r)dr}. \tag{7.5}$$

During any low-rate period, $t \in [t_i, s_{i+1})$, for $i \geq 1$ we have $\rho^*(t) = \frac{\lambda}{b}$.

Proof. For a given $t > 0$, define the function $\Psi(s)$, $0 \leq s < t$, by

$$\Psi(s) := \frac{(t-s)\lambda}{\int_s^t \mu(r)dr},$$

where $\frac{\lambda}{a} \leq \Psi(s) \leq \frac{\lambda}{b}$. It is not difficult to verify that $\Psi(s)$ is strictly increasing on $s \in [s_j, t_j)$ and strictly decreasing on $s \in [t_j, s_{j+1})$. Hence, it holds that: $\sup_{s\in[s_j,t_j)} \Psi(s) = \Psi(t_j)$ and $\sup_{s\in[t_j,s_{j+1})} \Psi(s) = \Psi(t_j)$. Then, for $t \in [s_i, t_i)$, $i \geq 2$, the effective load function (7.4) can be rewritten as

$$\rho^*(t) = \sup\left\{ \sup_{1\leq j\leq i-1} \sup_{s\in[s_j,t_j)} \Psi(s), \sup_{1\leq j\leq i-1} \sup_{s\in[t_j,s_{j+1})} \Psi(s), \sup_{s\in[s_i,t)} \Psi(s) \right\} \quad (7.6)$$

$$= \sup\left\{ \sup_{1\leq j\leq i-1} \Psi(t_j), \sup_{1\leq j\leq i-1} \Psi(t_j), \frac{\lambda}{a} \right\} = \sup_{1\leq j\leq i-1} \Psi(t_j). \quad (7.7)$$

by splitting up the supremum in (7.4) into suprema of a partition. Similarly, during low-periods we have for $t \in [t_i, s_{i+1})$, $i \geq 1$,

$$\rho^*(t) = \sup\left\{ \sup_{1\leq j\leq i} \Psi(t_j), \sup_{1\leq j\leq i} \Psi(t_j), \frac{\lambda}{b} \right\} = \frac{\lambda}{b}.$$

∎

Remark 7.2. *If $a > b > 0$, then $\rho^*(t)$ is (left and right) continuous and finite at $t = s_i$ (i.e., at the beginning of a high-rate period), but $\rho^*(t)$ has a jump at $t = t_i$ (i.e., at the beginning of a low-rate period).*

The supremum in equation (7.5) is achieved for a certain index j^*, with $j^* \leq i-1$. In general, if the high-rate periods are "sufficiently long" (relative to the low-rate periods, arrival and service rates), then the supremum is in general achieved for $j^* = i - 1$. In contrast, if the high-rate periods are too short, the supremum is achieved at a lower index $k^* \leq j^*$. A characterization of "how long" a high-rate period should be, will be given later.

As an illustration we have depicted the effective load in Figure 7.3, for the alternating high-low model with high and low periods of deterministic length 1, and with $\lambda = \frac{3}{2}$, $a = 4$, $b = \frac{4}{5}$. The instantaneous load $\rho(t)$ is 0.375 during high-rate periods, and 1.875 during low-rate periods. As was stated in Lemma 7.1, the effective load and the instantaneous load coincide during low service rate periods. The effective load at time t is strictly decreasing in t during high-rate periods (starting from the value λ/b at the beginning of a high-rate period). If the high-rate period is sufficiently long (relative to λ, a, and b), then the effective load drops below the value 1. The recovery time is the time needed (since the end of the last low-rate period) for the effective load to drop to 1. Heuristically speaking, we can say that the queue "becomes stable" at the time epoch u such that $\rho^*(u) = 1$.

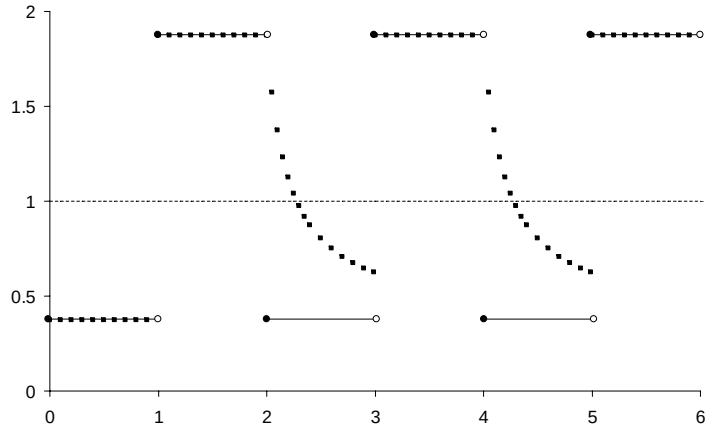

Figure 7.3: *Example of the effective load function $\rho^*(t)$ (marked with squares), with the instantaneous load $\rho(t)$ and the critical line 1.*

7.3.2 Distribution of the effective load

We now study the distribution of the effective load when the service rates $\mu(t)$ are not fixed, but follow a random process. Again, we do not aim at full generality, but limit ourselves to the case where $\mu(t)$ is determined by a modulating Markovian process, which leads to elegant closed-form results. (Methodologically, more general rate fluctuations, including the high-low model with generally distributed high-rate and low-rate periods, can be treated analogously.)

From the definition of $\rho^*(t)$ in (7.4) we observe that, for $x \in \mathbb{R}_+$: $\rho^*(t) > x$ is equivalent to $W_x(t) > 0$, where $W_x(t)$ is defined as the fluid content process at time t in the associated Markov modulated fluid queue [6, 90], where we replace the Poisson arrivals and the service times in the queue by fluid streams of rate λ (constant) and $x \cdot \mu(t)$. To this end, note that

$$\begin{aligned} \rho^*(t) > x &\Leftrightarrow \exists s \in [0,t) : \int_s^t \lambda(r)dr - x\int_s^t \mu(r)dr > 0 \\ &\Leftrightarrow \sup_{0\le s<t}\left\{\int_s^t \lambda(r)dr - x\int_s^t \mu(r)dr\right\} > 0 \\ &\overset{\text{def}}{\Leftrightarrow} W_x(t) > 0, \end{aligned} \tag{7.8}$$

where the supremum in equation (7.8) can be interpreted as a workload process (see e.g. [8]). The content $W_x(t)$ of the fluid queue is regulated by the background process

$M(t) \in \mathcal{M}$ as follows:

$$\frac{dW_x(t)}{dt} = \begin{cases} 0 & \text{if } W_x(t) = 0, \text{ and } \lambda < x\mu_{M(t)} \\ \lambda - x\mu_{M(t)}, & \text{otherwise.} \end{cases}$$

Note that the fluid queue and the original queue share exactly the same realization of the service rate process. The fluid queue, however, does not incorporate the fluctuations due to the randomness in the arrival and service processes.

The stability condition for the fluid queue is

$$\lambda - x\mu_0 + \sum_{i=1}^{m} \left(\prod_{j=0}^{i-1} \frac{\alpha_j}{\beta_{j+1}} \right) (\lambda - x\mu_i) < 0, \tag{7.9}$$

which is the same as that for the original queue (7.2) when $x = 1$. If (7.9) is satisfied, the stationary distribution of the fluid queue exists and can be determined through spectral analysis, see [90]. We will focus on the particular choice of the modulating process described in Section 7.2.2.

We construct a high-low system by setting $\mu_0 = b$ and $\mu_i = a$ for all $i \geq 1$, with $a > b$ and $m = \infty$. We further choose $\alpha_i \equiv \alpha$ and $\beta_i \equiv \beta > \alpha$. Note that the low-rate periods are exponentially distributed, but the high-rate periods are distributed as the busy period in an M/M/1 queue with arrival rate α and service rate β. This model can be seen as a counterpart of the on-off model discussed in Section 7.4 where the off-periods (low rate) have a general distribution and the on-periods (high rate) have an exponential distribution.

In this case Scheinhardt [90, pp. 26–28] shows that the stationary fluid content process W_x is given by, for any $y \geq 0$,

$$\mathbb{P}(W_x > y) = p_{0;x} \cdot \exp\left\{ -\left(\frac{\alpha}{\lambda - bx} - \frac{\beta}{(a-b)x} \right) y \right\},$$

where

$$p_{0;x} = \frac{1 - \alpha/\beta}{(ax - \lambda)/((a-b)x)}.$$

Invoking (7.8), for $y = 0$, we obtain the stationary distribution of the effective load ρ^* as:

$$\mathbb{P}(\rho^* > x) = \mathbb{P}(W_x > 0) = p_{0;x},$$

provided that $\frac{ax-\lambda}{(a-b)x} < \frac{\alpha}{\beta} < 1$, cf. [90].

Based on the distribution of the effective load we can calculate performance measures such as the long-run fraction of time that the system is (instantaneously) stable, which is less than the fraction of time that the system works at high service rates.

7.4 Analysis and intuition for on-off model

In this section we study the buffer content in a queue when no service is available for some time periods (called off-periods). We refine a result of Núñez-Queija [79] which considers the processor-sharing queue with service interruptions. Since the service requirements are exponentially distributed, the queue length process remains unchanged if we replace the processor-sharing discipline by another work-conserving service discipline. In particular, based on the explicit formulas from [79] we show that the conditional queue length distribution (given that the server is turned on) is defective in the quasi-stationary limit. The latter observation can be related to the notion of adjusted stability and to the recovery time.

As in the model [79] the on-periods A_i, $i \geq 1$, are assumed to be independently, identically and exponentially distributed with mean α^{-1}, and the service rate during on-periods is a. The off-periods B_i, $i \geq 1$, are identically and generally distributed as the random variable B with distribution function $B(t) := \mathbb{P}(B \leq t)$, $t \geq 0$. Let the Laplace-Stieltjes transform of B be given by $\widetilde{B}(s) := \mathbb{E}e^{-sB}$, for $\mathrm{Re}(s) \geq 0$; and let the k-th moment of B be given by

$$\beta_k := \int_{t=0}^{\infty} t^k dB(t).$$

7.4.1 Uniform Acceleration

We apply the Uniform Acceleration technique [73, 70, 74] to the model [79]. The queue length process in the 'accelerated system' is denoted by $Q^\eta(t)$, where the arrival and service rates are linearly scaled with a common parameter $\eta > 0$ (i.e., the arrival rate of the scaled system is given by $\eta\lambda$ and the service rate is given by $\eta\mu(t)$). Let (Q^η, μ) be a pair of random variables having the limiting distribution of $(Q^\eta(t), \mu(t))$. The joint distribution of (Q^η, μ) satisfies, cf. [79],

$$\mathbb{E}\left[z^{Q^\eta} \mid \mu = a\right] = \frac{a - \lambda(1 + \alpha\beta_1)}{a - \lambda\left(1 + \alpha\beta_1 \cdot \varphi_B(z, \eta\lambda)\right) z}, \tag{7.10}$$

$$\mathbb{E}\left[z^{Q^\eta} \mid \mu = 0\right] = \varphi_B(z, \eta\lambda) \cdot \mathbb{E}\left[z^{Q^\eta} \mid \mu = a\right], \tag{7.11}$$

where

$$\varphi_B(z, \eta\lambda) := \frac{1 - \widetilde{B}(\eta\lambda(1 - z))}{\beta_1 \eta\lambda\,(1 - z)}$$

is the probability generating function (pgf) of the number of arrivals according to a Poisson process with rate $\eta\lambda$, during the backward recurrence time of an off-period.

By differentiating the pgf (7.10) with respect to z and setting $z = 1$ we obtain the

conditional mean:

$$\mathbb{E}\left[Q^\eta \middle|\, \mu = a\right] = \frac{\lambda}{p_{ON}\cdot a - \lambda} + \frac{\alpha\beta_2}{2}\frac{p_{ON}\cdot\lambda^2}{p_{ON}\cdot a - \lambda}\eta, \tag{7.12}$$

where $p_{ON} = \frac{1}{1+\alpha\beta_1}$ is the long-run fraction of time that the server is available. Similarly, by (7.11) we have

$$\begin{aligned}\mathbb{E}\left[Q^\eta \middle|\, \mu = 0\right] &= \mathbb{E}\left[Q^\eta \middle|\, \mu = a\right] + \mathbb{E}N_{\eta\lambda}(B)\\ &= \mathbb{E}\left[Q^\eta \middle|\, \mu = a\right] + \eta\lambda\frac{\beta_2}{2\beta_1},\end{aligned} \tag{7.13}$$

where $\mathbb{E}N_{\eta\lambda}(B)$ is the expected number of Poisson arrivals with rate $\eta\lambda$ during the backward recurrence time of an off-period.

The conditional mean queue length (7.12) tends to infinity if the scaling parameter $\eta > 0$ tends to infinity. In other words, in the quasi-stationary limit, the mean queue length during on-periods is infinite *even when the usual stability criterion (7.1) is satisfied*. We will further condition the queue length during on-periods on whether or not the queue has recovered from previous instability periods. We observe that the conditional distribution of $(Q^\eta \mid \mu = a)$ is defective in the limit $\eta \to \infty$ with the defect probability

$$\lim_{\eta\to\infty} \mathbb{P}\left(Q^\eta = \infty \mid \mu = a\right) = \frac{\lambda}{a-\lambda}\alpha\beta_1 > 0. \tag{7.14}$$

To this end, note that in the quasi-stationary limit $\eta \to \infty$ we have the pgf

$$\lim_{\eta\to\infty} \mathbb{E}\left[z^{Q^\eta} \middle|\, \mu = a\right] = \frac{a - \lambda(1+\alpha\beta_1)}{a - \lambda z}, \tag{7.15}$$

by noting that $\lim_{\eta\to\infty}\varphi_B(z,\eta\lambda) = 0$ and taking the limit $\eta \to \infty$ in (7.10), and hence

$$\lim_{\eta\to\infty} \mathbb{P}\left(Q^\eta < \infty \mid \mu = a\right) = \frac{a - \lambda(1+\alpha\beta_1)}{a-\lambda} < 1.$$

Remark 7.3. *We can rewrite (7.15) as*

$$\frac{\lambda\alpha\beta_1}{a-\lambda}\times 0 + \frac{a-\lambda(1+\alpha\beta_1)}{a-\lambda}\times\frac{a-\lambda}{a-\lambda z}, \tag{7.16}$$

which can be interpreted as follows. With probability $\frac{\lambda\alpha\beta_1}{a-\lambda}$ the queue length is infinite in the quasi-stationary limit. With the complementary probability, the queue length is distributed as if the service rate is always a (i.e., as the queue length in the M/M/1 queue with load λ/a).

In order to refine the quasi-stationary limit, we scale the queue length. From the linearity of the mean queue length in η we see that the proper scaling is $Q^\eta(t)/\eta$. We then have the following:

Proposition 7.4. *The conditional distribution of the scaled queue length* $(\frac{1}{\eta}Q^\eta \mid \mu = a)$ *in the quasi-stationary limiting regime is given by*

$$\lim_{\eta\to\infty} \mathbb{E}\left[z^{\frac{1}{\eta}Q^\eta} \mid \mu = a\right] = \frac{a - \lambda\left(1 + \alpha\beta_1\right)}{a - \lambda\left(1 - \alpha\frac{1-\widetilde{B}(-\lambda\ln z)}{\lambda\ln z}\right)}. \tag{7.17}$$

Similarly,

$$\lim_{\eta\to\infty} \mathbb{E}\left[z^{\frac{1}{\eta}Q^\eta} \mid \mu = 0\right] = \frac{1 - \widetilde{B}\left(-\lambda\ln z\right)}{-\lambda\beta_1\ln z} \cdot \frac{a - \lambda\left(1 + \alpha\beta_1\right)}{a - \lambda\left(1 - \alpha\frac{1-\widetilde{B}(-\lambda\ln z)}{\lambda\ln z}\right)}. \tag{7.18}$$

Proof. Follows from (7.10) and the fact that

$$\lim_{\eta\to\infty} \varphi_B\left(z^{1/\eta}, \eta\lambda\right) = \frac{1 - \widetilde{B}\left(-\lambda\ln z\right)}{-\lambda\beta_1\ln z}.$$

∎

Differentiating (7.17) with respect to z and taking the limit $z \to 1$ leads to

$$\lim_{\eta\to\infty} \mathbb{E}\left[\frac{1}{\eta}Q^\eta \mid \mu = a\right] = \frac{\alpha\beta_2}{2}\frac{p_{ON}\cdot\lambda^2}{p_{ON}\cdot a - \lambda}, \tag{7.19}$$

and

$$\lim_{\eta\to\infty} \mathbb{E}\left[\frac{1}{\eta}Q^\eta \mid \mu = 0\right] = \frac{\alpha\beta_2}{2}\frac{p_{ON}\cdot\lambda^2}{p_{ON}\cdot a - \lambda} + \lambda\frac{\beta_2}{2\beta_1}, \tag{7.20}$$

which agrees with (7.12) and (7.13). The preceding can be interpreted as follows. From (7.16) we know that, in the limit, the non-scaled queue length during on-periods is non-defective with probability $\frac{a-\lambda(1+\alpha\beta_1)}{a-\lambda}$. Therefore, with that probability the scaled queue length during on-periods equals 0. With the complementary probability $\frac{\lambda\alpha\beta_1}{a-\lambda}$, the queue length "did not recover from instability" during an on-period. We therefore decompose (7.19) as

$$\frac{a-\lambda(1+\alpha\beta_1)}{a-\lambda} \times 0 + \frac{\lambda\alpha\beta_1}{a-\lambda} \times \frac{\beta_2}{2\beta_1}\frac{p_{ON}\cdot\lambda(a-\lambda)}{p_{ON}\cdot a - \lambda}. \tag{7.21}$$

Heuristically, we may say

$$\lim_{\eta\to\infty} \mathbb{E}\left[\frac{1}{\eta}Q^\eta \mid \mu = a \text{ but not yet recovered}\right] = \frac{\beta_2}{2\beta_1}\frac{p_{ON}\cdot\lambda(a-\lambda)}{p_{ON}\cdot a - \lambda}. \tag{7.22}$$

This decomposition of the queue length during on-periods can be done similarly for the entire distribution through the obtained expressions for the conditional pgfs.

Remark 7.5. *The above explains why constant-rate approximations for on-periods (high-rate periods) give poor results. The error can be made arbitrarily large by either increasing the second moment* β_2 *of the off-periods or the scaling parameter* η*.*

7.4.2 Intermediate discussion

The findings in the previous subsection have led us to a notion of adjusted stability as a refinement of the usual stability criterion (7.1). The fact that $(Q^\eta \mid \mu = a)$ is defective in the quasi-stationary limit is explained by the fact that Q^η explodes during an off-period when $\eta \to \infty$. Since the scaled system Q^η is stable in the long run, the system recovers from the explosion during an on-period. The queue becomes stable again (i.e., Q^η becomes finite) during an on-period if the on-period length is "sufficiently long". If the on-period length is not sufficiently long, then, in the quasi-stationary regime, Q^η remains infinite during the on-period.

7.5 Analysis for high-low model

In this section we extend the analysis of instability during high-rate periods to the high-low model. The discussion will center around a characterization of the recovery period (in subsection 7.5.1). We will think of the existence of these recovery periods as a refinement of the usual definition of stability. Particular attention will be given to the case with exponential high-rate and low-rate periods (in subsection 7.5.2), in which case closed-form results can readily be obtained. Ultimately, we will discuss the scaled version of the queue length in the quasi-stationary regime (in subsection 7.5.3).

7.5.1 Recovery period and adjusted stability

In this section we formalize the *recovery* period in the *high-low* model. Suppose at time s_i (i.e., at the beginning of i-th high-rate period) that for some $1 \leq k \leq i-1$ we have $\rho^*(t_k^-) < 1$ and $\rho^*(u) \geq 1$ for all $u \in [t_k, s_i)$, i.e., the moment t_k is the most recent moment where the effective load increased beyond 1.

If the system was unstable in the consecutive low-rate and high-rate periods B_k, A_{k+1}, B_{k+1}, A_{k+2}, ..., B_{i-2}, A_{i-1}, B_{i-1}, then all of these high-rate periods are not long enough compared to the preceding low-rate periods. More precisely, $\sum_{n=k}^{j-1} A_{n+1}$ is not long enough to remove the backlog accumulated in the associated fluid queue (cf. Section 7.3.2) during a period of length $\sum_{n=k}^{j-1} B_n$, for all $k < j \leq i-1$. At time s_i, define the accumulated low-rate and high-rate period lengths during the unstable interval length $[t_k, s_i)$ as

$$T_{\text{low}}(t_k, s_i) = \sum_{n=k}^{i-1} B_n, \text{ and } T_{\text{high}}(t_k, s_i) = \sum_{n=k}^{i-1} A_{n+1},$$

with $T_{\text{high}}(t_k, s_i) + T_{\text{low}}(t_k, s_i) = s_i - t_k$. We now investigate under which conditions the system becomes "effectively stable" during the i-th high-period A_i. We use $R(t_k, s_i)$

to denote the "recovery" time that is needed (after time s_i) to stabilize the queue, i.e., to reduce the effective load below 1. Equivalently, cf. Section 7.3.2, $R(t_k, s_i)$ is the time to drain the queue starting at s_i. Clearly, if $R(t_k, s_i) < A_i$, then the effective load drops below 1 during the i-th high-period, otherwise the queue remains effectively unstable throughout the i-th high-period. If $R(t_k, s_i) < A_i$, it must be that

$$\begin{aligned}&\lambda \left[T_{\text{high}}(t_k, s_i) + T_{\text{low}}(t_k, s_i)\right] + \lambda R(t_k, s_i)\\ &= \left[a \cdot T_{\text{high}}(t_k, s_i) + b \cdot T_{\text{low}}(t_k, s_i)\right] + a \cdot R(t_k, s_i).\end{aligned}$$

This gives

$$R(t_k, s_i) = \frac{\lambda - b}{a - \lambda} T_{\text{low}}(t_k, s_i) - T_{\text{high}}(t_k, s_i) \geq 0.$$

The following proposition summarizes the above.

Proposition 7.6. *Given that the queue was effectively unstable during the period $[t_k, s_i)$, i.e., during the periods B_k, A_{k+1}, B_{k+1}, ..., A_{i-1}, B_{i-1}, the queue stabilizes during the i-th high-period, if and only if*

$$\frac{\lambda - b}{a - \lambda} \sum_{j=k}^{i-1} B_j < \sum_{j=k}^{i-1} A_{j+1}. \tag{7.23}$$

Note that the term $(\lambda - b)$ in (7.23) is the growth rate of the fluid queue during low-rate periods, and $(a - \lambda)$ is the (potential) decrease rate during high-rate periods.

The number of high-rate periods needed for recovery

In this subsection, we determine the distribution of the random variable N, which is defined as the number of high-rate periods that is needed for recovery (re-stabilizing the system). Without loss of generality we take $k = 1$, i.e., t_1 is the most recent moment when the system became unstable. If $\{N = n\}$, for $n \geq 1$, then each of the first $n - 1$ high-periods are not long enough to stabilize the queue.

In general, the random variable N can be written as the *first ladder epoch* in a random walk (e.g., see [35]), i.e.,

$$N = \inf \left\{ n \geq 1 \mid S_n > 0 \right\}, \tag{7.24}$$

where $S_n = \sum_i^n V_i$ is the random walk with $V_i = A_{i+1} - cB_i$ and $c = \frac{\lambda - b}{a - \lambda}$ and $S_0 = 0$. Due to our assumptions, V_i are independently and identically distributed for $i \geq 1$. We get that

$$\begin{aligned}\{N = n\} &= \bigcap_{j=1}^{n-1} \{S_j \leq 0\} \cap \{S_n > 0\}\\ &= \bigcap_{j=1}^{n-1} \left\{ c \sum_{i=1}^{j} B_i \geq \sum_{i=1}^{j} A_{i+1} \right\} \cap \left\{ c \sum_{i=1}^{n} B_i < \sum_{i=1}^{n} A_{i+1} \right\}.\end{aligned}$$

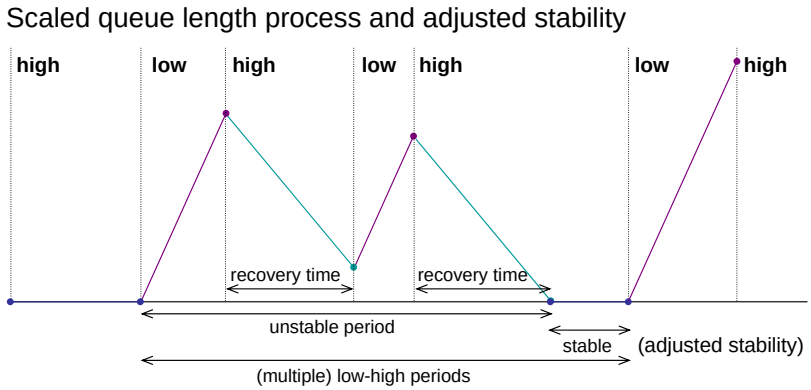

Figure 7.4: *Scaled queue length process and recovery periods.*

Note that indeed $\sum_{n=1}^{\infty} \mathbb{P}\left(N = n\right) = 1$, since $\mathbb{P}(N > n) = \mathbb{P}\left(\bigcap_{j=1}^{n} \{S_j \leq 0\}\right) \to 0$ as $n \to \infty$ whenever $\mathbb{E}V_i > 0$ (usual stability condition).

Adjusted stability

In Figure 7.4 we have depicted the workload process for $\eta \to \infty$ after linear scaling: $\lim_{\eta\to\infty} \frac{1}{\eta} Q^{\eta}(t)$. This model coincides with the associated fluid queue: compare with the discussion of the dynamics of the effective load in Figure 7.3. During the first low-rate period the scaled queue increases linearly. However, the length of the second high-rate period is not sufficiently long to remove the backlog built up during the preceding low-rate period. The third high-rate period is long enough to recover from the excess load in the previous low-rate periods.

Denote by π_{low} and π_{high} the fractions of time that the system serves at *low* and *high* service rate, i.e.,

$$\pi_{\text{low}} = \frac{\mathbb{E}B}{\mathbb{E}A + \mathbb{E}B} \text{ and } \pi_{\text{high}} = 1 - \pi_{\text{low}}.$$

We further define π_{stable} and π_{unstable}, as the fraction of time that the system is effectively stable and unstable, respectively. In addition, let π_{recovery} be the fraction of time that the system is in a recovery period:

$$\pi_{\text{unstable}} = \pi_{\text{low}} + \pi_{\text{recovery}},$$
$$\pi_{\text{stable}} = \pi_{\text{high}} - \pi_{\text{recovery}}.$$

We can interpret π_{stable} as a measure for reduced stability. Intuitively, we can argue that the system is perfectly stable if $\pi_{\text{unstable}} = 0$ (i.e., if there are no overload periods). Part

of the instability is directly due to periods with a positive drift, i.e., π_{low}, which would be a first measure for instability. From a practical perspective, however, the system does not show steady-state behavior during recovery periods. We next determine π_{recovery}.

Let $S_N > 0$ be the *first ladder height* of the random walk starting in $S_0 = 0$, then S_N satisfies

$$S_N = \sum_{i=1}^{N}(A_{i+1} - cB_i) \leq \sum_{i=1}^{N} A_{i+1},$$

where N is the first ladder epoch, cf. (7.24). Observe that S_N is exactly the time length that is stable, within the total period $\sum_{i=1}^{N}(A_{i+1} + B_i)$. Conditioned on $\{N = n\}$, we define $\pi_{\text{stable}}(n)$ as the fraction of stable-time during n consecutive low/high-periods. Hence, we have

$$\begin{aligned}\pi_{\text{stable}} &= \sum_{n=1}^{\infty} \pi_{\text{stable}}(n)\mathbb{P}(N=n) \equiv \sum_{n=1}^{\infty}\left(\frac{\mathbb{E}S_n}{\mathbb{E}\sum_{i=1}^{n}(A_{i+1}+B_i)}\right)\mathbb{P}(N=n)\\ &= \sum_{n=1}^{\infty}\left(\frac{\mathbb{E}\left[\sum_{i=1}^{n}(A_{i+1}-cB_i)\right]}{\mathbb{E}\left[\sum_{i=1}^{n}(A_{i+1}+B_i)\right]}\right)\mathbb{P}(N=n) = \frac{\mathbb{E}A - \frac{\lambda-b}{a-\lambda}\mathbb{E}B}{\mathbb{E}A+\mathbb{E}B},\end{aligned}$$

which gives

$$\pi_{\text{recovery}} = \frac{\lambda - b}{a - \lambda}\pi_{\text{low}}.$$

Note that in the special case that $\lambda = b < a$, i.e., $\rho(t) = 1$ during low-rate periods, and $\rho(t) < 1$ during high-rate periods we have that $\mathbb{P}(N = 1) = 1$. Then, the system becomes stable instantaneously at the beginning of each high-rate period (i.e., recovery time is zero), so that $\pi_{\text{recovery}} = 0$, $\pi_{\text{stable}} = \pi_{\text{high}}$, and $\pi_{\text{unstable}} = \pi_{\text{low}}$.

Remark 7.7. *If the usual stability condition is not satisfied, i.e.,*

$$(\lambda - b)\,\mathbb{E}B \geq (a - \lambda)\,\mathbb{E}A,$$

then $\pi_{\text{recovery}} = \pi_{\text{high}}$ *and* $\pi_{\text{unstable}} = 1$.

7.5.2 Recovery time and adjusted stability for exponential case

In this section we determine the distribution of N when V_i is given by

$$V_i := A_{i+1} - \frac{\lambda - b}{a - \lambda}B_i,$$

with A_i and B_i having exponential distributions with means $1/\alpha$ and $1/\beta$, respectively. To simplify the formulas in this section, we set $c := \frac{\lambda-b}{a-\lambda} = 1$. Note that cB_i has an exponential distribution too. Hence, it suffices to determine the distribution of $V_i := A_{i+1} - B_i$ with $c = 1$, for $i \geq 1$. Note that V_i is an independent and identically distributed sequence of random variables, taking values on $(-\infty, \infty)$. The distribution of N in the exponential case is given by the following proposition.

Proposition 7.8. *Let* $p := \frac{(a-\lambda)\mathbb{E}A}{(a-\lambda)\mathbb{E}A+(\lambda-b)\mathbb{E}B}$. *The distribution of the number of high-periods needed for recovery (re-stabilizing the system) is given by*

$$\mathbb{P}(N=n) = C_{n-1}p^n q^{n-1}, \quad \textit{for } n \geq 1,$$

where

$$C_n = \frac{1}{n+1}\binom{2n}{n} = \frac{(2n)!}{n!(n+1)!}$$

are Catalan numbers. The pgf $P_N(z) = \mathbb{E}z^N$ *is given by*

$$P_N(z) = \sum_{n=1}^{\infty} z^n \mathbb{P}(N=n) = \frac{1-\sqrt{1-4pqz}}{2q}, \quad \textit{for } |z| \leq 1.$$

Proof. See Section 7.7. ∎

In particular, the number of high-rate periods needed for recovery is finite with probability

$$P_N(1) = \frac{2p}{1+|2p-1|} = \frac{p \wedge q}{q},$$

where $p \wedge q = \min\{p, q\}$. Indeed, if $p \geq \frac{1}{2}$ then N is finite with probability 1. However, if $p = \frac{1}{2}$ then we have $\mathbb{E}N = \infty$ (see Proposition 7.9; and also see the relation with the symmetric Bernoulli walk [40]). The next proposition summarizes the mean and variance of N.

Proposition 7.9. *The expected number of high-rate periods needed for recovery is given by*

$$\mathbb{E}N = \frac{\mathbb{E}A}{\mathbb{E}A - \frac{\lambda-b}{a-\lambda}\mathbb{E}B}, \quad \textit{if } \mathbb{E}A > \frac{\lambda-b}{a-\lambda}\mathbb{E}B,$$

otherwise, if $\mathbb{E}A \leq \frac{\lambda-b}{a-\lambda}\mathbb{E}B$, *then* $\mathbb{E}N = \infty$. *In addition, the variance is given by*

$$\mathbb{V}arN = \frac{pq}{(p-q)^3} = \frac{\frac{\lambda-b}{a-\lambda}\mathbb{E}A\mathbb{E}B\left(\mathbb{E}A + \frac{\lambda-b}{a-\lambda}\mathbb{E}B\right)}{\left(\mathbb{E}A - \frac{\lambda-b}{a-\lambda}\mathbb{E}B\right)^3}.$$

Proof. By induction on n it follows that:

$$\frac{d^n}{dz^n}P_N(z) = n!C_{n-1}\frac{p^n q^{n-1}}{(1-4pqz)^{(2n-1)/2}}.$$

Then, use the fact that $\frac{d}{dz}P_N(z)\big|_{z=1} = \mathbb{E}N$ and $\frac{d^2}{dz^2}P_N(z)\Big|_{z=1} = \mathbb{E}N(N-1)$. ∎

7.5.3 Scaling of the queue length for the high-low model

We now extend the analysis to the high-low model. Here, we focus on the case where the A_i and B_i have exponential distributions with means $1/\alpha$ and $1/\beta$, respectively. The stationary distribution of Q^η is then known explicitly [77]:

$$\mathbb{P}(Q = i; \mu = j) = c_j p^i + d_j q^i, \tag{7.25}$$

for $j \in \{a, b\}$, where c_j and d_j are such that p and q are the two roots within the unit disc of the following equations

$$\begin{aligned} c_a p(\lambda + a + \alpha) &= c_a p^2 a + c_a \lambda + c_b p\beta, \\ c_b p(\lambda + b + \beta) &= c_b p^2 b + c_b \lambda + c_a p\alpha, \end{aligned}$$

and

$$\begin{aligned} d_a q(\lambda + a + \alpha) &= d_a q^2 a + d_a \lambda + d_b q\beta, \\ d_b q(\lambda + b + \beta) &= d_b q^2 b + d_b \lambda + d_a q\alpha. \end{aligned}$$

The precise form of these coefficients is not essential (they are characterized through the solution to a cubic equation). We are primarily interested in the queue length as $\eta \to \infty$. With standard algebra it follows that p and q tend to λ/a and 1 respectively. (Although p, q, c_j and d_j depend on η when applying uniform acceleration, we will not reflect this in the notation.) The corresponding constants then follow from the equations above and we get (after convenient rewriting):

$$\lim_{\eta\to\infty} \mathbb{P}(Q^\eta > x \mid \mu = a) = \frac{\lambda(\alpha+\beta) - b\alpha}{a\beta} + \left(1 - \frac{\lambda(\alpha+\beta) - b\alpha}{a\beta}\right)\left(\frac{\lambda}{a}\right)^x. \tag{7.26}$$

Naturally, we find $\lim_{\eta\to\infty} \mathbb{P}(Q^\eta > x \mid \mu = b) = 1$ for all x. The term $\frac{\lambda(\alpha+\beta)-b\alpha}{a\beta}$ can be interpreted as the fraction of high-rate service time that is needed for recovery. It can be shown that this indeed coincides with the probability that the associated fluid queue is non-empty.

If we scale the queue length with the parameter η, it can be shown that

$$\lim_{\eta\to\infty} q^{\frac{1}{\eta}} = \frac{\beta}{b-\lambda} - \frac{\alpha}{\lambda - a} =: \delta.$$

Substituting this into the distribution for Q we get

$$\lim_{\eta\to\infty} \mathbb{P}(\frac{1}{\eta}Q^\eta > x \mid \mu = a) = \frac{\lambda(\alpha+\beta) - b\alpha}{a\beta}\delta^x,$$

and

$$\lim_{\eta\to\infty} \mathbb{P}(\frac{1}{\eta}Q^\eta > x \mid \mu = b) = \delta^x,$$

as was to be expected.

Illustration

In Figures 7.5-7.7 we have depicted three different realizations of the scaled queue length process $\frac{1}{\eta}Q^{\eta}(t)$, for $\eta = 1$, $\eta = 10$ and $\eta = 100$, respectively. The realization for the high and low period lengths are the same in Figures 7.5-7.7 for comparison purposes. The service rate starts off in the higher value $a = 2$ and the process shows stationary behavior, since the instantaneous load $\rho(t)$ is less than 1 during (the first) high rate period(s). As soon as the service rate switches to the lower rate $b = \frac{1}{2}$, the queue starts building up (along a positive trend). Whenever the service rate switches back to the higher service rate, the queue starts decreasing along a negative trend. The fluctuations around the linear trend get smaller as η grows. From these figures, stationary behavior during high rate periods is observed when the queue has decreased "sufficiently". Ultimately, in the quasi-stationary limit $\eta \to \infty$, stationary behavior is observed when the negative drift hits the horizontal axis, which is also the time epoch where the buffer content in the associated fluid queue becomes empty. In the figures we also observe that the second high rate period is too short to recover from the excess load of the first low rate period. In contrast, the third high rate period is sufficiently long to recover from the excess load from the first two low rate periods. (We may say that the queue becomes stable again during the third high rate period.)

7.6 Conclusion and extensions

In this chapter we studied a queue with a Markov modulated service rate in which the system can be overloaded in some of the service states. With uniform stability, the quasi-stationary approximation can be seen as the leading term from a regular perturbation analysis of the system. Without uniform stability, the perturbation is singular, and our aim was to adapt the quasi-stationary approach to allow for temporary instability. The time needed to recover from an (excessively long) overload period is shown to be associated with the workload process in a fluid queue driven by the same Markov process as the original queue. More precisely, the recovery periods in the original queue correspond to a positive and decreasing buffer of the associated fluid queue. When the associated fluid queue is empty, the queue is 'effectively stable' in the original system. For the system with service rate alternating between high and low rate, we have discussed notions of effective load and adjusted stability that allow us to characterize the fraction of the high service rate periods during which the queue is recovering from instability due to overload periods.

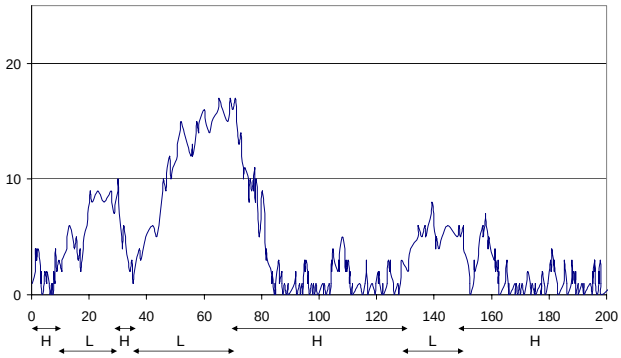

Figure 7.5: *A sample path of the scaled queue length process* $\frac{1}{\eta}Q^{\eta}(t)$*, for* $\eta = 1$*, in the high-low model with* $\lambda = 1$*,* $a = 2$*,* $b = \frac{1}{2}$

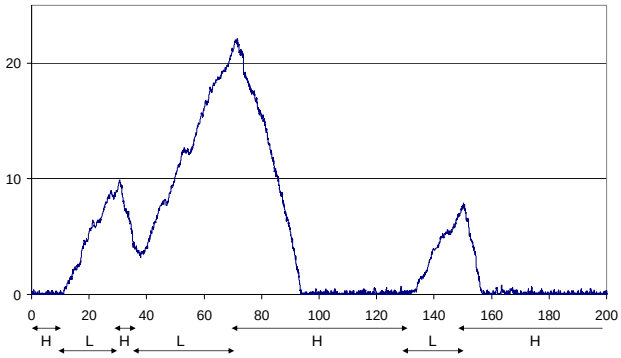

Figure 7.6: *A sample path of the scaled queue length process* $\frac{1}{\eta}Q^{\eta}(t)$*, for* $\eta = 10$*, in the high-low model with* $\lambda = 1$*,* $a = 2$*,* $b = \frac{1}{2}$

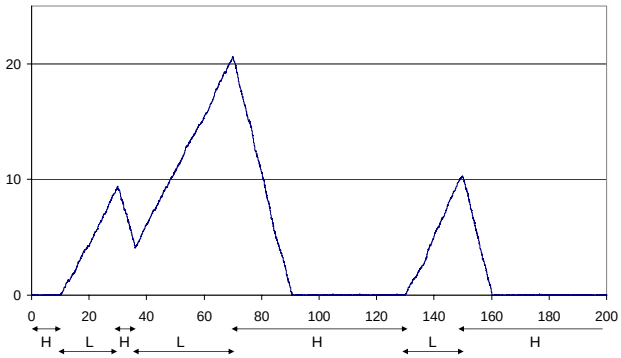

Figure 7.7: *A sample path of the scaled queue length process* $\frac{1}{\eta}Q^{\eta}(t)$*, for* $\eta = 100$*, in the high-low model with* $\lambda = 1$*,* $a = 2$*,* $b = \frac{1}{2}$

7.7 Proof of Proposition 7.8

Proof. In general first note that $\mathbb{P}(N = n)$ is given by the following multiple integral:

$$\int_{x_1=-\infty}^{x_1=0} \int_{x_2=-\infty}^{x_2=-x_1} \cdots \int_{x_{n-1}=-\infty}^{x_{n-1}=-\sum_{j=1}^{n-2} x_j} \int_{x_n=-\sum_{j=1}^{n-1} x_j}^{x_n=\infty} dF_{V_n}(x_n) dF_{V_{n-1}}(x_{n-1}) \cdots dF_{V_2}(x_2) dF_{V_1}(x_1),$$

where $\sum_{j=1}^{n-k} x_j \leq 0$ for all $k = 1, ..., n-1$, and with $F_{V_i}(x_i) = \mathbb{P}(V_i \leq x_i)$. It is also (recursively) defined by

$$\mathbb{P}(N = n) = \int_{x_1=-\infty}^{x_1=0} M_{n-1} dF_{V_1}(x_1), \tag{7.27}$$

where

$$M_k = \int_{x_{n-k+1}=-\infty}^{x_{n-k+1}=-\sum_{j=1}^{n-k} x_j} M_{k-1} dF_{V_{n-k+1}}(x_{n-k+1}), \tag{7.28}$$

for $k \geq 2$, and with

$$M_1 = \int_{x_n=-\sum_{j=1}^{n-1} x_j}^{x_n=\infty} dF_{V_n}(x_n). \tag{7.29}$$

In the exponential case it is not difficult to obtain the distribution function of V_i:

$$F_{V_i}(x) = 1 - \frac{\beta/c}{\alpha + \beta/c} e^{-\alpha x}, \quad \text{for } x \geq 0,$$
$$F_{V_i}(x) = \frac{\alpha}{\alpha + \beta/c} e^{\frac{\beta}{c} x}, \quad \text{for } x \leq 0.$$

Note that it is sufficient to take $c := \frac{\lambda - b}{a - \lambda} = 1$. Then by using induction on k, it follows that (7.28) has the following form:

$$M_k = \sum_{i=1}^{k} \frac{\delta_{i,k}}{(i-1)!} \frac{\alpha^{k-1} \beta^k}{(\alpha + \beta)^{2k-i}} e^{\alpha(\sum_{j=1}^{n-k} x_j)} (-\sum_{j=1}^{n-k} x_j)^{i-1},$$

for $2 \leq k \leq n-1$, and for some constants $\delta_{i,k} \in \mathbb{N}$, and which also holds for $k = 1$:

$$M_1 = \int_{x_n=-\sum_{j=1}^{n-1} x_j}^{x_n=\infty} dF_{V_n}(x_n) = \frac{\beta}{\alpha + \beta} e^{\alpha(\sum_{j=1}^{n-1} x_j)}.$$

Hence, the probability (7.27) is given by

$$\int_{x_1=-\infty}^{x_1=0} \left(\sum_{i=1}^{n-1} \delta_{i,n-1} \frac{\alpha^{n-2}\beta^{n-1}}{(\alpha+\beta)^{2n-2-i}} \frac{e^{\alpha x_1}(-x_1)^{i-1}}{(i-1)!} \right) \times \left(\frac{\alpha\beta}{\alpha+\beta} \right) e^{\beta x_1} dx_1 \quad (7.30)$$

$$= \frac{\alpha^{n-1}\beta^n}{(\alpha+\beta)^{2n-1}} \sum_{i=1}^{n-1} \delta_{i,n-1} \times \int_{x_1=-\infty}^{x_1=0} \left(\frac{(\alpha+\beta)e^{(\alpha+\beta)x_1}(-x_1(\alpha+\beta))^{i-1}}{(i-1)!} \right) dx_1 \quad (7.31)$$

$$= \frac{\alpha^{n-1}\beta^n}{(\alpha+\beta)^{2n-1}} \sum_{i=1}^{n-1} \delta_{i,n-1} = a_n p^n q^{n-1}, \quad (7.32)$$

where we have used that the integral in (7.31) equals 1, and where $a_n := \sum_{i=1}^{n-1} \delta_{i,n-1}$, $p = \frac{\beta}{\alpha+\beta}$, $q = \frac{\alpha}{\alpha+\beta}$. Note that if $c > 0$, then β should be replaced by β/c. The form of the probability in (7.32), implies that $a_n = C_{n-1}$ must be Catalan numbers (see the relation with Bernoulli walk; e.g., see [40]). Finally, by using the well-known generating function for the Catalan numbers $\sum_{n=1}^{\infty} C_{n-1}x^{n-1} = \frac{1-\sqrt{1-4x}}{2x}$, we readily obtain, for $|z| \leq 1$,

$$P_N(z) = \frac{1}{q} \sum_{n=1}^{\infty} C_{n-1} (zpq)^n = pz \sum_{n=1}^{\infty} C_{n-1} (zpq)^{n-1} = \frac{1-\sqrt{1-4pqz}}{2q},$$

which completes the proof. ∎

Alternatively, in case of exponential high-rate and low-rate periods, the proof can be given by interpreting the number of high-rate periods N needed for recovery as the first entrance time I_1 in the Bernoulli random walk [40]). Writing $S_n = S_0 + Y_1 + ... + Y_n$ with $Y_1, Y_2, ...$ independent and identically distributed as:

$$\mathbb{P}(Y_1 = 1) = p, \ \mathbb{P}(Y_1 = -1) = q = 1 - p,$$

and with an initial state $S_0 = 0$, then

$$I_1 = \inf \{n \in \mathbb{N} \mid S_n = 1\}$$

is the first entrance time of the Bernoulli walk at level $S_n = 1$. Note that I_1 always takes odd integer values $(2n-1)$, $n \in \mathbb{N}$, and it is well-known that:

$$\mathbb{P}(I_1 = 2n-1) = \mathbb{P}(S_1 \leq 0; ..., S_{2n-2} \leq 0; S_{2n-1} = 1) = C_{n-1}p^n q^{n-1},$$

with $\mathbb{P}(I_1 < \infty) = \frac{p \wedge q}{q}$, and $\mathbb{E}I_1 = \frac{1}{p-q}$ and $\mathbb{V}arI_1 = \frac{4pq}{(p-q)^3}$, if $p > \frac{1}{2}$. The Catalan number C_{n-1} is the number of paths of the random walk starting with $S_0 = 0$ and reaching the event I_1 in $(2n-1)$ steps. Finally, note that $2N - 1 \stackrel{d}{=} I_1$.

Chapter 8

Resource sharing and performance analysis of WLANs

In this chapter we give a flow-level performance analysis of Wireless LANs, in particular, for the IEEE 802.11E WLAN with QoS support. WLAN performance is largely determined by the maximum data rate at the physical layer and the MAC layer protocols (Medium Access Control) defined by the IEEE 802.11 standards [2, 3]. An extension of the most widely employed DCF protocol (Distributed Coordination Function) is the EDCA protocol (Enhanced Distributed Channel Access). Both the DCF and the EDCA protocols are random access schemes based on Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA). EDCA is aimed at providing QoS differentiation between various traffic classes, whereas DCF only supports best-effort services.

For best-effort WLANs (802.11B), several accurate analytical performance models have been developed (along the lines of Bianchi [18]) in order to study the system's saturated throughput as a function of the number of persistently active users. Foh and Zuckerman [41] and Litjens, Roijers, Van den Berg, Boucherie, and Fleuren [68] considered 802.11B with the practical situation of a dynamically varying number of active users due to the random initiation and completion of flow transfers. They obtain accurate approximations for the mean flow transfer time.

Analytical flow-level performance studies for WLANs with QoS support and a dynamically varying number of active users are not available. Performance studies of 802.11E WLANs are mainly based on simulation (e.g. [30, 49, 58, 72, 84, 100]). Relatively few papers present an analytical approach, generally considering a fixed number of persistent users; see e.g. Zhao, Guo, Zhang, and Zhu [110]. The simulation studies usually consider general scenarios (sometimes also capturing the impact of higher layer protocols like TCP), but without random user behavior.

In the present chapter, an analytical performance evaluation of 802.11E WLANs is given along the lines of the analysis in [68]. From the flow-level point-of-view, the 802.11E WLAN is considered as a queueing system with Poisson arrivals and *generalized* discriminatory processor-sharing (GDPS) service discipline with queue-dependent service capacity and queue-dependent weights. This queueing model reflects the EDCA MAC design principle of distributing the available queue-dependent transmission capacity among active traffic classes according to certain priority weights. In our modeling approach, the class weights and the system capacity depend on the number of active users and are obtained from a packet-level model that describes the MAC behavior of EDCA in detail in the situation with a fixed number of persistent users.

This chapter is organized as follows. Section 8.1 describes the DCF and the EDCA MAC protocol. Section 8.2 describes our analytical modeling approach, which is presented in more detail in Sections 8.3-8.5. We use the simple analytical and efficient decomposition method for approximating the mean file transfer times in *generalized* discriminatory PS models, see Chapter 5. The accuracy of our analytical approximations is validated by extensive simulation of WLAN systems in Section 8.6. Finally, the principal conclusions of our investigation as well as some topics for further research are outlined in Section 8.7.

8.1 IEEE 802.11B DCF and 802.11E EDCA

In the DCF mode for best-effort WLANs, when a station wants to transmit a packet, it first senses the channel to determine whether or not it is already in use by another station ("listen before talk"). If the channel is idle, and remains idle for a contiguous time period called DIFS (Distributed InterFrame Space), the station has to wait a random number of time slots before it is permitted to send the packet. This random back-off procedure is intended to reduce the probability of multiple stations sending at the same time resulting in a collision.

The BASIC access scheme is depicted in Figure 8.1. The back-off procedure draws a discrete random value for the back-off counter uniformly between 0 and $cw_r - 1$, where cw_r is the so-called contention window at the r-th re-attempt to send the packet. As long as the channel remains idle after a DIFS period, a station decrements its back-off counter by 1 for each time slot. When the back-off counter of a particular station reaches 0, the station transmits the packet. If the packet is received correctly, the destination responds by sending an acknowledgment (ACK) to the source. In case multiple packets are transmitted a collision occurs. If a station does not receive an ACK, it assumes that the packet was lost and it will retransmit the packet. The contention window cw_r is doubled for the first r^* re-attempts and a new back-off counter is drawn. The total number of re-attempts is limited to r_{max}. After a packet is successfully transmitted, the cw_r is reset

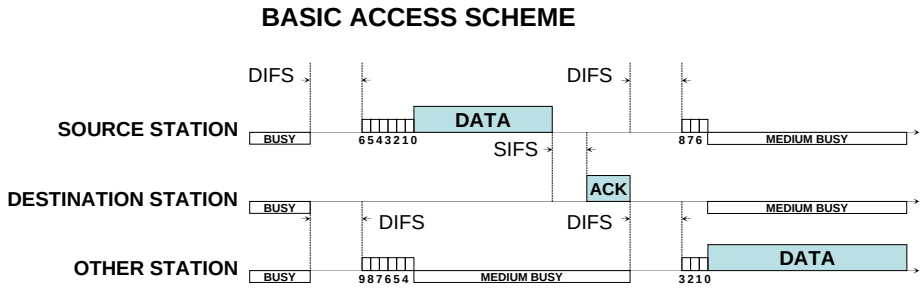

Figure 8.1: *BASIC access mode in the distributed coordination function*

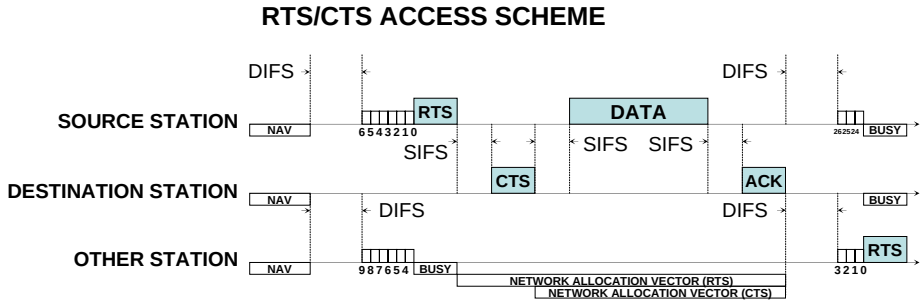

Figure 8.2: *RTS/CTS access mode in the distributed coordination function*

to its minimum contention window size cw_{min}. The DCF can also operate in RTS/CTS mode by first sending a small ReadyToSend (RTS) frame instead of the packet. When the source receives a small ClearToSend (CTS) from the destination, it transmits the packet; see Figure 8.2. RTS/CTS reduces the effect of collisions in busy systems (many active stations), but involves more overhead.

A major drawback of DCF is that it only supports best-effort services. A distributed access approach with QoS differentiation can be achieved with the Enhanced Distributed Channel Access (EDCA) – or Enhanced DCF – in 802.11E WLANs by using different parameter values for different traffic classes. The EDCA protocol defines several traffic classes, indexed by $i = 1, ..., K$. Similarly to best-effort DCF, when a user from traffic class i wants to transmit a data packet in EDCA protocol with BASIC access mode, it first senses the medium to determine whether or not the channel is already in use by another user. If the channel is idle, and remains idle for a contiguous time period called AIFS_i (AIFS value for class i) the user has to wait a random number of time slots before it is

permitted to send the packet. The discrete back-off counter is now uniformly sampled from $\{0, ..., cw_{r,i} - 1\}$, where $cw_{r,i}$ is the contention window size for a traffic class i user at the r-th re-attempt to send the packet. As long as the channel remains idle after an AIFS$_i$ period, a class i user decrements its back-off counter by 1 for each slot time.

In case of best-effort WLANs, the AIFS$_i$ values for all users are equal to the DIFS. Additional tunable parameter values are for example: minimum contention window size, TXOP$_{\text{limit}}$ (transmission opportunity limit), packet size, and the maximum contention window size. With a TXOP$_{\text{limit}}$ a user may send multiple packets as long as the last packet is completely transmitted before the TXOP$_{\text{limit}}$ time duration has expired. In general, QoS differentiation between various traffic classes is mainly achieved by contention window size and AIFS-based differentiation mechanisms.

8.2 Performance analysis and its modeling approach

We consider a single basic service set with users from K traffic classes contending for shared 802.11E WLAN radio access, and each traffic class has its own set of tunable parameter values (e.g. $cw_{\min,i}$ and AIFS$_i$). We follow an integrated packet/flow-level modeling approach similar to that used in [68]. The first stage of the modeling approach is an enhanced packet-level model that describes the MAC behavior of 802.11E QoS mechanisms in detail. The second stage is a flow-level model that describes the behavior of 802.11E, when the number of active users dynamically varies in time.

For the first stage, when $\mathbf{n} = (n_1, ..., n_K)$ is the number of persistent users in the system (with n_j fixed users of class j), the resulting output of the packet-level model is the expected aggregate throughput $R_i(\mathbf{n})$ for class i, and $R(\mathbf{n}) := \sum_{i=1}^{K} R_i(\mathbf{n})$ is the expected aggregate system throughput. From a single user's perspective, a class i user receives an expected throughput of $R_{\text{flow},i}(\mathbf{n}) = R_i(\mathbf{n})/n_i$, and for $n_i > 0, n_j > 0$, the relative received throughput between a single class i and class j user is defined as

$$\alpha_{ij}(\mathbf{n}) := \frac{R_{\text{flow},i}(\mathbf{n})}{R_{\text{flow},j}(\mathbf{n})} = \frac{R_i(\mathbf{n})/n_i}{R_j(\mathbf{n})/n_j}.$$

QoS differentiation in 802.11E WLANs is achieved by establishing that $\alpha_{ij}(\mathbf{n})$ generally differs from 1. The *egalitarian* processor-sharing queueing model, as accurately used in [68] for 802.11B file transfer time analysis, is not suitable for the extended 802.11E version with QoS support. However, the essential principle of distributing the queue-dependent bandwidth in a processor-sharing fashion still remains for 802.11E WLANs. In the latter case, the capacity is shared in a *discriminatory* fashion between the classes, as intended by the design of EDCA, and shared in an *egalitarian* fashion for users within the same class. On top of the discriminatory feature of 802.11E WLANs, the priority effect $\alpha_{ij}(\mathbf{n})$ is also dependent on the number of active users, as observed

from the packet-level analysis (and obviously $\alpha_{ij}(\mathbf{n})$ also depend on the type of QoS differentiation mechanisms).

Based on these observations, we propose a *generalized* discriminatory processor-sharing (GDPS) model with both queue-dependent service capacity and queue-dependent weights. An attractive feature of our modeling approach is that the general form of the flow-level model is independent of the packet-level model, in the sense that only the expected saturated class throughputs $R_i(\mathbf{n})$ from the packet-level model are required as input for the GDPS flow-level model. In particular, it is independent of the type of QoS mechanisms.

8.3 Packet-level: throughput analysis for persistent users

In this section, we first give our straightforward packet-level extension from [68], when only differentiation in the contention window sizes can be applied. Differentiation with $\text{TXOP}_{\text{limit}}$ and packet sizes are easily incorporated. AIFS-based extensions have been well studied in the literature. For example, Zhao, Guo, Zhang, and Zhu [110] proposed an extended Markov chain analysis which has been accurately validated. In principle, any accurate analytical packet-level model can be used as input of our non-persistent flow-level model.

In the second part of this section we briefly indicate some qualitative insights between the QoS mechanisms at the packet-level and the relative throughput measures $\alpha_{ij}(\mathbf{n})$. This subsection is included to place our *generalized* discriminatory processor-sharing model with queue-dependent weights $\alpha_{ij}(\mathbf{n})$ in the right setting. Finally, in the last part of this section, we present our flow-level modeling approach based on the GDPS model in significantly more detail.

The throughput analysis for 802.11B at the packet-level and with a fixed number of persistent users (as used in [68] and originally developed by Bianchi [18]) essentially remains the same for 802.11E if no AIFS-based differentiation is considered, since the Markov chain for the back-off counting process is described only from an isolated user point-of-view.

For 802.11E, we assume that all traffic classes i, $i = 1, ..., K$, have their own so-called packet error probability $\mathbf{P}_{e,i}$, independently of the other classes and independently of the number of collisions already involved. This is the same key assumption made in [18]. The influence of all other active users is captured by this packet error probability. Hence, the equilibrium distribution of the embedded jump chain for a class i user is similar to [68] and given by

$$\pi_i(r,b) = \frac{cw_{r,i} - b}{cw_{r,i}} \cdot \frac{2\left(1 - \mathbf{P}_{e,i}\right)\mathbf{P}_{e,i}^{r}}{\left(\mathbf{1} - \mathbf{P}_{e,i}^{r_{\max,i}+1}\right) + \left(1 - \mathbf{P}_{e,i}\right)\sum_{k=0}^{r_{\max,i}} cw_{k,i}\mathbf{P}_{e,i}^{k}},$$

where (r, b) denotes the back-off state ($0 \leq r \leq r_{\max,i}$ and $0 \leq b \leq cw_{r,i} - 1$), $r_{\max,i}$ denotes the maximum number of retries, and $cw_{r,i}$ is the contention window size of a class i user at the r-th re-attempt. Note that only the QoS parameters $cw_{\min,i}$ and $cw_{\max,i}$ appear in the equilibrium distribution of the Markov chain ($cw_{\min,i} := cw_{0,i} - 1$ and $cw_{\max,i} := cw_{r,i} - 1$ with $r = r_{\max,i}$). The packet error probability $\mathbf{P}_{e,i}$ is readily expressed by

$$\mathbf{P}_{e,i} = 1 - \left(1 - \mathbf{P}^*_{tr,i}\right)^{n_i - 1} \prod_{k=1, k\neq i}^{K} \left(1 - \mathbf{P}^*_{tr,k}\right)^{n_k}, \text{ for all } i = 1, ..., K,$$

where $\mathbf{P}^*_{tr,i}$ is the packet transfer probability (successful or not) for a class i user, i.e., $\mathbf{P}^*_{tr,i} = \sum_{r=0}^{r_{\max,i}} \pi_i(r, 0)$. It can be shown that a unique vector $\left(\mathbf{P}^*_{tr,i}, \mathbf{P}_{e,i}\right)_{i=1}^{K}$ exists.

Aggregate class throughputs

The expected aggregate class throughput $R_i(\mathbf{n}) \equiv R_i(n_1, ..., n_K)$ is given by:

$$R_i(\mathbf{n}) = \frac{\mathbf{P}_{suc,i} \cdot \mathbb{E}P_i}{\mathbf{P}_{idle} \cdot \tau + \sum_{i=1}^{K} \mathbf{P}_{suc,i} \cdot \mathbf{T}_{suc,i} + \mathbf{P}_{col} \cdot \mathbf{T}_{col}},$$

cf. [18, 68], where

$$\mathbf{P}_{idle} = \prod_{j=1}^{K} \left(1 - \mathbf{P}^*_{tr,j}\right)^{n_j}$$

is the probability that the channel is idle at a randomly selected time slot, and

$$\mathbf{P}_{suc,i} = n_i \mathbf{P}^*_{tr,i} \left(1 - \mathbf{P}^*_{tr,i}\right)^{n_i - 1} \prod_{j\neq i}^{K} \left(1 - \mathbf{P}^*_{tr,j}\right)^{n_j}$$

is the probability that exactly one class i user transmits a packet at a random time slot; Furthermore, the collision probability is given by

$$\mathbf{P}_{col} = 1 - \mathbf{P}_{idle} - \sum_{i=1}^{K} \mathbf{P}_{suc,i},$$

and τ is the time slot duration and the inter-event times $\mathbf{T}_{suc,i}$ (successful) and $\mathbf{T}_{col}$ (collision) are given by

$$\begin{aligned} \mathbf{T}^{BASIC}_{suc,i} &= PHY + MAC + r^{-1}_{WLAN}\mathbb{E}P_i + \delta + SIFS + ACK + \delta + DIFS, \\ \mathbf{T}^{BASIC}_{col} &= PHY + MAC + r^{-1}_{WLAN}\mathbb{E}\overline{P} + \delta + DIFS, \end{aligned}$$

under the BASIC access mode, and where r_{WLAN} is the channel rate, PHY is the physical header (plus preamble), SIFS is the Short InterFrame Space time, δ is the propagation

delay between sender and receiver (in seconds), $\mathbb{E}P_i$ is the expected net payload size for traffic class i (in kbits) and $\mathbb{E}\overline{P}$ is the expected net payload size of the largest packet involved in a collision. The MAC header and ACK size are converted to seconds. Differentiation with $\text{TXOP}_{\text{limit}}$ and packet size is easily incorporated in the inter-event times $\mathbf{T}_{suc,i}$ ($\mathbf{T}_{col}$) through the expected payload sizes.

The packet-level analysis for the RTS/CTS (ReadyToSend/ClearToSend) access mode is essentially the same; only the inter-event times $\mathbf{T}_{suc,i}^{RTS/CTS}$ and $\mathbf{T}_{col}^{RTS/CTS}$ are slightly differently computed:

$$\begin{aligned}\mathbf{T}_{suc,i}^{RTS/CTS} = \quad & RTS + 4\delta + 3SIFS + CTS + PHY + MAC + r_{WLAN}^{-1}\mathbb{E}P_i \\ & + ACK + DIFS, \\ \mathbf{T}_{col}^{RTS/CTS} = \quad & RTS + \delta + DIFS,\end{aligned}$$

also see Figure 8.2. An advantage of the RTS/CTS mode is that the time wasted by a collision is smaller as the RTS frame is significantly smaller than a data packet. A drawback is that more overhead is involved than in BASIC access.

8.4 Packet-level: qualitative insights

The packet-level model yields throughputs $R_i(\mathbf{n})$, that discriminate among active traffic classes and depend on the number of active users $\mathbf{n}$. Clearly, when $cw_{\min}$ differentiation is applied with $cw_{\min,j} < cw_{\min,i}$ and the other QoS parameter values for class i and j are kept equal, then it must hold that $\alpha_{ij}(\mathbf{n}) < 1$. The same differentiating effect $\alpha_{ij}(\mathbf{n}) < 1$ is also achieved when only $\text{AIFS}_j < \text{AIFS}_i$ is applied. However, the impact of the QoS differentiation parameters is generally different, and moreover, also heavily dependent on the number of active users $\mathbf{n}$ in the system. We illustrate this with two examples, as observed from the packet-level analysis.

8.4.1 Example 1

If $\text{AIFS}_j < \text{AIFS}_i$ is applied, then the high priority class j users always start and resume their back-off counting procedure sooner than the low priority class i users. When many high priority users are active in the system, a so-called 'starvation effect' can occur, i.e., $\alpha_{ij}(\mathbf{n}) \approx 0$. To this end, observe that after the end of the AIFS_j time duration, the back-off counters for all class i users always remain unchanged for at least a time duration of $\text{AIFS}_i - \text{AIFS}_j$, whereas all class j users resume to decrement their back-off counters after the end of the AIFS_j time. In fact, a class j user can gain access to the medium before the end of the AIFS_i time and hence leaving all back-off counters for class i users unchanged, while all class j users have decremented their back-off counters. Any successful packet transfer or any packet collision is beneficial for the high priority

class j (in terms of the frequency of access to the medium for class j users). In the same scenario of AIFS$_j <$ AIFS$_i$, but with few active users in the system, then this ratio $\alpha_{ij}(\mathbf{n})$ is usually much larger than zero but obviously still less than 1.

8.4.2 Example 2

The relative weight $\alpha_{ij}(\mathbf{n})$ can even be both less and greater than 1, depending on $\mathbf{n}$. For example if $cw_{\min,i} < cw_{\min,j}$ in combination with AIFS$_i >$ AIFS$_j$ is applied, then $cw_{\min}$ dominates the differentiation effect when the number of active users is small ($\alpha_{ij}(\mathbf{n}) > 1$), and the smaller AIFS$_j$ value dominates the differentiation effect when the number of active users is large ($\alpha_{ij}(\mathbf{n}) < 1$).

8.5 Flow-level: throughput analysis for persistent users

From the flow-level point-of-view, the number of active class i users in the system is not fixed, but varies dynamically in time due to initiation of file transfers and file transfer completions. We shall let N_i denote the (steady-state) random variable of the number of active class i users in the 802.11E WLAN network. Since the packet-level model yields queue-dependent capacity $R(\mathbf{n})$ and queue-dependent weights $\alpha_{ij}(\mathbf{n})$, the 802.11E WLAN network can be considered as a service center with a *generalized* DPS service discipline.

When no service differentiation is applied under 802.11E, the packet-level model will result in $\alpha_{ij}(\mathbf{n}) = 1$ (for all i, j and for all $\mathbf{n}$ such that $n_i > 0, n_j > 0$) and $R(\mathbf{n})$ only depends on the total number of active users $n_1 + ... + n_K$. In the latter case, the *generalized* DPS model is equivalent to Cohen's GPS model [32], which is applied to the flow-level performance analysis for best-effort WLANs [68].

We assume that the traffic classes generate data flows according to independent Poisson processes with rates $\lambda_i, i = 1, ..., K$, and a class i user requests the download transfer of a data file whose size is generally distributed with mean $\mathbb{E}X_i$ (in kbits). Each file is segmented into packets of a given size (with a final packet containing the file's remainder) which are processed at the WLAN's MAC layer. The offered data load of class i is denoted by $\rho_i \equiv \lambda_i \mathbb{E}X_i / r_{WLAN}$, and the total offered load by $\rho := \sum_{i=1}^{K} \rho_i$. To ensure stability and provide *statistically guaranteed* QoS, we limit the number of contending data flows by $n_{\max,i}$ for each class separately.

Analytical expressions for (G)DPS models are generally not available in tractable form. Therefore, we use an analytical approximation method from Chapter 5. In the remainder of this chapter, we keep the demonstration of the solution technique simple, by considering the example of two traffic classes, i.e., $K = 2$. The approximation can be generalized for an arbitrary number of traffic classes.

We apply the decomposition technique as follows. For specific traffic loads ρ_i and given the class capacities $R_1(n_1, n_2)$ and $R_2(n_1, n_2)$, we first approximate the equilibrium distribution of N_1 and N_2. If class 2 users are persistent (permanent) in the system, then the distribution of N_1 (conditional on n_2 fixed class 2 users) is easily computed with closed-form formulas (cf. Cohen's GPS model [32]), i.e., the conditional probabilities defined by

$$a(n_1, n_2) = \mathbb{P}(N_1 = n_1 \mid n_2 \text{ persistent class 2 users})$$

are readily computed as

$$a(0, n_2) = \left(1 + \sum_{n_1=1}^{n_{\max,1}} \left(\rho_1^{n_1} \prod_{k=1}^{n_1} \frac{r_{WLAN}}{R_1(k, n_2)}\right)\right)^{-1} =: G_1(n_2)^{-1}, \tag{8.1}$$

$$a(n_1, n_2) = \left(\rho_1^{n_1} \prod_{k=1}^{n_1} \frac{r_{WLAN}}{R_1(k, n_2)}\right) / G_1(n_2), \text{ for } n_1 = 1, ..., n_{\max,1}, \tag{8.2}$$

with $\sum_{n_1=0}^{n_{\max,1}} a(n_1, n_2) = 1$ for all $n_2 = 0, 1, ..., n_{\max,2}$. Analogously, the conditional probabilities defined by $b(n_2, n_1) = \mathbb{P}(N_2 = n_2 \mid n_1$ persistent class 1 users) are given by

$$b(0, n_1) = \left(1 + \sum_{n_2=1}^{n_{\max,2}} \left(\rho_2^{n_2} \prod_{k=1}^{n_2} \frac{r_{WLAN}}{R_2(n_1, k)}\right)\right)^{-1} =: G_2(n_1)^{-1}, \tag{8.3}$$

$$b(n_2, n_1) = \left(\rho_2^{n_2} \prod_{k=1}^{n_2} \frac{r_{WLAN}}{R_2(n_1, k)}\right) / G_2(n_1), \text{ for } n_2 = 1, ..., n_{\max,2}, \tag{8.4}$$

with $\sum_{n_2=0}^{n_{\max,2}} b(n_2, n_1) = 1$ for all $n_1 = 0, 1, ..., n_{\max,1}$.

The approximation of the unconditional and marginal distribution $\mathbb{P}(N_1 = n_1)$ is obtained by solving the linear system

$$\mathbb{P}(N_1 = i) = \sum_{k=0}^{n_{\max,2}} a(i, k)\mathbb{P}(N_2 = k), \quad \text{for } i = 0, 1, ..., n_{\max,1}, \tag{8.5}$$

$$\mathbb{P}(N_2 = j) = \sum_{k=0}^{n_{\max,1}} b(j, k)\mathbb{P}(N_1 = k), \quad \text{for } j = 0, 1, ..., n_{\max,2}. \tag{8.6}$$

This system can be solved efficiently and it can be shown that the solution is unique up to a multiplicative constant. A sufficient condition is that $R_i(\mathbf{n}) > 0$ whenever $n_i > 0$ (also see Assumption 3.1 in Chapter 3 and Chapter 5)

From the approximations for $\mathbb{P}(N_1 = n_1)$ and $\mathbb{P}(N_2 = n_2)$, the expectations $\mathbb{E}N_1$ and $\mathbb{E}N_2$ are easily computed, and by applying the well-known Little's formula, we

approximate the mean file transfer time $\mathbb{E}T_i$ for class i with $\mathbb{E}\widehat{T}_i$, given by the simple formula

$$\mathbb{E}\widehat{T}_i = \frac{\mathbb{E}N_i}{\lambda_i(1-\mathbb{P}(N_i = n_{\max,i}))}, \quad i = 1, 2. \tag{8.7}$$

8.6 Numerical results

In this section we present numerical results obtained from our analysis and compare them with WLAN simulation results. A detailed representation of the EDCA MAC layer has been implemented in a C/C++ program. Sufficient independent replications were run to obtain 95% confidence intervals with a relative precision no worse than 5%. The default WLAN parameter settings (without QoS differentiation) are given in Table 1, where the default DIFS value of 2 (in time slots) is equal to SIFS$+2\tau = 50\ \mu s$ (in micro seconds).

We present results for $cw_{\min}$ and AIFS-based differentiation only, and for BASIC access mode. We first present the throughput results for persistent users (stage 1). The considered QoS scenarios are labelled by $cw_{\min} = (cw_{\min,1}, cw_{\min,2})$ and $AIFS = (AIFS_1, AIFS_2)$. The simulation and analytical results for the aggregate system throughput $R(n_1, n_2)$ and the class throughputs $R_i(n_1, n_2)$, $i = 1, 2$, are shown only for $n_1 = n_2$ with total number $n_1 + n_2$, since it is difficult to visualize the difference between 2-dimensional functions (if mapped on $\mathbb{R}^3$). Graphs for $n_1 \neq n_2$ are similar. In stage 2, we present the mean file transfer time results for the corresponding non-persistent flow-level model.

Table 1
Default parameter settings for the MAC layer and the physical layer

parameter	value	parameter	value
PHY	192 bits	r_{WLAN}	1 Mbit/sec
MAC	272 bits	δ	1 μs
RTS	PHY+160 bits	τ	20 μs
CTS	PHY+112 bits	SIFS	10 μs
ACK	PHY+112 bits	DIFS	2 (time slots)
packet size	12 kbits		

parameter (all i)	value
$n_{\max,i}$	25
$cw_{\min,i}$	31
$r_{\max,i}$	3 (BASIC)
AIFS$_i$	DIFS

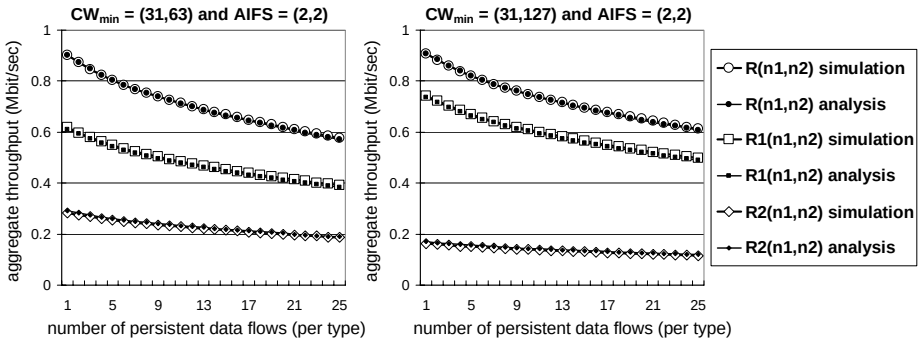

Figure 8.3: *Aggregate system and class throughputs as function of the number of persistent users $n_i = 1, ..., 25$; $n_1 = n_2$; for $cw_{\min} = (31, 63)$ and $cw_{\min} = (31, 127)$ (no AIFS differentiation, i.e., $AIFS = (2, 2)$).*

8.6.1 Packet-level: throughput results for persistent users

Figure 8.3 presents the throughput results where only $cw_{\min}$-based QoS differentiation is applied; respectively for $cw_{\min} = (31, 63)$ and $cw_{\min} = (31, 127)$. Figure 8.4 presents results for the scenario $AIFS = (2, 4)$, and the scenario $cw_{\min} = (31, 63)$ with $AIFS = (2, 4)$. The analytical results for AIFS-based differentiation are based on the Markov chain model [110]. The numerical results accurately represent 802.11E packet-level behavior.

The $cw_{\min}$ and AIFS parameters differ in several ways. The $cw_{\min}$ is used to reduce collision probabilities as well as providing QoS support. The $cw_{\min}$-based differentiation mechanism approximately maintains the bandwidth ratio $cw_{\min,2}/cw_{\min,1}$ for a large number of users. The QoS capabilities of the AIFS-based differentiation mechanism are particularly effective for busy systems. The class with the largest AIFS value will eventually suffer from starvation as the system gets busier. However, in a system with realistic, non-persistent traffic load, these large numbers of simultaneously active users are mostly not achieved.

8.6.2 Flow-level: transfer time results for non-persistent users

We consider the mean file transfer time $\mathbb{E}T_j$ for traffic class j as a function of the offered traffic load $\rho = \rho_1 + \rho_2$, with $\rho_2/\rho_1 = 2$ (the low priority class contributes twice as much to the total offered system load, compared to the high priority class). Any pair of loads (ρ_1, ρ_2) can be chosen; however, to avoid 3-dimensional graphs we only depict

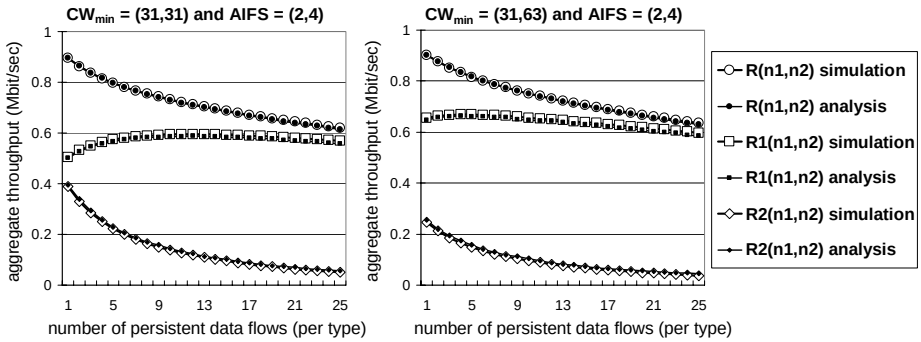

Figure 8.4: *Aggregate system and class throughputs as function of the number of persistent users* $n_i = 1, ..., 25$; $n_1 = n_2$; *for* $cw_{\min} = (31, 31), AIFS = (2, 4)$ *and* $cw_{\min} = (31, 63)$, $AIFS = (2, 4)$.

the transfer time results for the ratio of loads $\rho_2/\rho_1 = 2$. The approximation result will be 'at worst case' if $\rho_2/\rho_1 = 1$; and under a constant total load $\rho = \rho_1 + \rho_2$, it will improve for increasing ratio ρ_2/ρ_1, and also improve for decreasing ρ_2/ρ_1. To this end, observe that if $\rho_2/\rho_1 \to \infty$, then the corresponding scenario is simply a single-class (egalitarian) PS model with only active users from class 2. Analogously, if $\rho_2/\rho_1 \to 0$, then the corresponding scenario is a single-class (egalitarian) PS model with only active users from class 1.

We limit the number of users for each class in the system to $n_{\max,i} = 25$. Figure 8.5 shows results for BASIC access mode and with exponentially distributed file sizes with a mean of 120 kbits (10 packets of 12 kbits) for both classes.

When no service differentiation is applied, clearly $\mathbb{E}T_1 = \mathbb{E}T_2$, and the approximation is exact for egalitarian PS models. Also, the analytical approximation accurately represents the WLAN simulation result. For the other three scenarios with QoS differentiation, the approximation yields accurate results for a realistic region of parameter settings. The absolute approximation error is small for the high priority class 1, and relatively small for the low priority class 2, when the offered traffic load is moderate ($\rho < 0.7$). The region where the approximation breaks down is when the system is highly overloaded ($\rho >> 0.7$). In the latter case, N_i is often close to $n_{\max,i}$ and the number of blocked users tends to increase to infinity if the traffic load is above a certain threshold.

As expected, the analytical approximation for particularly the AIFS-based differentiation and extreme heavy-traffic is not accurate since $\alpha_{21}(\mathbf{n}) \approx 0$, if the number of active users is constantly large. However, from a practical flow-level point-of-view, an over-

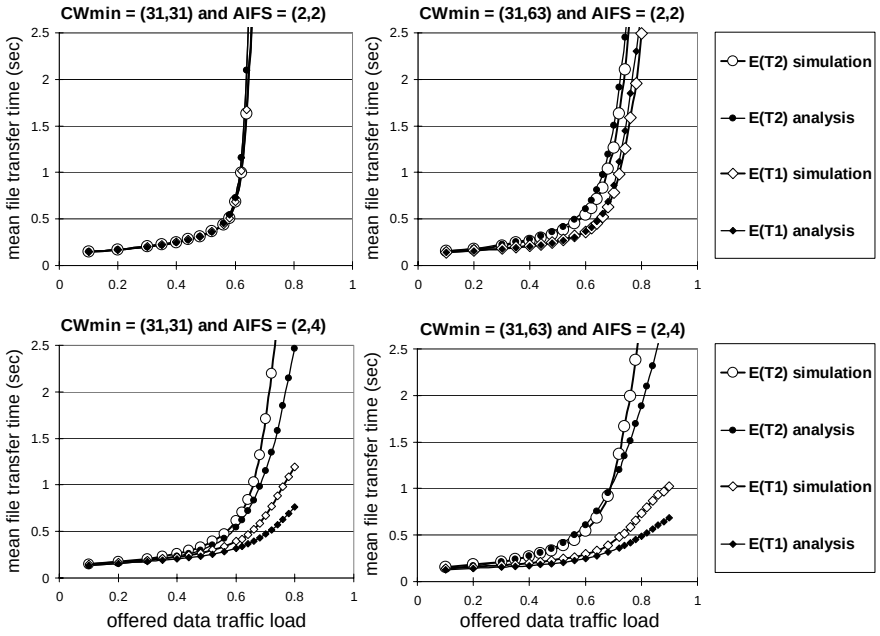

Figure 8.5: *Mean file transfer time $\mathbb{E}T_1$ and $\mathbb{E}T_2$ as a function of the offered data traffic load $\rho = \rho_1 + \rho_2$, with $\rho_1/\rho_2 = 2$, for the default scenario (no QoS differentiation) and 3 other QoS scenarios from stage 1.*

loaded WLAN system under BASIC access mode is not a realistic scenario setting. It is more efficient to use the RTS/CTS mode instead of the BASIC mode when the system is heavily loaded. From simulation results for RTS/CTS (not shown) we observed that the accuracy of the approximation is mostly slightly better than the results for BASIC mode.

Finally, we note that the analytical approximation method is insensitive to the file size distributions, i.e., the approximated mean file transfer times do *not* depend on the choice of file size distributions. The queue length distributions and mean flow transfer times in egalitarian PS models, such as Cohen's GPS [32] model, have the attractive property of insensitivity to the service time distributions. The discriminatory PS models (including GDPS) are non-product-form and do *not* have the insensitivity property, see e.g. [20]. However, the sensitivity only becomes significant when the priority weights are extremely asymmetric in combination with heavy-traffic.

8.7 Conclusions

In this chapter we have presented an integrated packet/flow-level modeling approach for performance evaluation of IEEE 802.11E WLANs with dynamically varying number of active users. The packet-level model describes the QoS mechanisms of the EDCA MAC layer in detail, and the flow-level model is based on the observation that the considered 802.11E system approximately behaves as a queueing system with a *generalized* discriminatory processor-sharing (GDPS) service discipline.

We used an analytical decomposition method from Chapter 5 for approximating mean file transfer times, since exact evaluation of (G)DPS models is not tractable. The 802.11E WLAN simulations show that the flow-level behavior of 802.11E is closely represented by a DPS model with queue-dependent service capacity and queue-dependent service weights. The approximation is accurate for a wide region of realistic parameter settings. The scenarios where the approximation breaks down (particularly for the low priority class) is when the *starvation effect* becomes significant, i.e., when QoS scenarios are considered such that the relative priority weights are extremely small or extremely large in combination with heavy-traffic (particularly from the high priority class users).

The (G)DPS modeling approach offers additional insights in the flow-level behavior of 802.11E. When a low priority user generates a data flow at a time instant when few high priority users are active, then the low priority user will have relatively small file transfer times. However, if more high priority users start to generate data flows, the low priority's performance decreases severely, as if the service process is suddenly 'frozen'. But on the other hand, due to the large share of bandwidth that the high priority users receive, the high priority users reside in the system for a relatively short period of time. When few high priority users are active, the low priority users still get a substantial share of the available bandwidth.

We conclude that the low priority's flow-level performance is characterized by a high variance, whereas the high priority's performance is much less variable. This shows that providing time-bounded QoS guarantees seems hard to realize, particularly for the low priority class. When statistical performance guarantees are given for the low priority class, these are at a generally small confidence level; unless the system is lightly loaded. Topics for further research include taking more enhancements in the physical layer into account (e.g. capture effects) and to optimize the EDCA performance under certain guaranteed QoS. Also, finding a tractable evaluation of the mean file transfer times conditional on the user's initial file size for IEEE 802.11E WLANs and (G)DPS is a challenging task.

Bibliography

[1] IEEE 802.11. *Wireless Local Area Networks – The Working Group for WLAN Standards*. `http://grouper.ieee.org/groups/802/11/`.

[2] IEEE 802.11B. *Supplement to Standard for Telecommunications and Information Exchange Between Systems – LAN/MAN Specific Requirements*. Part 11: Wireless MAC and PHY Specifications: higher speed physical layer extension in the 2.4 GHz band, P802.11B/D7.0, 1999.

[3] IEEE 802.11E. *Standard for Telecommunications and Information Exchange Between Systems – LAN/MAN Specific Requirements*. Part 11: Wireless MAC and PHY Specifications: MAC enhancements for Quality of Service (QoS), P802.11e 4.3, May 2003.

[4] E. Altman, K. Avrachenkov, and U. Ayesta. A survey on discriminatory processor sharing. *Queueing Systems*, 53(1-2):53–63, 2006.

[5] E. Altman, T. Jiménez, and D. Kofman. DPS queues with stationary ergodic service times and the performance of TCP in overload. In *Proceedings of IEEE INFOCOM*, Hong Kong, 2004.

[6] D. Anick, D. Mitra, and M.M. Sondhi. Stochastic theory of a data-handling system with multiple sources. *Bell Systems Technical J.*, 61:1871–1894, 1982.

[7] B.K. Asare and F.G. Foster. Conditional response times in the M/G/1 processor-sharing system. *Journal of Applied Probability*, 20:910–915, 1983.

[8] S. Asmussen. *Applied Probability and Queues, 2nd revised and extended edn.* Applications of Mathematics 51, Springer, New York, NY, 2003.

[9] B. Avi-Itzhak, H. Levy, and E. Brosh. *SQF: A Slowdown Queueing Fairness Measure*. RUTCOR Research Report, Rutgers University, January 2006.

[10] K. Avrachenkov, U. Ayesta, P. Brown, and R. Núñez-Queija. Discriminatory processor sharing revisited. In *Proceedings of IEEE INFOCOM*, Miami, USA, 2005.

[11] N. Bansal. Analysis of the M/G/1 processor-sharing queue with bulk arrivals. *Operations Research Letters*, 31(5):401–405, 2003.

[12] F. Baskett, K.M. Chandy, R.R. Muntz, and F.G. Palacios. Open, closed, and mixed networks of queues with different classes of customers. *Journal of the Association for Computing Machinery*, 22(2):248–260, 1975.

[13] J.V.L. Beckers, I. Hendrawan, R.E. Kooij, and R.D. van der Mei. Generalized processor sharing models for internet access lines. In *Proceedings of 9th IFIP conference on Performance Modeling and Evaluation of ATM & IP Networks*, Budapest, 2001.

[14] S. Ben Fredj, T. Bonald, A. Proutière, G. Régnié, and J.W. Roberts. Statistical bandwidth sharing: a study of congestion at flow level. In *Proceedings of ACM SIGCOMM'01*, pages 111–122, 2001.

[15] J.L. van den Berg. *Sojourn Times in Feedback and Processor Sharing Queues*. PhD thesis, Utrecht University, The Netherlands, 1990.

[16] J.L. van den Berg and O.J. Boxma. The M/G/1 queue with processor sharing and its relation to a feedback queue. *Queueing Systems*, 9:365–402, 1991.

[17] A. Bhattacharjee and D. Sengupta. On the coefficient of variation of the $\mathcal{L}$- and $\overline{L}$-classes. *Statistics & Probability Letters*, 27:177–180, 1996.

[18] G. Bianchi. Performance analysis of the IEEE 802.11 distributed coordination function. *IEEE Journal on Selected Areas in Communications*, 18(3):535–547, 2000.

[19] T. Bonald and L. Massoulié. Impact of fairness on internet performance. In *Proceedings of ACM Sigmetrics/Performance '01*, Cambridge, Massachusetts, USA, 2001.

[20] T. Bonald and A. Proutière. Insensitivity in processor-sharing networks. *Performance Evaluation*, 49:193–209, 2002.

[21] S.C. Borst. User-level performance of channel-aware scheduling algorithms in wireless data networks. In *Proceedings of IEEE INFOCOM*, 2003.

[22] S.C. Borst, R. Núñez-Queija, and A.P. Zwart. Sojourn time asymptotics in processor sharing queues. *Queueing Systems*, 53(1-2):31–51, 2006.

[23] A. Brandt and M. Brandt. A sample path relation for the sojourn times in G/G/1–PS systems and its applications. *Queueing Systems*, 52:281–286, 2006.

[24] L. Carlitz. Eulerian numbers and polynomials of higher order. *Duke Mathematical Journal*, 27:401–423, 1960.

[25] S.K. Cheung, R.J. Boucherie, and R. Núñez-Queija. Effective load and adjusted stability in queues with fluctuating service rates. *Report*, 2007.

[26] S.K. Cheung, B. Kim, and J. Kim. Slowdown in the M/M/1 discriminatory processor-sharing queue. *Report*, 2007.

[27] S.K. Cheung, J.L. van den Berg, and R.J. Boucherie. Decomposing the queue length distribution of processor-sharing models into queue lengths of permanent customer queues. *Performance Evaluation*, 62(1–4):100–116, 2005. Proceedings of IFIP's Performance 2005, Juan-les-Pins, France.

[28] S.K. Cheung, J.L. van den Berg, and R.J. Boucherie. Insensitive bounds for the moments of the sojourn time distribution in the M/G/1 processor-sharing queue. *Queueing Systems*, 53(1–2):7–18, 2006.

[29] S.K. Cheung, J.L. van den Berg, R.J. Boucherie, R. Litjens, and F. Roijers. An analytical packet/flow-level modelling approach for wireless LANs with Quality-of-Service support. In *Proceedings of the 19th International Teletraffic Congress*, Beijing, China, 2005.

[30] S. Choi, J. del Prado, S. Shankar N, and S. Mangold. IEEE 802.11E contention-based channel access (EDCF). In *Proceedings of ICC 2003*, Anchorage, USA, 2003.

[31] E.G. Coffman, R.R. Muntz, and H. Trotter. Waiting-time distributions for processor-sharing systems. *Journal of the Association for Computing Machinery*, 17:123–130, 1970.

[32] J.W. Cohen. The multiple phase service network with generalized processor sharing. *Acta Informatica*, 12:245–284, 1979.

[33] J.W. Cohen. *The Single Server Queue, 2nd edn.* North Holland, Amsterdam, 1982.

[34] F. Delcoigne, A. Proutière, and G. Régnié. Modeling integration of streaming and data traffic. *Performance Evaluation*, 55(3–4):185–209, 2004.

[35] R.A. Doney. Moments of ladder heights in random walks. *Journal of Applied Probability*, 17:248–252, 1980.

[36] R. Egorova, A.P. Zwart, and O.J. Boxma. Sojourn time tails in the M/D/1 processor sharing queue. *Probability in the Engineering and Informational Sciences*, 20(3):429–446, 2006.

[37] A.K. Erlang. "Sandsynlighedsregning og Telefonsamtaler" (in Danish; translated: "The theory of probabilities and telephone conversations"). *Nyt tidsskrift for matematik B*, 20:33–39, 1909.

[38] G. Fayolle, I. Mitrani, and R. Iasnogorodski. Sharing a processor among many job classes. *Journal of the Association for Computing Machinery*, 27:519–532, 1980.

[39] A. Federgruen and L. Green. Queueing systems with service interruptions. *Operations Research*, 34:752–768, 1986.

[40] W. Feller. *An Introduction to Probability Theory and Its Applications, volume I*. Wiley, third edition, New York, NY, 1966.

[41] C.H. Foh and M. Zukerman. Performance analysis of the IEEE 802.11 MAC protocol. In *Proceedings of European Wireless*, Florence, Italy, 2002.

[42] R.L. Graham, D.E. Knuth, and O. Patashnik. *Eulerian Numbers*, chapter §6.2 in Concrete Mathematics: A Foundation for Computer Science, pages 267–272. Addison-Wesley, Reading, MA, 2nd. edition, 1994.

[43] S. Grishechkin. On a relationship between processor-sharing queues and Crump-Mode-Jagers branching processes. *Advances in Applied Probability*, 24:653–698, 1992.

[44] F. Guillemin and J. Boyer. Analysis of the M/M/1 queue with processor sharing via spectral theory. *Queueing Systems*, 39(4):377–397, 2001.

[45] V. Gupta, M. Harchol-Balter, A.S. Wolf, and U. Yechiali. Fundamental characteristics of queues with fluctuating load. In *Proceedings of ACM Sigmetrics/Performance '06*. Saint Malo, France, June 2006.

[46] R.C. Hampshire, M. Harchol-Balter, and W.A. Massey. Fluid and diffusion limits for transient sojourn times of processor sharing queues with time varying rates. *Queueing Systems*, 53(1-2):19–30, June 2006.

[47] M. Harchol-Balter, K. Sigman, and A. Wierman. Asymptotic convergence of scheduling policies with respect to slowdown. In *Performance 2002*, 2002.

[48] Y. Hayel and B. Tuffin. Pricing for heterogeneous services at a discriminatory processor sharing queue. In *4rd IFIP-TC6 Networking Conference*, May 2005.

[49] D. He and Ch. Shen. *Simulation Study of IEEE 802.11E EDCF*. IST Project Moby Dick.

[50] V. Jacobson. Congestion avoidance and control. In *Proceedings of ACM SIGCOMM '88*, pages 314–329, Stanford, CA, August 1988.

[51] F.P. Kelly. *Reversibility and Stochastic Networks*. Wiley, Chichester, 1979.

[52] D.G. Kendall. Stochastic processes occuring in the theory of queues and their analysis by the method of the imbedded Markov chain. *Annals of Mathematical Statistics*, 24:338–354, 1953.

[53] G. van Kessel, R. Núñez-Queija, and S.C. Borst. Asymptotic regimes and approximations for discriminatory processor sharing. In *Performance Evaluation Review 32*, pages 44–46, 2004. Special issue on MAMA 2004, workshop on Mathematical Performance Modeling and Analysis.

[54] G. van Kessel, R. Núñez-Queija, and S.C. Borst. Differentiated bandwidth sharing with disparate flow sizes. In *Proceedings of IEEE INFOCOM*, Miami, USA, 2005.

[55] B. Kim and J. Kim. Comparison of DPS and PS systems according to DPS weights. *IEEE Communication Letters*, 10:558–560, 2006.

[56] J. Kim and B. Kim. Sojourn time distribution in the M/M/1 queue with discriminatory processor-sharing. *Performance Evaluation*, 58:341–365, 2004.

[57] J. Kim and B. Kim. The processor-sharing queue with bulk arrivals and phase-type services. *Performance Evaluation*, 64(4):277–297, 2007.

[58] J. Kim and C. Kim. Performance analysis and evaluation of IEEE 802.11E EDCF. In *Proceedings of Wireless Communications and Mobile Computing*, Florence, Italy, 2004.

[59] Y. Kitayev. The M/G/1 processor-sharing model: Transient behavior. *Queueing Systems*, 14:239–273, 1993.

[60] B. Klar. A note on the $\mathcal{L}$-class of life distributions. *Journal of Applied Probability*, 39(1):11–19, 2002.

[61] B. Klefsjö. A useful ageing property based on Laplace transform. *Journal of Applied Probability*, 20:615–626, 1983.

[62] L. Kleinrock. Analysis of a time-shared processor. *Naval Research Logistics Quarterly*, 11:59–73, 1964.

[63] L. Kleinrock. Time-shared systems: A theoretical treatment. *Journal of the Association for Computing Machinery*, 14:242–261, 1967.

[64] L. Kleinrock. *Queueing Systems, Vol I: Theory*. Wiley, New York, 1975.

[65] L. Kleinrock. *Queueing Systems, Vol II: Computer Applications*. Wiley, New York, 1976.

[66] G. Latouche and V. Ramaswami. *An Introduction to Matrix Analytic Methods in Stochastic Modeling*. Society for Industrial & Applied Mathematics, 1987.

[67] G. Lin. Characterizations of the $\mathcal{L}$-class of life distributions. *Statistics & Probability Letters*, 40:259–266, 1998.

[68] R. Litjens, F. Roijers, J.L. van den Berg, R.J. Boucherie, and M. Fleuren. Analysis of flow transfer times in IEEE 802.11 wireless LANs. *Annals of Telecommunications*, 59(11–12):1407–1432, 2004.

[69] J.D.C. Little. A proof of the queueing formula $L = \lambda W$. *Operations Research*, 9:383–387, 1961.

[70] A. Mandelbaum and W.A. Massey. Strong approximations for time dependent queues. *Mathematics of Operations Research*, 20(1):33–64, 1995.

[71] M. Mandjes and A.P. Zwart. Large deviations of sojourn times in processor sharing queues. *Queueing Systems*, 52:237–250, 2006.

[72] S. Mangold, S. Choi, P. May, O. Klein, G. Hiertz, and L. Stibor. IEEE 802.11E Wireless LAN for Quality of Service. In *European Wireless 2002*, Florence, Italy, 2002.

[73] W.A. Massey. Asymptotic analysis of the time dependent M/M/1 queue. *Mathematics of Operations Research*, 10:305–327, 1985.

[74] W.A. Massey and W. Whitt. Uniform acceleration expansions for Markov chains with time-varying rates. *Annals of Applied Probability*, 8(4):1130–1155, 1998.

[75] L. Massoulié and J.W. Roberts. Bandwidth sharing and admission control for elastic traffic. *Telecommunication Systems*, 15(1):185–201, 2000.

[76] J.A. Morrison. Response-time distribution for a processor-sharing system. *SIAM Journal of Applied Mathematics*, 45:152–167, 1985.

[77] M.F. Neuts. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. The Johns Hopkins University Press, 1981.

[78] R. Núñez-Queija. *Processor Sharing Models for Integrated Services Networks*. PhD thesis, Eindhoven University of Technology, The Netherlands, 2000.

[79] R. Núñez-Queija. Sojourn times in a processor sharing queue with service interruptions. *Queueing Systems*, 34(1–4):351–386, 2000.

[80] M. Nuyens and A. Wierman. *The Foregound-Background Processor-Sharing Queue: a survey*. Preprint.

[81] T.M. O'Donovan. Direct solutions of M/G/1 processor-sharing models. *Operations Research*, 22:1232–1235, 1974.

[82] T.J. Ott. The sojourn time distribution in the M/G/1 queue with processor sharing. *Journal of Applied Probability*, 21:360–378, 1984.

[83] A.K. Parekh and R.G. Gallager. A generalized processor sharing approach to flow control in integrated services networks: The single node case. *IEEE/ACM Transactions on Networking*, 1(3):344–357, 1993.

[84] T. Raimondi and M. Davis. Design rules for a class-based differentiated service QoS scheme in IEEE 802.11E Wireless LANs. In *Proceedings of MSWiM '04*, Venice, Italy, 2004.

[85] K.M. Rege and B. Sengupta. Queue length distribution for the discriminatory processor-sharing queue. *Operations Research*, 44:653–657, 1996.

[86] S. Ross. Average delay in queues with non-stationary Poisson arrivals. *Journal of Applied Probability*, 15:602–609, 1978.

[87] M. Sakata, S. Noguchi, and J. Oizumi. Analysis of a processor-shared queueing model for time-sharing systems. In *Proceedings of 2nd Hawaii International Conference on System Sciences*, pages 625–628, Jan. 1969.

[88] M. Sakata, S. Noguchi, and J. Oizumi. An analysis of the M/G/1 queue under round-robin scheduling. *Operations Research*, 19:371–385, 1971.

[89] R. Schassberger. A new approach to the M/G/1 processor sharing queue. *Advances in Applied Probability*, 16:202–213, 1984.

[90] W.R.W. Scheinhardt. *Markov-Modulated and Feedback Fluid Queues*. PhD thesis, University of Twente, Enschede, The Netherlands, 1998.

[91] B. Sengupta and D.L. Jagerman. A conditional response time of the M/M/1 processor-sharing queue. *AT & T Bell Laboratories Technical Journal*, 64:409–421, 1985.

[92] M. Shaked and J.G. Shanthikumar. *Stochastic Orders and Their Applications*. Academic Press Inc., 1994.

[93] L.R. Shenton and K.O. Bowman. The geometric distribution, cumulants, Eulerian numbers, and the logarithmic function. *Far East Journal of Theoretical Statistics*, 5:113–142, 2001.

[94] L.R. Shenton and K.O. Bowman. The geometric distribution's central moments and Eulerian numbers of the second kind. *Far East Journal of Theoretical Statistics*, 7:1–17, 2002.

[95] D. Stoyan. *Comparison Methods for Queues and Other Stochastic Models*. Wiley, Chichester, 1983.

[96] H. Takagi. *Queueing Analysis, Vol. 1: Vacation and Priority Systems*. Elsevier Science Publishers B.V., The Netherlands, 1991.

[97] H. Takagi. *Queueing Analysis, Vol. 2: Finite Systems*. Elsevier Science Publishers B.V., The Netherlands, 1993.

[98] H. Takagi. *Queueing Analysis, Vol. 3: Discrete-Time Systems*. Elsevier Science Publishers B.V., The Netherlands, 1993.

[99] H. Tijms. *A First Course in Stochastic Models*. John Wiley & Sons Inc., Chichester, 2003.

[100] H. Truong and G. Vannuccini. The IEEE 802.11E MAC for Quality Of Service in Wireless LANs. In *Proceedings of SSGRR '03*, L'Aquila, Italy, 2003.

[101] M.J.G. van Uitert. *Generalized Processor Sharing Queues*. PhD thesis, Eindhoven University of Technology, The Netherlands, 2003.

[102] A. Ward and W. Whitt. Predicting response times in processor-sharing queues. In P.W. Glynn, D.J. MacDonald, and S.J. Turner, editors, *Proceedings of the Fields Institute Conference on Communication Networks*, 2000.

[103] Eric W. Weisstein. *Eulerian Number*. From MathWorld – A Wolfram Web Resource, `http://mathworld.wolfram.com/EulerianNumber.html`.

[104] W. Whitt. The M/G/1 processor-sharing queue with long and short jobs. Unpublished manuscript, 1998.

[105] A. Wierman and M. Harchol-Balter. Classifying scheduling policies with respect to unfairness in an M/GI/1. In *Proceedings of ACM Sigmetrics '03*, pages 238–249, San Diego, 2003.

[106] S.F. Yashkov. A derivation of response time distribution for a M/G/1 processor-sharing queue. *Problems of Control and Information Theory*, 12:133–148, 1983.

[107] S.F. Yashkov. Processor-sharing queues: Some progress in analysis. *Queueing Systems*, 2:1–17, 1987.

[108] S.F. Yashkov. Mathematical problems in the theory of processor-sharing queueing systems. *Journal of Soviet Mathematics*, 58:101–147, 1992.

[109] S.F. Yashkov. On a heavy traffic limit theorem for the M/G/1 processor sharing queue. *Communications in Statistics – Stochastic Models*, 9(3):467–471, 1993.

[110] J. Zhao, Z. Guo, Q. Zhang, and W. Zhu. Performance study of MAC for service differentiation in IEEE 802.11. In *Proceedings of IEEE GLOBECOM*, Taipei, Taiwan, 2002.

[111] A.P. Zwart. *Queueing Systems with Heavy Tails*. PhD thesis, Eindhoven University of Technology, The Netherlands, 2001.

[112] A.P. Zwart and O.J. Boxma. Sojourn time asymptotics in the M/G/1 processor sharing queue. *Queueing Systems*, 35:141–166, 2000.

Index

Summary

In the past few decades, the processor-sharing (PS) model has received considerable attention in the queueing theory community and in the field of performance evaluation of computer and communication systems. The scarce resource is simultaneously shared among all users in these systems. PS models are used for modeling resource sharing mechanisms and have many applications in communication networks, as well as in logistics and manufacturing.

In Chapter 2 we consider the resource sharing mechanism in Wireless Local Area Networks (WLANs), and we discuss how this leads to processor-sharing models. In addition, we give an overview of the literature on two basic classes of PS models: the *egalitarian* processor-sharing (EPS) and the *discriminatory* processor-sharing (DPS) models. In EPS, all users are treated in the same manner, i.e., each user simultaneously receives the same share of the service capacity. In the DPS model, the capacity is not equally shared, and therefore it is possible to give certain users a higher priority at the expense of other users. Thorough knowledge of EPS and DPS models is of utmost importance for the many application areas of PS models such as in communication networks.

In Chapter 3 we obtain a decomposition result for the queue length distribution in EPS models with multiple customer types and the so-called property of balanced capacities. A crucial observation is that the marginal queue length distribution of each class is the same as the queue length distribution of a related EPS model with a random number of permanent customers. This decomposition result plays an important role in Chapter 4, where bounds are derived for the moments of the sojourn time distribution in the classical EPS queue with Poisson arrivals and generally distributed service requirements. The decomposition result also motivates an approximation method for the mean sojourn times in DPS models in Chapter 5.

In Chapter 4 we derive stochastic ordering results for the classical EPS queue with Poisson arrivals and generally distributed service requirements. In particular, we derive *insensitive* bounds for the moments of the sojourn time distribution given the initial service requirement of the customer. The upper bounds for the moments are in fact the moments of the so-called instantaneous sojourn time, which can be interpreted as the

sojourn time of a customer with a very small service requirement. The instantaneous sojourn time can also be interpreted as the sojourn time of customer with an arbitrary size of the service requirement, and where the number of other customers remains fixed after arrival of the tagged customer. Since the latter interpretation is a special case of an EPS model with permanent customers, the instantaneous sojourn time is also related to the queueing model with a random number of permanent customers from Chapter 3.

In Chapter 5 we propose a method for approximating the mean sojourn times in general discriminatory processor-sharing (GDPS) models with queue-dependent service capacity and queue-dependent priority weights. The method is based upon the queueing model with permanent customers and the decomposition result from Chapter 3. The approximation method is exact for PS models with balanced capacities.

In Chapter 6 we consider the DPS model with Poisson arrivals and exponentially distributed service requirements. For this queueing model, we derive the first and second moments of the slowdown, which is a measure for queueing fairness, i.e., it measures how fair the customers are treated by a service discipline. The fair and egalitarian PS model has a constant slowdown. In contrast, DPS models do not have a constant slowdown, and DPS models aim to give priority to certain customers at the expense of other customers. We illustrate how 'unfair' DPS queueing models are for different type of customers, depending on the system parameters.

In Chapter 7 we study a general queueing model with time-fluctuating service capacity. In particular, we consider queueing models with periods of temporary instability. For these models, we discuss a notion of effective load and introduce the notion of adjusted stability. These notions capture the effect of accumulated work, which can not be served during the periods of temporary instability, and therefore needs to be served before the queue recovers from overload and effectively shows steady-state behavior.

In Chapter 8 we present a performance analysis of WLANs with multiple traffic type and Quality-of-Service support. The analytical evaluation is based on a so-called integrated packet/flow-level modeling approach. The flow-level model, which captures dynamic user behavior, is a general discriminatory processor-sharing (GDPS) model. Using the decomposition- and approximation method from Chapter 5 we obtain accurate approximations for the mean file transfer times for a wide range of realistic parameter settings. These analytical results are validated with extensive and detailed WLAN simulations.

Samenvatting (Summary)

Het processor-sharing (PS) model heeft in de afgelopen decennia veel aandacht gekregen op de gebieden van de wachtrijtheorie en de prestatie analyse van computer en communicatie systemen. In deze modellen delen alle gebruikers simultaan de schaarse systeem capaciteit. In vele toepassingsgebieden in communicatie netwerken, alsmede in de logistiek en productie processen, worden PS modellen gebruikt ter modellering van mechanismen om capaciteit te verdelen.

In Hoofdstuk 2 wordt de capaciteitsverdeling van draadloze netwerken (Wireless Local Area Networks) beschouwd, en besproken hoe dit tot de processor-sharing modellen leidt. Tenslotte wordt een overzicht gegeven van de bestaande literatuur betreffende twee fundamentele typen PS modellen: het *egalitarian* processor-sharing (EPS) model en het *discriminatory* processor-sharing (DPS) model. In het EPS model worden alle gebruikers op een gelijke manier behandeld, d.w.z. alle gebruikers krijgen simultaan een gelijk deel van de capaciteit. In het DPS model wordt de capaciteit op een ongelijke manier verdeeld, en zodoende is het mogelijk om bepaalde type gebruikers een hogere prioriteit te geven ten koste van andere gebruikers. Kennis en inzicht van EPS en DPS modellen zijn uitermate belangrijk voor de vele toepassingsgebieden van PS modellen in bijvoorbeeld communicatie netwerken.

In Hoofdstuk 3 wordt een decompositie resultaat afgeleid voor de wachtrij verdeling van EPS modellen met meerdere typen klanten en met de zogenaamde eigenschap van gebalanceerde capaciteiten. Een belangrijke observatie is dat de marginale wachtrij verdeling voor elk type klant hetzelfde is als de wachtrij verdeling van een gerelateerd EPS model met een stochastisch aantal permanente klanten. Dit decompositie resultaat is van belang in Hoofdstuk 4 waar grenzen worden afgeleid voor de momenten van de verblijftijd in het klassieke EPS model met Poisson aankomsten en algemeen verdeelde bedieningsvraag. Het decompositie resultaat motiveert ook een approximatie methode om de verwachte verblijftijden voor DPS modellen te benaderen in Hoofdstuk 5.

In Hoofdstuk 4 worden stochastiche ordening resultaten afgeleid voor het klassieke EPS wachtrij systeem met Poisson aankomsten en algemeen verdeelde bedieningsvraag. In het bijzonder worden *ongevoelige* onder- en bovengrenzen afgeleid voor de momenten

van de verdeling van de verblijftijd gegeven de initïele bedieningsvraag van de klant. De bovengrenzen voor de momenten zijn in feite de momenten van de zogenaamde instantane verblijftijd, die geïnterpreteerd kan worden als de verblijftijd van een klant met een zeer kleine bedieningsvraag. De instantane verblijftijd kan ook geïnterpreteerd worden als de verblijftijd van een klant met een willekeurig grote bedieningsvraag, maar waarin het aantal andere klanten in het systeem niet meer verandert. Omdat de laatste interpretatie een speciaal geval is van een EPS model met permanente klanten, is de instantane verblijftijd ook gerelateerd aan het wachtrijmodel met een stochastisch aantal permanente klanten van Hoofstuk 3.

In Hoofdstuk 5 wordt een methode gegeven voor het benaderen van de verwachte verblijftijden in algemene DPS modellen (General DPS) met rijlengte-afhankelijke service capaciteit en rijlengte-afhankelijke prioriteitsgewichten. De methode is gebaseerd op het wachtrij systeem met permanente klanten en het decompositie resultaat van Hoofdstuk 3. De benaderingmethode is exact voor PS modellen met gebalanceerde capaciteiten.

In Hoofdstuk 6 wordt het DPS model beschouwd met Poisson aankomsten en exponentieel verdeelde bedieningsvraag. Voor dit wachtrijmodel worden de eerste twee momenten afgeleid van de zogenaamde ‘slowdown’, wat een maat is voor hoe eerlijk een service discipline is voor de klanten in het systeem. Het ‘eerlijke’ EPS model heeft een constante slowdown. De slowdown is niet constant in DPS modellen. DPS modellen beogen immers prioriteit te geven aan bepaalde klanten ten koste van andere klanten. We illustreren hoe ‘oneerlijk’ DPS wachtrijen zijn voor de verschillende typen klanten, afhankelijk van de systeem parameters.

In Hoofdstuk 7 wordt een algemeen wachtrijmodel beschouwd waarin de service capaciteit fluctueert in de loop der tijd. In het bijzonder worden modellen bestudeerd waarin de wachtrij tijdelijk instabiel is. De periodes waarin de wachtrij instabiel is, hebben een stochastische tijdsduur. Voor deze modellen wordt een notie van effectieve belasting behandeld en een notie van aangepaste stabiliteit geïntroduceerd. Deze noties nemen het effect mee dat het aangekomen werk tijdens instabiele perioden niet afgehandeld kan worden, en derhalve eerst weggewerkt moet worden voordat de wachtrijlengte zich herstelt en effectief gezien weer stabiel gedrag vertoont.

In Hoofdstuk 8 wordt een prestatie analyse gegeven van draadloze netwerken (Wireless LANs) met meerdere typen verkeer en Quality-of-Service ondersteuning. De analytische benaderingsmethode is gebaseerd op een zogenaamde geïntegreerde packet/flow-level modellering. Het flow-level model dat rekening houdt met dynamisch gebruikersgedrag, is het GDPS model. Met behulp van de decompositie- en approximatie methode van Hoofdstuk 5 worden accurate benaderingen verkregen voor de verwachte download- of verzendtijden van data voor een brede variatie van realistische parameterwaarden. Deze analytische resultaten worden gevalideerd met uitgebreide en gedetailleerde WLAN simulaties.

About the Author

SING-KONG (Sing Kwong) Cheung was born on July 2, 1979, in Amersfoort (The Netherlands). In the same city, he completed grammar school at the Vallei College in June 1997. The same year he started to study Econometrics at the Vrije Universiteit (VU) in Amsterdam. One of the most memorable experiences during this period includes a study-trip to South-India during November and December 2001, which he organized with three fellow Econometrics students on behalf of the INDECS committee ('INDian EConometric Studytrip'). After the organization of INDECS he completed his study in Econometrics with a specialization in Operations Research. He graduated in August 2002 at the Vrije Universiteit. His master's thesis was entitled "Efficient simulation of highly reliable Markovian dependability systems – based on a cross-entropy approach", supervised by Ad Ridder.

Later, in November 2002 he became a PhD student under supervision of Richard Boucherie and Hans van den Berg at the University of Twente. Sing-Kong visited the Telecommunications Mathematics Research Centre, Korea University (KU) in Seoul (South-Korea) for one month, and he also visited the Korean Advanced Institute of Science and Technology (KAIST) in Daejeon (South-Korea) for two weeks. Sing-Kong defends his PhD thesis at the University of Twente on June 1, 2007.