Fluid models for QoS provisioning in communication networks

Frank Roijers

Fluid models for QoS provisioning in communication networks

Frank Roijers

Fluid models for QoS provisioning in communication networks

Fluid models for QoS provisioning in communication networks
Frank Roijers
Proefschrift Universiteit van Amsterdam

Gedrukt door Ponsen & Looijen B.V.
ISBN: 978-90-6464-321-7

Fluid models for QoS provisioning in communication networks

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus prof. dr. D.C. van den Boom
ten overstaan van een door het college voor promoties ingestelde
commissie, in het openbaar te verdedigen in de Agnietenkapel
op woensdag 11 februari 2009, te 12:00 uur

door

Frank Roijers

geboren te Rotterdam

Promotiecommissie

Promotores:	prof. dr. M.R.H. Mandjes
	prof. dr. J.L. van den Berg
Overige leden:	prof. dr. R.J.M.M. Does
	prof. dr. ir. E.R. Fledderus
	prof. dr. ir. B.R.H.M. Haverkort
	prof. dr. C.A.J. Klaassen
	prof. dr. G.M. Koole
	dr. R. Núñez Queija

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

Voorwoord

Dit proefschrift is het resultaat van het promotieonderzoek dat ik de afgelopen 5 jaren heb uitgevoerd onder begeleiding van Hans van den Berg en Michel Mandjes. Ik ben hun zeer erkentelijk voor de heel prettige samenwerking, hun toewijding voor (dit) onderzoek, en hun kennis en kunde; zonder hen had dit resultaat er niet gelegen.

TNO wil ik danken voor de mogelijkheid en middelen om dit onderzoek uit te kunnen voeren. Bij aanvang, begin 2004, geschiedde dit onderzoek deels bij en in samenwerking met de groep PNA2 van het Centrum Wiskunde & Informatica (CWI), medio 2006 is deze rol overgenomen door het Korteweg-de Vries Instituut (KdVI) voor Wiskunde van de Universiteit van Amsterdam (UvA); beide wil ik danken voor de beschikbaar gestelde middelen en begeleiding. Verder dank ik prof. Bong Dae Choi voor de mogelijkheid om het Telecommunication Mathematics Research Center (TMRC) aan de Korea University te Seoul te bezoeken voor een periode van 2 maanden in 2008.

Delen van dit onderzoek zijn (gedeeltelijk) tot stand gekomen in SenterNovem projecten: hoofdstukken 3 en 4 in het 'ICT-doorbraakproject' EQUANET en de hoofdstukken 6, 7, 8 en 9 in het 'Innovatiesubsidie Samenwerkingsproject' EASY WIRELESS.

Tenslotte wil ik familie, vrienden en collega's danken voor hun interesse in dit onderzoek en hun steun gedurende de afgelopen jaren.

Contents

Chapter 1

Introduction

Until the early 1980's telephony was essentially the only communication service, but in the past two decades the technological developments in this field have sped up. Nowadays communication services, such as IP telephony, Internet television, and gaming, are widely available, and continuously evolve into more sophisticated ones. These developments are accompanied by an increasing demand for highly-performing communication networks to properly support these new services. High-speed Internet access became widely available via ADSL, and mobile services gradually evolved from speech, enabled by the GSM network, to broadband data access supported by UMTS and HSPA. More recent developments in wireless technologies, e.g. Wireless LAN, give rise to so-called wireless ad-hoc networks supporting communication services among users in areas without any underlying infrastructure.

Communication services generate traffic streams consisting of small packets, that have to be transported via the underlying network while obeying certain requirements in terms of, e.g., packet latency, throughput, and packet-loss ratio, as otherwise the Quality of Service (QoS) will not be satisfactorily. These traffic streams may be highly variable and unpredictable depending on the users' behaviors as well as on the characteristics of the services. In order to meet the desired requirements, various QoS-provisioning approaches can be deployed in the network, which vary in complexity and the extent in which they adjust the traffic streams.

The tool of stochastic modeling, and, in particular, queueing theory, is widely used to evaluate the performance of communication networks, aiming at optimization of network design and dimensioning. A first stage in this performance evaluation consists of formulating the system dynamics as a mathematical (stochastic) model. Subsequently, the model is analyzed to obtain the performance metrics of interest, providing insight in, e.g., attained service levels, maximum sustainable load for a given QoS demand, and the efficacy of various QoS-provisioning methods that may be deployed in the network.

This thesis is concerned with two principal performance modeling and analysis issues in communication network environments that are currently attracting substantial attention: i) the dimensioning of IP-network links, and ii) the impact of resource sharing on the performance in wireless ad-hoc networks. Although the research questions and the underlying communication networks differ, there is a strong resemblance in terms of the models and evaluation methodologies that are

used to analyze them. In particular, we concentrate on models in which the traffic streams generated by the network's users are modeled as *fluid*, i.e., we do as if they are generated in a continuous 'fluid' manner, thus abstracting from the transmission of individual packets. Fluid models have proven to be a powerful analysis approach for communication networks as they, in many occasions, yield relatively simple models that allow for tractable analysis, while still capturing the essential relevant system characteristics.

The objective of this chapter is to further elaborate on the research questions that are investigated in this thesis, and, in particular, to detail our contributions. In order to do so, we first sketch some characteristics of the behavior of users and services, and discuss the importance of providing QoS in communication networks (Section 1.1). Next, we explain the research questions in more depth (Section 1.2), followed by brief introductions of the models and evaluation approaches that we use to analyze these questions (Section 1.3). Finally, we state our contributions and give an outline of this thesis (Section 1.4).

1.1 Providing QoS in communication networks

Communication services generate traffic that has to be transported by an underlying communication network; in order to attain a suitable QoS level, the network must satisfy certain requirements. In Section 1.1.1 we focus on characteristics of communication services, their traffic profiles and the relation between the QoS and the performance at the network layer. In Section 1.1.2 we elaborate on the network-layer performance and introduce two QoS-provisioning approaches.

1.1.1 Communication services

Communication services, such as web browsing, peer-to-peer file sharing, Internet telephony (VOIP), Internet television (IP-TV), etc., are often classified into two groups: elastic and streaming services, see e.g. [110].

Elastic services, e.g., web browsing and file sharing, adapt their packet transmission rate while active. These services mostly use the well-known TCP protocol, that controls the transmission rate; the rate varies over time depending on the level of congestion in the network. In most cases the entire file has to be transferred before it can be used by a user; therefore users relate the QoS of elastic services directly to the throughput or the transfer time of a file transmission.

Streaming services, e.g., voice and video services, generate traffic in a non-adaptive manner with either a constant or a variable rate. These services generally have a real-time or interactive nature and consequently require that the traffic profile is

preserved throughout the transport. Insufficient network capacity results in packet loss and delays, which are significant causes of QoS degradation.

Obviously, the QoS and the performance at the network layer are strongly related; services will not properly function if the network performance is insufficient. Consequently, the desired QoS enforces a minimum performance level for the underlying network. The performance at the network layer is indicated by metrics such as throughput, latency, and the packet-loss ratio (PLR); the relations between these metrics and the QoS have been thoroughly investigated for many services, see e.g. [37, 67]. For example, voice services demand a given small packet latency, streaming video requires a specific small PLR, and file transfers require a certain minimum throughput.

1.1.2 QoS-provisioning techniques

The performance at the network layer depends strongly on the offered traffic and the available capacity. Clearly, provisioning additional capacity in a network improves the network's performance. Further observe that the performance can be *traded off* against the utilization of the network resources, i.e., increasing the traffic load (and, hence, the network utilization) will degrade the performance.

Basically, two strategies are applied to ensure the QoS in communication networks: QoS differentiation and overprovisioning. QoS *differentiation* implements traffic-differentiation mechanisms which preferentially treat traffic of highly-demanding services such that the available capacity can be used more economically. *Overprovisioning* is a 'QoS by provisioning' approach, i.e., allocating enough capacity to meet the QoS requirements of all services present; all traffic streams are treated in the same way, thus *all* meet the requirements of the most stringent service.

Both approaches have their own merits and, consequently, their own areas of applications. Overprovisioning has several advantages over QoS differentiation, see also, e.g. [46]. In the first place, the complexity of the network routers can be kept relatively low, as no advanced traffic-differentiation mechanisms, such as scheduling and prioritization, are needed. Secondly, traffic-differentiation mechanisms require that the parameters involved are 'tuned well', in order to meet the QoS needs of the different classes – this usually requires the selection of various parameters (for instance: weights in weighted fair queueing algorithms, etc.). Overprovisioning has drawbacks as well: the lack of any traffic-differentiation mechanism dictates that all flows should meet the most stringent QoS requirement, thus reducing the efficiency of the network (in terms of maximum achievable utilization). However, it is expected that this effect is mitigated if there is a high degree of aggregation, even in the presence of heterogeneous QoS requirements across users, as argued in, e.g., the introductions of [46] and in [71].

As said before both approaches have their own areas of applications based on

the trade-off between costs of extra capacity and the increasing complexity due to traffic-differentiation mechanisms. In situations without firm capacity restrictions one would keep the complexity as low as possible, and hence one opts for overprovisioning. Therefore overprovisioning is generally applied in networks which can achieve a high utilization of resources and where additional capacity is relatively cheap, e.g., core networks. QoS differentiation is preferred in situations where capacity is scarce (and therefore expensive), e.g., in mobile access networks or wireless ad-hoc networks. In these networks the number of active users typically remain small and a single user can jeopardize the QoS of all other users.

1.2 Research topics addressed in this thesis

In this thesis we address two particular research topics related to different types of communication networks. Each of them is dealt with in a separate part of the thesis:

I Dimensioning of IP-network links using overprovisioning,

II Impact of resource-sharing on the performance of wireless ad-hoc networks.

Below we describe our research topics in more detail. In Section 1.3 we will elaborate on the performance modeling and evaluation issues associated with them.

1.2.1 Dimensioning of IP-network links using overprovisioning

We consider a service provider who uses overprovisioning to offer the required QoS. Recall that the idea of overprovisioning is to provide sufficient capacity at the network links such that the QoS demands of all users are met; it is the most convenient QoS-provisioning approach in many practical situations, see Section 1.1.2.

Insight into the required link capacity can be obtained from a mathematical formulation of this issue. We present our mathematical formulation, without going into details here. Therefore, we let $A(T)$ denote the amount of traffic offered to the link during an arbitrary time interval of length T. Then, the required capacity C is the smallest C such that following inequality holds:

$$\mathbb{P}(A(T) \geq \mathrm{C}T) < \varepsilon, \tag{1.1}$$

where ε is typically small.

Note that the inequality allows overflows (defined as: $A(s) > \mathrm{C}s$) to occur for time periods s shorter than T; the reason is that network elements are equipped with buffers such that overflows for short periods of time do not directly result in QoS degradation. Further observe that the constraint is *probabilistic*, i.e., $\varepsilon > 0$, and

not *absolute*, i.e., $\varepsilon = 0$, which would require an extraordinary amount of capacity. In fact, the parameters T and ε reflect the QoS constraint; the choices of their values embody the trade-off between the desired QoS and the network utilization costs.

In the first part of this thesis we extensively investigate inequality (1.1) and the influence of different types of traffic behavior $A(T)$, the time scale T, and the QoS parameter ε. In particular, we consider dimensioning from a service provider's point of view, whose foremost difficulty is that he lacks detailed insight into the characteristics of the offered traffic. We develop a dimensioning procedure which relies on just coarse traffic-load measurements.

In addition, we investigate the properties of overflow periods, e.g., the duration which is how long will an overflow last when it occurs. Overflow periods affect the network performance and, consequently, the QoS, in particular, overflows that last for long periods of time severely degrade the QoS. The duration and the other properties are indicators of the impact of an overflow, and it relates to the transient behavior of overflows. Observe that inequality (1.1) effectively restricts the *frequency* of overflows of (at least) length T.

1.2.2 Impact of resource sharing on the performance of wireless ad-hoc networks

In Part II of this thesis we investigate the impact of resource sharing among nodes on the performance of wireless ad-hoc networks. Wireless ad-hoc networks consist of self-configuring wireless nodes and can be deployed instantly without a fixed infrastructure or predetermined configuration. An important feature of ad-hoc networks is multi-hop connectivity, i.e., if a node is not directly connected to its destination, it can use other nodes to relay its traffic. The underlying communication technology is usually based on shared medium access, i.e., neighboring nodes share a common radio channel with limited capacity, e.g., as in IEEE 802.11 Wireless LAN (see e.g. [63]). Nodes that have central locations in these networks are more likely to be used as relay nodes and can easily become performance bottlenecks.

The goal of the second part of this thesis is to evaluate the performance of a relay node in a wireless ad-hoc network. We consider the situation that a time-varying number of wireless source nodes are transmitting flows of data packets to destinations elsewhere in the network via the relay node. We are interested in the transfer time of a flow, i.e., the time that is required to entirely transmit a flow from a source node, via the relay node, to the destination. In particular, we extensively investigate the impact of the resource sharing between the source nodes and the relay node on the transfer time of a flow. In particular, we investigate the benefits of granting a different share of the capacity to the relay node than to each of the source nodes.

1.3 Models and performance-evaluation methodologies

In the performance evaluation of communication networks by mathematical modeling one can distinguish three main steps:

1 Modeling, i.e., formulating the dynamics of the communication network in terms of a stochastic model.

2 Analysis of the stochastic model to obtain insights into the performance metrics as a function of the relevant system parameters.

3 Numerical evaluation using the analytical expressions of phase 2 to obtain insight into the influence of the model parameters on the performance metrics.

This section briefly introduces fluid modeling (Section 1.3.1) and performance evaluation methodologies (Section 1.3.2) that are used throughout this thesis; more comprehensive descriptions are provided in Chapter 2. We emphasize that the third step, i.e., numerical evaluations, is extensively performed in this thesis for each of the research topics under consideration.

1.3.1 Fluid modeling

As mentioned in the beginning of this chapter, we use so-called *fluid modeling* of traffic sources for the performance analysis in this thesis. By fluid modeling we mean that traffic sources are modeled as if these generate traffic in a fluid manner, i.e., according to a *continuous process*. This approach differs from 'classical' traffic modeling and queueing networks where traffic sources generate discrete amounts of work (packets) which arrive according to a certain point process. The advantage of fluid modeling is that it abstracts from per-packet details resulting in a more tractable model (while retaining the essential behavior of the system), which, in many cases, allows for more explicit analysis. Fluid modeling of traffic sources can be applied to capture the behavior of i) individual traffic sources, or ii) the aggregated traffic of many sources.

Fluid modeling of *individual traffic sources* concentrates on initiations and departures of flows, where flows continuously generate traffic when active. The transmission rate of an ongoing flow depends on the type of service involved; streaming services are modeled with a non-adaptive (but potentially variable) rate, while the rate of elastic services is adaptive (it for instance reacts to the number of other active flows). This modeling approach is used to investigate flow-level performance metrics, i.e., performance metrics that characterize properties of individual flows, such as the transfer time of an elastic flow.

Fluid modeling of *the aggregated traffic of many sources* is particularly useful if there is a substantial level of aggregation; the amount of traffic offered by an individual user is then only a small fraction of the total aggregate and individual users are not distinguished from each other. Studies in literature reveal that, in many practical situations, large aggregates of traffic may be modeled by a *Gaussian* process; the amount of offered traffic during an interval follows a Normal distribution where the mean and the variance function depend on the length of the interval.

1.3.2 Performance-analysis methodologies

Besides 'standard' probability theory for the analysis of the performance metrics under consideration, we frequently use a number of methodologies which are briefly explained in this section, and in more detail in Chapter 2.

Transforms. On several instances we analyze the *transform* of the random variable under consideration; the analysis of the random variable itself is often hard, and considering the transform may result in more tractable expressions. For a non-negative random variable X the Laplace transform (LT) $\mathbb{E}\exp(-sX)$ and the moment generating function (MGF) $\mathbb{E}\exp(\theta X)$, are defined as

$$\mathbb{E}e^{-sX} = \int_{x=0}^{\infty} e^{-sx} f(x)\mathrm{d}x, \qquad \mathbb{E}e^{\theta X} = \int_{x=0}^{\infty} e^{\theta x} f(x)\mathrm{d}x,$$

where $f(x)$ is the probability density function of X. Transforms are often practical to work with as they have some properties that greatly simplify calculations; most notably, the LT of the sum of two independent random variables is the product of the LTs of the individual random variables (and the same applies to MGFs).

Large deviations. From a performance engineering point of view there is special interest in *rare events*, i.e., events that occur with a small probability. Communication networks often support services with demanding QoS requirements, which enforce stringent values for the performance metrics at the network layer, e.g., small delays or low packet loss ratio. Therefore, the network is designed such that the probability that such an event occurs, e.g., the excess traffic is above a certain value, is very small, typically in the order of 10^{-4} to 10^{-6}. Large deviations (LD) theory is dedicated to the analysis of rare events. In this thesis we rely on specific cases of LD theory, viz. the so-called *sample-mean* LD and *sample-path* LD.

Simulations. Simulations can be used to empirically estimate performance metrics. Random samples of, e.g., the arrival process and service requirements, are used as inputs for an *emulation* of a communication network while keeping track of the desired performance metrics. We use simulations as a method to validate the modeling phase (investigate whether the model accurately captures the system's behavior), and to validate the analysis of the model (in case that exact analysis is not possible

and assumptions or approximations were made in the analysis phase). A drawback of the use of simulations is that they can be time-consuming, in particular when considering rare events. In this thesis we use, besides straightforward simulations, also *importance sampling*, which is a method to simulate rare events more efficiently.

1.4 Contributions and outline

This thesis essentially consists of two parts, each of them addressing one of the two research topics introduced in Section 1.2. In this section we sketch the contributions of this thesis, at the beginning of each chapter we provide the contributions of that particular chapter with more detail.

1.4.1 Dimensioning of IP-network links using overprovisioning

In Chapter 3 we consider the dimensioning of an IP-network link that is carrying the traffic from multiple users. Current bandwidth provisioning procedures for IP-network links are mostly based on simple rules of thumb, using coarse traffic measurements made on a time scale of e.g., 5 or 15 minutes. A crucial question, however, is whether such coarse measurements give any useful insight into the capacity actually needed: QoS degradation experienced by the users is strongly affected by traffic rate fluctuations on much smaller time scales. We develop a bandwidth provisioning rule that is based on Expression (1.1), i.e., we determine the required capacity using the QoS measure that the probability that the traffic supply exceeds the available bandwidth, over some predefined (small) interval T, is below some small fixed number ε. In the dimensioning procedure we combine coarse traffic-load measurements with fluid traffic modeling that captures the behavior (i.e., rate fluctuations) on the shorter time scales. Furthermore, the provisioning formula explicitly gives the impact of the QoS parameters T and ε on the required capacity. The validity of the bandwidth provisioning rule is assessed through extensive measurements performed in several operational network environments. This chapter is based on [13].

In Chapters 4 and 5 we consider the $M/M/\infty$ queue which is used as a flow-level model for the occupancy of an IP-network link. We are particularly interested in *congestion periods*, which are defined as periods during which the offered traffic (number of active users) is continuously above a certain value C. For the so-called C-congestion periods we are interested in the following performance metrics: the duration D_C, the number of arrivals N_C, and the area A_C which is the amount of offered traffic in excess of the capacity. The motivation behind the analysis of congestion periods is that knowledge of the characteristics of congestion periods can be used to understand the performance of an IP-network link. Observe that Chapters 3, 4, and

5 all investigate overflow periods, but, there are some differences. Chapter 3 primarily focuses on the *frequency* of overflows, whereas Chapters 4 and 5 investigate the *'severity'* of an overflow by considering the duration and other related properties of congestion periods. Hence, Chapter 3 is essentially concerned with the stationary behavior, while Chapters 4 and 5 focus on the transient behavior.

In Chapter 4 we obtain explicit recursive expressions via which all moments and covariances of the quantities of congestion periods can be obtained. We derive explicit equations, for instance, we write $\mathbb{E}D_{\mathrm{C}}^2$ explicitly in terms of a starting condition $\mathbb{E}D_0^2$. We also present formulae for these starting conditions (which directly relate to the busy period in the $\mathrm{M}/\mathrm{M}/\infty$ queue). Further, we also define a C-*intercongestion* period, a period during which the number of customers is continuously *below* level C, and provide numerical evidence that the intercongestion period can be used as an approximation of a congestion period, which solves the difficulties of obtaining a starting condition. The presented results appeared as [117], which is an excerpt of the more comprehensive version [116].

Chapter 5 is also devoted to C-congestion periods of an $\mathrm{M}/\mathrm{M}/\infty$ queueing system, but now the goal is to shed light on the tail probabilities $\mathbb{P}(D_{\mathrm{C}} > x)$, $\mathbb{P}(A_{\mathrm{C}} > x)$, and $\mathbb{P}(N_{\mathrm{C}} > x)$ for x large. In the so-called many-flows scaling, we show that the tail asymptotics are essentially exponential in the scaling parameter. The proof techniques stem from large-deviations theory; we also identify the *most likely way* in which the event under consideration occurs. In the same scaling, we approximate the model by its Gaussian counterpart. We derive the tail asymptotics for the Gaussian counterpart. Then we use change-of-measure arguments to find upper bounds, uniform in the model parameters, on the probabilities of interest. These change-of-measures are applied to devise importance-sampling schemes, for fast simulation of rare-event probabilities. They turn out to yield a substantial speed-up in simulation effort, compared to straightforward simulations. This chapter is based on [91].

1.4.2 Impact of resource sharing on the performance of wireless ad-hoc networks

The main contribution of this part is the fluid-modeling approach of a central node in a wireless ad-hoc network; the central node is used by other nodes as a relay node to forward their traffic to destinations that cannot be reached directly due to their limited transmission range. This system is modeled by fluid sources that feed into a queue where the input rate into the queue as well as the output rate depend on the current state of the system, e.g., the number of active source nodes. The fluid model captures the essential features of a wireless ad-hoc network, in particular, the way resources are shared among the nodes.

In Chapter 6 we first explain wireless ad-hoc networks in more detail, and then we introduce the fluid model and the performance metrics. We also provide an

overview of the literature on performance modeling of wireless ad-hoc networks.

In Chapter 7, which is based on [14, 115], we analyze the fluid model introduced in Chapter 6. The primary aim is to obtain insightful expressions for the expected transfer time of a flow, i.e., the duration from the moment that a source node starts to transmit a flow till the moment that the relay node forwards the last packet. In particular, we consider the impact of the resource sharing between the relay node and the source nodes, where the relay node may obtain a different share of the capacity than each of the source nodes. In the analysis we first consider the special case of exponential flow sizes; then the model falls in the framework of so-called fluid queues with feedback. We exploit this framework to analyze the source-node dynamics, as well as the workload at the relay node. Interestingly, we observe from extensive numerical experimentation over a broad set of parameter values that the distribution of the number of active source nodes is (practically) *insensitive* to the flow-size distribution. By using this remarkable (empirical) result as an approximation assumption, we obtain explicit expressions for general flow-size distributions for the mean workload at the relay node and the mean overall flow transfer time.

Chapter 8 addresses the validation of the fluid-modeling approach for wireless ad-hoc networks. The fluid model and resulting expressions are validated by simulations of the actual communication system that include all details of the IEEE 802.11 Wireless LAN protocols (see, e.g. [63]). First, we consider the system where all nodes equally share the capacity which corresponds to the 'plain' IEEE 802.11 version. For the validation we compare simulations of the actual communication system with simulations of the fluid model to validate the fluid-modeling approach. We also compare these results with numerical evaluations of the expressions that were analytically obtained in Chapter 7. Next, we consider the situation where the relay node may obtain a different share of the capacity, which relates to the IEEE 802.11e version that allows for QoS differentiation. We first obtain a mapping of the IEEE 802.11e differentiation parameters to the fluid-model parameters, and then we validate by means of simulations that the fluid model accurately describes the system's behavior. This chapter appeared as [114].

In Chapter 9 we study a special case of the fluid model with exponentially distributed flow-sizes and equal sharing of the common capacity, i.e., each source node and the relay node receive the same share of the capacity. We characterize the distributions of performance metrics by their Laplace transforms. In addition, the corresponding tail probabilities of the performance metrics are studied using LD theory. Recall that the results in Chapter 7 are restricted to the *mean values* of the performance metrics. These results are published in [90].

Chapter 2

Fluid models and performance-evaluation methodologies

This chapter provides more comprehensive descriptions of the fluid-modeling approach used in this thesis. It first addresses the underlying fluid models (Section 2.1), and next (Section 2.2) the performance-evaluation methodologies that were concisely introduced in Section 1.3.

2.1 Fluid modeling

In this section we describe a number of fluid models of traffic sources (Section 2.1.1), and we introduce the concept of the fluid queue, i.e., a queue fed by fluid input (Section 2.1.2).

2.1.1 Fluid modeling of traffic sources

In Section 1.3.1 it was already claimed that traffic sources can be modeled as if these are 'fluid', i.e., as if the traffic sources continuously transmit traffic. In this section we justify this fluid-modeling approach, and provide a definition of a fluid source and highlight some of its properties. Subsequently, we present fluid models for: i) individual traffic sources, and ii) the aggregated traffic of many sources, cf. Section 1.3.1.

Rationale for fluid modeling

Fluid modeling of traffic sources by a continuous process is a widespread methodology, e.g. see [8, 10, 11, 18, 46, 100]. The idea behind this approach is to abstract from the discrete nature of packets, which we explain after we have introduced two different levels of abstraction: the packet level and the flow level, e.g., see [111].

The *packet level* considers traffic, potentially the aggregate of multiple users, at the granularity of individual packets. The associated time scale is in the order of the transmission time of a packet, i.e., order of milliseconds, and the traffic rate is highly variable due to the dynamics of both users and their services involved. Typical packet-level performance metrics include the packet delay and delay variation.

The *flow level* focuses on *flows*, to be understood as a succession of packets that comprise an instance of a service, e.g., a web page, voice call or IP-TV stream. The main events of interest are initiations of new flows and departures of finished flows (e.g., after completion of a flow transfer). Frequently used flow-level performance metrics include, e.g., the transfer time or average transmission rate of a flow; note that these strongly affect the QoS of elastic services.

Fluid modeling exploits the fact that the flow-level time scale substantially exceeds that of the packet-level; the packet-level fluctuations vanish if traffic is considered over longer time periods. In fluid modeling a source is modeled as if it continuously transmits traffic. A simple example is a streaming service, e.g., voice, that, when a user is active, generates packets of size B at constant time intervals of length $\triangle t$; this active state is modeled as a fluid source with continuous transmission rate $B/\triangle t$. This modeling approach can be applied to all kinds of traffic sources, e.g., individual flows or the aggregated traffic of many users, as long as the behavior is considered on a time scale that is long enough to smoothen out the packet-level fluctuations.

The advantage of the fluid-modeling approach is that it provides more tractable models compared to models that include packet-level details, and therefore allow, in many cases, for more tractable analysis. Fluid models still capture the essences of the underlying system; in particular, they include the effects of a varying number of active users related to the initiations of new flows and the completion of flow transfers. This modeling approach can be used to investigate properties of flows, e.g., durations of flow transfers. In addition, characteristics of the aggregate traffic at a time scale shorter than that associated with individual flows can be examined, e.g., overflow periods (cf. Section 1.2.1).

Fluid traffic source $A(t)$

Let $A(t) := \{A(t), t \in \mathbb{R}\}$ be a continuous-time stochastic process, where $A(t)$ denotes the *amount of traffic* generated in the interval $[0, t)$. Observe that this process is cumulative, in the sense that the amount of traffic $A(s, t)$ that arrives in interval $[s, t)$ is $A(s, t) = A(t) - A(s)$. Furthermore, in this thesis we only consider sources with stationary increments, i.e., $A(s, t) \stackrel{\mathrm{d}}{=} A(t - s)$, for all $s, t \in \mathbb{R}$.

Directly coupled with the amount of offered traffic $A(t)$ is the so-called *instantaneous rate* $R(t)$, i.e., the rate at which the source generates traffic at epoch t. Then $R(t) := \lim_{s \uparrow t} A(s, t)/(t - s)$. It is noted, however, that $R(t)$ is not always a well-defined notion, e.g., for some Gaussian sources (see page 14).

Fluid-flow modeling of *individual* traffic sources

It is a widely used approach to model an individual flow as a continuous fluid source, see e.g. [11, 18, 100]. The traffic source is often described by an ON-OFF-

process, where the source is in the ON-state when the user is active, i.e., when involved in a flow transfer, and in the OFF-state when inactive. The transmission rate, constant or variable, during the ON-state depends on the type of service that it represents.

Traffic source with a fixed transmission rate when active. Some streaming services, e.g., VOIP conversations, generate packets at constant intervals, when ON. As argued before, this behavior during the ON-state can be modeled as a source with a constant (positive) rate, and with rate 0 when OFF. This type of traffic source is used in Part I, e.g., in Chapter 3 to model ADSL users with generally distributed ON-durations.

Traffic source with a variable transmission rate when active. Elastic services, such as web-browsing or file-sharing, are rate-adaptive when ON. The traffic rate depends on the current state of the network, e.g., the number of active users. For example, in a link with a limited capacity C each source can transmit at rate $r_n := \mathrm{C}/n$ if in total n users are active. This type of traffic sources is used in Part II to model wireless nodes.

Markov-modulated fluid source. A Markov-modulated fluid source is defined by a continuous-time Markov chain N_t which transmits at rate r_{N_t} if the system is in state N_t. Then the amount of offered traffic $A(t)$ is obtained by

$$A(t) := \int_{\tau=0}^{t} r_{N_\tau} \mathrm{d}\tau.$$

A Markov-modulated fluid source can be used to model a broad class of traffic sources: individual traffic sources (including the two examples presented above in case of exponentially distributed ON-durations), but it can also model the superposition of multiple (multi-rate) sources. This modeling approach is used in Part II, in particular in Chapters 7 and 9, to model the arrival process of a superposition of multiple wireless source nodes. An overview of Markov fluid-models is presented in [51], the authors summarize the basic concepts and the potential use of these models.

Fluid models for the *aggregate traffic* of many sources

The traffic on high-speed transmission links in communication networks is, in many cases, the aggregate of many users each having a small access rate. Studies in literature, e.g., see [3, 8, 10, 46, 73, 97, 102], reveal that the aggregate traffic can often be modeled by a Gaussian process. The literature provides a theoretical foundation for the Gaussianity of network traffic based on Central Limit Theorem (CLT)-type of arguments. The Gaussianity is empirically validated by analyzing traffic traces of real networks.

The Gaussianity of the aggregate traffic requires a sufficient level of aggregation. In [73] the authors investigated the required aggregation, and distinguish between

horizontal and vertical aggregation. By horizontal aggregation we mean the minimal length of the time interval that one should consider before the traffic may be modeled by a Gaussian process; if the time scale is too short one typically observes packet-level behavior. Vertical aggregation involves the number of users that is required for the aggregate traffic to exhibit Gaussian behavior.

More recent studies on this topic include [46, 97]. In [97] the authors elaborate on the horizontal aggregation and examine whether the Gaussianity holds for various time scales. They observe that *when* the traffic is Gaussian for a certain time scale, this usually holds for a wide range of 'adjacent' time scales. The authors of [46] conclude that the required level of vertical aggregation is at least 50 Mbits. However, this can only be considered a temporarily useful rule of thumb; the continuous developments of new services and the increasing access rates cause the required level of vertical aggregation, expressed as the traffic volume, to grow over time.

Gaussian source. We have argued that the aggregate traffic of many sources can be modeled by a Gaussian process under certain conditions. A traffic source $A(\cdot)$ is called a *Gaussian source* if it is a Gaussian process with stationary increments, i.e., for all $s < t$,

$$A(s,t) \stackrel{\mathrm{d}}{=} N\left(\mu \cdot (t-s), v(t-s)\right),$$

where μ is the mean traffic rate and $v(\cdot)$ the variance function.

Gaussian sources have a number of characteristics that make them suitable to model network traffic, see e.g. [85, Section 2.6]. These characteristics are:

- *Stationary increments.* The distribution of the increment during interval $[s,t)$ only depends on the length of the interval, i.e., $t-s$, and is independent of the position of the interval.
- *Wide range of correlation structures.* The variance function $v(\cdot)$ allows for a wide range of correlation structures over time.
- *Extreme irregularity of the traffic rate.* The Gaussian source model can incorporate extreme traffic-rate fluctuations, where the instantaneous traffic rate can be obtained as $R(t) := \lim_{s \uparrow t} A(s,t)/(t-s)$. Recall that this traffic rate is not always well-defined for Gaussian sources.

An example of a Gaussian process is the so-called integrated Ornstein-Uhlenbeck (iOU) source which has variance function $v(t) = 2\lambda\mu^{-3}(t\mu - 1 + e^{-t\mu})$. Interestingly, this process is the so-called *Gaussian counterpart* of an $\mathrm{M}/\mathrm{M}/\infty$ process with arrival rate λ and service rate μ, i.e., the expectations and variances of the amount of offered traffic in interval $[0,t)$ coincide. This process has a well-defined rate function $R(t)$; in particular, it is the Gaussian counterpart of the number of active users in an $\mathrm{M}/\mathrm{M}/\infty$ system.

Another Gaussian process that is widely used for network-traffic modeling is fractional Brownian motion (fBm) with variance function $v(t) = t^{2H}$ where H is the Hurst-parameter. This process has a long-range dependent correlation structure for $H > 1/2$, behavior that was empirically discovered by, e.g., see [46, 80]. The fBm process is, after appropriate scaling, the limiting process of the superposition of many ON-OFF sources with heavy-tailed ON- or OFF-durations, see e.g. [33, 127].

2.1.2 Fluid queue

A queue fed by a (superposition of) fluid source(s) $A(t)$ is called a *fluid queue*. The queue typically has a limited service rate C. In case the instantaneous rate $R(t)$ of the traffic source exceeds the capacity C, the amount of work in excess of C is backlogged in the queue; the workload in the queue at time t is denoted by $W(t)$. The queue works at full rate C whenever it is backlogged ($W(t) > 0$), and when the traffic arrival rate equals or exceeds the service rate (i.e., $R(t) \geq$ C).

Next we will present a derivation of the steady-state buffer workload, which can be written as a functional of the arrival process $A(\cdot)$. As an illustrative example we consider a *slotted* cumulative arrival process $A(\cdot)$ that generates an amount of work X_n in the n-th time slot, i.e., $A(-n,-1) := X_{-n} + \cdots + X_{-1}$. We can relate the workload W_0 at slot 0 to the workload W_{-1} at slot -1:

$$W_0 = \max(0, W_{-1} + X_{-1} - \mathrm{C}).$$

After applying this step k times we obtain

$$W_0 = \max(W_{-k} + A(-k,-1) - k\mathrm{C},\ A(-k+1,-1) - (k-1)\mathrm{C},\ \ldots \\ \ldots, A(-1,-1) - \mathrm{C},\ 0).$$

This relation is known as *Lindley's recursion*. For stability of the queue it is required that the expected amount of work $\mathbb{E}X_i$ that arrives during a slot is less than the service rate, i.e., $\mathbb{E}X_i < \mathrm{C}$. For a stable queue there exists a k such that $W_{-k} = 0$. Now observe that when κ denotes the last slot that the queue was empty, i.e., $W_{-\kappa} = 0$ and $W_{-k} > 0$ for $0 < k \leq \kappa$, then $W_0 = A(-\kappa) - \kappa\mathrm{C}$. Hence, it is readily verified that

$$W_0 \stackrel{\mathrm{d}}{=} \sup_{n \geq 0} A(-n,-1) - n\mathrm{C}.$$

This means that the stationary distribution of the fluid queue is in distribution equivalent to the distribution of the maximum of the 'free process' $A(-n,-1) - n\mathrm{C}$. This procedure also applies to continuous-time processes and leads to

$$W_0 \stackrel{\mathrm{d}}{=} \sup_{t \geq 0} A(-t,0) - \mathrm{C}t. \tag{2.1}$$

This identity is often attributed to Reich [106]. Note that this concept can be used irrespective of whether or not the rate process $R(t)$ is well-defined.

In this thesis we use some variants of the fluid queue: a Markov-modulated fluid queue, a Markov-modulated fluid queue *with so-called feedback*, and a Gaussian queue, which are explained next.

Markov-modulated fluid queue. In the analysis of Chapters 5 and 9 we use a Markov-modulated source that feeds into a fluid queue. Let N_t follow a continuous-time Markov chain on $\{0, \ldots, n_{max}\}$ with generator matrix Q, and let at time t the traffic be generated at rate r_{N_t}. Seminal work on this model was presented in the papers [5, 77, 99]. A key contribution is the derivation of the stationary joint distribution of (N_t, W_t) which is defined as

$$F_n(x) := \lim_{t\to\infty} P(W_t \leq x; N_t = n) = \mathbb{P}(W \leq x; N = n).$$

The dynamics of the fluid queue is described as a linear system of differential equations: the workload satisfies the Kolmogorov forward equations: $F'(x)R = F(x)Q'$, where $R := \text{diag}\{r_0, \ldots, r_{n_{max}}\}$. The solution of this system is, under mild regularity conditions, given in the form of the following spectral expansion:

$$F(x) = \sum_{j=0}^{n_{max}} a_j v_j \exp(z_j x),$$

where (z_j, v_j) is an eigenvalue-eigenvector pair, i.e., a scalar and vector that solve $z_j v_j R = v_j Q$. Let ω denote the stationary distribution of N_t, then $\omega_n = F_n(\infty)$, and further observe that $\mathbb{P}(W \leq x) = \sum_n F_n(x)$.

The left graphs of Figure 2.1 present an M/M/∞-input process that feeds into a fluid queue, i.e., flows arrive according to a Poisson process (λ) with a maximum of n_{max} flows, and each flow generates traffic at rate 1 for an exponentially (μ) distributed duration. This system is a Markov-modulated fluid queue, and it is used in the analysis of Chapter 5. The top left graph displays the number of active sources, the bottom left graph the evolution of the buffer workload.

Markov-modulated fluid-queue with feedback. An extension of the above-explained model comprises the transmission rates to also depend on the workload at the fluid queue, so-called *feedback*. This behavior is modeled as a system with generator matrix Q in case of an empty fluid queue (i.e., $W_t = 0$), and it behaves according to a generator matrix $\bar{Q}$ in case the queue is non-empty ($W_t > 0$). Observe that N_t does not constitute a Markov chain. This model was examined in e.g. [88, 119, 120], and it used in the analysis of Chapter 7.

Gaussian queue. The fluid queue is called a *Gaussian queue* when the input process $A(t)$ is a Gaussian source. Gaussian queues are notoriously hard to analyze; in particular, exact results are only available for *Brownian motion* and *Brownian bridge* input

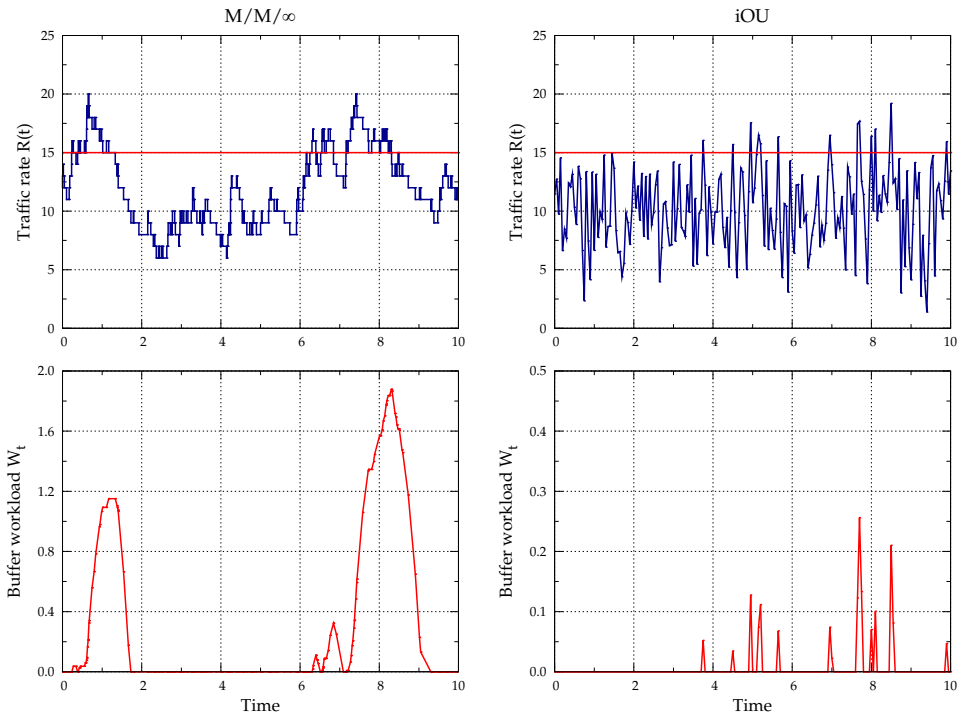

Figure 2.1: *Traffic rates $R(t)$ and evolution of the workload W_t of the fluid queue with $\lambda = 10$, $\mu = 1$ and $C = 15$. Left: M/M/∞ sources with unit rate. Right: iOU source.*

processes. For other Gaussian processes one has to rely on approximations, e.g. see [85, Section 5.4].

An interesting result is that a fluid queue fed by n exponential ON-OFF sources converges to a Gaussian queue with an iOU input process, under a special parametrization, e.g. see [78]. The Gaussian queue, and in particular, fed by iOU input processes is analyzed in Chapter 5. The right panels of Figure 2.1 present a Gaussian queue fed by an iOU source; the top right panel presents the instantaneous traffic rate, and the bottom right graph the evolution of the Gaussian queue.

2.2 Performance-evaluation methodologies

In the analysis of the performance metrics under consideration we use a number of methodologies which are explained in more detail in the following sections.

2.2.1 Transforms

On several occasions we analyze the *transform* of a random variable to obtain more solvable algebraic equations. Recall from Section 1.3.2 that, for a non-negative random variable X, the Laplace transform (LT) $\mathbb{E}\exp(-sX)$ and the moment generating function (MGF) $M(\theta) := \mathbb{E}\exp(\theta X)$ are defined as

$$\mathbb{E}e^{-sX} = \int_{x=0}^{\infty} e^{-sx} f(x)\mathrm{d}x, \qquad \mathbb{E}e^{\theta X} = \int_{x=0}^{\infty} e^{\theta x} f(x)\mathrm{d}x,$$

where $f(x)$ is the probability density function of the random variable X. Further recollect that the LT of the sum of two independent random variables is the product of the LTs of the individual random variables. For an overview of Laplace transforms and their properties we refer to the standard text of Kleinrock [75, Appendix I].

Transforms sometimes allow for exact inversion if they consist of (a combination of) standard transforms. In addition:

- The moments of a random variable can be obtained by differentiation of the transform; the k-th derivative of the MGF at 0 yields the k-th moment, in case of the LT the derivative has to be multiplied by a factor $(-1)^k$.

- Application of the Chernoff bound. The Chernoff bound is an upper bound for the tail probability of a random variable and it is derived by inserting the MGF into the Markov inequality. Recall that the classical Markov inequality yields $\mathbb{P}(X \geq a) \leq (\mathrm{E}X)/a$ for a non-negative random variable X; if we substitute X by $\exp(\theta X)$ we obtain the following upper bound:

  $$\mathbb{P}(X \geq a) = \mathbb{P}\left(e^{\theta X} \geq e^{\theta a}\right) \leq \mathbb{E}e^{\theta X - \theta a} = e^{-\theta a} M(\theta).$$

  Because this holds for *all* non-negative θ, we choose the *tightest* upper bound:

  $$\mathbb{P}(X \geq a) \leq \inf_{\theta \geq 0} e^{-\theta a} M(\theta). \tag{2.2}$$

  Unfortunately, the *Chernoff bound* is rather implicit in that it still requires an optimization over θ. Its major advantage, however, is that it does not require an inversion of $M(\theta)$.

- Transforms can be numerically inverted. In particular, Abate and Whitt have developed theory on the numerical inversion, for a survey, e.g., see [1, 2, 66].

In this thesis transforms are used throughout almost all chapters. For example: in Chapter 9 the performance metrics are characterized by their LTs, Chapters 3 and 5 rely on the Chernoff bound to obtain upper bounds on tail probabilities, and in Chapter 7 a particular performance metric is obtained by inversion of its LT.

2.2.2 Large deviations

From a performance engineering point of view there is special interest in *rare events*, i.e., events that occur with a small probability. For example, in communication network the packet loss ratio is often restricted to a certain value in the order of 10^{-4} to 10^{-6}. *Large deviations* (LD) *theory* focuses on rare events and can be used to estimate tail probabilities.

The scope of large deviations theory is on rare events that occur due to an accumulation of events, e.g., a large buffer content that occurs due to arrival of a large number of (large) flows; LD is less suitable to analyze quantities that can be caused by a single event (as typically occurs in heavy-tailed scenarios). For standard textbooks on large deviations we refer to [34]. In the context of this thesis, the book of Mandjes [85] and the work of Shwartz and Weiss [123] are of particular interest.

In this thesis we rely on specific instances of LD theory, i.e., the sample-mean LD and the sample-path LD.

Sample-mean large deviations. Consider a sequence of n i.i.d. samples $X_1, \ldots, X_n$ that are distributed as a random variable X with mean $\mu = \mathbb{E}X$. We are interested in the probability that the sample mean $n^{-1}\sum_{i=1}^{n} X_i$ deviates from μ. First we define

$$f(n) := \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i > a\right),$$

where $a > \mu$. An upper bound on this probability follows directly from the Chernoff bound (Expression (2.2)), i.e.,

$$\begin{aligned} f(n) &= \mathbb{P}\left(\exp\left(\theta\sum_{i=1}^{n} X_i\right) > e^{n\theta a}\right) \leq \inf_{\theta\geq 0} e^{-n\theta a}\mathbb{E}\left(\exp\left(\theta\sum_{i=1}^{n} X_i\right)\right) \\ &\leq \inf_{\theta\geq 0} e^{-n\theta a}\left(M(\theta)\right)^n = e^{-nI(a)}, \end{aligned}$$

where

$$I(a) := \sup_{\theta\geq 0}\left(\theta a - \log M(\theta)\right).$$

The function $I(a)$ is usually called the convex conjugated or Fenchel-Legendre Transform. An important result in large deviations is *Cramér's theorem*, which builds on the *large deviation principle* (LDP), see e.g. [34, 85, 123], from which can be concluded

$$\lim_{n\to\infty}\frac{1}{n}\log f(n) = -I(a).$$

Hence, the decay rate of $f(n)$ is of an exponential nature for increasing n and the decay rate is given by $I(a)$. Observe that the upper bound of the decay rate results from the Chernoff bound, whereas the lower bound is more difficult to prove.

Sample-path large deviations. Large deviations can also be applied to *stochastic processes*. The idea is rather similar as for the sample mean, but now we consider a sequence of stochastic processes $X_1(\cdot), \ldots, X_n(\cdot)$ and their mean sample path $n^{-1}\sum_{i=1}^{n} X_i(\cdot)$. We are interested in the probability that the mean sample path falls in a collection of paths S, i.e.,

$$\mathbb{P}\left(n^{-1}\sum_{i=1}^{n} X_i(\cdot) \in S\right) \approx \exp\left(-n \inf_{f\in S} \mathbb{I}(f)\right).$$

The functional $\mathbb{I}(\cdot)$ assigns 'costs' to any path, and $\inf_{f\in S}\mathbb{I}(f)$ represents the exponential decay rate of the mean sample path being in set S. Another interesting result is that *if* the sample-mean path is in set S, with overwhelming probability this path is close to path $f^\star \in S$, which is the path that minimizes the functional $\mathbb{I}(\cdot)$ over the set S; the path $f^\star$ is therefore called the *most-likely path*.

For sample-path LD we rely on two results: the framework of Shwartz and Weiss [123] for Markov-modulated traffic sources, and the generalized version of Schilder's theorem for Gaussian processes, see e.g. [85].

Sample-path large deviations for Markov-modulated sources. In the framework presented in Shwartz and Weiss [123] a crucial role is played by the *local rate function*. In case of a birth-death process, this function is defined as

$$I_x(u) := \sup_{\vartheta} \left(\vartheta u - \lambda_x(e^{\vartheta}-1) - \mu_x(e^{-\vartheta}-1)\right), \tag{2.3}$$

where λ_x and μ_x are the state-dependent birth- and death-rates. In fact, the local rate function measures the 'cost' of moving in direction u, when the mean process is in state x. The next step is to define the *action functional* $\mathbb{I}(f)$, which represents the 'cost' of the mean process $n^{-1}\sum_{i=1}^{n} X_i(\cdot)$ following a path $f(\cdot)$:

$$\mathbb{I}(f) := \int_{-\infty}^{\infty} I_{f(s)}(f'(s))\mathrm{d}s.$$

Analogously to the sample mean LD a large deviations principle can be stated which says that the decay rate of probability that the mean sample follows a path in the set S is

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(n^{-1}\sum_{i=1}^{n} X_i(\cdot) \in S\right) = -\inf_{f\in S} \mathbb{I}(f).$$

This framework is used in Chapter 5.

Sample-path large deviations for Gaussian sources. The *generalized version* of Schilder's theorem considers sample-path large deviations for Gaussian processes. We consider the probability that the sample mean of a sequence $X_1(\cdot), X_2(\cdot), \ldots$ of i.i.d. Gaussian

processes is in a set S, then, informally, Schilder's theorem provides us the corresponding decay rate:

$$\mathbb{P}\left(n^{-1}\sum_{i=1}^{n} X_i(\cdot) \in S\right) \approx \exp\left(-n \inf_{f\in S} \mathbb{I}(f)\right) = \exp\left(-\frac{n}{2} \inf_{f\in S} \|f\|_R^2\right),$$

where $\|\cdot\|_R$ is a norm for paths within a well-defined path space R, the so-called *reproducing kernel Hilbert space*; for more details see e.g. [85, Section 4.2].

Unfortunately the rate function of a given path, $\mathbb{I}(f)$, can usually not be given in closed-form. Notably, in case of Brownian motion sources the rate function can be calculated explicitly, due to the independence of the increments. In Chapter 5 the case of iOU sources is considered, which also leads to rather explicit results due to its specific characteristics, e.g., the Markovian nature of the rate process. For other Gaussian sources the rate functions cannot be given explicitly, to the best of our knowledge.

2.2.3 Simulations

Another methodology to estimate a performance metric is by using *simulations*. In this thesis we use simulations as a validation tool for i) the *analysis*-phase, and ii) the *modeling*-phase. Besides the 'direct simulations', we also use importance sampling (IS), in particular, in case of rare-event simulation.

Simulation for validation purposes

Validation of the analysis. Exact analysis of a stochastic model is not always possible and one has to use assumptions or approximations to obtain explicit or insightful expressions. These assumptions and approximations can be validated by numerically comparing the analytically obtained expressions with simulations of the fluid model.

Validation of the model. In the modeling phase the communication network and users' behavior are formulated in terms of a stochastic model. Typically, details of the system are simplified in order to obtain a tractable model. These modeling assumptions can be validated by comparing results of the real communication system to the results from the model; this is done by simulating both the real telecommunication system in all its details, as well as the (fluid) model. The appropriateness of the (fluid) model can be assessed by comparing the outputs of these simulations.

Simulation methodologies

We use two simulation methods, so-called direct simulations and importance sampling, both are described below shortly.

Direct simulations. Direct simulations can be applied to both the actual (communication) system as well as to the (fluid) model of the system. In simulations one draws samples of the arrival process and service requirements to use these as inputs of an emulation of the behavior of the model or actual communication system while keeping track of the desired performance metrics. An introductory textbook on simulation techniques is by Law and Kelton [79].

Suppose we are interested in the probability that a random variable obtains a value in set $\mathscr{E}$, i.e. $\mathbb{P}(E \in \mathscr{E})$. This probability is estimated by observing many occurrences $E_1, \ldots, E_n$ of the random variable, and then the probability is estimated by $\alpha(\mathscr{E})$ in the following manner:

$$\alpha(\mathscr{E}) := \mathbb{P}(E \in \mathscr{E}) = \mathbb{E}\left[I(\mathscr{E})\right] = \frac{1}{n}\sum_{i=1}^{n} I_i(\mathscr{E}),$$

where $I_i(\mathscr{E})$ is the indicator function which is 1 if $E_i \in \mathscr{E}$ and 0 otherwise.

An indicator of the accuracy of an estimator is provided by the width of the confidence intervals, which depends directly on the variance of the estimator. A rule of thumb, that is often used, is that the width of the confidence interval should be less than, say, 10% of the estimator itself. From the definition of the confidence interval it can be seen that the number of required replications is inversely proportional to the variance of a single experiment. This means that direct simulation can be very time-consuming, in particular, if the performance metric corresponds to a rare event. A methodology to reduce the variance, and thereby the simulation time, is importance sampling.

Importance sampling. Importance sampling is a so-called variance-reduction methodology. It uses the idea of change-of-measure, which means that samples are drawn from an other probability measure, e.g., $\mathbb{Q}$, under which the event under consideration is more likely to occur. Consequently, the increased occurrences of this event are compensated for by the *likelihood ratio* $\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{Q}$. The probability we are interested in is then estimated as follows:

$$\alpha(\mathscr{E}) = \mathbb{E}_{\mathbb{Q}}\left[I(\mathscr{E})\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right].$$

In principle any probability measure $\mathbb{Q}$ (for which $\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{Q}$ is well-defined) can be chosen, but an inappropriate choice can result in a variance increase.

Importance sampling is used in Chapter 5 to estimate rare-event probabilities of the quantities related to congestion periods.

Part I

QoS-aware dimensioning of IP-network links

Chapter 3

QoS-aware provisioning of IP-network links

3.1 Introduction

Current bandwidth provisioning procedures for IP-network links are mostly based on simple rules of thumb, using coarse traffic measurements made on a time scale of e.g., 5 or 15 minutes. A crucial question, however, is whether such coarse measurements give any insight into the capacity actually needed: QoS degradation experienced by the users is strongly affected by traffic rate fluctuations on a much smaller time scale, e.g., seconds (file transfers, web browsing) or even less (interactive, real-time applications). Therefore, a thorough understanding is required of the relation between the characteristics of the offered traffic, the link speed, and the resulting QoS (cf. Section 1.1.2). To overcome this problem, we develop a dimensioning procedure which combines coarse traffic-load measurements with traffic modeling to obtain this relation. This enables us to select a link capacity for which the fraction of time that the aggregate rate of the offered traffic exceeds the link speed is less than a predefined (small) value.

A common procedure is to i) use MRTG [101] to get coarse measurement data (e.g., 5 min. intervals), ii) determine the average traffic rate during these 5 min. intervals, and iii) estimate the required capacity by some quantile of the 5 min. measurement data – a commonly used value is the 95% quantile. This procedure is sometimes 'refined' by focusing on certain parts of the day (e.g., office hours, in the case of business customers), or by adding safety and growth margins. The main drawback of this approach, as mentioned before, is that it is not clear how the coarse measurement data relates to the traffic behavior at time scales relevant for QoS.

3.1.1 Contribution

The goal of our work is to develop accurate and reliable provisioning procedures that require a minimal measurement effort. In particular, we derive an 'interpolation' formula that predicts the bandwidth requirement on relatively short time scales (say the order of 1 sec.), *by using large time scale measurements* (e.g., in the order of 5 min.). In our approach we express QoS in terms of the probability (to be interpreted as fraction of time) that, on a predefined time scale T, the traffic supply exceeds the

available bandwidth. The bandwidth C should be chosen such that this probability does not exceed some given bound ε. The time scale T and performance target ε are case-specific: they are parameters of our model, and can be chosen on the basis of the specific needs of the most demanding application involved. We remark that in this setting buffers are not explicitly taken into account; evidently, there is a relation between the time scale T in which the traffic rate exceeds the link rate and the buffer size needed to absorb the excess traffic.

Our approach relies on minimal modeling assumptions. Notably, we assume that the underlying traffic model is Gaussian – empirical evidence for this assumption can be found in e.g., [46, 73]. For the special case of peak-rate constrained traffic (peak rate r), we can use (the Gaussian counterpart of) $\mathrm{M/G/\infty}$ type of input processes [3], leading to an elegant, explicit formula for the required bandwidth; $\mathrm{M/G/\infty}$ corresponds to a flow arrival process that is Poisson with rate λ and flow durations that are i.i.d. as some random variable D (with $\delta := \mathbb{E}D$). We find that, measuring a load $\rho \equiv \lambda\delta r$ (in Mbit/s), the required bandwidth (to meet the QoS criterion) has the form $\rho + \alpha\sqrt{\rho}$. It is clear that the ρ can be estimated by coarse traffic measurements (e.g., 5 or 15 minutes measurements). The α depends on the characteristics of the individual flows, and its estimation requires detailed (i.e., on time scale T) measurements. In many situations, however, there are reasons to believe that the α is fairly constant in time; the estimate needs to be updated only when one expects that the flow characteristics have changed (for instance due to the introduction of new applications).

We expect that the provisioning approach advocated in this chapter extends to several other networking environments. In situations with large numbers of more or less i.i.d. users, the Gaussian assumption will apply due to Central-Limit type of arguments, and hence the procedure followed goes through. In this chapter we present a validation of this approach in an IP setting; future work includes an assessment of the provisioning guidelines in other types of networks.

Apart from its simplicity, our bandwidth provisioning formula $\rho + \alpha\sqrt{\rho}$ has a number of attractive features. In the first place it is *transparent*, in that the impact of changing the 'QoS parameters' (that is, T and ε) on α is explicitly given. Secondly, the provisioning rule is to some extent *insensitive*: α does not depend on λ, but just on characteristics of the individual flows, i.e., the flow duration D and the peak rate r. This property enables a simple estimate of the additionally required bandwidth if in a future scenario traffic growth is mainly due to a change in λ (e.g., due to growth of the number of subscribers). Furthermore, the analytical expression for α provides valuable insight into the impact of changes in D and r. Our bandwidth provisioning rule has been empirically investigated through the analysis of extensive traffic measurements in various network environments with different aggregation levels, user populations, etc.

3.1.2 Literature

There is a vast body of literature on bandwidth provisioning, see for instance [111]. With respect to traffic modeling, it was empirically shown that Poisson packet arrivals do not accurately capture the dependencies present in network traffic [103]. Gaussian approximations do incorporate these dependencies; their use was advocated in several papers, e.g., [3, 46, 73, 100] and Section 2.1.1. The use of flow level traffic models, like the $M/G/\infty$ model (in which flows arrive according to a Poisson process), is justified in, e.g., [11, 18, 102]. In [102] it is pointed out that the $M/G/\infty$ traffic model is extremely flexible, in that it allows all types of dependence structures: by choosing the flow durations Pareto-type one can construct long-range dependent traffic, whereas exponential-type flows lead to short-range dependent traffic. The use of $M/G/\infty$ input is also investigated extensively in [4]; this paper also includes the analysis of a number of dimensioning rules.

The study by Fraleigh *et al.* [46] is related to ours, in that it uses bandwidth provisioning based on traffic measurements to deliver QoS. An important difference, however, is that in their case the performance metric is packet delay (rather than our link rate exceedance criterion). Also, in [46] measurements are used to fit the Gaussian model, and subsequently this model is used to estimate the bandwidth needed; this is an essential difference with our work, where our objective is to minimize the required measurement input/effort, and bandwidth provisioning is done on the basis of only coarse measurements. Another closely related paper is [43], where several bandwidth provisioning rules are empirically validated.

3.1.3 Outline

The remainder of this chapter is organized as follows. In Section 3.2 we describe in detail the objectives of this chapter and the proposed modeling approach; next, we provide the analysis leading to our bandwidth provisioning rule. Numerical results of our modeling and analysis are presented and discussed in Section 3.3. In Section 3.4, the bandwidth provisioning rule is assessed through extensive measurements performed in several operational network environments, and Section 3.5 describes a bandwidth provisioning procedure based on the provisioning rule. Finally, conclusions and topics for further research are given in Section 3.6.

3.2 Objectives, modeling and analysis

The typical network environment that we focus on corresponds to an IP network with a considerable number of users generating mostly TCP traffic (from, e.g., web

browsing, downloading music and video, etc.). Then the main objective of bandwidth provisioning is to take care that the links are more or less 'transparent' to the users, in that the users should not (or almost never) perceive any degradation of their QoS due to a lack of bandwidth. Clearly, this objective (cf. Section 1.2.1) will be achieved when the link rate is chosen such that only during a small fraction of time ε the aggregate rate of the offered traffic (measured on a sufficiently small time scale T) exceeds the link rate. The values to be chosen for the QoS parameters T and ε typically depend on the specific needs of the application(s) involved. Clearly, the more interactive the application, the smaller T and ε should be chosen.

In more formal terms our objective can be stated as follows: the fraction ('probability') of sample intervals of length T in which the aggregate offered traffic exceeds the available link capacity C should be below ε, for prespecified values of T and ε. In other words, with $A(t)$ denoting the amount of traffic offered in $[0, t]$,

$$\mathbb{P}(A(T) \geq \mathrm{C}T) \leq \varepsilon, \tag{3.1}$$

which was earlier introduced as Expression (1.1). For provisioning purposes, the crucial question is: for given T and ε, find the *minimally* required bandwidth $\mathrm{C}(T, \varepsilon)$ to meet the target.

In the remainder of this section we derive explicit, tractable expressions for our target probability $\mathbb{P}(A(T) \geq \mathrm{C}T)$, see Expression (3.1). We do this for a traffic input process $\{A(t), t \geq 0\}$ (cf. Section 2.1.1), for which the only explicit assumption imposed is that $\{A(t), t \geq 0\}$ has *stationary increments*, i.e., for any s, t and $u > 0$ we require that the amount of traffic $A(s + u) - A(s)$ arrived in $[s, s + u]$ has the same distribution as the amount of traffic $A(t + u) - A(t)$ arrived in $[t, t + u]$. In other words: the amount of traffic offered in a certain window depends on the window length only, and does *not* depend on the 'position' of the window. This stationarity will likely hold on time-scales that are not too long (up to, say, hours); on longer time-scales there is no stationarity due to day-patterns, and growth (or decline) of the number of subscriptions (time-scale of weeks, months, ...).

Once we have an expression for (an upper bound to) $\mathbb{P}(A(T) \geq \mathrm{C}T)$, we can find the minimal C required to make sure that this probability is kept below ε. We thus find the required bandwidth $\mathrm{C}(T, \varepsilon)$ – it is expected that this function decreases in both T and ε (as increasing T or ε makes the service requirement less stringent).

3.2.1 General traffic

For the upper bound on $\mathbb{P}(A(T) \geq \mathrm{C}T)$ we apply the *Chernoff bound*, which was derived as Expression (2.2) in Section 2.2.1, and obtain:

$$\mathbb{P}(A(T) \geq \mathrm{C}T) \leq \min_{\theta \geq 0} \left(\mathbb{E} e^{\theta A(T) - \theta \mathrm{C}T} \right). \tag{3.2}$$

Note that $\mathrm{C}(T,\varepsilon)$ could be chosen as the smallest number C such that the right hand side of (3.2) is smaller than ε:

$$\mathrm{C}(T,\epsilon) := \min\left\{\mathrm{C} : \min_{\theta\geq 0}\left(\mathbb{E}e^{\theta A(T)-\theta \mathrm{C}T}\right) \leq \varepsilon\right\}.$$

Rearranging terms, we find that equivalently we are looking for the smallest C such that there is a $\theta \geq 0$ such that

$$\mathrm{C} \geq \frac{\log \mathbb{E}e^{\theta A(T)} - \log\varepsilon}{\theta T}.$$

This C is obviously equal to the minimum of the right-hand side over $\theta \geq 0$:

$$\mathrm{C}(T,\varepsilon) = \min_{\theta\geq 0}\frac{\log \mathbb{E}e^{\theta A(T)} - \log\varepsilon}{\theta T}. \tag{3.3}$$

3.2.2 Explicit formula for Gaussian traffic

Assuming that $A(T)$ contains the contributions of many individual users, it is justified (based on the Central Limit Theorem) to assume that $A(T)$ is *Gaussian* if T is not too small, see Section 2.1.1 or e.g. [46, 73]. In other words $A(T) \sim \mathrm{Norm}(\rho T, v(T))$, for some load ρ (in Mbit/s), and variance $v(T)$ (in Mbits^2). For this Gaussian case we now show that we can determine the right hand side of (3.3) explicitly.

The first step is to compute the moment generating function involved (this is done by isolating the square):

$$\mathbb{E}e^{\theta A(T)} = \exp\left(\theta\rho T + \frac{1}{2}\theta^2 v(T)\right).$$

The calculation of the minimum in (3.3) is now straightforward:

$$\mathrm{C}(T,\varepsilon) = \rho + \min_{\theta\geq 0}\left(\frac{\frac{1}{2}\theta v(T)}{T} - \frac{\log\varepsilon}{\theta T}\right) = \rho + \frac{1}{T}\sqrt{(-2\log\varepsilon)\cdot v(T)}; \tag{3.4}$$

the minimum is attained at $\theta = \sqrt{(-2\log\varepsilon)/v(T)}$.

Evidently, $\mathrm{C}(T,\varepsilon)$ can also be found by first computing the Chernoff bound for Gaussian traffic

$$\mathbb{P}(A(T) \geq \mathrm{C}T) \leq \exp\left(-\frac{1}{2}\frac{(\mathrm{C}-\rho)^2T^2}{v(T)}\right); \tag{3.5}$$

then it is easily checked that (3.4) is the smallest C such that (3.5) is below ε.

As for any input process with stationary increments $v(\cdot)$ cannot increase faster than quadratically (in fact, a quadratic function $v(\cdot)$ corresponds to perfect positive correlation), $\sqrt{v(T)}/T$ is decreasing in T, and hence also the function $\mathrm{C}(T,\varepsilon)$ – the longer T, the easier it is to meet the QoS requirement. Also, the higher ε, the easier it is to meet the requirement, which is reflected by the fact that the function decreases in ε.

REMARK 3.2.1 (EFFECTIVE BANDWIDTH). *There is some reminiscence between formula (3.4) and the effective bandwidth concept proposed earlier in the literature, see, e.g. [39], [40], [70], but there are major differences as well. One of the key attractive properties of effective bandwidths is their 'additivity': if there are two sources, both are assigned a bandwidth, parameterized by the* QoS *criterion, such that their sum represents the bandwidth needed by their superposition:*

$$\mathrm{C}_{1+2}(\varepsilon) = \mathrm{C}_1(\varepsilon) + \mathrm{C}_2(\varepsilon).$$

Importantly, it can be argued that interpreting (3.4) as an effective bandwidth would lead to a bandwidth allocation that is too pessimistic from a cost perspective: noting that

$$\sqrt{v_1(T) + v_2(T)} \leq \sqrt{v_1(T)} + \sqrt{v_2(T)},$$

the amount of bandwidth to be provisioned for the aggregate input could be substantially less than the sum of the individually required bandwidths. ♢

REMARK 3.2.2 (EQUIVALENT CAPACITY FROM GUÉRIN *et al.* [52]). *We remark that Expression (3.4) is of the same spirit as the 'Gaussian' equivalent capacity formula advocated in [52], but some remarks need to be made.*

- *Time scale. The (classical) formula proposed in [52] is of the form*

$$\mathrm{C}(\varepsilon) = \rho + \sqrt{-2\log\varepsilon - \log(2\pi)} \cdot \sigma, \tag{3.6}$$

  where σ^2 is the variance of the 'instantaneous traffic rate' R:

$$\sigma^2 := \mathbb{V}\mathrm{ar}R = \lim_{T\downarrow 0} \mathbb{V}\mathrm{ar}\left(\frac{A(T)}{T}\right) = \lim_{T\downarrow 0} \frac{v(T)}{T^2}.$$

  Hence, $\mathrm{C}(\varepsilon)$ as derived in [52] relates to the time scale $T = 0$, and is in this sense less general than our $\mathrm{C}(T, \varepsilon)$. We remark that for many Gaussian processes σ does not exist; think of fractional Brownian motion with $H < 1$.

- *Exceedance probability: [52]'s approximation vs. Chernoff bound. It is easily verified that (3.6) essentially relies on the approximation*

$$\mathbb{P}(R > \mathrm{C}) \approx \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(\mathrm{C}-\rho)^2}{\sigma^2}\right); \tag{3.7}$$

  the actual value of $\mathbb{P}(R > \mathrm{C})$ is the (complementary) Gaussian distribution function

$$\mathbb{P}(R > \mathrm{C}) = \int_{\mathrm{C}}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x-\rho)^2}{\sigma^2}\right) \mathrm{d}x.$$

  In [52] it is reported that $\mathrm{C}(\varepsilon)$ – based on the approximation (3.7) of the exceedance probability $\mathbb{P}(R > \mathrm{C})$ – is 'a good approximation' of the equivalent capacity, but no

(mathematical) motivation was given. Relying on the approximation $\mathbb{P}(N > x) \sim x^{-1}(2\pi)^{-1/2}\exp(-x^2/2)$, *where* $N \sim \mathrm{Norm}(0,1)$, *we find that* $\mathbb{P}(R > \mathrm{C})$ *reads*

$$\int_{\mathrm{C}}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x-\rho)^2}{\sigma^2}\right) \mathrm{d}x \approx \left(\frac{\sigma}{\mathrm{C}-\rho}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(\mathrm{C}-\rho)^2}{\sigma^2}\right).$$

Hence, if $\mathrm{C} - \rho \approx \sigma$, *then we indeed find approximation (3.7). However, we could not find a rationale for* $\mathrm{C} - \rho$ *being of the same order as* σ. *In fact, we could construct cases in which approximation (3.7) is extremely optimistic, in that it substantially underestimates* $\mathbb{P}(R > \mathrm{C})$. *Hence, its use is not appropriate for provisioning purposes.*

This motivates why our approach above uses the (provably conservative) Chernoff bound (i.e., Expression (3.5)) rather than approximations of the type of Expression (3.7). Also, numerical experiments indicated that there is usually just a modest difference between the capacities based on the Chernoff bound and capacities based on inversion of the (complementary) Gaussian distribution function, where, evidently, the former are always conservative. ♢

The formula for $\mathrm{C}(T,\varepsilon)$ indicates that, given that we are able to estimate the load ρ and the variance $v(T)$ on the 'advertised' time scale T, we have found a straightforward provisioning rule. In the next subsection, we focus on the special case of (the Gaussian counterpart of) $\mathrm{M}/\mathrm{G}/\infty$ input; in that (still quite general) case the expressions simplify further.

3.2.3 M/G/∞ traffic

Whereas the above provisioning formula holds for general Gaussian traffic, we now focus on an important sub-class: Gaussian traffic that has the variance function of the $\mathrm{M}/\mathrm{G}/\infty$ input process, also called the *Gaussian counterpart* of the $\mathrm{M}/\mathrm{G}/\infty$ input process [3]. In the $\mathrm{M}/\mathrm{G}/\infty$ input model, jobs arrive according to a Poisson process with rate λ, and stay in the system during a period that is distributed as the random variable D (i.i.d.), while in the system they generate traffic at rate r. Hence, $\rho = \lambda\delta r$, with $\delta = \mathbb{E}D$. Notice that the $\mathrm{M}/\mathrm{G}/\infty$ traffic model is particularly appropriate in scenarios in which a peak-rate limitation is imposed, see also [3, 102]. As we will see later, by choosing D appropriately, it covers a broad range of correlation structures.

Denote by f_X and F_X the density and the distribution function, respectively, of the random variable X. Let D^r be the residual distribution of D, i.e., $1 - F_D(x) = \delta f_{D^r}(x)$. Now the variance of $A(T) = r\int_0^T N(t)\mathrm{d}t$ (with $N(t)$ denoting the number of flows present at time t) can be explicitly calculated, as follows. As the number of flows present at both time s and time t has a Poisson distribution with mean $\lambda\,\mathbb{E}D\,\mathbb{P}(D^r > |t-s|)$, we obtain that

$$\mathbb{C}\mathrm{ov}(N(t), N(s)) = \lambda\,\mathbb{E}D\,\mathbb{P}(D^r > |t-s|).$$

Hence

$$v(T) = r^2 \mathbb{V}\mathrm{ar} \int_0^T N(t)\mathrm{d}t = r^2 \int_0^T \int_0^T \mathbb{C}\mathrm{ov}(N(t), N(s))\,\mathrm{d}s\mathrm{d}t$$
$$= \rho r \int_0^T \int_0^T \mathbb{P}(D^r > |t-s|)\mathrm{d}s\mathrm{d}t,$$

which can be simplified to the sum of three single integrals, see also [92, 94]:

$$\lambda r^2 \left(2T \int_0^T x(1 - F_D(x))\mathrm{d}x - \delta \int_0^T x^2 f_{D^r}(x)\mathrm{d}x + \delta T^2 (1 - F_{D^r}(T)) \right). \quad (3.8)$$

Hence, cf. Expression (3.4), the required bandwidth $\mathrm{C}(T, \varepsilon)$ can be expressed as

$$\mathrm{C}(T, \varepsilon) = \rho + \alpha\sqrt{\rho}. \quad (3.9)$$

Importantly, α depends exclusively on the characteristics of the individual flows, i.e., the distribution of the flow duration D and the peak rate r (and QoS requirements T and ϵ), but does *not* depend on the flow arrival rate λ – this will turn out to be a key property in our experimental investigations on provisioning that are presented in Section 3.4. Now we evaluate Expression (3.8) for different distributions D, covering both the long-range dependent and short-range dependent case.

Exponential flow durations

For exponentially distributed flow lengths D, the variance $v(T)$ reads

$$v(T) = 2\rho\delta^2 r(e^{-T/\delta} - 1 + T/\delta),$$

such that

$$\alpha = \left(\frac{T}{\delta}\right)^{-1} \sqrt{(-2\log\varepsilon) \cdot 2r(e^{-T/\delta} - 1 + T/\delta)}.$$

Observe that $v(T)$ is, for T large, linear, corresponding to short-range dependent input. Also observe, that α depends on T only through the ratio T/δ.

Pareto flow durations

For Pareto-distributed flow lengths D, i.e., obeying

$$F_D(x) = 1 - \left(\frac{b}{x+b}\right)^a, \quad x \geq 0, \quad (3.10)$$

and $\delta = b/(a-1)$ (where $a > 1$ and $b > 0$), substantial calculus gives (assume for ease $a \neq 2$, $a \neq 3$)

$$v(T) = \frac{2\rho r}{(3-a)(2-a)} \cdot \left(b^{a-1}\left(T+b\right)^{3-a} - (3-a)bT - b^2\right);$$

$$\alpha = \frac{1}{T}\sqrt{(-2\log\varepsilon)\cdot\frac{2r}{(3-a)(2-a)}\cdot(b^{a-1}(T+b)^{3-a}-(3-a)bT-b^2)}.$$

(Notice that [3] uses $F_D(x) = 1-(b/x)^a$ for $x \geq b$, but, as flow sizes do not obey some natural lower bound b, we have chosen to use the more natural 'shifted version' (3.10) instead.) If $a < 2$, $v(T)$ grows 'superlinearly' for large T (in fact, it grows as T^{3-a}), corresponding to long-range dependent input; for $a > 2$, we see that $v(T)$ is essentially linear, cf. [32].

Discussion on the M/G/∞ input model

1. If T is small (i.e., small compared to δ), then α becomes insensitive in the flow duration D. This can be seen as follows. From Expression (3.8) it can be derived that $v(T)/T^2 \to \rho r$ if $T \downarrow 0$. Then Expression (3.4) yields $\mathrm{C}(T,\varepsilon) \approx \rho + \sqrt{(-2\log\varepsilon)\cdot\rho r}$, exclusively depending on ρ, for T small.

   This result can be derived differently, by noting that for $T \downarrow 0$, the performance criterion boils down to requiring that the number of active users does not exceed C/r. It is well-known that the number of active users has a Poisson distribution with mean $\lambda\delta$; this explains the insensitivity.

2. The case of exponential flow lengths can be easily extended to, e.g., *hyperexponentially* distributed flows; a random variable X is hyperexponentially distributed [129, p. 446] if with probability $p \in (0,1)$ it is distributed exponentially with mean δ_1, and else exponentially with mean δ_2. Then the hyperexponential case is just the situation with two flow types feeding independently into the link (each type has its own exponential flow length distribution); note that the variance of the total traffic is equal to the sum of the variances of the traffic generated by each of the different exponential flow types.

3. The above approach assumes that traffic arrives as 'fluid': it is generated at a constant rate r. It is perhaps more realistic to assume that, during the flow's 'life time', traffic arrives as a Poisson stream of packets (of size s); the rate of the Poisson process is γ, where γs is equal to r. Denoting the above, fluid-based, variance function by $v_{\mathrm{f}}(T \mid r)$, and the packet-based variance function by $v_{\mathrm{p}}(T \mid \gamma, s)$, it can be verified that

   $$v_{\mathrm{p}}(T \mid \gamma, s) = v_{\mathrm{f}}(T \mid r) + \rho s T, \tag{3.11}$$

   irrespective of the flow duration distribution. Importantly, the provisioning formula $\mathrm{C}(T,\varepsilon) = \rho + \alpha\sqrt{\rho}$ remains valid (for an α that does not depend on λ).

3.3 Numerical results

This section presents numerical results obtained by using the analytical model of Section 3.2. The goal is to illustrate a few key features of our bandwidth provisioning formula. We use the traffic parameters and QoS parameters displayed in Table 3.1, unless specified otherwise.

Table 3.1: *Default parameter settings for the numerical experiments.*

arrival rate and flow size			QoS parameters			model-specific parameters		
ρ	10	Mbit/s	T	1	sec.	r	1	Mbit/s
D	exponential		ε	0.01	-	γ	83.3	packets/s
δ	1	sec.				s	1500	Bytes

Experiment 1: Fluid model vs. packet-level model.
Figure 3.1 shows the required capacity obtained by the packet-level and fluid model as a function of T, for various mean flow durations δ. It is seen that for large values of the time scale T, both models obtain the same required capacity. This can be understood by looking at the extra term $\rho s T$ of Expression (3.11), which influence on $\mathrm{C}(T, \varepsilon)$ becomes negligible for increasing T, cf. Expression (3.4).

For $T \downarrow 0$ the required capacity obtained by the packet-level model behaves like

$$\mathrm{C}(T, \varepsilon) \sim \rho + \sqrt{\rho s} \frac{1}{\sqrt{T}} \sqrt{-2 \log \varepsilon}$$

and hence $\mathrm{C}(T, \varepsilon) \uparrow \infty$ as $T \downarrow 0$, whereas the required capacity of the fluid model converges to $\rho + \sqrt{(-2 \log \varepsilon) \cdot \rho r}$, as was already argued in Section 3.2.3.

The fast increase in the required capacity in the packet-level model for a decreasing time scale T was also observed in e.g., [46]. Note that, in fact, the required capacity is not influenced by the *absolute value* of T, but rather by the *ratio* of T/δ. The right graph of Figure 3.1 shows the same results as the left graph, but now on a linear axis and only for $T \in [0, 1]$.

In the remainder of this section we restrict ourselves to the flow-level model, as we will focus on situations with values of $T/\delta > 0.1$ for which the required capacity is almost identical in both models.

Experiment 2: Impact of the flow duration distribution.
Next we investigate the impact of the flow duration distribution on the required capacity. Figure 3.2 contains four graphs with results for hyperexponentially distributed flow durations D. Each graph shows, for a particular value of the mean flow size δ, the required capacity as a function of the offered load ρ, for different Coefficients of Variation (CoV) of D. These graphs show that the required capac-

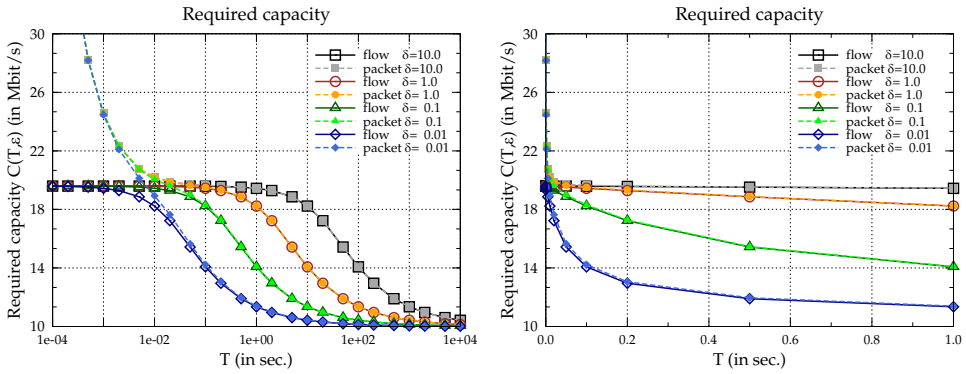

Figure 3.1: *Experiment 1: Comparison of the required capacity for the flow-level and packet-level model as a function of the time scale T. Left: logarithmic axis. Right: linear axis.*

ity is almost insensitive to the CoV for the long ($\delta = 10$ sec.) and short flow durations ($\delta = 0.01$ sec.). For the other cases ($\delta \in \{0.1, 1\}$ sec.) the required capacity is somewhat more sensitive to the CoV. The graphs show that for hyperexponentially distributed flow durations *less* capacity is required if the CoV increases.

It should also be noticed that the required capacity for $T = 0$, also shown in Figure 3.2, corresponds to the often used $M/G/\infty$ bandwidth provisioning approach, cf. the discussion in Section 3.2.3 and the discussion on Experiment 1. The numerical results show that particularly for short flow durations significantly less capacity is required than suggested by the classical $M/G/\infty$ approach; for longer flows this effect is less pronounced.

Experiment 3: Impact of QoS parameter ε.
Figure 3.3 shows the required capacity as a function of the QoS requirement ε, which specifies the fraction of intervals in which the offered traffic may exceed the link capacity. A larger value of ε means relaxing the QoS requirement, and hence less capacity is needed. Obviously, for $\varepsilon \to 1$ the required capacity converges to the long term average load $\rho = 10$. For $\varepsilon \downarrow 0$ the required capacity increases rapidly to infinity (according to $\sqrt{-2\log\varepsilon}$).

Experiment 4: Impact of the CoV of the flow-duration distribution.
To investigate the impact of the flow duration characteristics, we computed the required capacity for exponential, hyperexponential, and Pareto distributed flow durations with different CoV values, see the left panel of Figure 3.4. The graph shows that the required capacity is almost insensitive to the flow duration distribution. The left graph also confirms the earlier observations that the capacity is almost insensitive to the CoV of the flow duration distribution. Note, that for hyperexponentially distributed flow durations the required capacity slightly decreases for increasing CoV,

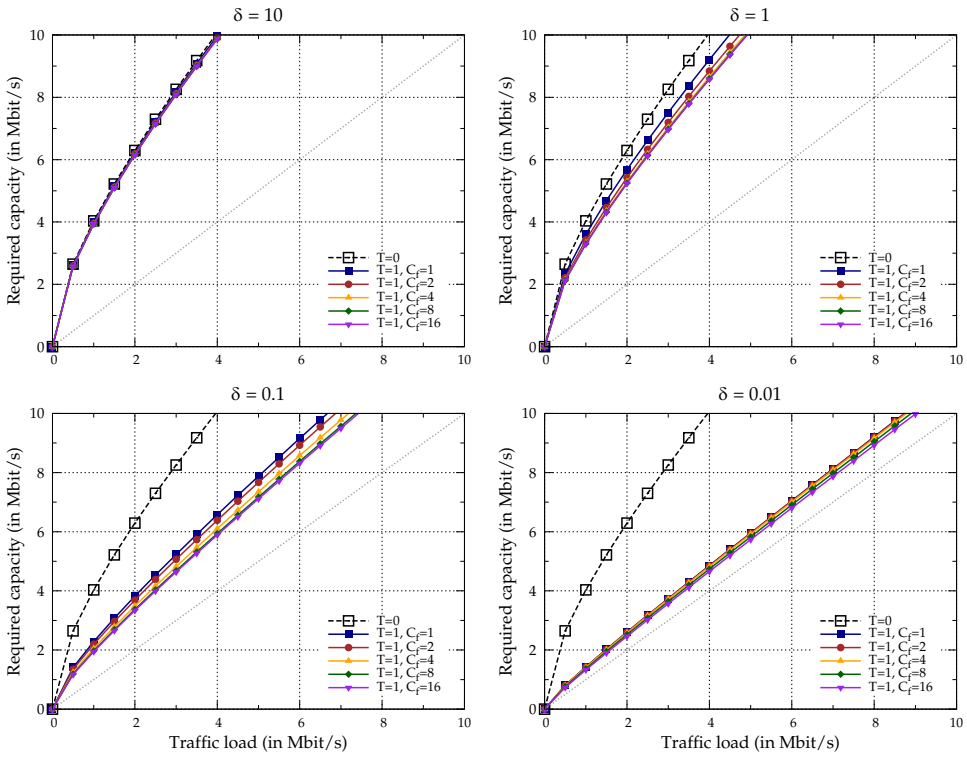

Figure 3.2: *Experiment 2: Required capacity for hyper-exponential flow durations with different means and* CoVs.

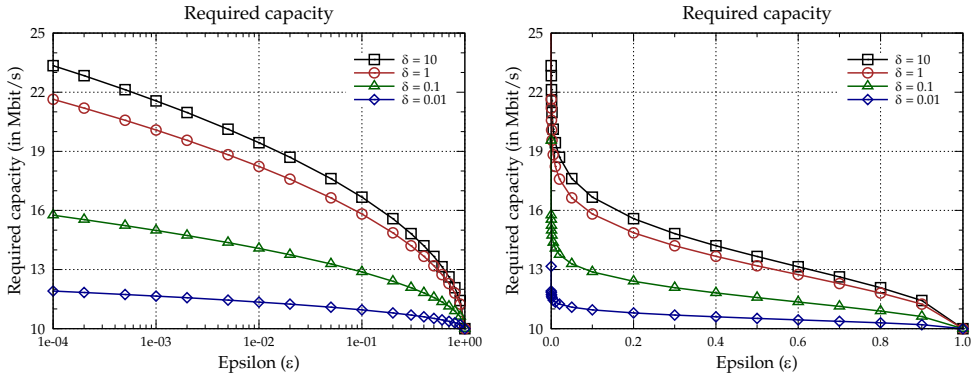

Figure 3.3: *Experiment 3: Comparison of the required capacity for the flow-level and packet-level model as a function of the* QoS *requirement ε. Left: logarithmic axis. Right: linear axis.*

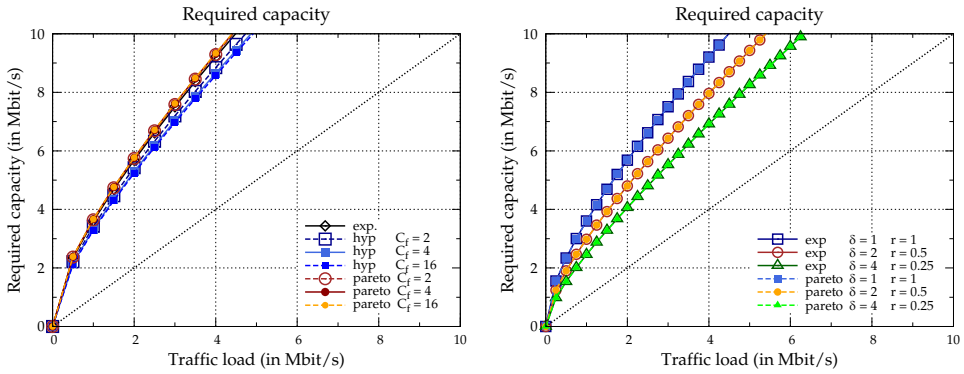

Figure 3.4: *Left. Experiment 4: Required capacity for different flow duration distribution and* CoVs*. Right. Experiment 5: Required capacity for different access rates.*

while for Pareto distributed flow durations the required capacity slightly increases for increasing CoV.

Experiment 5: Impact of the access rate.
Finally, the right panel of Figure 3.4 studies the effect of the access rate r on the required capacity. Three values of the access rate r and the mean flow duration δ are chosen such that the mean flow size $\delta \cdot r$ remains constant. As expected, the required capacity increases considerably when r becomes larger (i.e. the traffic burstiness grows). The results in this graph for hyperexponential and Pareto flow sizes confirm the conclusions from Experiments 2 and 4 that the required capacity is almost insensitive to the flow duration distribution.

3.4 Experimental verification

In this section we will analyze measurement results obtained in operational network environments in order to validate the modeling approach and bandwidth provisioning rule presented in Section 3.2. In particular, we will investigate the relation between measured traffic load values $\hat{\rho}$ during 5 min. periods (long enough to assume stationarity) and the traffic fluctuations at a 1 sec. time scale within these periods.

Clearly, if (A) our $M/G/\infty$ traffic modeling assumptions of Section 3.2.3 apply *and* if (B) differences in the load ρ are caused by changes in the flow arrival rate λ (i.e., the flow size characteristics remain unchanged during the measurement period), then, as a function of ρ, for given (T, ε), the required bandwidth C_ρ should

satisfy

$$C_\rho = \rho + \alpha\sqrt{\rho}, \tag{3.12}$$

for some fixed value of α. To assess the validity of this relation, we have carried out measurements in three different network environments: i) a national IP network providing Internet access to residential ADSL users, ii) a college network, and iii) a campus network.

In the ADSL network environment the main assumptions made in Section 3.2.3 in order to justify use of the $M/G/\infty$ traffic model seem to be satisfied, i.e., the flow peak rates are limited due to the ADSL access rates (which are relatively small compared to the network link rates), and the traffic flows behave more or less independently of each other (the IP network links are generously provisioned and, hence, there hardly is any interaction among the flows). The other network environments have essentially different characteristics. In particular, in the college and campus network the ratio of the access rate and link rate is relatively large, which, obviously, may lead to violation of our traffic modeling assumptions.

In Section 3.5 we demonstrate how to estimate the α in (3.12) directly from measurements of the aggregate traffic at time scale T. An alternative to this approach would be to fit the flow-size distribution, such that α can be computed by inserting this into the explicit formulae of Section 3.2.3. Recall that Experiment 2 of Section 3.3 showed that the CoV of the flow-size has just a modest impact on α. Also, fitting the full flow-size distribution has the evident drawback that it requires per-flow measurements. Therefore, we prefer direct estimation of α.

The measurement scenarios and results will be described and discussed in more detail in the following subsections.

3.4.1 ADSL network environment

We first focus on the ADSL network environment with residential users, see Figure 3.5. An ADSL connection consists of an ADSL modem on both sides of the local loop between the subscriber and the local exchange. On the local exchange side, up to 500 modems are contained in Digital Subscriber Line Access Multiplexers (DSLAM).

The DSLAMs are connected to the core IP infrastructure by means of optical STM-1 (155 Mbit/s) links. The aggregated traffic of all the ADSL subscribers of a certain Internet Service Provider (ISP) is carried over a high-capacity link between the core infrastructure and the ISP. Depending on the size of the ISP, this can vary between a single STM-1 link and multiple Gigabit Ethernet links. At the time of the measurements, none of the network links were saturated, and hence the traffic was not affected by any shortage of capacity in the ADSL network.

We choose the sample size $T = 1$ sec., motivated by the fact that this can be assumed to be the time scale that is most relevant for the Quality of Service percep-

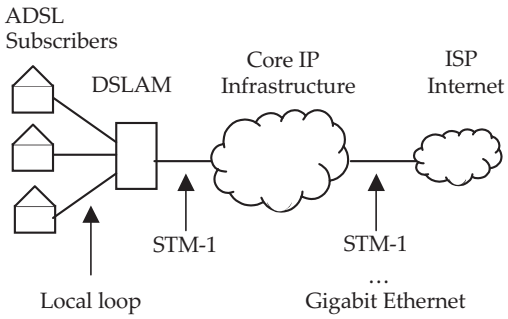

Figure 3.5: *Overview of an* ADSL *infrastructure.*

tion of end-users of typical applications like web browsing. Elementary transactions, such as retrieving single web pages, are normally completed in intervals roughly in the order of 1 sec. If the network performance is seriously degraded during one or several seconds, then this will affect the quality as perceived by the users.

The method that was chosen to measure the traffic at the 1 sec. time scale was to use the internal traffic counters (interface MIBs) of the DSLAMs. These counters keep track of the accumulated number of bytes that are transported on each port in each direction. In this experiment, the counters for the STM-1 ports in the downstream direction (toward the subscribers) were used. The counters were read-out using SNMP.

The measurements were done during several evenings (between 5 PM and 11 PM), as this is the busiest period of the day for ADSL traffic. The measurements were performed on a large number of DSLAMs, in locations ranging from small villages to major cities. Time was split into 5 minute chunks, over which the load ρ is determined for each STM-1 link. In addition, for each 5 min. period, the 99% quantile of the 1 sec. measurements was determined. This quantile was assumed to indicate the minimum capacity C that is needed to fulfill the QoS requirement $\mathbb{P}(A(T) \geq \mathrm{C}T) \leq \varepsilon$, with $\varepsilon = 1\%$ and $T = 1$ sec.

The left graph in Figure 3.6 results from measurements on 11 STM-1 links at various locations. Each location is represented by a distinct color. For orientation purposes, the dotted line shows the unity $(y = x)$ relation. It is remarkable how the 99% quantiles almost form a solid curve. We fitted a function $\rho + \alpha\sqrt{\rho}$, such that roughly 95% of the 99% quantiles are lower or equal to this function. The reason for fitting an upper bound, instead of finding the function that gives the minimum least square deviation, is that eventually we intend to use this function for capacity planning: then it is better to *over*estimate the required bandwidth than to underestimate it. The graph shows an extremely nice fit for the function $\mathrm{C} = \rho + 1.0\sqrt{\rho}$ (with C and ρ expressed in Mbit/s).

At the time of the measurements, the busiest STM-1's did not carry more traffic

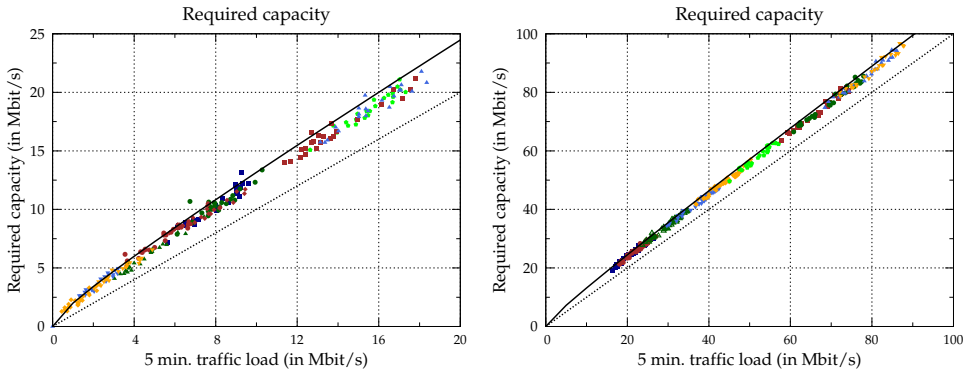

Figure 3.6: *Left:* 99% *maximum* 1 *sec. traffic as function of* 5 *min. traffic mean. Right: synthesized traffic measurements for higher traffic volumes.*

than 20 Mbit/s during the busiest hours, so we could not verify that the found upper bound also holds for higher traffic volumes. To overcome this problem at least partly, we synthesized artificial traffic measurements by taking the superposition of the traffic measured on several (unrelated) STM-1's. The right graph of Figure 3.6 shows the results of this experiment. As expected on theoretical grounds, the fitted function $C_\rho = \rho + 1.0\sqrt{\rho}$ remains valid.

3.4.2 College and campus network

We have performed similar experiments in two other network environments, viz. a college network and a campus network, with essentially different characteristics than the ADSL network. In particular, in these alternative network environments the ratio of the access rate and link rate is relatively small, and, hence, one would expect that the $M/G/\infty$ modeling assumption underlying the analysis in Section 3.2 is not valid anymore. The question is whether (or up to what extent) the bandwidth requirement formula (3.9) still applies.

In the first scenario, we have measured a 1 Gbit/s link connecting a college network to the Internet. This link is shared by about 1000 students and teachers, each having a 100 Mbit/s FastEthernet connection (a ratio of $1 : 10$). In the second scenario, we have measured a 300 Mbit/s (trunked) link connecting an university campus (residential) network to the Internet. This link is shared by some 2000 students, each of them having a 100 Mbit/s connection (a ratio of 1:3). Thus, theoretically, it takes only 10 or 3 users, respectively, to saturate the observed network links.

The left graph of Figure 3.7 shows the measurement results for the college network. As expected, the cloud of 99% quantiles of the 1 sec. traffic rate samples within 5 min. intervals does not form such a nice 'curve' as in the previous (ADSL) scenario,

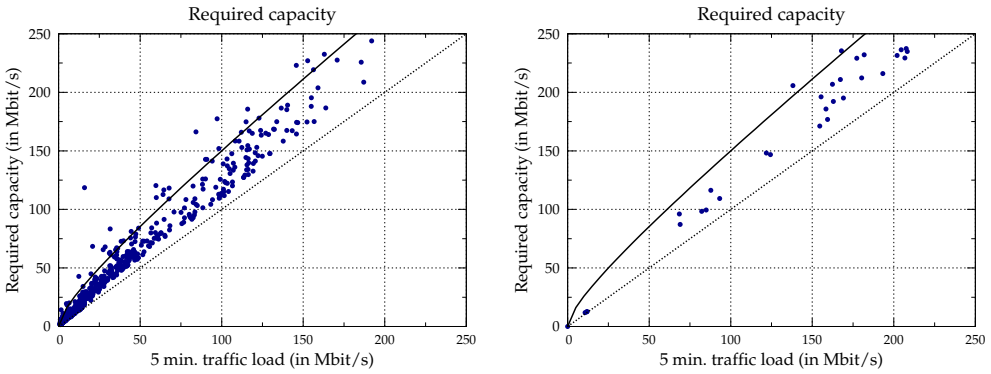

Figure 3.7: *99% maximum 1 sec. traffic as function of 5 min. traffic mean. Left: college network. Right: campus residential network.*

but the typical square-root behavior can still be recognized.

For the university campus network, the 'peak versus average load' is plotted in the right graph of Figure 3.7. Note that the traffic in both directions has been aggregated during the measurements, which explains that the link load as plotted in the graph is sometimes higher than the link capacity (which is one-way). Although the number of measurements available for the campus network is relatively low, we conclude from the graph that the relation between the average link loads and the 99% quantiles of 1 sec. samples shows a similar behavior as in the college network.

From the above results it is concluded that, as expected, for these alternative scenarios our model developed in Section 3.2 does clearly not apply as well as for the ADSL scenario. Indeed, it may be expected that this is caused by the high access link speed, which leads to a possibly high variability in the rate at which traffic is generated by the users in the alternative scenarios (while the $M/G/\infty$ model assumes that sources generate traffic at a fixed rate). Apparently, under these highly variable traffic conditions the 5 min. average traffic rate does not provide sufficient information to estimate the traffic behavior on much smaller time scales (i.e. more detailed information than just ρ is needed), and, consequently, other underlying traffic models should be applied.

3.5 Bandwidth provisioning procedure

Our formula (3.9) for the required bandwidth $C(T, \varepsilon)$ can be used to develop bandwidth provisioning procedures. Obviously, a first step in this procedure is to verify whether the main $M/G/\infty$ modeling assumptions are satisfied in the network envi-

ronment under consideration, such that formula (3.9) can indeed be applied. A next step is then to estimate ρ and α. Clearly, ρ can be estimated through coarse traffic measurements, as it is just the average load; α, however, contains (detailed) traffic characteristics on time scale T (viz., the variance $v(T)$). In particular, as noticed in Section 3.2.3, α depends on the flow peak rate r and on the parameters of the flow duration D, but, importantly, α does *not* depend on the flow arrival rate λ; α can be considered as a characteristic of the individual flows.

This 'dichotomy' between ρ and α gives rise to efficient provisioning procedures. Consider the following two typical situations:

- Situations in which there is a set of links, that differ (predominantly) in the *number* of connected users; across the links, the individual users have essentially the same type of behavior (in terms of the distribution D and the access rate r). Then the α can be estimated by performing detailed measurements at (a part of) the existing links. When a new link is connected, one could obtain an estimate $\hat{\rho}$ of the load by performing coarse measurements (e.g., every 5 min., by using the MRTG tool [101]). Then the provisioning rule $\hat{\rho} + \hat{\alpha}\sqrt{\hat{\rho}}$ can be used. An example is the ADSL scenario described above, in which one could use $\hat{\alpha} \approx 1.0$ to dimension a new link.

- Growth scenarios in which it is expected that the increase of traffic is (mainly) due to a growing number of subscribers (i.e., the λ), while the user behavior remains unchanged. Here it suffices to perform infrequent detailed measurements at time scale T, yielding an estimate $\hat{\alpha}$ of α. If a future load $\hat{\rho}$ is envisaged, the required bandwidth can be estimated by the provisioning rule $\hat{\rho} + \hat{\alpha}\sqrt{\hat{\rho}}$.

The estimate of α has to be updated after a certain period (perhaps in the order of months). This should correspond to the time at which it is expected that the 'nature' of the use of resources changes (due to, e.g., new applications, etc.).

The explicit formulas for α derived in Section 3.2 are also useful when examining the impact of changes in the user behavior or the QoS parameters. For instance, the impact of an upgrade of the access speed r can be evaluated. Also one could assess the effect of imposing a stronger of weaker performance criterion ε: when replacing ε_1 by ε_2, the α needs to be multiplied by $\sqrt{\log \varepsilon_2 / \log \varepsilon_1}$.

The measurement period of 5 min. mentioned above for estimating the load ρ is motivated by the fact that this is the time scale on which measurements in an operational network can be (and are) performed on a routine basis. A higher frequency would be desirable, but this would put a high load on the processing capacity of routers, the transport capacity of management links, etc., particularly if there are many routers and ports involved. On the other hand, measurements performed at lower frequencies (for instance 1 to several hours) are too coarse, as traffic is not

likely to be stationary over such long periods. Therefore, 5 min. will usually be a suitable trade-off, as this is feasible to measure, and at the same time a reasonable period during which the traffic can still be assumed stationary.

3.6 Concluding remarks

In this chapter we have considered bandwidth provisioning for IP network links. Our goal was to develop accurate and reliable provisioning procedures that require a minimal measurement effort. First we derived a formula for the minimally required link bandwidth $\mathrm{C}(T,\varepsilon)$, such that the aggregate traffic rate (measured on a time scale T) exceeds the link rate only during a small fraction ε of time. In particular, for the situation that the traffic is generated by peak rate constrained flows that arrive according to a Poisson process and remain active for some random time D (i.e., $\mathrm{M/G/\infty}$ input traffic) the resulting bandwidth provisioning rule is of the form $\mathrm{C}(T,\varepsilon) = \rho + \alpha\sqrt{\rho}$; here ρ is the traffic load, which can be estimated easily from coarse traffic measurements (typically in the order of 5 min.). Importantly, the coefficient α is determined by characteristics of the individual flows, and does *not* depend on the flow arrival rate λ. We have shown that this property opens up the possibility of elementary (yet adequate) provisioning procedures.

The explicit expression of α shows the impact of the flow size, peak rate and other traffic and system parameters on the required link bandwidth. In particular, α lies somewhere between 0 and $\sqrt{(-2\log\varepsilon)r}$; its exact value depends mainly on the ratio of the time scale of interest T and the mean flow duration. Extensive numerical results show that $\mathrm{C}(T,\varepsilon)$ is quite insensitive to the flow size distribution (apart from its mean value).

The above provisioning rule has been empirically validated through the analysis of extensive traffic measurements in three practical scenarios: i) an IP network connecting private and small business ADSL users to the Internet, ii) a college network and iii) a campus network. A particularly good correspondence with our theoretical results was found from the measurements in the IP network scenario, where the flow rates are bounded by relatively small ADSL access rates. The measurement results for the other scenarios showed, as expected, less good correspondence: the $\mathrm{M/G/\infty}$ modeling assumptions are not really satisfied there; in particular the flow rates may be strongly variable due to the relatively high access rates (compared to the network link rates) in these scenarios.

Topics for further research. It remains for further research whether other underlying traffic models could be used to improve the results for network environments like the college and campus network. An attractive alternative traffic model is the fractional Brownian motion (fBm) model, as used in e.g., [46].

Another topic for further research is the investigation of the validity of the stationarity assumption in our modeling approach. In particular, up to which time scale $\hat{t}$ can the traffic arrival process be assumed stationary? It is clear that the estimates of the average rate ρ should be measured at a time scale smaller than $\hat{t}$. It should also be investigated in more detail under which conditions (and to what extent) the Gaussian traffic assumption is valid, cf. Section 3.2.2.

As a last topic for further research we mention the QoS criterion used in this chapter, i.e., the fraction of time ε that the aggregate offered traffic rate (measured at time scale T) is restricted by the link rate. In particular, we could obtain more insight in the relation between this QoS criterion, which we used as link bandwidth provisioning objective, and the actual QoS that the users are offered. In other words: to what extent does this criterion actually determine the *duration* of a congestion period (this will depend on the traffic characteristics, in particular the flow-level dynamics)? What are appropriate choices of T and ε for different (TCP) application types (file downloading, interactive web browsing, etc.)?

Chapter 4

Moments of congestion periods

4.1 Introduction

In the previous chapter we investigated as a QoS metric the *fraction of time* that a link is in overload; in order to avoid a QoS degradation, the network should be dimensioned such that the fraction of time that the link occupancy exceeds the link speed is kept small. It is noted, however, that QoS as experienced by the users of the network is not only affected by the *frequency* of overload periods, but also by their *durations*; in this (and the following) chapter we focus on these *durations* of overload periods. For that purpose we consider the $M/M/\infty$ queue, and in particular, we investigate so-called C-*congestion periods*, which are defined as periods during which the offered traffic (number of users) is continuously above a certain value C.

We consider an $M/M/\infty$ queueing system where customers arrive according to a Poisson process with arrival rate λ and have an exponential service requirement with mean μ^{-1}. There are an infinite number of identical servers and customers start service immediately upon arrival. The $M/M/\infty$ queueing system can be used as a flow-level model for the occupancy of a link in a communication network, see e.g. Chapter 3.

A C-congestion period is defined as the period during which the number of users present is continuously above level C. In other words: a C-congestion period is the period starting at the epoch that an arriving customer finds C customers in the system, until the first time that a departing customer leaves behind C customers. The duration of a C-congestion period is denoted by D_C. Other interesting quantities which are related to a C-congestion period, are the number of users that arrive during the congestion period, denoted by N_C, and the total amount of work in excess of level C during the C-congestion period, which is the so-called area A_C above level C.

4.1.1 Literature

Keilson [68] studied passage times of a birth-death process by decomposing a passage time into the convolution of congestion periods; due to the general nature of birth-death processes the results are rather implicit. There are several papers that

have studied the congestion period in M/M/∞-queueing systems. Guillemin and Simonian [58] present closed-form expressions for the means of $D_{\rm C}$, $N_{\rm C}$ and $A_{\rm C}$. They also obtained the Laplace transforms (LTs) for the above-mentioned quantities (expressed in terms of special functions), and analyzed the first passage time of level C starting in steady-state. A continued fraction analysis of the duration is presented in [55]. Preater [104] elaborates on the results of Guillemin and Simonian; by using an alternative derivation he finds a more attractive form of the LT of the congestion period. He also presents the *joint* LT of the congestion period triple $\Theta_{\rm C}(D_{\rm C}, N_{\rm C}, A_{\rm C})$ of the duration, number of arrivals and the area. In another paper [105] Preater examines the height of a congestion period, e.g., the maximum level that is reached during a congestion period. Knessl and Yang [76] study $\mathbb{P}(D_{\rm C} > t)$ in several asymptotic regimes. Both Guillemin and Simonian [58] and Preater [104] observe that, when C grows large, a C-congestion period of an M/M/∞ queue behaves similarly to the busy period of the M/M/1-queue. The LT of the duration and number of arriving customers in the busy period of an M/M/1-queue can easily be obtained, and see [57] for an analysis of the area of a busy period. Robert [109] presents an approximation of the order of the mean passage time from level n to level 0 for large n. Progress on systems with heterogeneous servers has been reported recently in Tsybakov [130].

Another related subject of frequent study is the busy period of the M/G/∞ queueing system, which in fact coincides with the congestion period of level 0 (i.e., the 0-congestion period), with generally distributed service times. One of the earliest works on the busy period is by Takács [126]. He presents the LST of the busy cycle duration of a so-called type II counter, which is similar to an M/G/∞ queue. This result is used by others, e.g. Stadje [125] and Liu and Shi [83]. Liu and Shi [83] consider the busy period in $\mathrm{GI}^{\rm X}$/G/∞-queueing systems with batch arrivals and for several special cases they obtain expressions for the first and second moment of both the busy period and busy cycle. A joint LT for both the duration and number of arrivals was already presented by Shanbhag [122].

Although the Laplace transforms of $D_{\rm C}, N_{\rm C}$ and $A_{\rm C}$ are known [58, 104], differentiating these is fairly non-straightforward due to the rather implicit nature of the functions involved. This explains the absence of explicit formulae for higher moments (the means are known) and covariances (between $D_{\rm C}, N_{\rm C}$ and $A_{\rm C}$). Also, so far no attention was paid to C-*intercongestion periods*, which are the periods during which the number of customers in the system is continuously *below* C.

Strikingly little is known about the tail probabilities $\mathbb{P}(D_{\rm C} > x)$, $\mathbb{P}(A_{\rm C} > x)$, and $\mathbb{P}(N_{\rm C} > x)$ (which will be investigated in Chapter 5). By majorizing the M/M/∞ queue by an appropriate M/M/1 queue, cf. [58, 104], upper bounds on the tails can be derived relatively easily, but it is not *a priori* clear how tight these bounds are. We mention here also a related result by Guillemin and Pinchon [56] on the area of a busy period of an M/M/1 queue, stating that its tail distribution decays essentially in a Weibullian way.

4.1.2 Contribution

In this chapter we investigate the moments of the duration, number of arrivals and area swept above C (i.e., D_C, N_C and A_C) for C-congestion periods in an M/M/∞ queue. Recursive relations are derived through which all the moments of the above-mentioned values can be obtained. In particular it is demonstrated that there is a recursive relation between the congestion periods of two adjacent levels, e.g., level C and level C-1: any quantity of level C can be expressed in terms of the same quantity of a $(C-1)$-congestion period. Iterating these, we can express the quantities related to a C-congestion period in terms of the quantities related to a 0-congestion period (which is, as observed above, a busy period of the M/M/∞ queue). For instance, we write $\mathbb{E}D_C^2$ explicitly in terms of $\mathbb{E}D_0^2$. Furthermore, similar recursions are derived for the covariances between the quantities D_C, N_C and A_C.

Thus, in order to solve for the higher moments, we have to find the starting values for our recursion; in our example: to find an expression for $\mathbb{E}D_C^2$ we have to find an explicit formula for $\mathbb{E}D_0^2$. The derivation of these starting values can be done through the differentiation of the LT of these busy-period related quantities. In particular, explicit expressions for the first and second moments are presented. In addition to this, we find the covariances $\mathbb{C}\text{ov}(D_C, N_C)$, $\mathbb{C}\text{ov}(D_C, A_C)$ and $\mathbb{C}\text{ov}(N_C, A_C)$. With $\mathbb{E}D_C, \mathbb{E}N_C$ and $\mathbb{E}A_C$ being known, this reduces to finding the 'joint expectations' $\mathbb{E}[D_C N_C], \mathbb{E}[D_C A_C]$ and $\mathbb{E}[N_C A_C]$. Again, we first express these in terms of the busy-period quantities (for example, $\mathbb{E}[D_C N_C]$ is phrased in terms of $\mathbb{E}[D_0 N_0]$), and then the busy-period related starting condition is solved. Theoretically, all moments (joint expectations) of the quantities of a C-congestion period can be obtained by differentiating the LT of the quantities (from Preater's [104] congestion triple), but practically this is far from trivial. It is considerably easier to obtain the moments (and joint expectations) of the busy-period quantities and to insert these as the starting conditions into the recursive relations.

Analogously to a C-congestion period, a C-intercongestion period is defined as the period that the number of users is continuously *below* level C. The analysis and results for the quantities duration, number of arrivals, and the area below C are presented, which are also recursive relations for the moments and covariances. Again, the recursion can be solved in terms of the quantities of level 0. Importantly, these relate to the period that the system has *less* than 0 customers; hence, all moments and joint expectations of the quantities are 0. The recursion has attractive numerical properties: it is more stable than those of the C-congestion periods. In addition, similarly to Preater's derivation of the LT of a congestion triple [104], the LT of the intercongestion triple is derived.

Guillemin and Simonian [58] and Preater [104] already observed that, for large C, the busy period of an M/M/1-queue can be used to approximate the behavior of a C-congestion period of an M/M/∞ queue. The approximation works well for large

C, but, not for C close to the average number of users in the system ρ. We present results indicating that the quantities of a C-congestion period can be approximated accurately by a $\rho - (\mathrm{C} - \rho)$-intercongestion period (which has, as indicated above, favorable numerical properties). The approximation works particularly well for C close to ρ, and can consequently be used complementary to the above-mentioned M/M/1-based approximation.

4.1.3 Outline

The outline of this chapter is as follows. Section 4.2 introduces the notation and illustrates how a transient period of an M/M/∞ queue can be subdivided into C-congestion periods. Section 4.3 presents the recursion schemes for the first and second moment of the $D_{\mathrm{C}}, N_{\mathrm{C}}$ and A_{C}. The recursions are solved resulting in closed-form expressions which still contain the starting condition: for instance, $\mathbb{E}D_{\mathrm{C}}^2$ is explicitly written in terms of $\mathbb{E}D_0^2$. Similarly, Section 4.4 yields the derivation of the covariances of the quantities in terms of the covariances relating to the busy period: $\mathbb{E}[D_{\mathrm{C}}N_{\mathrm{C}}]$ is presented in terms of $\mathbb{E}[D_0N_0]$. In Section 4.5 the first and second moments of D_0 as well as the joint expectation $\mathbb{E}[D_0N_0]$ are obtained. These busy-period quantities are then the 'starting conditions' of the recursions of Sections 4.3 and 4.4. For the first and second moments of N_0 and A_0 and for the joint expectations $\mathbb{E}[D_0A_0]$ and $\mathbb{E}[N_0A_0]$ we refer the reader to Section 5 of [116]. Section 4.6 presents the definition, analysis and results for C-intercongestion periods. Section 4.7 provides some numerical results and illustrates that an $\rho - (\mathrm{C} - \rho)$-intercongestion period can be used as an accurate approximation of a C-congestion period when C is close to ρ. Section 4.8 concludes this chapter.

4.2 Model and preliminaries

4.2.1 Definitions

Consider an M/M/∞ queue with arrival rate λ and mean service requirement μ^{-1}. For convenience we introduce the notation $\nu_n := \lambda + n\mu$. The average workload of the system is denoted by $\rho = \lambda/\mu$. Let the Markov process $\Lambda_t \in \{0, 1, 2, \ldots\}$ denote the number of customers in the system at time t. Let

$$D_j(i) := \quad \inf\{t > 0 : \Lambda_t = j \mid \Lambda_0 = i\}, \qquad i > j, \tag{4.1}$$

$$N_j(i) := \quad \#\{t : \Lambda_t - \Lambda_{t-} = 1, 0 < t \leq D_j(i)\}, \qquad i > j,$$

$$A_j(i) := \quad \int_{t=0}^{D_j(i)} (\Lambda_t - j)\mathrm{d}t, \qquad i > j, \tag{4.2}$$

where $\Lambda_{t-} := \lim_{\epsilon \downarrow 0} \Lambda_{t-\epsilon}$. Then, $D_j(i)$ is the first passage time of state j from state i, $N_j(i)$ the number of arrivals during this first passage time $D_j(i)$, and $A_j(i)$ is the

area above j during the same period of time. Note that Guillemin and Simonian (GS) [58] have a slightly different interpretation of the number of arrivals[1].

An important sub-class of these passage times is the class of C-congestion periods. A C-congestion period is the duration until the first return to level C after an arriving customer raised the number of users above level C. So, a C-congestion period is the period that the system is continuously above level C. Duration D_{C} is defined by (4.1) where $i = \mathrm{C}+1$ and $j = \mathrm{C}$. For short-hand notation we introduce $D_{\mathrm{C}} := D_{\mathrm{C}}(\mathrm{C}+1)$, $N_{\mathrm{C}} := N_{\mathrm{C}}(\mathrm{C}+1)$ and $A_{\mathrm{C}} := A_{\mathrm{C}}(\mathrm{C}+1)$. The special case where $\mathrm{C} = 0$ is called the 'busy period'.

4.2.2 Decomposition of a passage time into congestion periods

By its definition $D_j(i)$ is a stopping time of the Markov process Λ_t. It can be decomposed as the sum of the hitting times D_{i-1} and $D_j(i-1)$. The strong Markov property states that these hitting times are independent. The first component is already a congestion period and the second term can be decomposed repeatedly in a similar way and finally results in the following equality in distribution:

$$D_j(i) \;=\; \sum_{k=j}^{i-1} D_k, \tag{4.3}$$

where the D_k for $k = j, \ldots, i-1$ are independent. Expression (4.3) resembles Expression (5.1.1) of Keilson [68].

The number of arrivals $N_j(i)$ and the area $A_j(i)$ can also be decomposed, based on the decomposition of the duration $D_j(i)$, resulting in

$$N_j(i) \;=\; \sum_{k=j}^{i-1} N_k, \tag{4.4}$$

$$A_j(i) \;=\; \sum_{k=j}^{i-1} \left(A_k + (k-j)D_k\right). \tag{4.5}$$

Proof Equation (4.4) follows directly due to (4.3). Equation (4.5) is obtained because area $A_j(i)$ can be decomposed in a similar way as $D_j(i)$ and $N_j(i)$, but caution is required because of the definition of the area. $A_j(i)$ can be decomposed into the terms A_{i-1} and $A_j(i-1)$, but A_{i-1} only consists of the area above level $i-1$, ignoring the area between $i-1$ and j for the duration D_{i-1}. The missing area for A_{i-1} is $(i-1-j)D_{i-1}$ and correction of all terms A_k leads to (4.5). □

[1] GS [58] include the arrival that starts a C-congestion period. Formally this arrival did not occur within the C-congestion period as the customer entered the system when only C customers where present. Preater [104] also ignores the arrival that initiates the congestion period.

This subdivision of the passage times into the sum of independent congestion periods simplifies the analysis. All the moments can be directly derived from the moments of the individual congestion periods, e.g., for the duration it yields

$$\mathbb{E}D_j(i) = \sum_{k=j}^{i-1} \mathbb{E}D_k \quad \text{and} \quad \mathbb{E}D_j^2(i) = \sum_{k=j}^{i-1} \mathbb{E}D_k^2 + 2\sum_{k=j}^{i-2}\sum_{l=k+1}^{i-1} \mathbb{E}D_k\mathbb{E}D_l.$$

4.2.3 Analysis of a C-congestion period

In this section a recursive relation for the quantities of a C-congestion period are derived using straightforward analysis.

A C-congestion period is initiated by a customer who finds C other customers in the system upon arrival. The number of customers is increased to $\mathrm{C}+1$ and the system will remain at this level for an exponentially $\nu_{\mathrm{C}+1}$ distributed time, as both the interarrival time and the service times are exponentially distributed. The next transition of the system is caused either by the arrival of a new customer or by the departure of one of the $\mathrm{C}+1$ customers present. With probability $(\mathrm{C}+1)\mu/\nu_{\mathrm{C}+1}$ the next transition is a departure, which immediately ends the currently ongoing C-congestion period. With probability $\lambda/\nu_{\mathrm{C}+1}$ the next transition is initiated by an arrival, which increases the number of customers to $\mathrm{C}+2$; then the remaining duration of the C-congestion period is the duration of a transient period $D_{\mathrm{C}}(\mathrm{C}+2)$.

Let T_{C} be the duration that the system remains at level C, which is exponentially $\nu_{\mathrm{C}+1}$ distributed, and define random variable P_{C} as

$$P_{\mathrm{C}} = \begin{cases} 1 & \text{with probability } \lambda/\nu_{\mathrm{C}} \\ 0 & \text{with probability } \mathrm{C}\mu/\nu_{\mathrm{C}}. \end{cases}$$

Notice that, as P_{C} is Bernoulli distributed all moments are the same: $\mathbb{E}P_{\mathrm{C}}^k = \lambda/\nu_{\mathrm{C}}$ for all k. Now, for the duration of a C-congestion period D_{C} the above reasoning leads to:

$$D_{\mathrm{C}} \;=\; T_{\mathrm{C}+1} + P_{\mathrm{C}+1}D_{\mathrm{C}}(\mathrm{C}+2) = T_{\mathrm{C}+1} + P_{\mathrm{C}+1}\left(D_{\mathrm{C}+1} + D'_{\mathrm{C}}\right). \tag{4.6}$$

Here X' denotes an independent, statistically identical copy of X. By the memoryless property of the exponential distribution all the random variables, e.g., $T_{\mathrm{C}+1}$, $P_{\mathrm{C}+1}$ and $D_{\mathrm{C}}(\mathrm{C}+2)$, are mutually independent. Expression (4.6) is a recursive relation which illustrates that the duration of a C-congestion period can be expressed in terms of the duration of a $(\mathrm{C}-1)$-congestion period. By repeated iterations the duration can be expressed in terms of D_0, which is the duration of a busy period.

For the quantities N_{C} and A_{C} the following similar relations can be derived:

$$N_{\mathrm{C}} = \quad P_{\mathrm{C}+1}\left(1 + N_{\mathrm{C}}(\mathrm{C}+2)\right) \quad = P_{\mathrm{C}+1}\left(1 + N_{\mathrm{C}+1} + N'_{\mathrm{C}}\right). \tag{4.7}$$

$$A_{\mathrm{C}} = \quad T_{\mathrm{C}+1} + P_{\mathrm{C}+1}A_{\mathrm{C}}(\mathrm{C}+2) \quad = T_{\mathrm{C}+1} + P_{\mathrm{C}+1}\left(A_{\mathrm{C}+1} + D_{\mathrm{C}+1} + A'_{\mathrm{C}}\right). \tag{4.8}$$

Note that by definition of N_{C} the (possible) arrival that ends $T_{\mathrm{C}+1}$ and initiates the passage time $D_{\mathrm{C}}(\mathrm{C}+2)$ is not accounted for in $N_{\mathrm{C}}(\mathrm{C}+2)$ and has to be accounted for separately. The second equality of (4.8) follows directly from (4.5).

4.3 Quantities of a C-congestion period

In this section we present the mean and second moment of the duration, number of arrivals and the area swept above C. For the quantity duration the mean, the second moment and also higher moments are written out. Next, the moments of the number of arrivals and the mean of the area are rather trivial, the second moment of the area is more complicated as definition (4.8) includes a term D_{C} which requires the joint expectation of the quantities D_{C} and A_{C}.

4.3.1 Duration of a C-congestion period

For the derivations of the moments of the duration we use result (4.6) of Section 4.2.3. Although the expected duration of a congestion period is already given in Guillemin and Simonian [58], the derivation of the mean duration is presented to become acquainted with the methodology of the recursions.

Mean duration of a C*-congestion period*. Taking the expectation on both sides of Expression (4.6) yields

$$\mathbb{E}D_{\mathrm{C}} \;=\; \mathbb{E}\left[T_{\mathrm{C}+1} + P_{\mathrm{C}+1}\left(D_{\mathrm{C}+1} + D'_{\mathrm{C}}\right)\right] = \frac{1}{\nu_{\mathrm{C}+1}} + \frac{\lambda}{\nu_{\mathrm{C}+1}}\left(\mathbb{E}D_{\mathrm{C}+1} + \mathbb{E}D_{\mathrm{C}}\right).$$

By isolating $\mathbb{E}D_{\mathrm{C}+1}$ at the left side, we obtain the following expression:

$$\mathbb{E}D_{\mathrm{C}+1} = \frac{(\mathrm{C}+1)\mu\mathbb{E}D_{\mathrm{C}} - 1}{\lambda}. \tag{4.9}$$

Expression (4.9) is a difference equation and illustrates that the mean duration of a $(\mathrm{C}+1)$-congestion period depends on the mean duration of C-congestion period. By iteration $\mathbb{E}D_{\mathrm{C}+1}$ (or preferably $\mathbb{E}D_{\mathrm{C}}$) can be expressed in terms of $\mathbb{E}D_0$, which is the expected duration of a busy period. This yields the following closed-form expression

$$\mathbb{E}D_{\mathrm{C}} = \frac{\mathrm{C}!}{\rho^{\mathrm{C}}}\mathbb{E}D_0 - \frac{\mathrm{C}!}{\lambda\rho^{\mathrm{C}}}\sum_{j=1}^{\mathrm{C}}\frac{\rho^j}{j!} = \frac{\mathrm{C}!}{\lambda\rho^{\mathrm{C}}}\sum_{j=\mathrm{C}+1}^{\infty}\frac{\rho^j}{j!}, \tag{4.10}$$

where $\mathbb{E}D_0$ is obtained via renewal arguments. Let π_0 denote the fraction of time that the system is empty, T_{idle} the duration that the system is empty, and T_{busy} the duration that the system is busy. As $\pi_0 = e^{-\rho}$, $\mathbb{E}T_{idle} = 1/\lambda$, $\mathbb{E}D_0 = \mathbb{E}T_{busy}$ and $\pi_0 = \mathbb{E}T_{idle}/(\mathbb{E}T_{busy} + \mathbb{E}T_{idle})$ it follows that $\mathbb{E}D_0 = (e^{\rho} - 1)/\lambda$.

Second moment of duration of a C*-congestion period.* The second moment of the duration can also be obtained by taking the second moment of Expression (4.6). Then we obtain

$$\begin{aligned}\mathbb{E}D_{\mathrm{C}}^2 &= \mathbb{E}\left[T_{\mathrm{C}+1} + P_{\mathrm{C}+1}\left(D_{\mathrm{C}+1} + D_{\mathrm{C}}\right)\right]^2 \\ &= \mathbb{E}\left[T_{\mathrm{C}+1}^2 + 2T_{\mathrm{C}+1}P_{\mathrm{C}+1}\left(D_{\mathrm{C}+1} + D_{\mathrm{C}}\right) + P_{\mathrm{C}+1}^2\left(D_{\mathrm{C}+1}^2 + 2D_{\mathrm{C}+1}D_{\mathrm{C}} + D_{\mathrm{C}}^2\right)\right] \\ &= \frac{2}{\nu_{\mathrm{C}+1}^2} + \frac{2\lambda}{\nu_{\mathrm{C}+1}^2}\left(\mathbb{E}D_{\mathrm{C}+1} + \mathbb{E}D_{\mathrm{C}}\right) + \frac{\lambda}{\nu_{\mathrm{C}+1}}\left(ED_{\mathrm{C}+1}^2 + 2\mathbb{E}D_{\mathrm{C}+1}\mathbb{E}D_{\mathrm{C}} + \mathbb{E}D_{\mathrm{C}}^2\right).\end{aligned}$$

as $\mathbb{E}[D_{\mathrm{C}+1}D_{\mathrm{C}}] = \mathbb{E}D_{\mathrm{C}+1}\mathbb{E}D_{\mathrm{C}}$ by the strong Markov property. Rearranging leads to the following difference equation:

$$\mathbb{E}D_{\mathrm{C}+1}^2 = \frac{\mathrm{C}+1}{\rho}\mathbb{E}D_{\mathrm{C}}^2 - \frac{2}{\lambda\nu_{\mathrm{C}+1}} - \frac{2}{\nu_{\mathrm{C}+1}}\left(\mathbb{E}D_{\mathrm{C}+1} + \mathbb{E}D_{\mathrm{C}}\right) - 2\mathbb{E}D_{\mathrm{C}+1}\mathbb{E}D_{\mathrm{C}}.$$

This equation can be solved in terms of $\mathbb{E}D_0^2$, the second moment of the duration of a busy period which is treated in Section 4.5.2, and yields

$$\begin{aligned}\mathbb{E}D_{\mathrm{C}}^2 = \frac{\mathrm{C}!}{\rho^{\mathrm{C}}}\mathbb{E}D_0^2 &- 2\frac{\mathrm{C}!}{\rho^{\mathrm{C}}}\sum_{j=1}^{\mathrm{C}}\frac{\rho^j}{j!}\frac{1}{\nu_j}[\mathbb{E}D_{j-1} + \mathbb{E}D_j] \\ &-2\frac{\mathrm{C}!}{\rho^{\mathrm{C}}}\sum_{j=1}^{\mathrm{C}}\frac{\rho^j}{j!}\mathbb{E}D_{j-1}\mathbb{E}D_j - \frac{2}{\lambda}\frac{\mathrm{C}!}{\rho^{\mathrm{C}}}\sum_{j=1}^{\mathrm{C}}\frac{\rho^j}{j!}\frac{1}{\nu_j},\end{aligned} \tag{4.11}$$

Higher moments of the duration of a C*-congestion period.* Higher moments can also be obtained using the recursive relation (4.6), although calculations are more tedious. By using the binomium theorem for both $\mathbb{E}D_{\mathrm{C}}^n$ and $\mathbb{E}D_{\mathrm{C}}^n(\mathrm{C}+2)$ we obtain:

$$\begin{aligned}\mathbb{E}D_{\mathrm{C}}^n &= \sum_{l=0}^{n}\binom{n}{l}\mathbb{E}T_{\mathrm{C}+1}^{n-l}\,\mathbb{E}[P_{\mathrm{C}+1}D_{\mathrm{C}}(\mathrm{C}+2)]^l \\ &= \frac{n!}{\nu_{\mathrm{C}+1}^n} + \frac{\lambda}{\nu_{\mathrm{C}+1}}\sum_{l=1}^{n-1}\binom{n}{l}\frac{(n-l)!}{\nu_{\mathrm{C}+1}^{n-l}}\sum_{k=0}^{l}\binom{l}{k}\mathbb{E}D_{\mathrm{C}+1}^k\mathbb{E}D_{\mathrm{C}}^{l-k} \\ &\quad + \frac{\lambda}{\nu_{\mathrm{C}+1}}\sum_{l=0}^{n}\binom{n}{l}\,\mathbb{E}D_{\mathrm{C}+1}^l\,\mathbb{E}D_{\mathrm{C}}^{n-l}.\end{aligned}$$

Rearranging leads to a difference equation which can be solved in terms of $\mathbb{E}D_0^n$:

$$\begin{aligned}\mathbb{E}D_{\mathrm{C}}^n &= \frac{\mathrm{C}!}{\rho^{\mathrm{C}}}\mathbb{E}D_0^n - \frac{\mathrm{C}!}{\rho^{\mathrm{C}}}\sum_{j=1}^{\mathrm{C}}\frac{\rho^j}{j!}\sum_{l=1}^{n-1}\binom{n}{l}\frac{(n-l)!}{\nu_j^{n-l}}\sum_{k=0}^{l}\binom{l}{k}\mathbb{E}D_j^k\mathbb{E}D_{j-1}^{l-k} \\ &\quad -\frac{\mathrm{C}!}{\rho^{\mathrm{C}}}\sum_{j=1}^{\mathrm{C}}\frac{\rho^j}{j!}\sum_{l=1}^{n-1}\binom{n}{l}\mathbb{E}D_j^l\mathbb{E}D_{j-1}^{n-l} - \frac{n!}{\lambda}\frac{\mathrm{C}!}{\rho^{\mathrm{C}}}\sum_{j=1}^{\mathrm{C}}\frac{\rho^j}{j!}\frac{1}{\nu_j^{n-1}}.\end{aligned} \tag{4.12}$$

From Expression (4.12) it can be observed that the n-th moment of level C depends on all moments $\mathbb{E}D_{\text{C}}^m$ for $m < n$ and $\mathbb{E}D_k^m$ for $k < \text{C},\ m \leq n$. This illustrates that for $\mathbb{E}D_{\text{C}}^n$ all moments $\mathbb{E}D_0^m$ for $m = 1, \ldots, n$ have to be known. This is a drawback as closed-form expressions for the second and higher moments are not presented in literature. An expression for $\mathbb{E}D_0^2$ will be derived in Section 4.5.2. The method can also be used for higher moments, but the calculations become substantially more tedious.

4.3.2 Number of arriving customers during a C-congestion period

The mean and second moment are obtained by taking the expectation of Expression (4.7) and the square of Expression (4.7) respectively.

Mean number of arriving customers in a C-congestion period. Taking the expectation of Expression (4.7) and rearranging leads to a difference equation in terms of $\mathbb{E}N_0$, which is the number of arrivals during a busy period. $\mathbb{E}N_0$ is easily obtained as $\mathbb{E}N_0 = \lambda\mathbb{E}D_0 = e^\rho - 1$ and the solution of the difference equation is the following closed-form expression:

$$\mathbb{E}N_{\text{C}} = \frac{\text{C}!}{\rho^{\text{C}}}\mathbb{E}N_0 - \frac{\text{C}!}{\rho^{\text{C}}}\sum_{j=0}^{\text{C}-1}\frac{\rho^j}{j!} = \frac{\text{C}!}{\rho^{\text{C}}}\sum_{j=\text{C}+1}^{\infty}\frac{\rho^j}{j!}. \tag{4.13}$$

Second moment of the number of arriving customers. The second moment is derived in terms of $\mathbb{E}N_0^2$ in a similar manner and yields

$$\mathbb{E}N_{\text{C}}^2 = \frac{\text{C}!}{\rho^{\text{C}}}\mathbb{E}N_0^2 - \frac{\text{C}!}{\rho^{\text{C}}}\sum_{j=1}^{\text{C}}\frac{\rho^j}{j!}\left(1 + 2\mathbb{E}N_j\mathbb{E}N_{j-1} + 2\mathbb{E}N_j + 2\mathbb{E}N_{j-1}\right). \tag{4.14}$$

For the derivation of $\mathbb{E}N_0^2$ we refer to Section 5.3 of [116].

4.3.3 Area swept above C during a C-congestion period

The mean and second moment can be obtained by using Expression (4.8).

Mean area swept above C. Taking the expectation of (4.8) leads to a difference equation that can be solved iteratively in terms of $\mathbb{E}A_0$. $\mathbb{E}A_0$, the area above 0 during a busy period, can be obtained by observing that the system is a renewal process of cycles consisting of busy and idle period. The average workload ρ during a cycle should all be obtained during a busy period. Then $\rho = \mathbb{E}A_0/(\mathbb{E}D_0 + 1/\lambda)$ and thus $\mathbb{E}A_0 = \rho e^\rho$. Finally, we obtain the following closed-form expression for $\mathbb{E}A_{\text{C}}$:

$$\mathbb{E}A_{\text{C}} = \frac{\text{C}!}{\rho^{\text{C}}}\mathbb{E}A_0 - \sum_{j=1}^{\text{C}}\frac{\text{C}!}{j!}\left(\mathbb{E}D_j + \frac{1}{\lambda}\right) = \frac{1}{\lambda}\frac{\text{C}!}{\rho^{\text{C}}}\sum_{j=\text{C}+1}^{\infty}\frac{\rho^j}{j!}. \tag{4.15}$$

Second moment of the area swept above C. Taking the second moment of (4.8) and isolating $\mathbb{E}A_{\mathrm{C}+1}^2$ leads to a difference equation. The difference equation includes the 'joint expectation' $\mathbb{E}[D_{\mathrm{C}+1}A_{\mathrm{C}+1}]$, which results from the term $\mathbb{E}[A_{\mathrm{C}}(\mathrm{C}+2)]^2$. By definition (4.2) $A_{\mathrm{C}+1}$ is dependent on $D_{\mathrm{C}+1}$, hence $\mathbb{E}[D_{\mathrm{C}+1}A_{\mathrm{C}+1}] \neq \mathbb{E}D_{\mathrm{C}+1}\mathbb{E}A_{\mathrm{C}+1}$; an expression for $\mathbb{E}[D_{\mathrm{C}+1}A_{\mathrm{C}+1}]$ is required and will be derived in Section 4.4. Then, the difference equation can be solved in terms of $\mathbb{E}A_0^2$ (see Section 5.4 of [116]) and the solution yields

$$\begin{aligned}\mathbb{E}A_{\mathrm{C}}^2 &= \frac{\mathrm{C}!}{\rho^{\mathrm{C}}}\mathbb{E}A_0^2 - \frac{\mathrm{C}!}{\rho^{\mathrm{C}}}\sum_{j=1}^{\mathrm{C}}\frac{\rho^j}{j!}\left(\mathbb{E}D_j^2 + 2\mathbb{E}A_j\mathbb{E}A_{j-1} + 2\mathbb{E}[D_jA_j] + 2\mathbb{E}D_j\mathbb{E}A_{j-1}\right)\\ &\quad -2\frac{\mathrm{C}!}{\rho^{\mathrm{C}}}\sum_{j=1}^{\mathrm{C}}\frac{\rho^j}{j!}\frac{1}{\nu_j}\left(\mathbb{E}A_j + \mathbb{E}D_j + \mathbb{E}A_{j-1}\right) - \frac{2}{\lambda}\frac{\mathrm{C}!}{\rho^{\mathrm{C}}}\sum_{j=1}^{\mathrm{C}}\frac{\rho^j}{j!}\frac{1}{\nu_j}. \qquad (4.16)\end{aligned}$$

Observe that Expression (4.16) requires, besides $\mathbb{E}A_0^2$, the terms $\mathbb{E}D_j^2$ and $\mathbb{E}[D_jA_j]$ for $1 \leq j \leq \mathrm{C}$. Recall that $\mathbb{E}D_j^2$ is given by (4.12), and $\mathbb{E}[D_jA_j]$ can be obtained from Section 4.4, so expressions are available for all the required terms.

4.4 Joint expectations of the C-congestion period quantities

In this section the joint expectations $\mathbb{E}[D_{\mathrm{C}}N_{\mathrm{C}}]$, $\mathbb{E}[D_{\mathrm{C}}A_{\mathrm{C}}]$ and $\mathbb{E}[N_{\mathrm{C}}A_{\mathrm{C}}]$ are derived. The covariances between the quantities can easily be found as, e.g., $\mathbb{C}\mathrm{ov}(D_{\mathrm{C}}, N_{\mathrm{C}}) = \mathbb{E}[D_{\mathrm{C}}N_{\mathrm{C}}] - \mathbb{E}D_{\mathrm{C}}\mathbb{E}N_{\mathrm{C}}$. Furthermore, the joint expectation $\mathbb{E}[D_{\mathrm{C}}A_{\mathrm{C}}]$ is required to determine the second moment of the area above k for all $k \geq \mathrm{C}$, see Section 4.3.3.

Joint expectation of the duration and number of arrivals. By (4.6) and (4.7) we have

$$\begin{aligned}\mathbb{E}[D_{\mathrm{C}}N_{\mathrm{C}}] &= \mathbb{E}\left[\left(T_{\mathrm{C}+1} + P_{\mathrm{C}+1}D_{\mathrm{C}}(\mathrm{C}+2)\right)P_{\mathrm{C}+1}\left(1 + N_{\mathrm{C}}(\mathrm{C}+2)\right)\right]\\ &= \frac{\lambda}{\nu_{\mathrm{C}+1}}\big(\mathbb{E}T_{\mathrm{C}+1} + \mathbb{E}T_{\mathrm{C}+1}\mathbb{E}N_{\mathrm{C}+1} + \mathbb{E}T_{\mathrm{C}+1}\mathbb{E}N_{\mathrm{C}} + \mathbb{E}[D_{\mathrm{C}+1}N_{\mathrm{C}+1}]\\ &\quad +\mathbb{E}D_{\mathrm{C}+1}\mathbb{E}N_{\mathrm{C}} + \mathbb{E}D_{\mathrm{C}}\mathbb{E}N_{\mathrm{C}+1} + \mathbb{E}[D_{\mathrm{C}}N_{\mathrm{C}}] + \mathbb{E}D_{\mathrm{C}+1} + \mathbb{E}D_{\mathrm{C}}\big).\end{aligned}$$

This difference equation can be solved in terms of $\mathbb{E}[D_0N_0]$, the derivation of which is presented in Section 4.5.3, and yields

$$\begin{aligned}\mathbb{E}[D_{\mathrm{C}}N_{\mathrm{C}}] &= \frac{\mathrm{C}!}{\rho^{\mathrm{C}}}\mathbb{E}[D_0N_0] - \frac{\mathrm{C}!}{\rho^{\mathrm{C}}}\sum_{j=1}^{\mathrm{C}}\frac{\rho^j}{j!}\frac{1}{\nu_j}(1 + \mathbb{E}N_j + \mathbb{E}N_{j-1})\\ &\quad -\frac{\mathrm{C}!}{\rho^{\mathrm{C}}}\sum_{j=1}^{\mathrm{C}}\frac{\rho^j}{j!}\left(\mathbb{E}D_j\mathbb{E}N_{j-1} + \mathbb{E}D_{j-1}\mathbb{E}N_j + \mathbb{E}D_j + \mathbb{E}D_{j-1}\right). \qquad (4.17)\end{aligned}$$

Joint expectation of the duration and the area swept above C. By (4.6) and (4.8) we have

$$\mathbb{E}[D_{\mathrm{C}}A_{\mathrm{C}}] \quad = \quad \mathbb{E}[(T_{\mathrm{C}+1} + P_{\mathrm{C}+1}D_{\mathrm{C}}(\mathrm{C}+2))\,(T_{\mathrm{C}+1} + P_{\mathrm{C}+1}A_{\mathrm{C}}(\mathrm{C}+2))].$$

Isolating $\mathbb{E}[D_{\mathrm{C}+1}A_{\mathrm{C}+1}]$ yields

$$\begin{aligned}\mathbb{E}[D_{\mathrm{C}+1}A_{\mathrm{C}+1}] = \frac{\mathrm{C}+1}{\rho}\mathbb{E}[D_{\mathrm{C}}A_{\mathrm{C}}] - \big(&\mathbb{E}D_{\mathrm{C}+1}^2 + \mathbb{E}D_{\mathrm{C}}\mathbb{E}A_{\mathrm{C}+1} + \mathbb{E}D_{\mathrm{C}+1}\mathbb{E}A_{\mathrm{C}} \\ &+\mathbb{E}D_{\mathrm{C}+1}\mathbb{E}D_{\mathrm{C}}) - \frac{1}{\nu_{\mathrm{C}+1}}\left(2\mathbb{E}D_{\mathrm{C}+1} + \mathbb{E}D_{\mathrm{C}} + \mathbb{E}A_{\mathrm{C}+1} + \mathbb{E}A_{\mathrm{C}}\right) - \frac{2}{\lambda\nu_{\mathrm{C}+1}}.\end{aligned}$$

Notice that expression includes a term $\mathbb{E}D_{\mathrm{C}+1}^2$ that results from the decompositions of $D_{\mathrm{C}}(\mathrm{C}+2)$ and $A_{\mathrm{C}}(\mathrm{C}+2)$ that both consist of a term $D_{\mathrm{C}+1}$. The difference equation can be solved in terms of $\mathbb{E}[D_0A_0]$, which is deduced in Section 5.6 of [116], and yields

$$\begin{aligned}\mathbb{E}[D_{\mathrm{C}}A_{\mathrm{C}}] = \frac{\mathrm{C}!}{\rho^{\mathrm{C}}}\mathbb{E}[D_0A_0] - \frac{\mathrm{C}!}{\rho^{\mathrm{C}}}\sum_{j=1}^{\mathrm{C}}\frac{\rho^j}{j!}\big(&\mathbb{E}D_j^2 + \mathbb{E}D_{j-1}\mathbb{E}A_j + \mathbb{E}D_{j-1}\mathbb{E}A_j \\ &+\mathbb{E}D_j\mathbb{E}D_{j-1}) - \frac{\mathrm{C}!}{\rho^{\mathrm{C}}}\sum_{j=1}^{\mathrm{C}}\frac{\rho^j}{j!}\frac{1}{\nu_j}\left(2\mathbb{E}D_j + \mathbb{E}D_{j-1} + \mathbb{E}A_j + \mathbb{E}A_{j-1}\right) \\ &-\frac{2}{\lambda}\frac{\mathrm{C}!}{\rho^{\mathrm{C}}}\sum_{j=1}^{\mathrm{C}}\frac{\rho^j}{j!}\frac{1}{\nu_j}.\end{aligned} \tag{4.18}$$

Observe that the solution requires the second moments $\mathbb{E}D_j^2$ for $1 \le j \le \mathrm{C}$, which are given by (4.11).

Joint expectation of the number of arrivals and the area swept above C. By (4.7) and (4.8) we have

$$\mathbb{E}[N_{\mathrm{C}}A_{\mathrm{C}}] = \mathbb{E}[P_{\mathrm{C}+1}(1 + N_{\mathrm{C}}(\mathrm{C}+2))(T_{\mathrm{C}+1} + P_{\mathrm{C}+1}A_{\mathrm{C}}(\mathrm{C}+2))]$$

The solution in terms of $\mathbb{E}[N_0A_0]$, see Section 5.7 of [116], yields

$$\begin{aligned}\mathbb{E}N_{\mathrm{C}}A_{\mathrm{C}} = \frac{\mathrm{C}!}{\rho^{\mathrm{C}}}\mathbb{E}[N_0A_0] - \frac{\mathrm{C}!}{\rho^{\mathrm{C}}}\sum_{j=1}^{\mathrm{C}}\frac{\rho^j}{j!}\big(&\mathbb{E}[D_jN_j] + \mathbb{E}N_j\mathbb{E}A_{j-1} + \mathbb{E}N_{j-1}\mathbb{E}A_j \\ &+ \mathbb{E}D_j\mathbb{E}N_{j-1} + \mathbb{E}A_j + \mathbb{E}D_j + \mathbb{E}A_{j-1}\big) \\ &-\frac{\mathrm{C}!}{\rho^{\mathrm{C}}}\sum_{j=1}^{\mathrm{C}}\frac{\rho^j}{j!}\frac{1}{\nu_j}\left(1 + \mathbb{E}N_j + \mathbb{E}N_{j-1}\right).\end{aligned} \tag{4.19}$$

4.5 Moments and joint expectations of the busy-period quantities

In Sections 4.3 and 4.4 expressions were obtained for the moments and joint expectations for the quantities of a C-congestion period. The expressions are all solved in terms of the busy-period quantities (i.e., 0-congestion period quantities). The goal of this section is to derive these busy-period quantities. This section only presents the first and second moments of the duration and the joint expectation of the duration and number of arrivals; the derivations of the other busy-period quantities $\mathbb{E}N_0$, $\mathbb{E}A_0$, $\mathbb{E}[D_0A_0]$ and $\mathbb{E}[N_0A_0]$ are presented Section 5 of [116].

The moments of the quantities are obtained by differentiating the Laplace transform (LT) of the congestion triple (D_0, N_0, A_0) that was obtained by Preater [104]. Section 4.5.1 presents Preater's LT and additionally a lemma that simplifies the calculations that are presented in the succeeding subsections.

Theoretically, all moments and joint expectations of the quantities of level C can be obtained by differentiating Preater's LT of the congestion triple, but this task appeared to be far from trivial. Therefore we decided to first express them in terms of moments and joint expectations of the 0-congestion period; subsequently, we derive these 0-congestion period quantities through (relatively easy, but still tedious) differentiations.

4.5.1 Preater's LT of the 0-congestion triple (D_0, N_0, A_0)

Analogously to Guillemin and Simonian [58], Preater uses $\mu = 1$, and so $\lambda = \rho$. To obtain the Laplace transform of the C-congestion triple, Preater first considers the LT of the duration of a C-congestion period. By two different derivations he obtains the LT in two different expressions: the first is a continued fraction, the second is a fraction of the functions I_{C+1} and I_C (see (4.21)). The equality of these two expressions is the most important result of his *Proposition 2.2*. In his *Theorem 3.1* he derives the (joint) LT of the congestion triple by the first derivation and the result is also in the form of a continued fraction. Using the equality of his Proposition 2.2, the continued fraction can be rewritten as a fraction of I_{C+1} and I_C. The LT for $C = 0$ resulting from his Proposition 2.2 and Theorem 3.1 is stated below.

Preater's Theorem 3.1 and Proposition 2.2 combined for $C = 0$. *The vector (D_0, N_0, A_0) has Laplace transform*

$$\Theta_0^*(s,t,u) \quad := \quad \mathbb{E}\exp(-sD_0 - tN_0 - uA_0) = \frac{1}{u+1}\frac{I_1(a-b,b)}{I_0(a-b,b)}. \tag{4.20}$$

where

$$a := a(s,t,u) = \frac{s+\rho}{u+1}, \qquad b := b(s,t,u) = \frac{\rho e^{-t}}{(u+1)^2},$$
$$I_{\mathrm{C}}(a,b) := \int_0^1 e^{-bx}(1-x)^{a-1}x^{\mathrm{C}}\mathrm{d}x. \tag{4.21}$$

Differentiating (4.21) is a tedious job, but can be simplified considerably by the next lemma.

LEMMA 4.5.1.

$$I_0(a,b) = e^{-b}\sum_{k=0}^{\infty}\frac{1}{a+k}\frac{b^k}{k!}, \qquad I_1(a,b) = I_0(a,b) - I_0(a+1,b). \tag{4.22}$$

Proof

$$\begin{aligned} I_0(a,b) &= \int_0^1 e^{-bx}(1-x)^{a-1}\mathrm{d}x = e^{-b}\int_0^1 e^{bx}x^{a-1}\mathrm{d}x \\ &= e^{-b}\sum_{k=0}^{\infty}\frac{b^k}{k!}\int_0^1 x^{k+a-1}\mathrm{d}x = e^{-b}\sum_{k=0}^{\infty}\frac{1}{a+k}\frac{b^k}{k!}. \\ I_1(a,b) &= \int_0^1 e^{-bx}(1-x)^{a-1}x\mathrm{d}x = e^{-b}\int_0^1 e^{bx}x^{a-1}(1-x)\mathrm{d}x \\ &= e^{-b}\sum_{k=0}^{\infty}\frac{b^k}{k!}\int_0^1 (x^{k+a-1}-x^{k+a})\mathrm{d}x \overset{\text{by (4.22)}}{=} I_0(a,b) - I_0(a+1,b). \end{aligned}$$

□

Furthermore, we introduce the following notation:

$$\xi(\rho) := \sum_{k=0}^{\infty}\frac{1}{(k+1)^2}\frac{\rho^k}{k!}.$$

Notice that $\xi(\rho) < \infty$.

4.5.2 Moments of the duration of the busy period

By (4.20) and using (4.22) we have

$$D_0^*(s) = \Theta_0^*(s,0,0) = 1 - \frac{f(s)}{n(s)}, \tag{4.23}$$

where

$$f(s) := \sum_{k=0}^{\infty}\frac{1}{s+k+1}\frac{\rho^k}{k!} \qquad \text{and} \qquad n(s) := \sum_{k=0}^{\infty}\frac{1}{s+k}\frac{\rho^k}{k!}.$$

Let $n^{(m)}(s)$ denote the m-th derivative of $n(s)$ (hence $n(s) = n^{(0)}(s)$). Then it can be shown that

$$n^{(m)}(s) = \sum_{k=0}^{\infty} \frac{(-1)^m \cdot m!}{(s+k)^{m+1}} \frac{\rho^k}{k!} \qquad \text{and} \qquad \lim_{s \to 0} n^{(m)}(s) \sim \frac{(-1)^m m!}{s^{m+1}}.$$

The first equation is obtained by repeated derivation of $n(s)$. The second statement is obtained by proving that the $(k \geq 1)$-terms can be bounded by a finite term as follows:

$$\sum_{k=1}^{\infty} \frac{\rho^k}{k^m \cdot k!} \leq \sum_{k=1}^{\infty} \frac{\rho^k}{k!} < \sum_{k=0}^{\infty} \frac{\rho^k}{k!} = e^{\rho}.$$

Then the second statement is proven by observing that the second statement is exactly the $(k = 0)$-term which goes to infinity for s close to 0.

First moment. Although the first moment is already obtained in Section 4.3.1, we also present its derivation for the sake of completeness. It is well known that $\mathbb{E}D_0 = -(D_0^*)'(0)$. Differentiation of (4.23) yields

$$(D_0^*)'(s) = \frac{\mathrm{d}}{\mathrm{d}s}\left(1 - \frac{f(s)}{n(s)}\right) = \frac{n'(s)f(s)}{n^2(s)} - \frac{f'(s)}{n(s)}.$$

We conclude that $\mathbb{E}D_0 = f(0) - 0 = (e^{\rho} - 1)/\rho$, which coincides with the results earlier obtained in Section 4.3.1 for $\mu = 1$.

Second moment. Now $\mathbb{E}D_0^2 = (D_0^*)''(0)$. The second derivative is

$$\begin{aligned}(D_0^*)''(s) &= \frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{n'(s)f(s)}{n^2(s)} - \frac{f'(s)}{n(s)}\right) \\ &= -\frac{f''(s)}{n(s)} + 2\frac{n'(s)f'(s)}{n^2(s)} - 2\frac{(n'(s))^2 f(s)}{n^3(s)} + \frac{n''(s)f(s)}{n^2(s)}.\end{aligned}$$

The first of these four terms goes to 0, and the second to $-2f'(0)$. The third term goes to $-\infty$, and, as $n''(s) \sim 2/s^3$, the fourth term goes to $+\infty$. Define for ease

$$g_n(s) := \sum_{k=1}^{\infty} \frac{1}{(k+s)^n} \frac{\rho^k}{k!};$$

for any $n \in \mathbb{N}$, it holds that $g_n(0) < e^{\rho} < \infty$. Simple manipulations yield

$$\begin{aligned}&\lim_{s \downarrow 0}\left(\frac{n''(s)}{n^2(s)} - 2\frac{(n'(s))^2}{n^3(s)}\right) \\ &\quad = \lim_{s \downarrow 0} \frac{(s^{-1} + g_1(s))(2s^{-3} + 2g_3(s)) - 2(s^{-2} + g_2(s))^2}{(s^{-1} + g_1(s))^3} \;=\; 2g_1(0).\end{aligned}$$

Thus

$$\begin{aligned}\mathbb{E}D_0^2 &= 2g_1(0)f(0) - 2f'(0)\\ &= 2\left(\sum_{k=1}^{\infty}\frac{1}{k}\frac{\rho^k}{k!}\right)\frac{e^\rho-1}{\rho} + 2\sum_{k=0}^{\infty}\frac{1}{(k+1)^2}\frac{\rho^k}{k!} \;=\; 2e^\rho\xi(\rho). \end{aligned} \tag{4.24}$$

Relation between (4.24) and the results of Liu and Shi [83]. Liu and Shi [83] obtained the following expression for the second moment of the busy period of an M/G/∞ queue:

$$\mathbb{E}D_0^2 = \frac{2}{\lambda P_0^2}\int_0^\infty [P_0(t) - P_0]\,\mathrm{d}t$$

where $P_0(t) = \exp\left\{-\rho\int_0^t e^{-x}\mathrm{d}x\right\} = \exp\{-\rho(1-e^{-t})\}$ and P_0 is the probability that the system is idle, thus $P_0 = e^{-\rho}$. Then, by using that $\exp\{\rho e^{-t}\} = \sum_{k=0}^{\infty}(\rho e^{-t})^k/k!$, we have

$$\begin{aligned}\frac{2}{\rho P_0^2}\int_0^\infty [P_0(t)-P_0]\,\mathrm{d}t &= \frac{2e^{2\rho}}{\rho}\int_0^\infty e^{-\rho}\left[e^{\rho e^{-t}}-1\right]\mathrm{d}t\\ &= \frac{2e^\rho}{\rho}\int_0^\infty\sum_{k=1}^{\infty}\frac{(\rho e^{-t})^k}{k!}\mathrm{d}t \;=\; \frac{2e^\rho}{\rho}\sum_{k=1}^{\infty}\frac{\rho^k}{k!}\int_0^\infty e^{-kt}\mathrm{d}t\\ &= \frac{2e^\rho}{\rho}\rho\sum_{k=0}^{\infty}\frac{\rho^k}{(k+1)^2\,k!} \;=\; 2e^\rho\xi(\rho).\end{aligned}$$

We conclude that Expression (4.24) and the result of Liu and Shi [83] coincide.

4.5.3 Joint expectation $\mathbb{E}[D_0N_0]$ of the busy period

The joint expectation $\mathbb{E}[D_0N_0]$ can be obtained by differentiating the Laplace transform (4.20) to both s and t:

$$\mathbb{E}[D_0N_0] \;=\; \lim_{s\downarrow 0,t\downarrow 0}\frac{\mathrm{d}^2}{\mathrm{d}s\mathrm{d}t}\mathbb{E}e^{-sD_0-tN_0}.$$

By (4.20) and (4.22) we have

$$\mathbb{E}e^{-sD_0-tN_0} \;=\; \Theta_0^*(s,t,u) = 1 - \frac{I_0(a-b+1,b)}{I_0(a-b,b)} = 1 - \frac{f(s,t)}{n(s,t)},$$

where $a = a(s,t) = s+\rho$ and $b = b(s,t) = \rho e^{-t}$ and by the definition we have

$$\begin{aligned} n(s,t) &:= \sum_{k=0}^{\infty}\frac{\rho^k}{k!}\frac{e^{-kt}}{s+\rho(1-e^{-t})+k};\\ f(s,t) &:= \sum_{k=0}^{\infty}\frac{\rho^k}{k!}\frac{e^{-kt}}{s+\rho(1-e^{-t})+k+1}.\end{aligned}$$

Define $n'_s := \mathrm{d}n(s,t)/\mathrm{d}s$, $n'_t := \mathrm{d}n(s,t)/\mathrm{d}t$ and $n''_{st} := \mathrm{d}^2n(s,t)/\mathrm{d}t\mathrm{d}s$. Analogously, the derivatives f'_s, f'_t and f''_{st} are defined. Then

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}t\mathrm{d}s}\mathbb{E}e^{-sD_0-tN_0} &= \frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{fn'_s}{n^2}-\frac{f'_s}{n}\right]\\ &= -\frac{f''_{st}}{n}+\frac{f'_sn'_t+f'_tn'_s}{n^2}+\frac{fn''_{st}}{n^2}-\frac{2fn'_sn'_t}{n^3}.\end{aligned}$$

For $s,t\to 0$, the first term goes to 0, and the second term yields $-\rho f'_s(0,0)-f'_t(0,0)$. The third and fourth terms result in $(\infty-\infty)$, and require careful analysis, similar as was done for $\mathbb{E}D_0^2$. This eventually yields $2\rho^2\xi(\rho)f(0,0)$. Then we obtain the following expression for $\mathbb{E}[D_0N_0]$:

$$\mathbb{E}[D_0N_0] = -\rho f'_s(0,0)-f'_t(0,0)+2\rho^2\xi(\rho)f(0,0). \tag{4.25}$$

Notice that $\mathbb{E}[D_0N_0]$ is bounded for finite ρ; $f(0,0)\le\sum_{k=0}^{\infty}\rho^k/k!=e^{\rho}$ and similar bounds can be obtained for $f'_s(0,0)$ and $f'_t(0,0)$.

4.5.4 Moments for the mean service times other than 1

The expressions for the moments of D_0 (Section 4.5.2) and the joint expectation of D_0N_0 (Section 4.4) are derived for the case of mean service time $\mu=1$. To adapt the derived expressions for service times $\mu\neq 1$, it suffices to see that varying λ or μ for fixed ρ is only a scaling of time. The scaling of time does not influence the number of arrivals, but it does influence the duration and area. The expressions for the moments can be adapted to $\mu\neq 1$ by a factor $(\mu^{-1})^n$ where n is the order of the moment, e.g., in self-evident notation:

$$\begin{aligned}\mathbb{E}D_{\mathrm{C}}^n &= \mu^{-n}\mathbb{E}D_{\mathrm{C}|\{\mu=1\}}^n, & \mathbb{E}D_{\mathrm{C}}N_{\mathrm{C}} &= \mu^{-1}\mathbb{E}D_{\mathrm{C}}N_{\mathrm{C}|\{\mu=1\}},\\ \mathbb{E}N_{\mathrm{C}}^n &= \mathbb{E}N_{\mathrm{C}|\{\mu=1\}}^n, & \mathbb{E}D_{\mathrm{C}}A_{\mathrm{C}} &= \mu^{-2}\mathbb{E}D_{\mathrm{C}}A_{\mathrm{C}|\{\mu=1\}},\\ \mathbb{E}A_{\mathrm{C}}^n &= \mu^{-n}\mathbb{E}A_{\mathrm{C}|\{\mu=1\}}^n, & \mathbb{E}N_{\mathrm{C}}A_{\mathrm{C}} &= \mu^{-1}\mathbb{E}N_{\mathrm{C}}A_{\mathrm{C}|\{\mu=1\}}.\end{aligned}$$

4.6 C-intercongestion periods

Besides the duration of a C-congestion period, we are also interested in the time that the system is below level C, a so-called C-intercongestion period. This section consists of the definitions of a C-intercongestion period, the derivation of a LT of the intercongestion triple, and the derivation of the first and second moments of the quantities and the covariances between the quantities.

4.6.1 Definitions

Analogously to the definitions of a C-congestion period in Section 4.2.1 we define

$$\begin{aligned}
\mathcal{D}_j(i) &:= \inf\{t > 0 : \Lambda_t = j \mid \Lambda_0 = i\}, && i < j,\\
\mathcal{N}_j(i) &:= \#\{t : \Lambda_t - \Lambda_{t-} = 1, 0 < t \le \mathcal{D}_j(i)\}, && i < j,\\
\mathcal{A}_j(i) &:= \int_{t=0}^{\mathcal{D}_j(i)} (j - \Lambda_t)\mathrm{d}t, && i < j,
\end{aligned}$$

where $\mathcal{D}_j(i)$ is the duration of the transient period to go from state i to state j for $i < j$, $\mathcal{N}_j(i)$ is the number of arrivals during $\mathcal{D}_j(i)$, and $\mathcal{A}_j(i)$ is the area *under* C during $\mathcal{D}_j(i)$. For convenience we use the following short notation: $\mathcal{D}_{\mathrm{C}} := \mathcal{D}_{\mathrm{C}}(\mathrm{C}-1)$, $\mathcal{N}_{\mathrm{C}} := \mathcal{N}_{\mathrm{C}}(\mathrm{C}-1)$ and $\mathcal{A}_{\mathrm{C}} := \mathcal{A}_{\mathrm{C}}(\mathrm{C}-1)$. $\mathcal{D}_j(i)$ is a hitting time which allows for the following decomposition $\mathcal{D}_j(i) = \sum_{k=i+1}^{j} \mathcal{D}_k$. Due to the definition of $\mathcal{N}_j(i)$ and $\mathcal{A}_j(i)$ these can be similarly decomposed $\mathcal{N}_j(i) = \sum_{k=i+1}^{j} \mathcal{N}_k$ and $\mathcal{A}_j(i) = \sum_{k=i+1}^{j} (\mathcal{A}_k + (j-k)\mathcal{D}_k)$.

Finally, we derive a recursive structure for a C-intercongestion period. Using random variables T_{C} and P_{C}, which have the same definition as in Section 4.2.1, the duration $\mathcal{D}_{\mathrm{C}}$ ($\mathrm{C} \ge 2$) can be subdivided into the independent durations $T_{\mathrm{C}-1}$ and $\mathcal{D}_{\mathrm{C}}(\mathrm{C}-2)$ as follows:

$$\mathcal{D}_{\mathrm{C}} = T_{\mathrm{C}-1} + (1 - P_{\mathrm{C}-1})\mathcal{D}_{\mathrm{C}}(\mathrm{C}-2) \qquad \text{for } \mathrm{C} \ge 2.$$

Similarly decompositions of $\mathcal{N}_{\mathrm{C}}$ and $\mathcal{A}_{\mathrm{C}}$ gives the following recursive equations:

$$\begin{aligned}
\mathcal{N}_{\mathrm{C}} &= P_{\mathrm{C}-1} + (1 - P_{\mathrm{C}-1})\mathcal{N}_{\mathrm{C}}(\mathrm{C}-2) && \text{for } \mathrm{C} \ge 2,\\
\mathcal{A}_{\mathrm{C}} &= T_{\mathrm{C}-1} + (1 - P_{\mathrm{C}-1})\mathcal{A}_{\mathrm{C}}(\mathrm{C}-2) && \text{for } \mathrm{C} \ge 2.
\end{aligned}$$

This set of equations again gives rise to recursions for D_{C}, $\mathcal{N}_{\mathrm{C}}$ and $\mathcal{A}_{\mathrm{C}}$, which can be solved in terms of D_0, $\mathcal{N}_0$ and $\mathcal{A}_0$.

4.6.2 Laplace transforms of the duration and the intercongestion triple

The derivation of the Laplace transforms is done analogously to the derivation of the LTs of the congestion period done by Preater [104]. First, the LT of the duration will be derived in two different ways which results in two different forms. The equality of these two forms is exploited in the derivation of the LT of the intercongestion triple. In this section we follow Preater's assumption that $\mu = 1$.

Laplace transform of the intercongestion period duration

LEMMA 4.6.1. *Let x_n be a non-negative, bounded sequence satisfying*

$$x_n := \frac{a + bn}{n + c - x_{n-1}}, \qquad n \ge 1,$$

where $a, c > 0$, $b \geq 0$. Then

$$x_0 = \mathcal{L}(a,b,c) := -1 - c + \frac{a+b}{-2-c+\frac{a+2b}{-3-c+\frac{a+3b}{-4-c+\cdots}}}. \tag{4.26}$$

Proof The proof is derived by mimicking Lemma 2.1 of [104] for

$$x_{n-1} := -n - c + \frac{a+bn}{x_n}, \qquad n \geq 1.$$

Writing x_0 as a continued fraction yields (4.26). □

PROPOSITION 4.6.2. *The Laplace transform of the duration $\mathcal{D}_{\mathrm{C}}$ is*

$$\begin{aligned} \mathcal{D}^*_{\mathrm{C}}(s) &= \mathrm{C}^{-1}\mathcal{L}(\lambda\mathrm{C}, \lambda, \lambda + s + \mathrm{C} - 1) \\ &= \frac{\lambda}{\mathrm{C}} \frac{I_{\mathrm{C}}(s,\lambda)}{I_{\mathrm{C}-1}(s,\lambda)} \frac{\sum_{k=0}^{\mathrm{C}-1} \binom{\mathrm{C}-1}{k} \frac{\lambda^k}{k!} I_{2k}(s+\mathrm{C}-1-k,\lambda)}{\sum_{k=0}^{\mathrm{C}} \binom{\mathrm{C}}{k} \frac{\lambda^k}{k!} I_{2k}(s+\mathrm{C}-k,\lambda)}. \end{aligned} \tag{4.27}$$

Proof (a) An n-intercongestion period starts with a sojourn time T_{n-1} at level $n-1$. At the end of T_{n-1} with probability $(n-1)/(\lambda+n-1)$ a customer departs, starting a $(n-1)$-intercongestion period followed by another sojourn at level $n-1$. At the end of each sojourn time T_{n-1} a new $(n-1)$-intercongestion period can be started by a departure or the n-intercongestion can be ended by the arrival of a new customer. With obvious notation, we find the following equality in distribution:

$$\mathcal{D}_n = T^{(0)}_{n-1} + \sum_{i=1}^{G_{n-1}-1} \left(\mathcal{D}^{(i)}_{n-1} + T^{(i)}_{n-1} \right), \qquad n \geq 0,$$

where all variables on the right are independent, T_n is exponentially $(\lambda+n)$ distributed and G_n is geometrically $(p_n := \lambda/(\lambda+n))$ distributed. Then,

$$T^*_n(s) := \mathbb{E}e^{-sT_n} = \frac{\lambda+n}{s+\lambda+n},$$

and

$$\begin{aligned} \mathcal{D}^*_n(s) &= \frac{p_{n-1}T^*_{n-1}(s)}{1-(1-p_{n-1})T^*_{n-1}(s)\mathcal{D}^*_{n-1}(s)} \\ &= \frac{\lambda}{n-1+s+\lambda-(n-1)\mathcal{D}^*_{n-1}(s)}. \end{aligned} \tag{4.28}$$

Let $x_n = (n+\mathrm{C})\mathcal{D}^*_{n+\mathrm{C}}(s)$. Then (4.28) fulfils the setting of Lemma 4.6.1 with $a = \lambda\mathrm{C}$, $b = \lambda$ and $c = \lambda + s + \mathrm{C} - 1$. Hence, the first equality in (4.27) follows from (4.26).

(b) We follow the lines of the proof of Proposition 2.2 of Preater [104]. Let X_t be a stationary version of the M/M/∞ occupation process, so X_t is Poisson (ρ) distributed ($\rho = \lambda$ as $\mu = 1$). Preater defined $\pi_n(t) := \mathbb{P}(X_t = n | X_0 = 0)$, which is

Poisson $(\lambda(1-e^{-t}))$ distributed and has LT $\pi_n^*(s) = (\lambda^n/n!)I_n(s,\lambda)$. Additionally we define $\chi_n(t) := \mathbb{P}(X_t = n|X_0 = n)$ and denote its LT by $\chi_n^*(s)$. By conditioning on the number $k \leq n$ of the initial n customers that were present at epoch 0, and that are still present at epoch t. We obtain

$$\chi_n(t) \quad = \quad \sum_{k=0}^{n} \binom{n}{k} (1-e^{-t})^{n-k} (e^{-t})^k \pi_k(t).$$

Then its LT can be obtained as follows:

$$\begin{aligned}
\chi_n^*(s) &= \int_0^\infty e^{-st} \sum_{k=0}^{n} \binom{n}{k} (1-e^{-t})^k (e^{-t})^{n-k} \frac{(\lambda(1-e^{-t}))^k}{k!} e^{-\lambda(1-e^{-t})} \mathrm{d}t \\
&= \sum_{k=0}^{n} \binom{n}{k} \frac{\lambda^k}{k!} \int_0^1 (1-u)^{s+n-k-1} u^{2k} e^{-\lambda u} \mathrm{d}u \\
&= \sum_{k=0}^{n} \binom{n}{k} \frac{\lambda^k}{k!} I_{2k}(s+n-k,\lambda),
\end{aligned}$$

by using the substitution $u := 1 - e^{-t}$ in the second step.

Next, we introduce the first passage time $\tau_n := \inf\{t \geq 0 : X_t = n|X_0 = 0\}$. Then for $n \geq 0$

$$\int_0^t \chi_n(t-x)\mathbb{P}(\tau_n \in \mathrm{d}x) \quad = \quad \pi_n(t).$$

Taking Laplace transforms on both sides results in $\mathbb{E}e^{-s\tau_n} = \pi_n^*(s)/\chi_n^*(s)$. We thus obtain

$$\mathcal{D}_{\mathrm{C}}^*(s) \quad = \quad \frac{\mathbb{E}e^{-s\tau_{\mathrm{C}}}}{\mathbb{E}e^{-s\tau_{\mathrm{C}-1}}} = \frac{\lambda}{\mathrm{C}} \frac{I_{\mathrm{C}}(s,\lambda)}{I_{\mathrm{C}-1}(s,\lambda)} \frac{\sum_{k=0}^{\mathrm{C}-1} \binom{\mathrm{C}-1}{k} \frac{\lambda^k}{k!} I_{2k}(s+\mathrm{C}-1-k,\lambda)}{\sum_{k=0}^{\mathrm{C}} \binom{\mathrm{C}}{k} \frac{\lambda^k}{k!} I_{2k}(s+\mathrm{C}-k,\lambda)}.$$

which proves the second equality in (4.27). □

A 'sanity check' of (4.27) is the special case $\mathrm{C} = 1$; for $\mathrm{C} = 1$ the intercongestion period reduces to an exponentially (λ) distributed idle period. In Appendix B of [116] it is shown that then (4.27) indeed reduces to $\lambda/(\lambda+s)$.

Laplace transform of C-intercongestion triple $(\mathcal{D}_{\mathrm{C}}, \mathcal{N}_{\mathrm{C}}, \mathcal{A}_{\mathrm{C}})$

THEOREM 4.6.3. *Let* $\mathrm{C} \in \mathbb{N}$. *The vector* $(\mathcal{D}_{\mathrm{C}}, \mathcal{N}_{\mathrm{C}}, \mathcal{A}_{\mathrm{C}})$ *has* LT

$$\Omega_{\mathrm{C}}^*(s,t,u) \quad := \quad \mathbb{E}\exp\left(-s\mathcal{D}_n - t\mathcal{N}_n - u\mathcal{A}_n\right) = \mathrm{C}^{-1}\mathcal{L}(a'\mathrm{C}, a', b')$$

where $a' := \lambda e^{-t}$; $b' := s + \lambda + u + \mathrm{C} - 1$.
In particular,

$$\Omega_{\mathrm{C}}^*(s-u,t,u) \quad = \quad \frac{\lambda}{\mathrm{C}} \frac{I_{\mathrm{C}}(s,\lambda)}{I_{\mathrm{C}-1}(s,\lambda)} \frac{\sum_{k=0}^{\mathrm{C}-1} \binom{\mathrm{C}-1}{k} \frac{\lambda^k}{k!} I_{2k}(s+\mathrm{C}-1-k,\lambda)}{\sum_{k=0}^{\mathrm{C}} \binom{\mathrm{C}}{k} \frac{\lambda^k}{k!} I_{2k}(s+\mathrm{C}-k,\lambda)}.$$

Proof

$$\begin{aligned}
\Omega_n^*(s,t,u) &= \mathbb{E}\exp\left(-s\mathcal{D}_n - t\mathcal{N}_n - u\mathcal{A}_n\right) \\
&= \mathbb{E}\exp\left(-(s+u)\mathcal{D}_n - t\mathcal{N}_n - u(\mathcal{A}_n - \mathcal{D}_n)\right) \\
&= \mathbb{E}\exp\Big(- (s+u)T_{n-1}^{(0)} - t \\
&\qquad - \sum_{i=1}^{G_{n-1}-1} \Big[(s+u)T_{n-1}^{(i)} + (s+u)\mathcal{D}_{n-1} + t\mathcal{N}_{n-1} + u\mathcal{A}_{n-1}\Big]\Big) \\
&= T_{n-1}^*(s+u)p_n e^{-t}\left[1-(1-p_n)T_{n-1}^*(s+u)\Omega_{n-1}^*(s+u,t,u)\right]^{-1} \\
&= \frac{\lambda e^{-t}}{n-1+s+u+\lambda-(n-1)\Omega_{n-1}^*(s+u,t,u)} \qquad (4.29)
\end{aligned}$$

Let $x_n = (n+\mathrm{C})\Omega_{n+\mathrm{C}}^*(s-nu,t,u)$. Then (4.29) falls in the framework of Lemma 4.6.1 with $a = a'\mathrm{C}$; $\quad b = a'$; $\quad c = b'$. □

4.6.3 Moments of the C-intercongestion period quantities

The derivations of the moments and joint expectations of the intercongestion-period quantities are analogous to the derivation of the congestion-period quantities in Sections 4.3 and 4.4, although there is a large difference in obtaining the starting conditions. As the system can never have less than 0 customers, all quantities corresponding to level 0 are 0 themselves, e.g., $\mathbb{E}\mathcal{D}_0^n = 0$, $\mathbb{E}\mathcal{N}_0^n = 0$, $\mathbb{E}\mathcal{A}_0^n = 0$, $\mathbb{E}\mathcal{D}_0\mathcal{N}_0 = 0$, $\mathbb{E}\mathcal{D}_0\mathcal{A}_0 = 0$, $\mathbb{E}\mathcal{N}_0\mathcal{A}_0 = 0$.

Moments of the duration of an C*-intercongestion period.*

$$\mathbb{E}\mathcal{D}_{\mathrm{C}} = \frac{1}{\lambda}\frac{(\mathrm{C}-1)!}{\rho^{\mathrm{C}-1}}\sum_{j=0}^{\mathrm{C}-1}\frac{\rho^j}{j!}, \qquad (4.30)$$

$$\begin{aligned}
\mathbb{E}\mathcal{D}_{\mathrm{C}}^2 &= \frac{(\mathrm{C}-1)!}{\rho^{\mathrm{C}-1}}\frac{2}{\lambda^2} + 2\frac{(\mathrm{C}-1)!}{\rho^{\mathrm{C}}}\sum_{j=1}^{\mathrm{C}-1}\frac{\rho^j}{(j-1)!}\frac{1}{\nu_j}\left(\mathbb{E}\mathcal{D}_{j+1} + \mathbb{E}\mathcal{D}_j\right) \\
&\quad + 2\frac{(\mathrm{C}-1)!}{\rho^{\mathrm{C}}}\sum_{j=1}^{\mathrm{C}-1}\frac{\rho^j}{(j-1)!}\left(\mathbb{E}\mathcal{D}_{j+1}\mathbb{E}\mathcal{D}_j\right) + \frac{2}{\lambda}\frac{(\mathrm{C}-1)!}{\rho^{\mathrm{C}-1}}\sum_{j=1}^{\mathrm{C}-1}\frac{\rho^j}{j!}\frac{1}{\nu_j}. \qquad (4.31)
\end{aligned}$$

Moments of the number of arrivals during a C*-intercongestion period.*

$$\begin{aligned}
\mathbb{E}\mathcal{N}_{\mathrm{C}} &= \frac{(\mathrm{C}-1)!}{\rho^{\mathrm{C}}}\sum_{j=1}^{\mathrm{C}-1}\frac{\rho^j}{(j-1)!} + 1, \\
\mathbb{E}\mathcal{N}_{\mathrm{C}}^2 &= \frac{(\mathrm{C}-1)!}{\rho^{\mathrm{C}-1}} + 2\frac{(\mathrm{C}-1)!}{\rho^{\mathrm{C}}}\sum_{j=1}^{\mathrm{C}-1}\frac{\rho^j}{(j-1)!}\mathbb{E}\mathcal{N}_{j+1}\mathbb{E}\mathcal{N}_j + \frac{(\mathrm{C}-1)!}{\rho^{\mathrm{C}-1}}\sum_{j=1}^{\mathrm{C}-1}\frac{\rho^j}{j!}\frac{1}{\nu_j}.
\end{aligned}$$

Moments of the area swept under C *during a* C*-intercongestion period.*

$$\begin{aligned}
\mathbb{E}\mathcal{A}_{\mathrm{C}} &= \frac{1}{\lambda}\frac{(\mathrm{C}-1)!}{\rho^{\mathrm{C}-1}} + \frac{(\mathrm{C}-1)!}{\rho^{\mathrm{C}-1}}\sum_{j=1}^{\mathrm{C}-1}\frac{\rho^j}{(j-1)!}\mathbb{E}\mathcal{D}_j + \frac{1}{\lambda}\frac{(\mathrm{C}-1)!}{\rho^{\mathrm{C}-1}}\sum_{j=1}^{\mathrm{C}-1}\frac{\rho^j}{j!},\\
\mathbb{E}\mathcal{A}_{\mathrm{C}}^2 &= \frac{\mathrm{C}-1}{\rho^{\mathrm{C}-1}}\frac{2}{\lambda^2} + \frac{(\mathrm{C}-1)!}{\rho^{\mathrm{C}}}\sum_{j=1}^{\mathrm{C}-1}\frac{\rho^j}{(j-1)!}\left(\mathbb{E}\mathcal{D}_j^2 + 2\mathbb{E}[\mathcal{D}_{j-1}\mathcal{A}_{j-1}]\right.\\
&\quad \left. + 2\mathbb{E}\mathcal{A}_{j-1}\mathbb{E}\mathcal{A}_j + 2\mathbb{E}\mathcal{D}_{j-1}\mathbb{E}\mathcal{A}_j\right) + \frac{2}{\lambda}\frac{(\mathrm{C}-1)!}{\rho^{\mathrm{C}-1}}\sum_{j=1}^{\mathrm{C}-1}\frac{\rho^j}{j!}\\
&\quad + \frac{(\mathrm{C}-1)!}{\rho^{\mathrm{C}}}\sum_{j=1}^{\mathrm{C}-1}\frac{\rho^j}{(j-1)!}\frac{1}{\nu_j}\left(\mathbb{E}\mathcal{A}_{j-1} + \mathbb{E}\mathcal{D}_{j-1} + \mathbb{E}\mathcal{A}_j\right).
\end{aligned}$$

Joint expectations of a C*-intercongestion period.* The joint expectations can be obtained in a similar fashion; they can be found in Section 6.4 of [116].

4.7 Intercongestion period as an approximation of a congestion period

From a numerical perspective a drawback of the congestion period recursions is that the starting condition corresponds to a busy period; for high loads the system will hardly ever be empty, and hence the busy-period quantities will tend to grow large, thus resulting in numerical instability. The intercongestion period recursions do not have this problem; as remarked before, all moments of the quantities of level 0 are 0, and consequently the recursions are numerically stable.

Congestion and intercongestion periods are similar in the sense that a C-congestion period is the duration that the system is *above* level C and a C-intercongestion period is the duration that the system is *below* level C. For C close to ρ, the birth rate (λ) and the death rate $(\rho\mu)$ are (almost) identical, hence $P_\rho \approx 1 - P_\rho$ in distribution. As $T_n \approx T_{\rho-(n-\rho)}$ in distribution for n close to ρ, it follows that a ρ-congestion and a ρ-intercongestion period exhibit similar stochastic behavior; a ρ-congestion period can be approximated by a ρ-intercongestion period. More generally, a C-congestion periods can be approximated by $(\rho-(\mathrm{C}-\rho))$-intercongestion periods by the observation $P_n \approx 1 - P_{\rho-(n-\rho)}$ for n close to ρ. In particular, this approximation is expected to work well for C close to the average load ρ.

Figures 4.1 and 4.2 present numerical results of the proposed approximation for arrival rate $\lambda = 1$ and mean service time $\mu^{-1} = 1000$, so the average load ρ is 1000. The moments of the congestion period quantities are obtained by simulations; the recursions are numerically unstable as the busy periods are very large due to the

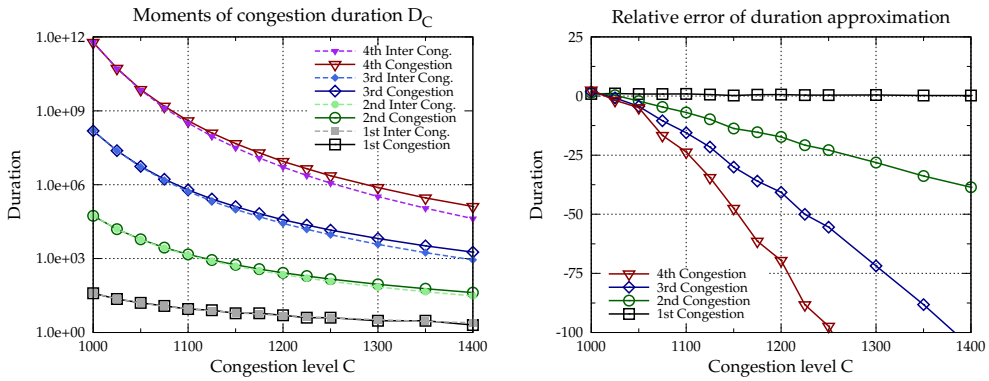

Figure 4.1: *Approximation of the simulated* C*-congestion period (*CP*) duration by an analytically derived* $(2\rho - \mathrm{C})$*-intercongestion period (*ICP*). Left: duration moments. Right: Relative error between simulated* C*-congestion period and derived* $(2\rho - \mathrm{C})$*-intercongestion period*

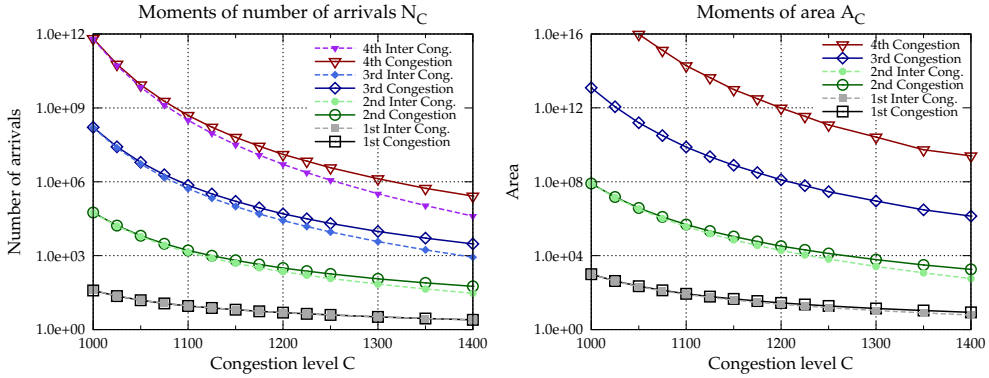

Figure 4.2: *Approximation of a simulated* C*-congestion period (*CP*) by an analytically derived* $(2\rho - \mathrm{C})$*-intercongestion period (*ICP*). Left: number of arrivals. Right: area swept above* C*.*

high average load. The intercongestion period quantities are obtained analytically by the recursive relations presented in Section 4.6. The left graph of Figure 4.1 shows the first four moments of C-congestion period duration and an approximation of the duration; the approximation is the duration of a $(2\rho - \mathrm{C})$-intercongestion period. The right graph presents the relative error of the approximation. We see that for values of C in the neighborhood of ρ the approximation is close to the simulated results. Especially for 'lower' moments the approximation is accurate; as could be expected, for higher moments the relative error becomes larger. In the range of $\mathrm{C} = (\rho, \ldots, 1100)$ the error of the second moment is less than 7%; here it is important to notice that the system will hardly ever have more than 1100 customers (probability is in the order of 0.001). Figure 4.2 presents the results for the number of arrivals and

the area for the same scenario. The results for these quantities are also accurate, so the intercongestion period seems to be a very good approximation for the congestion period, in particular for C close to average load ρ.

Another approximation was proposed earlier by Guillemin and Simonian [58]. They argue that a C-congestion period converges (after a specific scaling) to an M/M/1 busy period for large C. They propose to use the death-rate $C\mu$ of the M/M/∞ queue as the death-rate for the M/M/1 queue, which results in an accurate approximation for C large compared to ρ. For C close to ρ the approximation is not so good; the behavior of the M/M/∞ congestion period differs significantly from the M/M/1 busy period. However, as concluded earlier, the approximation of a congestion period by an intercongestion period is very accurate for C close to the average load ρ. We remark that the regime in which C is close to ρ is from a practical point of view perhaps the most relevant regime: networks are usually dimensioned such that C is exceeded only a small fraction of time. Hence, our main conclusion is that our approximation (for C close to ρ) nicely complements the one proposed by Guillemin and Simonian.

4.8 Concluding remarks

We studied the quantities duration, number of arrivals, and area for C-congestion periods of an M/M/∞ queue. We presented a derivation using recursive relations thus obtaining all moments and 'joint expectations' of the above quantities. The starting conditions of the recursions correspond to the busy period (a 0-congestion period); it is noted that the derivation of the higher moments and the joint expectations of these busy-period quantities were far from trivial, and followed from tedious calculations.

Furthermore, we introduced C-intercongestion periods, which are the intervals in which the system is *below* level C. Analogously to C-congestion periods, recursive relations are presented for the moments and joint expectations of the quantities. These are also solved in terms of a starting condition, but in contrast with C-congestion periods, the starting conditions of C-intercongestion period quantities are easily obtained: all moments and joint expectations of 0-intercongestion period quantities are 0. For the C-intercongestion period we also derived the Laplace transforms of the duration and the so-called intercongestion triple.

Finally, it was shown that an intercongestion period can be used in an approximation of a congestion period, in particular for C close to the average load ρ. This approximation is especially useful as the calculations of the intercongestion period are numerically more stable than those of the congestion periods. The proposed approximation complements other approximations proposed in literature, as these tend to be less accurate for C close to the average load ρ.

Chapter 5

Tail asymptotics of congestion periods

5.1 Introduction

In this chapter we, once again, consider C-congestion periods of an M/M/∞ queue, but now we focus on the tail asymptotics of the quantities D_{C}, A_{C}, and N_{C}. Knowledge of the probabilistic characteristics of a C-congestion period is useful, for instance when designing packet-based networks, similar to what we did in Chapter 3. These networks are typically designed such that the impact of overflows is limited, or, in other words, C should be chosen such that long congestion periods are rare. More precisely, we wish to find a value for C such that the probability that the duration of the congestion period exceeds a given threshold is kept very small, typically in the order of 10^{-4} to 10^{-6}. These probabilities relate to the so-called tail of a probability distribution; in this chapter we characterize the tail behavior of the distributions of the quantities using large-deviations theory (cf. Section 2.2.2).

5.1.1 Contribution

Our contribution is to shed light on the tail probabilities $\mathbb{P}(D_{\mathrm{C}} > x)$, $\mathbb{P}(A_{\mathrm{C}} > x)$, and $\mathbb{P}(N_{\mathrm{C}} > x)$, for which hardly anything is known (cf. Section 4.1.1). In more detail, our contributions are the following.

Asymptotics under many-flows scaling. We scale the arrival rate Λ by a parameter n, i.e., we let $\Lambda \equiv n\lambda$, but leave the mean service time μ^{-1} unchanged, so that the system load becomes $\mathrm{P} = n\rho$, where $\rho := \lambda/\mu$. Starting a congestion period at level $\mathrm{C} \equiv nc$, our aim is to find the asymptotics of the probabilities $\mathbb{P}(D_{nc} > x)$, $\mathbb{P}(A_{nc} > nx)$, and $\mathbb{P}(N_{nc} > nx)$, for n large and $x > 0$ given. We succeed in doing so by using *sample-path large-deviations techniques*, relying predominantly on the theory developed in [123]. It turns out that the probability $\mathbb{P}(D_{nc} > x)$ decays roughly exponentially in n (that is, we show that $-n^{-1} \cdot \log \mathbb{P}(D_{nc} > x)$ tends to a positive, finite limit); analogous results hold for $\mathbb{P}(A_{nc} > nx)$ and $\mathbb{P}(N_{nc} > nx)$. Assuming that $c > \rho$ (which we shall do throughout this chapter), we explicitly identify the corresponding decay rates. As a by-product, we also identify the *most likely path*, which is essentially the most probable way in which the events under consideration occur: given that the rare event happens, then with overwhelming probability

it does so via a path in the direct neighborhood of the most likely path. Clearly, the many-flows scaling is particularly suitable for systems with a considerable level of multiplexing.

Asymptotics of the Gaussian counterpart. We approximate the $\text{M}/\text{M}/\infty$ model under the many-flows scaling by an appropriate Gaussian process, the so-called *Gaussian counterpart* of the $\text{M}/\text{M}/\infty$ system; see for further background on this type of approximation [3], and [54, Section 2]. We argue that this counterpart is the so-called *integrated Ornstein-Uhlenbeck* (iOU) model [85]. Now we can analyze the rare events under considerations by applying *sample-path large deviations results*, viz. the generalized version of Schilder's theorem [7, 35, 85]. Owing to the fact that the iOU process has a well-defined rate process (unlike for instance fractional Brownian motion), the corresponding large-deviations rate function can be expressed in a considerably more explicit way than in the standard version of the generalized version of Schilder's theorem.

Relying on this explicit sample-path large-deviations result, we determine the tail asymptotics of $\mathbb{P}(D_{nc} > x)$ and $\mathbb{P}(A_{nc} > nx)$ for n large for the Gaussian counterpart; the quantity N_{nc} does not have a meaningful Gaussian counterpart. As could be expected, these Gaussian asymptotics become increasingly accurate when c approaches ρ from above, that is, in a heavy-traffic setting. Again we also find the corresponding *most likely paths*.

Uniform bounds. All results mentioned above relate to the $\text{M}/\text{M}/\infty$ model under the many-flows scaling, and are in terms of (relatively crude) asymptotics. For practical purposes, however, it would be helpful to have bounds — particularly *upper* bounds — on the probabilities of interest, that are valid for all parameter settings (i.e., not just in an asymptotic regime). Using change-of-measure arguments, and relying on the celebrated Chernoff bound, we are able to derive such uniform upper bounds; these are in closed-form.

Importance sampling algorithms. Estimating the probabilities $\mathbb{P}(D_{\text{C}} > x)$, $\mathbb{P}(A_{\text{C}} > x)$, and $\mathbb{P}(N_{\text{C}} > x)$ by direct, naïve simulation is inherently difficult, particularly for large x, because of the rarity of the event under consideration. This motivates the search for 'fast-simulation' techniques [24]. The change-of-measures, mentioned above in the context of the uniform bounds, suggest parameters that can be used in importance-sampling procedures. In a numerical study, we compare the estimates (as obtained under the many-flows scaling), as well as the uniform upper bounds, with results obtained from importance-sampling-based simulations. The importance-sampling schemes turn out to yield a substantial speed-up compared to direct, naïve simulations. They are very useful for practical purposes, as the uniform upper bounds tend to overestimate the probabilities of interest.

5.1.2 Outline

Section 5.2 introduces the model, i.e., the M/M/∞ queue, the C-congestion period, and formally defines the quantities of interest, i.e., D_{C}, N_{C}, and A_{C}. In Section 5.3 we present the analysis of the tail probabilities under the many-flows scaling, whereas Section 5.4 addresses the Gaussian counterpart. Where Sections 5.3 and 5.4 present logarithmic asymptotics of the scaled model, in Section 5.5 we establish uniform, closed-form (upper) bounds on the probabilities of interest. Further, this section also describes change-of-measures that can be used in importance-sampling-based simulation schemes. In Section 5.6 we numerically evaluate the decay rates of Sections 5.3 and 5.4, and compare these with the uniform bounds, as well as with simulation results (obtained by the importance-sampling procedure sketched in Section 5.5). Section 5.7 concludes.

5.2 Model and preliminaries

Model. We consider a resource at which flows arrive according to a Poisson process with intensity Λ, and at which the jobs stay for an exponentially distributed time with mean μ^{-1}. We are thus in the setting of the (classical) M/M/∞ model. The following properties are well-known: i) in stationarity the number of trunks occupied has a Poisson distribution with mean $\mathrm{P} := \Lambda/\mu$; ii) the number of arriving flows in an interval of length t is Poisson distributed with mean Λt, and each of them has arrived on an epoch uniformly distributed over the interval $[0, t]$, independently of the other arrivals.

Congestion periods. We define the key quantities studied in this chapter. To this end, we first need some additional notation. First, let $X(t)$ denote the number of flows present at time t; $X(\cdot)$ constitutes a continuous-time Markov chain on $\{0, 1, \ldots\}$, with upward transition rate λ, and downward transition rate (from state k) $k\mu$. $A(t)$ is the defined as the work generated by the flows in the interval $[0, t]$, which is essentially the integral of $X(\cdot)$:

$$A(t) := \int_0^t X(s)\mathrm{d}s.$$

We also need the discrete-time embedding of the above described continuous-time process. We let Y_m be the number of flows present after m jumps, where a jump is an arrival or departure. It is clear that $(Y_m)_{m\in\mathbb{N}}$ is a discrete-time Markov chain, with upward transition probability $\lambda/(\lambda + k\mu)$ and downward transition probability $k\mu/(\lambda + k\mu)$ (from state k).

A first observation is that the process $A(t)$ is rather convenient to work with, owing to its nice structure. In particular, using elementary arguments and relying

on properties i) and ii) mentioned above, it can be verified that, for $\vartheta < \mu$,

$$\log \mathbb{E}(e^{\vartheta A(t)} \mid X(0) = \mathrm{C}+1) = (\mathrm{C}+1)\log\left(\frac{\mu}{\mu-\vartheta} - \frac{\vartheta}{\mu-\vartheta}e^{-(\mu-\vartheta)t}\right) + \frac{\Lambda t\vartheta}{\mu-\vartheta} - \frac{\Lambda\vartheta}{(\mu-\vartheta)^2}(1-e^{-(\mu-\vartheta)t}). \tag{5.1}$$

It is clear that $A(t)$ is smaller than the amount of work that has arrived in $[0,t]$ when the full flow would have been 'injected' instantaneously. This reasoning yields that

$$\log \mathbb{E}(e^{\vartheta A(t)} \mid X(0) = \mathrm{C}+1) \leq (\mathrm{C}+1)\log\left(\frac{\mu}{\mu-\vartheta}\right) + \frac{\Lambda t\vartheta}{\mu-\vartheta}, \tag{5.2}$$

which is in agreement with (5.1). More specifically, it is readily checked that $\mathbb{E}A(t) = \mathrm{P}t$ and

$$\mathbb{V}\mathrm{ar}A(t) = \frac{2\Lambda}{\mu^3}\left(t\mu - 1 + e^{-t\mu}\right). \tag{5.3}$$

We study the tail behavior of the following three random variables:

$$D_{\mathrm{C}} := \inf\{t \geq 0 : X(t) = \mathrm{C} \mid X(0) = \mathrm{C}+1\};$$

$$A_{\mathrm{C}} := (A(D_{\mathrm{C}}) - \mathrm{C}D_{\mathrm{C}} \mid X(0) = \mathrm{C}+1);$$

$$N_{\mathrm{C}} := \frac{1}{2}\inf\{m \in \mathbb{N} : Y_m = \mathrm{C} \mid Y_0 = \mathrm{C}+1\} - \frac{1}{2}.$$

We refer to D_{C} as the duration of the *congestion period* above level C. A_{C} can be interpreted as a proxy for the amount of traffic lost during a congestion period (in systems in which the number of lines is truncated at C); informally, this is the *area* under the graph of $X(s) - \mathrm{C}s$ during a congestion period. Furthermore, it is readily verified that N_{C} corresponds to the *number* of arrivals during a congestion period (which equals the number of departures during a congestion period, decreased by 1). We throughout assume that $\mathrm{P} < \mathrm{C}$.

We recall that the LTs of the distributions of D_{C}, A_{C}, and N_{C} were found by Guillemin and Simonian [58] in terms of special functions, whereas Preater [104] elegantly derived their joint LT. Chapter 4 already presented explicit expressions for expected values and variances of D_{C}, A_{C}, and N_{C}, and their covariances.

Performance metrics. As an alternative to deriving the distribution functions of D_{C}, A_{C}, and N_{C} from the Laplace transforms, we apply a scaling that allows explicit asymptotic analysis. In this scaling one identifies $\Lambda \equiv n\lambda$ and $\mathrm{C} \equiv nc$, where n is large; likewise $\mathrm{P} \equiv n\rho$. We can equivalently write that the total traffic arrival process $A^n(t)$ corresponds to the sum of n independent and identically distributed arrival processes, each distributed as the process $A(t)$ introduced above, but with Λ

replaced by λ, and C replaced by c. Similarly $X^n(t)$ is defined as the aggregate rate process, and $(Y^n_m)_{m\in\mathbb{N}}$ as the aggregate rate at jump epochs.

Our first goal is to asymptotically characterize the probabilities $\mathbb{P}(D_{nc} > x)$, $\mathbb{P}(A_{nc} > nx)$, and $\mathbb{P}(N_{nc} > nx)$, for n large. The scaling applied is usually referred to as the 'many-flows scaling' [21, 48, 85], and is particularly appropriate if the level of multiplexing is reasonably large. We recall that it is assumed that $c > \rho$, so that the events under consideration are increasingly rare when n grows large. We rely on large-deviations theory to show that the above probabilities decay essentially exponentially in n, and to explicitly determine the corresponding exponential decay rates, i.e., for $x > 0$,

$$\delta(x) := \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(D_{nc} > x),$$

and likewise also the decay rate corresponding to a large area, for $x > 0$,

$$\alpha(x) := \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(A_{nc} > nx),$$

and the decay rate corresponding to many arriving flows per congestion period, for $x > 0$,

$$\nu(x) := \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(N_{nc} > nx).$$

5.3 Large deviations analysis of congestion period

In this section we consider the M/M/∞ model under the many-flows scaling that was described above, and apply sample-path large deviations to compute the decay rates (n large) of $\mathbb{P}(D_{nc} > x)$, $\mathbb{P}(A_{nc} > nx)$, and $\mathbb{P}(N_{nc} > nx)$. In the first subsection we review the main results for sample-path large deviations of Markovian systems. Then we subsequently determine the decay rates $\delta(x)$, $\alpha(x)$, and $\nu(x)$.

5.3.1 Sample-path large-deviations theory

In our exposition, we rely extensively on the framework presented in Shwartz and Weiss [123]. In this framework a crucial role is played by the *local rate function*. In case of the M/M/∞ process, this function is defined as (cf. Expression (2.3))

$$I_x(u) := \sup_{\vartheta} \left(\vartheta u - \lambda(e^{\vartheta} - 1) - \mu x(e^{-\vartheta} - 1)\right).$$

In fact, the local rate function measures the 'cost' of moving in direction u, when the (scaled) process is in state x, in the following sense. Suppose x flows are present.

Then the position of the scaled process $X^n(\cdot)/n$ after ε time units (ε small) is, in expectation, roughly $x + (\lambda - \mu x)\varepsilon$, and hence the 'most likely' derivative of moving is $u(x) := \lambda - \mu x$. Indeed, it is verified that $I_x(u(x)) = 0$: there is no 'cost' involved in moving into this most-likely direction. It is checked that any other direction yields strictly positive costs. We further remark that the function $I_x(u)$ can be calculated explicitly (the first order condition being a quadratic equation), but this is, for the purposes of the present study, not necessary.

Having the local rate function at our disposal, we can define the *action functional*. Informally, this action functional $\mathbb{I}(f)$ represents the 'cost' of the scaled process $X^n(\cdot)/n$ following a path $f(\cdot)$:

$$\mathbb{I}(f) := \int_{-\infty}^{\infty} I_{f(s)}(f'(s))\mathrm{d}s.$$

It is a matter of elementary calculus to check that, considering just the time after time 0, the path $\varphi(s) := \rho + (\varphi_0 - \rho)e^{-\mu s}$ (for some $\varphi_0 > 0$) yields cost 0: as $\varphi'(s) = (\lambda - \varphi_0\mu)e^{-\mu s}$,

$$\mathbb{I}(\varphi) = \int_0^{\infty} \sup_{\vartheta} \big(\vartheta(\lambda - \varphi_0\mu)e^{-\mu s} \\ -\lambda(e^{\vartheta} - 1) - (\lambda + (\varphi_0\mu - \lambda)e^{-\mu s})(e^{-\vartheta} - 1)\big)\,\mathrm{d}s = 0;$$

this answer makes sense, as this path is essentially the 'average path' starting at φ_0 at time 0 to the system's equilibrium value ρ.

Using this framework, the following *sample-path large-deviations principle* can be stated:

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} X^n(\cdot) \in \mathscr{S}\right) = -\inf_{f\in\mathscr{S}} \mathbb{I}(f). \tag{5.4}$$

Informally, one find the most likely path f in the set $\mathscr{S}$, say $f^\star$; given that the event $\{X^n(\cdot)/n \in \mathscr{S}\}$ occurs, the realization will be close to $f^\star$. Intentionally, (5.4) has been stated in a slightly imprecise way: in fact one has two inequalities, respectively for open and closed sets (in the path-space). These issues are not crucial in the scope of this study, and we refer to [123] for these and related details.

In discrete time, i.e., for the process Y_m^n, a similar framework can be set up, see for instance Bucklew [23]. Then the local rate function is given by

$$J_x(u) := \sup_{\vartheta} \left(\vartheta u - \log\left(\frac{\lambda}{\lambda + \mu x} e^{\vartheta} + \frac{\mu x}{\lambda + \mu x} e^{-\vartheta}\right)\right).$$

Again, this function can be evaluated in a more explicit manner, but we will refrain from doing this. Similar to before, we can define the action functional as

$$\mathbb{J}(f) := \int_{-\infty}^{\infty} J_{f(s)}(f'(s))\mathrm{d}s.$$

Again we have a sample-path large-deviations principle:

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} Y^n(\cdot) \in \mathscr{S}\right) = -\inf_{f\in\mathscr{S}} \mathbb{J}(f). \tag{5.5}$$

5.3.2 Congestion period

We cast our problem of identifying the decay rate of $\mathbb{P}(D_{nc} > x)$ into the large-deviations framework of the previous subsection. Immediately from the sample-path large-deviations result (5.4), we have that

$$\delta(x) = -\inf_{f\in T} \mathbb{I}(f),$$

with $T := \{f \mid \forall s \in [0,x] : f(s) \geq c, f(0) = c\}$. Heuristically reasoning, as we are looking for the 'cheapest' path in T, it cannot be that the optimal path is such that $f(x) > c$, as otherwise even a longer congestion period could be obtained 'for free'. Based on this argumentation, it is seen that $\inf_{f\in T} \mathbb{I}(f) = \inf_{f\in\bar{T}} \mathbb{I}(f)$, with

$$\bar{T} := \{f \mid \forall s \in [0,x] : f(s) \geq c, f(0) = f(x) = c\}.$$

We therefore further study the following variational problem:

$$\delta(x) = -\inf_{f\in\bar{T}} \int_0^x I_{f(s)}(f'(s))\mathrm{d}s.$$

PROPOSITION 5.3.1. *For $x \geq 0$,*

$$\delta(x) = -x\delta^\star; \quad \delta^\star := (\sqrt{\lambda} - \sqrt{\mu c})^2.$$

Proof We prove this result by subsequently establishing a lower bound and an upper bound. Define the path f_c through $f_c(s) = c$ for all $s \in [0,x]$. As $f_c \in \bar{T}$, it follows that

$$\begin{aligned}
\delta(x) &\geq -\mathbb{I}(f_c) = -x \sup_\vartheta \left(-\lambda(e^\vartheta - 1) - \mu c(e^{-\vartheta} - 1)\right) \\
&= -x\left(\lambda - 2\sqrt{\lambda\mu c} + \mu c\right) = -x(\sqrt{\lambda} - \sqrt{\mu c})^2;
\end{aligned}$$

the optimizing ϑ equals $\vartheta^\star := \frac{1}{2}\log(\mu c/\lambda) = \frac{1}{2}\log(c/\rho) > 0$. Hence we have proven the lower bound. On the other hand,

$$\begin{aligned}
\delta(x) &\leq -\inf_{f\in\bar{T}} \int_0^x \left(\vartheta^\star f'(s) - \lambda(e^{\vartheta^\star} - 1) - \mu f(s)(e^{-\vartheta^\star} - 1)\right) \mathrm{d}s \\
&\overset{\text{(i)}}{=} \sup_{f\in\bar{T}} \int_0^x \left(\lambda(e^{\vartheta^\star} - 1) + \mu f(s)(e^{-\vartheta^\star} - 1)\right) \mathrm{d}s \\
&= \sup_{f\in\bar{T}} \int_0^x \left(\sqrt{\lambda\mu c} - \lambda + f(s)\sqrt{\frac{\lambda\mu}{c}} - \mu f(s)\right) \mathrm{d}s \\
&\overset{\text{(ii)}}{\leq} \int_0^x \left(\sqrt{\lambda\mu c} - \lambda + c\sqrt{\frac{\lambda\mu}{c}} - \mu c\right) \mathrm{d}s = -x(\sqrt{\lambda} - \sqrt{\mu c})^2,
\end{aligned}$$

recalling for i) that $f(0) = f(x) = c$ for all $f \in \bar{T}$, and for ii) Lemma 5.A.1 (to be found in the appendix). This yields the upper bound: all $f \in \bar{T}$ yield at most decay rate $-x\delta^\star$, as desired. □

REMARK 5.3.2. *Also [123, Section 13.5.6] focuses on congestion periods, albeit with a slightly different definition. They consider the random variable*

$$B_{\mathrm{C}} := \sup\{t \geq 0 : A(t) \geq \mathrm{C}t \mid X(0) = \mathrm{C} + 1\}.$$

Again invoking the sample-path large-deviations result (5.4), the decay rate of $\mathbb{P}(B_{nc} > x)$ can be rewritten as $-\inf_{f\in\mathscr{B}} \mathbb{I}(f)$, where

$$\mathscr{B} := \left\{ f \;\middle|\; \forall s \in [0,x] : \int_0^s f(r)\mathrm{d}r \geq cs \;, f(0) = c \right\}.$$

In [123, Eq. (13.65)] it is claimed that this decay rate equals $-x\delta^\star$, i.e., $\delta(x)$. This, however, seems an error, and the correct decay rate should be [87]

$$-\sup_{\vartheta}(\vartheta cx - c\log\phi(\vartheta,x) - \psi(\vartheta,x)), \tag{5.6}$$

where, cf. (5.1),

$$\phi(\vartheta,t) := \frac{\mu}{\mu-\vartheta} - \frac{\vartheta}{\mu-\vartheta}e^{-(\mu-\vartheta)t} \tag{5.7}$$

$$\psi(\vartheta,t) := \frac{\lambda t\vartheta}{\mu-\vartheta} - \frac{\lambda\vartheta}{(\mu-\vartheta)^2}(1-e^{-(\mu-\vartheta)t}). \tag{5.8}$$

The proof is based on the fact that it turns out that the most likely path in

$$\bar{\mathscr{B}} := \left\{ f \;\middle|\; \int_0^x f(s)\mathrm{d}s \geq cx \;, f(0) = c \right\}$$

lies in $\mathscr{B}$; notice that $\bar{\mathscr{B}} \supseteq \mathscr{B}$. It is a direct implication of Cramér's theorem that the decay rate of the optimal path in $\bar{\mathscr{B}}$ indeed equals (5.6). Hence, the decay rate of $\mathbb{P}(B_{nc} > x)$ is (5.6), which is larger than $-x\delta^\star$. In other words: the event is less rare than suggested by [123, Equation (13.65)]; there is a cheaper path than $f_c(\cdot)$, namely a path that is strictly larger than c on $(0,x)$. For additional details, we refer to Case 3 in Theorem 3.1 in [87]. ♢

5.3.3 Area

We now turn our attention to the tail asymptotics of the area A_{nc}. Again applying the sample-path large-deviations result (5.4), we obtain

$$\alpha(x) = -\inf_{f\in\mathscr{A}} \mathbb{I}(f), \tag{5.9}$$

where $\mathscr{A}$ is the set of paths that lead to an area of at least x:

$$\mathscr{A} := \left\{ f \;\middle|\; \exists t > 0 : \int_0^t f(s)\mathrm{d}s \geq x + ct,\ \forall s \in [0,t] : f(s) \geq c, f(0) = c \right\}.$$

In the following lemma we prove that the set $\bar{\mathscr{A}}$, given by

$$\bar{\mathscr{A}} := \left\{ f \;\middle|\; \exists t > 0 : \int_0^t f(s)\mathrm{d}s \geq x + ct\ , f(0) = c \right\},$$

which is evidently larger than $\mathscr{A}$, contains the optimal path in $\mathscr{A}$.

LEMMA 5.3.3. *The following identity holds:*

$$\inf_{f\in\mathscr{A}} \mathbb{I}(f) = \inf_{f\in\bar{\mathscr{A}}} \mathbb{I}(f).$$

Proof As mentioned above, $\mathscr{A} \subseteq \bar{\mathscr{A}}$. Hence, in order to prove the stated, it suffices to show that the minimizer in the larger set, $\bar{\mathscr{A}}$, is element of the smaller set, $\mathscr{A}$.

This follows directly from a reasoning analogous to Section 13.2 of [123]. To this end, first observe that

$$\inf_{f\in\bar{\mathscr{A}}} \mathbb{I}(f) = \inf_{t>0} \inf_{f\in\bar{\mathscr{A}}_t} \mathbb{I}(f), \text{ where } \bar{\mathscr{A}}_t := \left\{ f \;\middle|\; \int_0^t f(s)\mathrm{d}s \geq x + ct\ , f(0) = c \right\}.$$

For a model intimately related to our M/M/∞ model (viz. the model with exponential on-off sources) [123] identifies, using calculus-of-variations techniques, the optimizing $t^\star$, as well as the corresponding most likely path $f^\star$ in $\bar{\mathscr{A}}_{t^\star}$. This path $f^\star$ turns out to be a symmetric hyperbolic cosine, i.e., $t^\star$ is such that $f^\star(0) = f^\star(t^\star) = c$, $f^\star(s) > c$ for all $s \in (0, t^\star)$, and

$$\int_0^{t^\star} f^\star(s)\mathrm{d}s = x + ct^\star.$$

Mimicking the analysis in [123], it is elementary to check that the same properties hold for the M/M/∞ model. This implies that $f^\star \in \mathscr{A}$, which proves the stated. □

We have reduced the problem of finding $\alpha(x)$ to finding the most likely path in $\bar{\mathscr{A}}$. Before actually computing this decay rate, which we will do in Proposition 5.3.5, we first establish another auxiliary result that reveals a relation between the decay rate corresponding to the most likely path in $\bar{\mathscr{A}}$ on one hand, and the decay rate of tail probabilities in a related queueing system.

To this end, consider a queue fed a Poisson stream of jobs (rate $n\lambda$), each staying in the system for an exponentially distributed time (mean μ^{-1}), generating traffic at a unit rate while in the system, where the buffer is emptied at a constant rate nc. Let

Q^n denote the steady-state buffer content of this queue; as before, it is assumed that $\rho < c$. The following distributional equality is well-known:

$$Q^n \stackrel{\mathrm{d}}{=} \sup_{t\geq 0} A^n(t) - nct,$$

a relation usually attributed to Reich [106]. Define, for $\vartheta < \mu$,

$$\log N_t(\vartheta) = \rho(\phi(\vartheta, t) - 1) + \psi(\vartheta, t), \tag{5.10}$$

where $\phi(\vartheta, t)$ and $\psi(\vartheta, t)$ are given in (5.7).

LEMMA 5.3.4. *The following identity holds:*

$$\begin{aligned} -\inf_{f\in\mathscr{A}} \mathbb{I}(f) - (\rho - c) - c\log\frac{c}{\rho} &= \lim_{n\to\infty} \frac{1}{n}\log \mathbb{P}(Q^n > nx) \\ &= -\inf_{t\geq 0}\sup_{\vartheta>0} \left(\vartheta(x+ct) - \log N_t(\vartheta)\right). \end{aligned} \tag{5.11}$$

Proof The first equality follows from reasoning as in Section 13.2 of [123]. The decay rate of the steady-state probability $\mathbb{P}(Q^n > nx)$ can be rewritten as $-\inf \mathbb{I}(f)$, where the infimum is over all f that start off in ρ at time $-\infty$, and for which the busy period in which overflow (over level x) is reached starts at time 0; if level x is reached at some time $t > 0$, this means that

$$\int_0^t f(s)\mathrm{d}s = b + ct,$$

and in addition $f(s) \geq c$ for all $s \in [0, t]$, and $f(0) = c$. Now an elementary splitting argument yields that this decay rate can be decomposed into

$$-\inf_{f\in\mathscr{A}^-} \mathbb{I}(f) - \inf_{f\in\mathscr{A}} \mathbb{I}(f),$$

where $\mathscr{A}^- := \{f \mid f(-\infty) = \rho, f(0) = c\}$. Using arguments as in Section 13.1 of [123],

$$\inf_{f\in\mathscr{A}^-} \mathbb{I}(f) = (\rho - c) + c\log\frac{c}{\rho}.$$

This, and an application of Lemma 5.3.3, proves the first equality.

The second equality follows from Botvich and Duffield [21], as follows. Let $N_t(\vartheta)$ be the moment generating function of the work generated by a *single* Poisson stream of jobs (that is, with rate λ), each staying in the system for an exponentially distributed time (with mean μ^{-1}):

$$\log N_t(\vartheta) = \rho(\phi(\vartheta, t) - 1) + \psi(\vartheta, t);$$

here it is used that the number of flows present at time 0 has a Poisson distribution with mean ρ. According to [21], the decay rate of $\mathbb{P}(Q^n > nx)$ equals

$$-\inf_{t\geq 0}\sup_{\vartheta>0}\left(\vartheta(x+ct)-\log N_t(\vartheta)\right). \tag{5.12}$$

This implies the second equality. □

Now the decay rate $\alpha(x)$ follows immediately from Lemma 5.3.4, in conjunction with Equation (5.9).

PROPOSITION 5.3.5. *For $x \geq 0$,*

$$\alpha(x) = -\inf_{t\geq 0}\sup_{\vartheta>0}\left(\vartheta(x+ct)-\log N_t(\vartheta)\right) + (\rho - c) + c\log\frac{c}{\rho}.$$

As opposed to $\delta(x)$ and (as we will see later) $\nu(x)$, there is no explicit, closed-form available for $\alpha(x)$. It is, however, possible to explicitly characterize $\alpha(x)$ for $x \downarrow 0$ and $x \to \infty$. We define

$$\begin{aligned}
\alpha_0^\star &:= 2\sqrt{2}\cdot\sqrt{\lambda\left(1-\frac{\rho}{c}+\frac{\rho}{c}\log\frac{\rho}{c}\right)};\\
\alpha_\infty^\star &:= \mu - \frac{\lambda}{c};\\
\beta_\infty^\star &:= \frac{(c-\rho)^2}{\rho} + c - \rho - c\log\frac{c}{\rho}.
\end{aligned}$$

PROPOSITION 5.3.6. *The asymptotic behavior of $\alpha(x)$ is given by*

$$\alpha(x) = -\alpha_0^\star\sqrt{x} - O(x) \quad \textit{as } x \downarrow 0;$$

$$\alpha(x) = -\beta_\infty^\star - \alpha_\infty^\star x + o(1) \quad \textit{as } x \to \infty.$$

Proof The behavior around $x = 0$ follows directly from Mandjes and Kim [86] (see the remark on the open model in Section 3), in conjunction with Lemma 5.3.4. It is readily verified that, in the notation used in that remark, $\vartheta_0 = \log(c/\rho)$, and then it is a matter of evaluating the expressions.

The behavior for $t \to \infty$ follows immediately from the expression for $N_t(\vartheta)$ for t large, and Theorem 3 of Botvich and Duffield [21]. The latter result states that the decay rate of $\mathbb{P}(Q^n > nx)$ equals $-\bar{\beta}_\infty^\star - \alpha_\infty^\star x + o(1)$ for x large, where $\alpha_\infty^\star$ solves

$$\lim_{t\to\infty}\frac{1}{t}\log N_t(\vartheta) = c\vartheta,$$

i.e., $\alpha_\infty^\star = \mu - \lambda/c = \mu(1-\rho/c)$, and

$$\bar{\beta}_\infty^\star := -\lim_{t\to\infty}\left(\log N_t(\alpha_\infty^\star) - c\,\alpha_\infty^\star\, t\right) = \frac{(c-\rho)^2}{\rho}.$$

Now an application of Lemma 5.3.4 yields the stated. □

5.3.4 Number of flows

We cast our problem of identifying the decay rate of $\mathbb{P}(N_{nc} > nx)$ into the large-deviations framework introduced earlier. We now use the discrete-time sample-path large deviations. Immediately from (5.5),

$$\nu(x) = -\inf_{f\in N} \mathbb{J}(f),$$

with $N := \{f \mid \forall s \in [0,2x] : f(s) \geq c, f(0) = c\}$. Analogously to the duration of the congestion period we are looking for the 'cheapest' path, which cannot be optimal if $f(2x) > c$, as otherwise even a longer congestion period could be obtained 'for free'. Hence $\inf_{f\in N} \mathbb{J}(f) = \inf_{f\in \bar{N}} \mathbb{J}(f)$, with

$$\bar{N} := \{f \mid \forall s \in [0,2x] : f(s) \geq c, f(0) = f(2x) = c\}.$$

We therefore further study the following variational problem:

$$\nu(x) = -\inf_{f\in \bar{N}} \int_0^{2x} J_{f(s)}(f'(s))\mathrm{d}s.$$

PROPOSITION 5.3.7. *For $x > 0$,*

$$\nu(x) = -x\nu^\star; \quad \nu^\star := 2\log\frac{\lambda+\mu c}{2\sqrt{\lambda\mu c}} = \log\frac{(\lambda+\mu c)^2}{4\lambda\mu c}.$$

Proof We prove this result by subsequently establishing a lower bound and an upper bound. Define the path f_c through $f_c(s) = c$ for all $s \in [0,2x]$. As $f_c \in \bar{N}$, it follows that

$$\begin{aligned}\nu(x) &\geq -\mathbb{J}(f_c) = -2x \cdot \sup_\vartheta \left(-\log\left(\frac{\lambda}{\lambda+\mu c}e^{\vartheta} + \frac{\mu c}{\lambda+\mu c}e^{-\vartheta}\right)\right)\\ &= 2x\cdot\log\left(\frac{\lambda\sqrt{\mu c/\lambda} + \mu c/\sqrt{\mu c/\lambda}}{\lambda+\mu c}\right) = -2x\cdot\log\frac{\lambda+\mu c}{2\sqrt{\lambda\mu c}};\end{aligned}$$

the optimizing ϑ equals $\vartheta^\star := \frac{1}{2}\log(\mu c/\lambda) = \frac{1}{2}\log(c/\rho) > 0$. Hence we have proven the upper bound. On the other hand,

$$\begin{aligned}\nu(x) &\leq -\inf_{f\in\bar{T}} \int_0^{2x}\left(\vartheta^\star f'(s) - \log\left(\frac{\lambda}{\lambda+\mu f(s)}e^{\vartheta^\star} + \frac{\mu f(s)}{\lambda+\mu f(s)}e^{-\vartheta^\star}\right)\right)\mathrm{d}s\\ &\overset{\text{(i)}}{=} -\inf_{f\in\bar{T}} \int_0^{2x}\left(-\log\left(\frac{\lambda}{\lambda+\mu f(s)}e^{\vartheta^\star} + \frac{\mu f(s)}{\lambda+\mu f(s)}e^{-\vartheta^\star}\right)\right)\mathrm{d}s\\ &= \sup_{f\in\bar{T}} \int_0^{2x}\log\left(\frac{1}{\lambda+\mu f(s)}\left(\lambda e^{\vartheta^\star} + \mu f(s)e^{-\vartheta^\star}\right)\right)\mathrm{d}s\\ &= \sup_{f\in\bar{T}} \int_0^{2x}\log\left(\frac{\sqrt{\lambda\mu c}\,(1+f(s)/c)}{\lambda+\mu f(s)}\right)\mathrm{d}s\\ &\overset{\text{(ii)}}{\leq} 2x\log\left(\frac{2\sqrt{\lambda\mu c}}{\lambda+\mu c}\right).\end{aligned}$$

Here i) is due to the fact that $f(0) = f(x) = c$ for all $f \in \bar{N}$, and ii) due to Lemma 5.A.2. This yields the lower bound: all $f \in \bar{N}$ yield at most decay rate $-x\nu^\star$, as desired. □

5.4 Large deviations analysis of the Gaussian counterpart

So far, we have considered the asymptotics of $\mathbb{P}(D_{nc} > x)$, $\mathbb{P}(A_{nc} > nx)$, and $\mathbb{P}(N_{nc} > nx)$, using sample-path large-deviations. In this section we approximate $A^n(\cdot)$ by its so-called *Gaussian counterpart* $\bar{A}^n(\cdot)$, that is, the superposition of n Gaussian processes, each with mean and variance given through

$$\mathbb{E}\bar{A}(t) = \rho t, \quad v(t) := \mathbb{V}\mathrm{ar}\bar{A}(t) = \frac{2\lambda}{\mu^3}\left(t\mu - 1 + e^{-t\mu}\right),$$

cf. (5.3). This specific Gaussian process is known as the *integrated Ornstein-Uhlenbeck* (iOU) process. The procedure of replacing stochastic processes by their Gaussian counterpart was proposed and extensively motivated by Addie, Mannersalo, and Norros [3]; see for a further justification in the M/M/∞ case also [54, Section 2] and [132].

With $\bar{A}(\cdot)$ corresponding to a single iOU process with the mean and variance define above, it is observed that $\bar{A}(\cdot)$ is a genuine Gaussian counterpart of our original Markovian system, in the sense that the following two properties hold:

- In the first place, the 'rate process' $\bar{X}(t) := \bar{A}'(t)$ is well-defined (which is not the case for several other Gaussian processes such as fractional Brownian motion). This is a stationary Gaussian process (where $\bar{A}(\cdot)$ was a Gaussian process with stationary increments). It is readily verified that $\mathbb{E}\bar{X}(t) = \rho$, and

  $$\mathbb{V}\mathrm{ar}\big(\bar{X}(t)\big) = \lim_{\varepsilon \downarrow 0} \frac{v(t+\varepsilon) - v(t)}{\varepsilon^2} = \rho;$$

  these results are in agreement with the fact that in the original (that is, non-Gaussian) model $X(t)$ has a Poisson distribution with mean (and hence also variance) ρ.

- In the second place the Gaussian process has a Markovian structure, in the sense that, for $0 < u < T$ and $s > 0$,

  $$(\bar{A}(T, T+s) \mid \bar{X}(T), \bar{A}(0,u)) \stackrel{\mathrm{d}}{=} (\bar{A}(T, T+s) \mid \bar{X}(T)).$$

  This follows by showing that both sides of the previous display have the same mean and variance. We briefly present the procedure for the mean; the variance

can be done analogously. To this end, recall

$$\mathbb{E}(Y_1 \mid Y_2 = y_2, Y_3 = y_3) = \mathbb{E}Y_1 + \begin{pmatrix} \mathbb{C}\mathrm{ov}(Y_1, Y_2) \\ \mathbb{C}\mathrm{ov}(Y_1, Y_3) \end{pmatrix}^{\mathrm{T}} \Sigma^{-1} \begin{pmatrix} y_2 - \mathbb{E}Y_2 \\ y_3 - \mathbb{E}Y_3 \end{pmatrix};$$

$$\mathbb{V}\mathrm{ar}(Y_1 \mid Y_2 = y_2, Y_3 = y_3) = \mathbb{V}\mathrm{ar}(Y_1) + \begin{pmatrix} \mathbb{C}\mathrm{ov}(Y_1, Y_2) \\ \mathbb{C}\mathrm{ov}(Y_1, Y_3) \end{pmatrix}^{\mathrm{T}} \Sigma^{-1} \begin{pmatrix} \mathbb{C}\mathrm{ov}(Y_1, Y_2) \\ \mathbb{C}\mathrm{ov}(Y_1, Y_3) \end{pmatrix}$$

where

$$\Sigma := \begin{pmatrix} \mathbb{V}\mathrm{ar}(Y_2) & \mathbb{C}\mathrm{ov}(Y_2, Y_3) \\ \mathbb{C}\mathrm{ov}(Y_2, Y_3) & \mathbb{V}\mathrm{ar}(Y_3) \end{pmatrix}.$$

As a consequence, $\mathbb{E}(\bar{A}(T, T+s) \mid \bar{X}(T) = x, \bar{A}(0,u) = y)$ equals

$$\rho s + \begin{pmatrix} \mathbb{C}\mathrm{ov}\big(\bar{A}(T, T+s), \bar{X}(T)\big) \\ \mathbb{C}\mathrm{ov}\big(\bar{A}(T, T+s), \bar{A}(0,u)\big) \end{pmatrix}^{\mathrm{T}} \times \begin{pmatrix} \rho & \mathbb{C}\mathrm{ov}\big(\bar{X}(T), \bar{A}(0,u)\big) \\ \mathbb{C}\mathrm{ov}\big(\bar{X}(T), \bar{A}(0,u)\big) & v(u) \end{pmatrix}^{-1} \begin{pmatrix} x - \rho \\ y - \rho u \end{pmatrix}.$$

It is a matter of straightforward computations to verify that

$$\begin{aligned} \mathbb{C}\mathrm{ov}\big(\bar{A}(T, T+s), \bar{X}(T)\big) &= \frac{\lambda}{\mu^2}(1 - e^{-\mu s}); \\ \mathbb{C}\mathrm{ov}\big(\bar{A}(T, T+s), \bar{A}(0,u)\big) &= \frac{\lambda}{4\mu^3}\, e^{-\mu T}(1 - e^{-\mu s})(e^{\mu u} - 1); \\ \mathbb{C}\mathrm{ov}\big(\bar{X}(T), \bar{A}(0,u)\big) &= \frac{\lambda}{4\mu^2}\, e^{-\mu T}(e^{\mu u} - 1). \end{aligned}$$

Now tedious calculus yields that

$$\mathbb{E}(\bar{A}(T, T+s) \mid \bar{X}(T) = x, \bar{A}(0,u) = y) = \rho s + \frac{1 - e^{-\mu s}}{\mu} \cdot (x - \rho),$$

which is in agreement with $\mathbb{E}(\bar{A}(T, T+s) \mid \bar{X}(T) = x)$ (observe that, in particular, u and y cancel). Similarly, $\mathbb{V}\mathrm{ar}\big(\bar{A}(T, T+s) \mid \bar{X}(T) = x, \bar{A}(0,u) = y\big)$ coincides with $\mathbb{V}\mathrm{ar}\big(\bar{A}(T, T+s) \mid \bar{X}(T) = x\big)$, and equals $v(s) + (\lambda/\mu^3)(1 - e^{-\mu s})^2 = (2\lambda/\mu^3)(s\mu - 3/2 + 2e^{-s\mu} - e^{-2\mu s}/2)$.

Useful relations. We give a number of additional useful relations:

$$\mathbb{E}(\bar{A}(0,t) \mid \bar{X}(0)=x) = \rho t + \frac{(1-e^{-t\mu})}{\mu}(x-\rho).$$

$$\mathbb{V}\mathrm{ar}(\bar{A}(0,t) \mid \bar{X}(0)=x) = \frac{2\lambda}{\mu^3}\left(t\mu - \frac{3}{2} + 2e^{-t\mu} - \frac{1}{2}e^{-2t\mu}\right).$$

Now

$$\begin{aligned}\mathbb{E}(\bar{X}(t) \mid \bar{X}(0)=x) &= \mathbb{E}\left(\lim_{\epsilon\downarrow 0}\frac{\bar{A}(0,t+\epsilon)-\bar{A}(0,t)}{\epsilon} \mid \bar{X}(0)=x\right)\\ &= \rho + e^{-t\mu}(x-\rho)\end{aligned}$$

entails that $\mathbb{E}(\bar{X}(\epsilon) \mid \bar{X}(0)=x) = x + \epsilon(\lambda - \mu x) + O(\epsilon^2)$, for $\epsilon \downarrow 0$, and likewise

$$\begin{aligned}\mathbb{V}\mathrm{ar}\big(\bar{X}(t) \mid \bar{X}(0)=x\big) &= \mathbb{V}\mathrm{ar}\left(\lim_{\epsilon\downarrow 0}\frac{\bar{A}(0,t+\epsilon)-\bar{A}(0,t)}{\epsilon} \mid \bar{X}(0)=x\right)\\ &= \rho(1-e^{-2t\mu})\end{aligned}$$

leads to $\mathbb{V}\mathrm{ar}\big(\bar{X}(\epsilon) \mid \bar{X}(0)=x\big) = 2\lambda\epsilon + O(\epsilon^2)$.

5.4.1 Sample-path large deviations theory

The computation of the decay rates $\bar{\delta}(x)$ of the congestion period, $\bar{\alpha}(x)$ of the area, and $\bar{\nu}(x)$ of the number of customers, can, as before, be done relying on a sample-path large-deviations result. In the setting of Gaussian processes, this result is known as (the generalized version of) *Schilder's theorem* [7, 35, 85]. It is noted that this result is of a rather implicit nature, in that there is in general no closed-form expression for the action functional (that is, we do not have an explicit expression for the 'cost' of a given path f). Owing to the fact that the iOU process has a well-defined rate process, however, the corresponding action functional can, for this specific Gaussian process, be expressed explicitly. The goal of this subsection is to identify this action functional — we do so by first heuristically deriving the sample-path large-deviations result, which will be rigorized in the second part of this subsection.

Heuristic approach. With $\bar{X}^n(t) := (\bar{A}^n)'(t)$, we focus on the likelihood that the sample mean $n^{-1}\bar{X}^n(\cdot)$ follows the function $f(\cdot)$ on the interval $[0,T]$, given the initial condition $n^{-1}\bar{X}^n(0) = x$. The function $f(\cdot)$ is evidently such that $f(0) = x$. Then we require, after discretizing time for $k = 1, \ldots, T/\triangle t$, that

$$\frac{1}{n}\bar{X}^n(k\triangle t) = f(k\triangle t)$$

for all k; the finer the grid, the better the approximation. Hence we consider for $\triangle t$ small

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{n}\bar{X}^n(t) \approx f(t), \forall t \in [0,T] \;\middle|\; \frac{1}{n}\bar{X}^n(0) = x\right)\\ &\quad\approx\quad \mathbb{P}\left(\frac{1}{n}\bar{X}^n(k\triangle t) \approx f(k\triangle t), \forall k \in \{1,...,T/\triangle t\} \;\middle|\; \frac{1}{n}\bar{X}^n(0) = x\right).\end{aligned}$$

By the Markovian property of the rate process, the previous display reads

$$\begin{aligned}&\prod_{k=1}^{T/\triangle t} \mathbb{P}\left(\frac{1}{n}\bar{X}^n(k\triangle t) \approx f(k\triangle t) \;\middle|\; \frac{1}{n}\bar{X}^n((k-1)\triangle t) \approx f((k-1)\triangle t)\right)\\ &\quad=\quad \prod_{k=1}^{T/\triangle t} \mathbb{P}\left(\frac{1}{n}\bar{X}^n(\triangle t) \approx f(k\triangle t) \;\middle|\; \frac{1}{n}\bar{X}^n(0) \approx f((k-1)\triangle t)\right),\end{aligned}$$

for paths $f(\cdot)$ with $f(0) = x$. Relying on standard large-deviations results for the Normal distribution, we thus obtain the decay rate

$$\begin{aligned}&\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}\bar{X}^n(t) \approx f(t), \forall t \in [0,T] \;\middle|\; \frac{1}{n}\bar{X}^n(0) = x\right)\\ &\quad=\quad \lim_{\triangle t \downarrow 0} \sum_{k=1}^{T/\triangle t} \frac{\left(f(k\triangle t) - \mathbb{E}\left(\bar{X}(\triangle t) \mid \bar{X}(0) = f((k-1)\triangle t)\right)\right)^2}{2\mathbb{V}\mathrm{ar}\big(\bar{X}(\triangle t) \mid \bar{X}(0) = f((k-1)\triangle t)\big)}.\end{aligned}$$

Applying the approximations of $\mathbb{E}(\bar{X}(\epsilon) \mid \bar{X}(0) = x)$ and $\mathbb{V}\mathrm{ar}\big(\bar{X}(\epsilon) \mid \bar{X}(0) = x\big)$ for small ϵ, as given above, this further reduces to

$$\begin{aligned}&\lim_{\triangle t \downarrow 0} \frac{1}{2} \sum_{k=1}^{T/\triangle t} \frac{(f(k\triangle t) - f((k-1)\triangle t) - \triangle t(\lambda - \mu f((k-1)\triangle t))^2}{2\lambda \triangle t}\\ &\quad=\quad \lim_{\triangle t \downarrow 0} \frac{1}{4\lambda} \sum_{k=1}^{T/\triangle t} \triangle t \left(\frac{(f(k\triangle t) - f((k-1)\triangle t))}{\triangle t} - (\lambda - \mu f((k-1)\triangle t))\right)^2\\ &\quad=\quad \lim_{\triangle t \downarrow 0} \frac{1}{4\lambda} \sum_{k=1}^{T/\triangle t} \triangle t \left(f'((k-1)\triangle t) - \lambda + \mu f((k-1)\triangle t)\right)^2\\ &\quad=\quad \frac{1}{4\lambda} \int_0^T \left(f'(t) - \lambda + \mu f(t)\right)^2 \mathrm{d}t.\end{aligned}$$

Hence the candidate rate function of a path f is

$$\bar{\mathbb{I}}(f) = \int_0^T \bar{I}_{f(s)}(f'(s))\mathrm{d}s, \quad \text{where} \quad \bar{I}_x(u) := \frac{(u - \lambda + \mu x)^2}{4\lambda}.$$

So far we have considered paths on $[0,T]$, that start in x at time 0. Extending the argument to paths on $(-\infty, \infty)$, the candidate for the rate function would become

$$\bar{\mathbb{I}}(f) = \int_{-\infty}^{\infty} \bar{I}_{f(s)}(f'(s))\mathrm{d}s. \tag{5.13}$$

The remainder of this subsection is devoted to a formal approach to establishing (5.13) by applying the generalized version of Schilder's theorem.

Sample-path large-deviations principle. For any Gaussian process with stationary increments $\bar{A}(\cdot)$, the generalized version of Schilder's theorem states that, under mild conditions,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}\bar{A}^n(\cdot) \in \mathscr{S}\right) = -\inf_{F\in\mathscr{S}} \mathbb{K}(F); \tag{5.14}$$

as before, we should formally distinguish between open en closed sets $\mathscr{S}$. In general, the action functional $\mathbb{K}(F)$ is only explicitly given for paths $F(\cdot)$ that are mixtures of covariance functions: if, for $\alpha_i, s_i \in \mathbb{R}$, and $\Gamma(s,t) := \mathbb{C}\mathrm{ov}\big(\bar{A}(s), \bar{A}(t)\big)$, the path $F(s)$ is of the form $\sum_{i=1}^d \alpha_i \Gamma(s, s_i)$, then

$$\mathbb{K}(F) = \sum_{i=1}^d \sum_{j=1}^d \alpha_i \alpha_j \Gamma(s_i, s_j).$$

Also for $F(\cdot)$ that are not given as a mixture of covariance functions, one can determine $\mathbb{K}(F)$ by approximating $F(\cdot)$ by a mixture of covariance functions, and by using a limiting procedure — we leave out details here.

In case $\bar{A}(\cdot)$ has a derivative, then one could also consider large deviations probabilities that relate to the *rate process* $\bar{X}^n(\cdot)$ rather than the cumulative traffic process $\bar{A}^n(\cdot)$. With $f(s) := F'(s)$,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}\bar{X}^n(\cdot) \in \mathscr{S}\right) = -\inf_{f\in\mathscr{S}} \mathbb{K}(F). \tag{5.15}$$

In order to rigorize (5.13), we show that for $F(\cdot)$ being a linear combination of covariance functions, we have that indeed

$$\mathbb{K}(F) = \frac{1}{4\lambda} \int_{-\infty}^{\infty} \left(f'(t) - \lambda + \mu f(t)\right)^2 \mathrm{d}t. \tag{5.16}$$

It is elementary to show that, using the shorthand notation $\Gamma_i(t) := \Gamma(t, s_i)$, the right-hand side of the previous display equals

$$\frac{1}{2\lambda} \sum_{i=1}^d \sum_{j=1}^d \alpha_i \alpha_j \int_{-\infty}^{\infty} (\Gamma_i''(t) + \mu\Gamma_i'(t))(\Gamma_j''(t) + \mu\Gamma_j'(t))\mathrm{d}t.$$

Then (5.16) indeed follows from the next lemma.

LEMMA 5.4.1. *For all $i, j = 1, \ldots, d$,*

$$\int_{-\infty}^{\infty} (\Gamma_i''(t) + \mu\Gamma_i'(t))(\Gamma_j''(t) + \mu\Gamma_j'(t))\mathrm{d}t = 2\lambda\Gamma(s_i, s_j).$$

Proof See appendix. □

5.4.2 Congestion period

The decay rate of the congestion period can be found in the same fashion as in Section 5.3.2, the only major difference being that now we should rely on the sample-path large-deviations result for iOU traffic.

PROPOSITION 5.4.2. *For $x \geq 0$,*

$$\bar{\delta}(x) = -x\bar{\delta}^\star; \quad \bar{\delta}^\star := \frac{(\mu c - \lambda)^2}{4\lambda}.$$

Proof We prove this result by first establishing a lower bound. Clearly, $\bar{\delta}(x) \geq \bar{\mathbb{I}}(f_c)$, recalling that $f_c(s) = c$ for all $s \in [0, x]$. Evaluating $\bar{\mathbb{I}}(f_c)$, we find the lower bound.

Now focus on the upper bound. Straightforward algebra yields that

$$\bar{\delta}(x) = -\inf_{f\in\bar{T}} \int_0^x \left(\frac{(f'(s))^2}{4\lambda} - \frac{1}{2}f'(s) + \frac{1}{4}\frac{f(s)f'(s)}{\lambda} + \frac{\lambda}{4} + \frac{\mu^2}{4\lambda}f^2(s) - \frac{1}{2}\mu f(s) \right) \mathrm{d}s.$$

Evidently, the fact that $f(0) = f(x) = c$ (for all $f \in \bar{T}$) entails that

$$\int_0^x \frac{1}{2}f'(s)\mathrm{d}s = \int_0^x \frac{1}{4}\frac{f(s)f'(s)}{\lambda}\mathrm{d}s = 0,$$

which immediately leads to

$$\bar{\delta}(x) \leq -\inf_{f\in\bar{T}} \int_0^x \left(\frac{\mu^2}{4\lambda}f^2(s) - \frac{1}{2}\mu f(s) \right) \mathrm{d}s - \frac{\lambda}{4}x.$$

The right-hand side of the previous display is smaller than $-x\bar{\tau}^\star$, as follows, after elementary algebra, from the inequality

$$\frac{\mu^2}{4\lambda}(y^2 - c^2) = \frac{\mu^2}{4\lambda}(y + c)(y - c) \geq \frac{\mu^2}{4\lambda} \cdot 2c \cdot (y - c) = \frac{\mu}{2} \cdot \frac{c}{\rho}(y - c) \geq \frac{\mu}{2}(y - c),$$

for all $y \geq c$. This completes the upper bound.

REMARK 5.4.3. *We now consider the so-called heavy-traffic regime $c = \rho + \epsilon$ for ϵ small, and we show that $\delta^\star$ and $\bar{\delta}^\star$ are very much alike. In other words: in heavy-traffic the Gaussian approximation is particularly accurate. The formal calculation is as follows. It is easily checked that*

$$\delta^\star = \mu\left(\frac{\epsilon^2}{4\rho} + \frac{\epsilon^3}{8\rho^2}\right) + O(\epsilon^4), \quad \text{and} \quad \bar{\delta}^\star = \mu\left(\frac{\epsilon^2}{4\rho}\right),$$

as $\epsilon \downarrow 0$. See for related results also [54, Section 5.2]. ♢

5.4.3 Area

We now consider the decay rate $\bar{\alpha}(x)$ of the area exceeding an amount x, which is obtained in a similar manner as was done for the M/M/∞ model in Section 5.3.3. Again exploiting the relation with the tail probabilities of an appropriately chosen queueing system, and the large-deviations results by Botvich and Duffield [21], we obtain the following proposition. As its proofs is identical to that of Proposition 5.3.5, we leave it out. Realize that

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}\bar{X}^n(0) \geq c\right) = -\frac{(c-\rho)^2}{2\rho}.$$

PROPOSITION 5.4.4. *For $x \geq 0$,*

$$\bar{\alpha}(x) = -\inf_{t\geq 0} \frac{(x+(c-\rho)t)^2}{2v(t)} + \frac{(c-\rho)^2}{2\rho}.$$

There is no explicit, closed-form available for $\alpha(x)$. It is, however, possible to explicitly characterize $\alpha(x)$ in the asymptotic regimes $x \downarrow 0$ and $x \to \infty$.

PROPOSITION 5.4.5. *The asymptotic behavior of $\alpha(x)$ is given by*

$$\bar{\alpha}(x) = -\sqrt{2\lambda/3} \cdot \left(\frac{c-\rho}{\rho}\right)^{3/2} \cdot \sqrt{x} + O(x) \quad \text{as } x \downarrow 0;$$

$$\bar{\alpha}(x) = -\frac{\mu(c-\rho)}{\rho} x - \frac{(c-\rho)^2}{2\rho} \quad \text{as } x \to \infty.$$

Proof First consider the regime $x \downarrow 0$. The infimum will be reached for $t(x)$ close to 0, in which case the variance function $v(t)$ behaves as

$$\rho t^2 - \lambda t^3/3 + O(t^4).$$

By virtue of Equation (6.6) of [85],

$$\inf_{t\geq 0} \frac{(x+(c-\rho)t)^2}{2v(t)} = \frac{(c-\rho)^2}{2\rho} + \sqrt{2\lambda/3} \cdot \left(\frac{c-\rho}{\rho}\right)^{3/2} \cdot \sqrt{x} + O(x).$$

Now focus on $x \to \infty$. We again use Theorem 3 of Botvich and Duffield [21], which implies that

$$\inf_{t\geq 0} \frac{(x+(c-\rho)t)^2}{2v(t)} = \frac{(c-\rho)^2}{\rho} + \frac{\mu(c-\rho)}{\rho} x + o(1)$$

for x large. Now an application of Lemma 5.3.4 yields the stated. □

REMARK 5.4.6. *There is not a Gaussian equivalent of an 'arrival'. We therefore do not have a Gaussian counterpart of the decay rate $\nu(x)$. See, however, the appendix of [54], where it is pointed out how an 'artificial' Gaussian counterpart can be constructed (which lacks a straightforward interpretation).*

5.5 Uniform bounds

In the previous sections we computed the decay rates of the probabilities of interest, but these do not provide us with estimates of the probabilities themselves. In particular, a statement of the type $n^{-1}\log f(n) \to -\zeta$ just says that $f(n) = g(n)\exp(-\zeta n)$ for some subexponential function $g(\cdot)$, that is $\log g(n) = o(n)$ as $n \to \infty$; the function $g(n)$ can still be of the form $\exp(n^{1-\delta})$ for a small positive constant δ. For practical purposes *conservative* (but preferably tight) estimates of the probabilities of interest are useful. In this section such approximations are derived. They indicate that the logarithmic estimates are rather precise.

5.5.1 Congestion period

We consider the original, that is, *unscaled*, model. The exponential part in the following bound should be compared with the logarithmic asymptotics found in Section 5.3.2.

PROPOSITION 5.5.1. *Uniformly in* $x \geq 0$,

$$\mathbb{P}(D_{\mathrm{C}} > x) \leq \left(\sqrt{\frac{\mathrm{C}}{\mathrm{P}}}\right)^{\mathrm{C}+1} \cdot \exp\left(-(\sqrt{\Lambda} - \sqrt{\mathrm{C}\mu})^2 x\right).$$

Proof It is clear that $\mathbb{P}(D_{\mathrm{C}} > x) \leq \mathbb{P}(A(x) > \mathrm{C}x \mid X(0) = \mathrm{C}+1)$. The Markov inequality yields that, for any $\vartheta > 0$,

$$\mathbb{P}(A(x) > \mathrm{C}x \mid X(0) = \mathrm{C}+1) \leq \mathbb{E}(e^{\vartheta A(x)} \mid X(0) = \mathrm{C}+1)e^{-\vartheta \mathrm{C}x}.$$

Applying (5.2), we obtain

$$\mathbb{P}(A(x) > \mathrm{C}x \mid X(0) = \mathrm{C}+1) \leq \left(\frac{\mu}{\mu-\vartheta}\right)^{\mathrm{C}+1} \exp\left(\vartheta\left(\frac{\Lambda}{\mu-\vartheta} - \mathrm{C}\right)x\right).$$

Now plug in $\vartheta = \vartheta^\star := \mu - \sqrt{\Lambda\mu/\mathrm{C}} > 0$. □

A slightly different bound can be found as follows. We include it here, as it gives us insight into the way importance-sampling algorithms might be devised. Suppose we wish to estimate $\mathbb{P}(D_{\mathrm{C}} > x)$ by simulation, applying importance sampling with arrival rate $\Lambda^\star := \sqrt{\Lambda\mu\mathrm{C}}$ and service rate $\mu^\star := \sqrt{\Lambda\mu/\mathrm{C}}$, irrespective of the number of flows present; call the new measure $\mathbb{Q}$. It is elementary that, in self-evident notation,

$$\mathbb{P}(D_{\mathrm{C}} > x) = \mathbb{E}_{\mathbb{Q}} LI,$$

where L is the so-called likelihood ratio, and I the indicator function of the event under consideration. In more detail, the likelihood ratio can be expressed as follows. Let τ_i denote the i-th jump of the congestion period, i.e., τ_i is 1 if the i-th jump is upward and 0 if it is downward. (With τ_0 we mean the jump to level $\mathrm{C}+1$ that starts the congestion period.) Let Z_i denote the *state* (i.e., the number of flows present) between the i-th and $(i+1)$-st jump, and S_i the *time* between these jumps. Then, with N denoting the *last* jump before time x, realizing that the $(N+1)$-st jump epoch is the first jump epoch at which we are certain that it can decided whether indeed $D_{\mathrm{C}} > x$,

$$L = \prod_{i=0}^{N} \frac{(\Lambda+\mu Z_i)\exp(-(\Lambda+\mu Z_i)S_i)}{(\Lambda^\star+\mu^\star Z_i)\exp(-(\Lambda^\star+\mu^\star Z_i)S_i)} \prod_{i=0}^{N-1} (p(Z_i))^{\tau_{i+1}}(q(Z_i))^{1-\tau_{i+1}},$$

where

$$p(k) := \left(\frac{\Lambda}{\Lambda+\mu k}\right) \Big/ \left(\frac{\Lambda^\star}{\Lambda^\star+\mu^\star k}\right), \;\; q(k) := \left(\frac{\mu x}{\Lambda+\mu k}\right) \Big/ \left(\frac{\mu^\star x}{\Lambda^\star+\mu^\star k}\right).$$

Elementary calculus yields that this likelihood equals

$$\frac{\Lambda+\mu Z_N}{\sqrt{\Lambda\mu\mathrm{C}}+\sqrt{\Lambda\mu/\mathrm{C}}Z_N} \times \exp\left(-\sum_{i=0}^{N}\left(\Lambda+\mu Z_i-\sqrt{\Lambda\mu\mathrm{C}}-\sqrt{\Lambda\mu/\mathrm{C}}Z_i\right)S_i\right)\left(\sqrt{\frac{\mathrm{P}}{\mathrm{C}}}\right)^{Z_N-(\mathrm{C}+1)}.$$

Relying on Lemma 5.A.1, it is elementary to show that, as long as $Z_i > \mathrm{C}$,

$$\Lambda+\mu Z_i-\sqrt{\Lambda\mu\mathrm{C}}-\sqrt{\Lambda\mu/\mathrm{C}}Z_i \geq \Lambda - 2\sqrt{\Lambda\mu\mathrm{C}}+\mu\mathrm{C}.$$

Now observe that, during a run in which $I=1$, $Z_i > \mathrm{C}$ for all $i \in \{0,\ldots,N\}$, and $\sum_{i=0}^{N} S_i > x$ as well as $Z_N > \mathrm{C}$. Also, due to $\mathrm{P} < \mathrm{C}$,

$$\frac{\Lambda+\mu Z_N}{\sqrt{\Lambda\mu\mathrm{C}}+\sqrt{\Lambda\mu/\mathrm{C}}Z_N} \leq \sqrt{\frac{\mathrm{C}}{\mathrm{P}}}.$$

We thus find the upper bound

$$\mathbb{P}(D_{\mathrm{C}} > x) \;\; \leq \;\; \frac{\mathrm{C}}{\mathrm{P}} \cdot \exp\left(-(\sqrt{\Lambda}-\sqrt{\mathrm{C}\mu})^2 x\right). \tag{5.17}$$

Note that for $\mathrm{C} > 2$ this bound is *sharper* than the one we presented in Proposition 5.5.1.

5.5.2 Area

A similar argument can be used to find a uniform upper bound on $\mathbb{P}(A_{\mathrm{C}} > x)$.

PROPOSITION 5.5.2. *Uniformly in* $x \geq 0$,

$$\mathbb{P}(A_{\mathrm{C}} > x) \leq \left(\frac{\mathrm{C}}{\mathrm{P}}\right)^2 \cdot \exp\left(-\left(\mu - \frac{\Lambda}{\mathrm{C}}\right)x\right).$$

Proof First observe that

$$\mathbb{P}(A_{\mathrm{C}} > x) \leq \mathbb{P}(\exists t \geq 0 : A(t) > x + \mathrm{C}t).$$

Let us find an upper bound for the right-hand side of the previous display. Suppose we perform importance sampling under a measure $\mathbb{Q}$ that is such that the arrival rate is $\Lambda^\star := \mu\mathrm{C}$ and service rate $\mu^\star := \Lambda/\mathrm{C}$. It is clear that the resulting system is such that, under the new measure, $A(t)$ indeed crosses level $x + \mathrm{C}t$ with probability 1, as the mean rate under $\mathbb{Q}$ is $\Lambda^\star/\mu^\star = \mathrm{C}^2\mu/\Lambda = \mathrm{C}^2/\mathrm{P} > \mathrm{C}$.

A fundamental equality is, with $\mathbb{E}_{\mathbb{Q}}$ denoting expectation under $\mathbb{Q}$,

$$\mathbb{P}(\exists t \geq 0 : A(t) > x + \mathrm{C}t) = \mathbb{E}_{\mathbb{Q}}L,$$

where L is the so-called likelihood ratio. In more detail, the likelihood ratio can be expressed as follows. Using the same definitions as in the previous section,

$$L = \prod_{i=0}^{N} \frac{(\Lambda + \mu Z_i)\exp(-(\Lambda + \mu Z_i)S_i)}{(\Lambda^\star + \mu^\star Z_i)\exp(-(\Lambda^\star + \mu^\star Z_i)S_i)} \prod_{i=0}^{N-1} (p(Z_i))^{I_{i+1}} (q(Z_i))^{1-I_{i+1}},$$

Elementary calculus yields that

$$L = \frac{\Lambda + \mu Z_N}{\mu\mathrm{C} + \Lambda Z_N/\mathrm{C}} \times \exp\left(-\left(\mu - \frac{\Lambda}{\mathrm{C}}\right)\left(A\left(\sum_{i=0}^{N} S_i\right) - \mathrm{C}\cdot\sum_{i=0}^{N} S_i\right)\right)\left(\frac{\mathrm{P}}{\mathrm{C}}\right)^{Z_N - (\mathrm{C}+1)}.$$

Due to $\mathrm{P} < \mathrm{C}$, it holds that

$$\frac{\Lambda + \mu Z_N}{\mu\mathrm{C} + \Lambda Z_N/\mathrm{C}} \leq \frac{\mathrm{C}}{\mathrm{P}}.$$

As we know that $Z_N \geq \mathrm{C}$, and, by definition of N,

$$A\left(\sum_{i=0}^{N} S_i\right) - \mathrm{C}\cdot\sum_{i=0}^{N} S_i \geq x,$$

the upper bound follows. □

5.5.3 Number of flows

Finally, we use the change-of-measure technique to find a uniform upper bound on $\mathbb{P}(N_{\mathrm{C}} > m)$. We start, however, by a result that can be proven in a more elementary way.

PROPOSITION 5.5.3. *Uniformly in* $x \in \mathbb{N}$,

$$\mathbb{P}(N_{\mathrm{C}} > x) \leq \left(\frac{4\Lambda\mu\mathrm{C}}{(\Lambda + \mu\mathrm{C})^2} \right)^x.$$

Proof First observe that, stochastically, $Y_{2x} \leq Y'_{2x} := \mathrm{C} + \sum_{i=1}^{2x} Z_i$, where the Z_i are i.i.d., and $Z_i = 1$ with probability $\Lambda/(\Lambda + \mu\mathrm{C})$ and -1 otherwise. By the Markov inequality, it follows that, for any $\vartheta \geq 0$,

$$\mathbb{P}(N_{\mathrm{C}} > m) \leq \mathbb{P}\left(\sum_{i=1}^{2x} Z_i \geq 0 \right) \leq \left(\mathbb{E}e^{\vartheta Z} \right)^{2x},$$

with Z distributed as the Z_i. Now minimize the last expression over all $\vartheta \geq 0$, and the desired follows. □

As mentioned above, a similar bound can be found by an importance-sampling argumentation. Simulate the discrete-time process (i.e., the jump process) that results after changing Λ into $\Lambda^\star := \sqrt{\Lambda\mu\mathrm{C}}$ and $\mu^\star := \sqrt{\Lambda\mu/\mathrm{C}}$ until either the process drops below the value C, or $2x$ transitions have been performed. It is readily checked that the likelihood at this stopping epoch equals

$$L = \left(\sqrt{\frac{\mathrm{P}}{\mathrm{C}}} \right)^{Z_N - (\mathrm{C}+1)} \prod_{i=0}^{N} \frac{\sqrt{\Lambda\mu\mathrm{C}} + \sqrt{\Lambda\mu/\mathrm{C}} Z_i}{\Lambda + \mu Z_i}.$$

Applying Lemma 5.A.2, it is elementary to show that

$$\frac{\sqrt{\Lambda\mu\mathrm{C}} + \sqrt{\Lambda\mu/\mathrm{C}} Z_i}{\Lambda + \mu Z_i} \leq 2 \frac{\sqrt{\Lambda\mu\mathrm{C}}}{\Lambda + \mu\mathrm{C}}.$$

Using that, if $I = 1$, then $N \geq 2x$ and $Z_N > \mathrm{C}$, we find the same upper bound as above, but now multiplied with $\sqrt{\mathrm{C}/\mathrm{P}}$, i.e., slightly weaker.

5.6 Numerical results

In this section we demonstrate our asymptotics and bounds through a number of numerical experiments. In these experiments we choose $\mu = 1$, and $c = 1$, and we

compare the situation $\lambda = 0.5$ with $\lambda = 0.9$. The primary goal of this section is to present a comparison between the rough asymptotics of Sections 5.3-5.4, the bounds of Section 5.5, and the 'real' values.

A number of remarks need to be made here.

- The results in Sections 5.3-5.4 are in terms of decay rates, and in order to compare them we do as if the decay is 'purely exponential'. For instance for the congestion duration, the resulting approximation, based on Proposition 5.3.1, is, with as before $\Lambda = n\lambda$ and $\mathrm{C} = nc$,

$$\mathbb{P}(D_{\mathrm{C}} > x) \approx \exp\left(-(\sqrt{\Lambda} - \sqrt{\mu \mathrm{C}})^2 x\right), \tag{5.18}$$

  cf. Proposition 5.5.1. In case of the area, this approximation is somewhat trickier to derive; we now sketch how the approximation for $\mathbb{P}(A_{\mathrm{C}} > x)$ can be found. Focusing for the moment on the regime $x \to \infty$, Proposition 5.3.6 entails that

$$\mathbb{P}(A_{nc} > nx) \approx \exp\left(-n\beta^\star_\infty - n\alpha^\star_\infty x\right).$$

  Noticing that

$$\beta^\star_\infty = \frac{1}{n}\left(\frac{(\mathrm{C}-\mathrm{P})^2}{\mathrm{P}} + \mathrm{C} - \mathrm{P} - \mathrm{C}\log\frac{\mathrm{C}}{\mathrm{P}}\right);\ \ \alpha^\star_\infty = \mu - \frac{\Lambda}{\mathrm{C}},$$

  we obtain the approximation

$$\mathbb{P}(A_{\mathrm{C}} > x) \approx \exp\left(-\frac{(\mathrm{C}-\mathrm{P})^2}{\mathrm{P}} - \mathrm{C} + \mathrm{P} + \mathrm{C}\log\frac{\mathrm{C}}{\mathrm{P}} - \left(\mu - \frac{\Lambda}{\mathrm{C}}\right)x\right);$$

  cf. Proposition 5.5.2. In the regime $x \downarrow 0$ an analogous argumentation yields

$$\mathbb{P}(A_{\mathrm{C}} > x) \approx \exp\left(-2\sqrt{2}\cdot\sqrt{x\Lambda\left(1 - \frac{\mathrm{P}}{\mathrm{C}} + \frac{\mathrm{P}}{\mathrm{C}}\log\frac{\mathrm{P}}{\mathrm{C}}\right)}\right).$$

  The Gaussian counterparts can be dealt with similarly.

- To obtain 'real' values of the probabilities of interest, we used importance-sampling-based simulations, with the change-of-measures suggested in Section 5.5. We also performed direct simulations (that is, simulations under the original measure), where we empirically observed that under importance sampling substantially less simulation effort is needed to obtain an estimate of given precision; for higher values of x direct simulation becomes prohibitively time-consuming. The outcomes of these direct simulations (in the graphs corresponding to the label '$\mathrm{M}/\mathrm{M}/\infty$') coincide with the importance-sampling-based estimates, as should be the case.

For the congestion duration D_{C} we consider the tail probabilities for $n = 20, 50, 100$ in Figure 5.1. In the numerical results we compare Proposition 5.3.1 for the $\mathrm{M}/\mathrm{M}/\infty$ process, Proposition 5.4.2 for the Gaussian counterpart, the uniform upper bound given by Expression (5.17), and results from importance-sampling simulations as well as direct simulations of the $\mathrm{M}/\mathrm{M}/\infty$ process.

In each of the graphs we see that the uniform upper bound of Expression (5.17) is slightly larger than Proposition 5.3.1, namely a factor C/P. The result from the Gaussian counterpart (Proposition 5.4.2) is in between the results for the $\mathrm{M}/\mathrm{M}/\infty$ and the simulation results for low load, but for high load it is close to Proposition 5.3.1 and Expression (5.17). Observe that for higher loads (right graphs) the probabilities of a long congestion duration are higher, which is evident as these occurrences become less rare. By comparing the graphs of Figure 5.1 it is seen that increasing the scaling parameter n indeed leads to smaller probabilities.

Given our analytical results, the curve with simulated probabilities should eventually be (that is, for n large) parallel to the curves obtained from Proposition 5.3.1 and Expression (5.17), which is evidently not yet the case for $n = 100$. A similar slow convergence has been observed for the tail asymptotics of the sojourn time in Processor-Sharing (PS) queues in e.g. [95]. It may also play a role that, just as is the case for the sojourn-time distribution in the $\mathrm{M}/\mathrm{M}/1$ PS queue, the asymptotics are likely to be *not* of a 'purely exponential' form (as suggested by (5.18)); instead there may be in addition a polynomial factor $\delta x^{-\gamma}$ (for some $\gamma, \delta > 0$), and potentially also a Weibullian factor $\exp(-\alpha x^{\beta})$ (for some $\alpha > 0$ and $\beta \in (0, 1)$), cf. [20, 44].

The figures show that the results obtained from Proposition 5.3.1 and Expression (5.17) can, in practical situations, only be used as (very rough) indications of the probability of interest. In case quick, reliable estimates are required (for instance for dimensioning purposes), we advise to rely on the described (efficient) importance sampling scheme.

The results for the area A_{C} are displayed in Figure 5.2; we only present the result for $n = 20$, as the effect of increasing n is similar as for the duration. The graphs compare the results of Proposition 5.3.6 for the $\mathrm{M}/\mathrm{M}/\infty$ process, Proposition 5.4.5 for the Gaussian counterpart, the uniform upper bound of Proposition 5.5.2 and the simulation results, both from direct simulations and importance sampling. Recall that Propositions 5.3.6 and 5.4.5 include both the behavior of x close to 0, and x large, respectively; therefore in the graphs there are two curves for each proposition, and it is emphasized that these curves are not valid for the entire range of x. The uniform upper bound corresponds to the 'highest' curve, as expected. For low loads all curves are relatively close, and the simulation results are in between the other mentioned results; the latter property is in contrast with the results for the duration and the number of arrivals, for which the probabilities from the simulation are always the smallest. It is seen that in the low-load case the part of Proposition 5.3.6

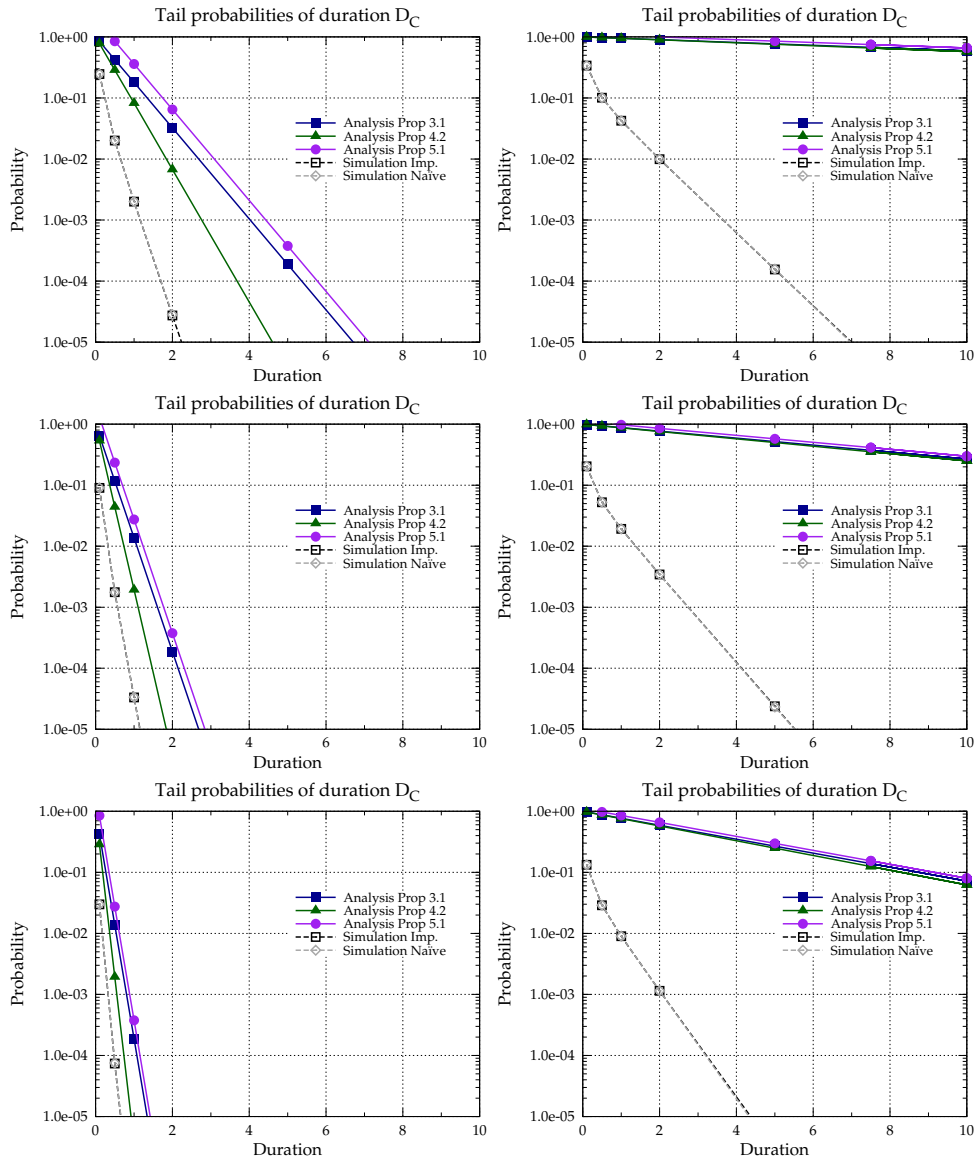

Figure 5.1: *Congestion duration for* $\mu = 1$*, and* $c = 1$*. Left:* $\lambda = 0.5$*. Right:* $\lambda = 0.9$*. Top:* $n = 20$*. Middle:* $n = 50$*. Bottom:* $n = 100$*.*

corresponding to $x \to \infty$ is already highly accurate for moderate x.

In Figure 5.3 the tail probabilities of the number of arrivals N_C are considered, again for $n = 20$. We compare the results from Proposition 5.3.7, the uniform upper

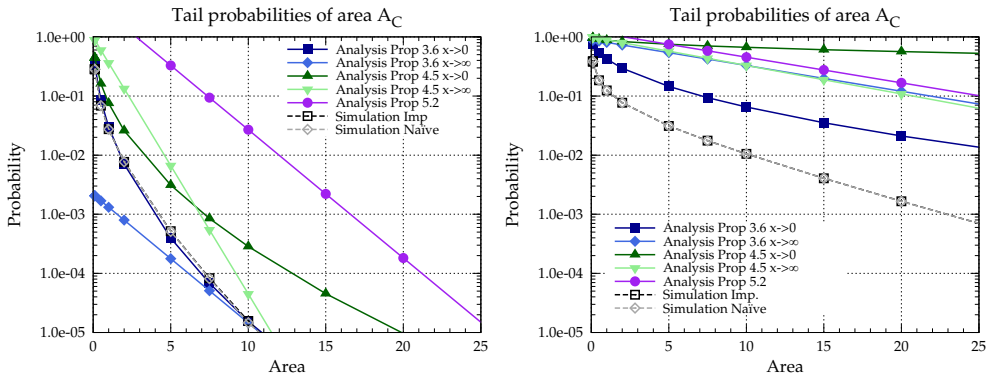

Figure 5.2: *Area for* $n = 20$, $\mu = 1$, *and* $c = 1$. *Left:* $\lambda = 0.5$. *Right:* $\lambda = 0.9$.

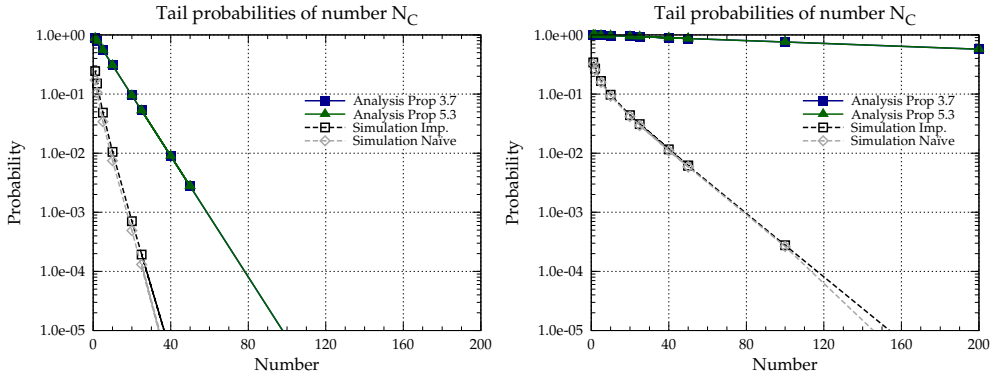

Figure 5.3: *Number of arrivals for* $n = 20$, $\mu = 1$, *and* $c = 1$. *Left:* $\lambda = 0.5$. *Right:* $\lambda = 0.9$.

bound from Proposition 5.5.3, and simulation results; recall that in this case there is no meaningful Gaussian counterpart. Propositions 5.3.7 and 5.5.3 lead to the same expression, viz.,

$$\mathbb{P}(N_{\mathrm{C}} > x) \approx \left(\frac{4\Lambda\mu\mathrm{C}}{(\Lambda + \mu\mathrm{C})^2} \right)^x,$$

as is easily checked. Furthermore, observe that the propositions grossly overestimate the real (that is, simulated) probabilities, as is immediately clear after inspection of the simulation results (just as was the case for the congestion duration). We again advise to use the proposed importance sampling scheme to quickly generate reliable estimates of the probability of interest.

5.7 Concluding remarks

This chapter considered tail asymptotics of congestion-period-related quantities. Large-deviations theory is applied to explicitly calculate the exponential decay rates under the many-flows scaling, both for the actual M/M/∞ model and its Gaussian counterpart. Then uniform upper bounds on the tail probabilities are derived, which also reveal an efficient change-of-measure to be used in importance-sampling simulations. There are several directions for future research, of which we now mention a few.

We derived tail asymptotics under the many-flows scaling (x fixed, n large). These are presumably tightly related to the asymptotics of $\mathbb{P}(D_{\mathrm{C}} > x)$, $\mathbb{P}(A_{\mathrm{C}} > x)$, and $\mathbb{P}(N_{\mathrm{C}} > x)$ for x large. As the LTs of these three random variables are known [58], one may attempt to obtain from these the corresponding tail asymptotics, cf. also [54, 55]. In addition, we saw that our asymptotic results and bounds not necessarily lead, for given n, x, to accurate approximations of the tail probabilities of interest, and therefore one may investigate techniques to improve on this (for values of n, x of practical interest), cf. the results in [53, Section 5].

We saw that for iOU Gaussian processes the generalized-Schilder-based large deviations rate function of a path f could be computed explicitly. One may wonder for which class within the family of Gaussian processes similar explicit expressions can be derived; one may expect that these should be such that, like is the case for iOU, the corresponding rate process is well-defined, but it is not *a priori* clear what additional conditions should be imposed.

Experientially we observed that the proposed importance-sampling algorithms led to a substantial speed up: in order to obtain estimates with a predefined level of precision, the simulation time needed was reduced significantly. We expect that the proposed change-of-measures are actually asymptotically efficient. A proof of this property is beyond the scope of the present study.

5.A Useful relations

5.A.1 A few elementary inequalities

The following, useful, lemmas are straightforward to prove.

LEMMA 5.A.1. *For all $\alpha, \beta > 0$ with $\alpha < \beta$, and $y > c$,*

$$-\sqrt{\alpha\beta} + \alpha - \frac{y}{c}\sqrt{\alpha\beta} + \frac{y}{c}\beta \geq (\sqrt{\alpha} - \sqrt{\beta})^2.$$

LEMMA 5.A.2. *For all $\alpha, \beta > 0$ with $\alpha < \beta$, and $y > c$,*

$$\frac{1 + y/c}{\alpha + \beta y/c} \leq \frac{2}{\alpha + \beta}.$$

5.A.2 Rate function

We here present the proof of Lemma 5.4.1.

Proof of Lemma 5.4.1

$$\Gamma_i(t) = \frac{\lambda}{\mu^3} \times \begin{cases} 1 - e^{-|t|\mu} + e^{-(|t|+s_i)\mu} - e^{-s_i\mu} & \text{for } t \leq 0, \\ e^{-t\mu} + e^{-s_i\mu} - e^{-(s_i-t)\mu} - 1 + 2t\mu & \text{for } t \in (0, s_i), \\ e^{-t\mu} + e^{-s_i\mu} - e^{-(t-s_i)\mu} - 1 + 2s_i\mu & \text{for } t \geq s_i. \end{cases}$$

$$\Gamma_i'(t) = \frac{\mathrm{d}}{\mathrm{d}t}\Gamma_i(t) = \frac{\lambda}{\mu^3} \times \begin{cases} -\mu e^{-|t|\mu} + \mu e^{-(|t|+s_i)\mu} & \text{for } t \leq 0, \\ -\mu e^{-t\mu} - \mu e^{-(t_i-s)\mu} + 2\mu & \text{for } t \in (0, s_i), \\ \mu e^{-t\mu} - \mu e^{-(t-s_i)\mu} & \text{for } t \geq s_i. \end{cases}$$

$$\Gamma_i''(t) = \frac{\mathrm{d}^2}{\mathrm{d}t^2}\Gamma_i(t) = \frac{\lambda^2}{\mu^3} \times \begin{cases} -\mu^2 e^{-|t|\mu} + \mu^2 e^{-(|t|+s_i)\mu} & \text{for } t \leq 0, \\ \mu^2 e^{-t\mu} - \mu^2 e^{-(s_i-t)\mu} & \text{for } t \in (0, s_i), \\ -\mu^2 e^{-t\mu} + \mu^2 e^{-(t-s_i)\mu} & \text{for } t \geq s_i. \end{cases}$$

Integrating by parts assuming $0 < s_i < s_j$ yields

$$\begin{aligned} &\int_{-\infty}^{0} (\Gamma_i''(t) + \mu\Gamma_i'(t))(\Gamma_j''(t) + \mu\Gamma_j'(t))\mathrm{d}t \\ &\quad = \frac{\lambda^2}{\mu^2} \int_{-\infty}^{0} \left(-2e^{-|t|\mu} + 2e^{-(|t|+s_i)\mu}\right)\left(-2e^{-|t|\mu} + 2e^{-(|t|+s_j)\mu}\right)\mathrm{d}t = \frac{2\lambda^2}{\mu^3}; \end{aligned}$$

$$\begin{aligned} &\int_{0}^{s_i} (\Gamma_i''(t) + \mu\Gamma_i'(t))(\Gamma_j''(t) + \mu\Gamma_j'(t))\mathrm{d}t \\ &\quad = \frac{\lambda^2}{\mu^2} \int_{0}^{s_i} \left(-2e^{-(s_i-t)\mu} + 2\right)\left(-2e^{-(s_j-t)\mu} + 2\right)\mathrm{d}t \\ &\quad = \frac{2\lambda^2}{\mu^3}\left(2s_i\mu - 2 - e^{(s_i-s_j)\mu} + 2e^{-s_i\mu} + 2e^{-s_j\mu} - e^{-(s_i+s_j)\mu}\right) \end{aligned}$$

Observe that $\Gamma_i''(t) + \mu\Gamma_i'(t) = 0$ for $t > s_i$, hence,

$$\begin{aligned} &\int_{s_i}^{s_j} (\Gamma_i''(t) + \mu\Gamma_i'(t))(\Gamma_j''(t) + \mu\Gamma_j'(t))\mathrm{d}t \\ &\quad = \int_{s_j}^{\infty} (\Gamma_i''(t) + \mu\Gamma_i'(t))(\Gamma_j''(t) + \mu\Gamma_j'(t))\mathrm{d}t = 0. \end{aligned}$$

Finally, upon combining the above, it is straightforward that

$$\int_{-\infty}^{\infty} (\Gamma_i''(t) + \mu\Gamma_i'(t))(\Gamma_j''(t) + \mu\Gamma_j'(t))\mathrm{d}t = 2\lambda\Gamma(s_i, s_j).$$

□

Part II

Resource sharing in wireless ad-hoc networks

Chapter 6

Resource sharing in wireless ad-hoc networks

6.1 Introduction

In this part of the thesis we focus on wireless ad-hoc networks (cf. Section 1.2.2). We investigate the impact of the resource-sharing policy on the performance of an ad-hoc network where multiple 'source nodes' transmit their flows via a common relay node. The principal goal of this chapter is to introduce the fluid-modeling approach of wireless ad-hoc networks that will be studied extensively in the next chapters. The present chapter serves as an introduction to this part.

Outline of this chapter. In order to introduce the fluid model and the performance metrics, we first explain wireless ad-hoc networks in more detail, and, in particular, the resource sharing among nodes (Section 6.2). Next, we provide an overview of the literature on the performance modeling of wireless ad-hoc networks (Section 6.3). Subsequently, we describe our fluid model in detail, together with its performance metrics (Section 6.4), and we present some preliminary analysis (Section 6.5). Finally, we introduce validation scenarios that are used in the numerical evaluations in Chapters 7 and 8 (Section 6.6).

6.2 Wireless ad-hoc networks

Developments in wireless communication technology open up the possibility of operating wireless ad-hoc networks. These networks can be deployed without a fixed infrastructure or predetermined configuration, and one of the key-features is multi-hop connectivity. For this reason ad-hoc networks are particularly suitable in situations where a fixed communication infrastructure, wireline or wireless, does not exist or malfunctions, e.g., due to a disaster, for instance see [12, 49, 50, 124]. The communication technology is usually based on shared medium access (for example, IEEE 802.11 Wireless LAN, see [63]), i.e., neighboring nodes share a common underlying radio capacity.

As mentioned above, wireless ad-hoc networks have two important characteristics: i) multi-hop connectivity, i.e., nodes that cannot directly communicate with

their destinations use other nodes as relay nodes; ii) nodes contend for access to a shared wireless medium in a distributed fashion. A consequence of the first characteristic is that certain nodes, in particular nodes that have central locations, are likely to become *relay nodes* having considerably higher traffic loads than other nodes. The second characteristic entails that there is a lack of coordination between the nodes which may result in non-optimal sharing of the medium capacity. Therefore, a relay node can easily become a *performance bottleneck*. For example, when a relay node obtains the same share of the medium capacity as each of its neighboring nodes, the input rate of traffic into the relay node exceeds the output rate when more than one neighboring node sends traffic via the relay node. This results in the accumulation of traffic at the relay node and consequently in increasing delays.

Currently, IEEE 802.11 Wireless LAN [62, 63, 64, 65] is the most popular wireless ad-hoc networking technology; our fluid model (cf. Section 6.4) is inspired by this technology, but the model is also suitable for other technologies with distributed resource-sharing. The performance of WLAN is largely determined by the maximum data rate at the physical (PHY) layer and the Medium Access Control (MAC) layer protocols defined by the IEEE 802.11 standards. WLAN nodes have to contend for access to the wireless medium according to the Distributed Coordination Function (DCF). The DCF is a random access scheme based on Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA), which uses random backoffs in order to manage packet retransmissions in case of a destructive collision. In literature it is shown that the DCF tends to share the wireless medium capacity *equally* among the contending nodes, cf. [16, 82]. Clearly, DCF is particularly appropriate in the context of ad-hoc networks as it operates in a fully distributed fashion. The DCF has, however, also significant drawbacks. Most notably, it only facilitates equally sharing of the capacity among active nodes. It cannot grant different shares of the available capacity to different nodes, hence, it is not capable to relieve the burden on relay nodes.

In 2005 a QoS-enabled version of the DCF was standardized, viz. the Enhanced Distributed Channel Access (EDCA), which is part of amendment IEEE 802.11e, see e.g. [63, 65]. The EDCA provides several parameters enabling QoS differentiation among the traffic originating from services with different QoS requirements. In principle, it can be used to enforce unequal sharing of the capacity among nodes. In particular, in our study we use this technology to grant a larger share of the capacity to the relay node. We refer to Section 8.2 for more elaborate descriptions of the DCF and EDCA. For more details on QoS aspects in ad-hoc networks, we refer to the survey [15].

Recall that our goal is to investigate the impact of the resource-sharing policy on the performance of wireless nodes in an ad-hoc network. We do so by considering a *fluid model* of a single relay node that is fed by multiple source nodes. Source nodes

become active at random time instances and start transmitting flows to a destination via the relay node; after a source node has transmitted its flow to the relay node it becomes inactive.

We consider the general situation in which the relay node may obtain a larger share of the capacity than each of the source nodes. The 'resource-sharing ratio' m indicates the share of the overall capacity C that the relay node obtains relative to the share that is allocated to each of the source nodes. More precisely, if n source nodes are active, then the relay node obtains service rate $m\text{C}/(m+n)$ while each source node receives service rate $\text{C}/(m+n)$. If the aggregate rate of traffic flowing into the relay node exceeds the service rate of the relay node, work is backlogged at the relay node in a buffer (of infinite size). It is stressed that the relay node can only claim its full share $m\text{C}/(m+n)$ if either the number of active source nodes exceeds m (that is, $n \geq m$), or if the relay node is backlogged. Otherwise the relay node is allocated a share $\text{C}/2$, while each source node obtains $\text{C}/(2n)$ (and hence the buffer remains empty). An important consequence of these allocation rules is that the system is *work-conserving*. We will explain the model in more detail in Section 6.4.

6.3 Literature

This overview of the literature mainly focuses on analytical studies on the performance of wireless ad-hoc networks. It is divided into three classes: packet-level behavior, fluid modeling of flows in single-hop WLAN, and fluid modeling of multi-hop flows in wireless ad-hoc network. Our interest is on the modeling of multi-hop flows in wireless ad-hoc networks; as will become clear, this topic was hardly covered in literature. The packet-level studies are mentioned as their outcomes are used in our fluid models.

Packet-level behavior of WLAN *and wireless ad-hoc networks.* The packet-level behavior of IEEE 802.11 WLAN in a *single-hop network*, i.e., the contention for a transmission opportunity among active nodes, has been investigated extensively. A detailed mathematical performance model of the DCF has been developed and analyzed by Bianchi [16], and slightly improved by Wu *et al.* [134]. In these papers the authors assume a constant number of persistently contending stations and rely on a relatively simple Markov chain analysis, neglecting only minor dependencies among the behavior of different nodes, and is used to obtain the saturated throughput of the wireless medium. Comparison with simulation shows that the analytical results are in general remarkably accurate.

The modeling approach introduced by Bianchi [16] is extended in many directions, e.g. in [82] various modeling enhancements on the PHY and MAC layer are included in the DCF performance model. This modeling approach is also used to ob-

tain packet contention delays for saturated sources [25], non-saturated sources [41], and heterogenous non-saturated sources [84]. In [41] the average contention delay is obtained by exploiting the state probabilities of Bianchi's Markov chain.

The Markov chain analysis, presented in [16], is extended to include the QoS-differentiation capabilities of the EDCA in e.g. [27, 136]. In particular, these models include the impact of variation of the AIFS parameter (one of the four EDCA parameters) on the saturation throughputs. In [113] the IEEE 802.11e QoS-differentiation parameters (EDCA parameters) CW_{min}, CW_{max}, AIFS, and the $TXOP_{limit}$ are systematically evaluated by means of simulations.

Analytical models for packet-level performance in *multi-hop ad-hoc networks* are presented in, e.g. [26, 60]. These papers consider situations in which the nodes experience different channel-conditions as the number of neighbors and their distances to these neighbors vary per node. The performance measures of interest are the overall aggregate throughput and the throughput per node including the impact of hidden nodes; the authors do not consider the performance of multi-hop flows.

Fluid modeling of flows in single-hop WLAN. Flow-level behavior in a single-hop network is studied in [45, 82, 133]. These papers consider situations where the number of active nodes varies dynamically in time according to the initiation and completion of file transfers at random time instants. They propose and analyze simplified analytical models yielding approximations for the expected flow (file) transfer time. In particular, in [82] the analysis is based on the modeling assumption that, from the flow-level point of view, WLAN can be regarded as a Processor Sharing (PS) type of queueing system where flows are modeled as if they *continuously* send traffic instead of sending individual *packets*. A closed-form expression for the mean flow-transfer time is obtained by considering the system as a Processor Sharing queue with state-dependent service rates. In [27] the model is extended to include EDCA's service differentiation; the authors use a queueing system with Discriminatory Processor Sharing (DPS) service discipline to model flow level behavior; results are validated by simulations.

The analyses in [45, 82, 133] ignore the effects of higher-layer protocols, in particular TCP, on the traffic behavior. In the papers [98, 112, 118] the transfer times of TCP flows over WLAN are investigated using an analytical packet/flow-level approach analogously to the one in [82]. They first determine the aggregate system throughput for a fixed number of persistent TCP flows, which is obtained using essentially an analysis similar to the one of [16]. The resulting throughputs, which are obtained for each number of persistent flows, are used as the service capacities in a Processor Sharing queue with state-dependent service rates modeling the situation with a time-varying number of non-persistent TCP flows; the main result is an expression for the mean TCP flow transfer time. In addition, [112] also analyzes the second moment of the transfer time of a TCP flow.

Fluid modeling of multi-hop flows in wireless ad-hoc networks. As mentioned at the beginning of this section, to the best of our knowledge there were no flow-level models for multi-hop flows in ad-hoc networks that yield analytical results, before we presented our fluid model (cf. Section 6.4, which appeared as [14]). We first list our own contributions on the fluid modeling of multi-hop flows, and next we discuss related literature on this topic.

In [14] we introduced the 'standard' fluid model that plays a central role in this part of the thesis, which corresponds to the special case of resource-sharing ratio $m = 1$ (relating to the IEEE 802.11 DCF), cf. Sections 6.1 and 6.4. For this fluid model (described in more detail in Section 6.4.3) analytical expressions are presented for a number of performance metrics (see also Section 8.3.2), in particular, we analyze the time required to entirely transmit a flow from a source node to its destination.

In [90], Chapter 9 of this thesis, we derive the Laplace transforms and we characterize the tail probabilities of the performance metrics of interest, still for the case the resource-sharing ratio $m = 1$. We focus on the case of exponentially distributed flow-sizes.

The fluid model of [14, 90] is extended to a general resource-sharing policy $m \in [0, \infty)$ in [115], Chapter 7 of this thesis. This extension entails that a larger share of the medium capacity can be granted to the relay node than to each of the neighboring nodes, in order to improve the overall flow transfer time. We present analytical expressions for the performance metrics of this fluid model. It is stressed that this general case is significantly more difficult than the 'standard' fluid model with $m = 1$ as in [14, 90] due to the fact that the resource sharing between the source nodes and the relay node is influenced by the workload at the relay node, and it is not solely determined by the number of active source nodes as is the case for $m = 1$.

The focus of [114], Chapter 8 of this thesis, is on the validation of the fluid model. By system simulations incorporating all details of the IEEE 802.11b and IEEE 802.11e Wireless LAN technology it was demonstrated that the fluid model accurately describes the resource sharing among the source nodes and their common relay node.

Our fluid-modeling approach has been adopted in several other studies. In [9] the 'standard' fluid model, i.e., with equal sharing of the capacity $m = 1$ as was introduced in [14], is considered for the case of regularly-varying flow sizes (that is, *heavy-tailed* flows). In particular, the tail asymptotics of the overall flow transfer time are derived by sample-path arguments; it is proven that the tail behaves roughly as the residual flow size. In [107] a versatile infinite-state Markov reward model is proposed to investigate the impact of various resource-sharing strategies for exponentially distributed flow-sizes, and in [108] the authors specialize their framework towards the IEEE 802.11e model; in both papers the authors numerically compute their performance metrics, such as the distribution of the number of active source

nodes and the workload at the relay node; they do not study flow-level performance metrics, e.g., flow-transfer times.

The model of [115] with general resource-sharing ratio $m \in [0, \infty)$ is also the subject of study in [93]. The authors derive the transforms for the performance metrics (analogously to [90]). In [28] an ad-hoc network with one- and two-hop flows sharing the same radio capacity is considered; in cohort to our model, the flows are relayed via different network nodes. The authors show that this situation can be modeled and analyzed by a Discriminatory Processor Sharing (DPS) model providing (for exponentially distributed flow-sizes) closed-form expressions for the mean transfer time of one- and two-hop flows.

Finally, we stress that the vast majority of ad-hoc network performance studies available in the literature are based on simulation, see e.g. [47, 59]. These papers usually capture great detail of the ad-hoc network protocols, but have the intrinsic drawback that they do not provide any deeper understanding of the impact of the parameters on the realized performance. Moreover, simulation runtime may become prohibitively large, hampering, e.g., sensitivity analysis or parameter optimization. Analytical performance models usually capture less detail in order to retain tractability, but do provide insight into the behavior of the system in a more explicit fashion.

6.4 Fluid model

This section presents the fluid model and the performance metrics that are analyzed in this part of the thesis. In Section 6.4.1 we describe the considered ad-hoc network scenario, and Section 6.4.2 describes the resource sharing among the network nodes. Finally, the performance metrics are defined in Section 6.4.3.

6.4.1 Ad-hoc network scenario

We consider a network with a large number of source nodes which may become active and start transmitting flows of data (files) to destinations via a common relay node, see Figure 6.1. Flow transfers are initiated according to a Poisson process with rate λ ('flow arrival rate'). Flow sizes (in terms of fluid or bits) are generally distributed with distribution F with mean f, second moment f_2 (assumed to be finite), and Coefficient of Variation (CoV) C_F, i.e., $\mathrm{C}_F^2 := \mathrm{Var}(F)/(\mathbb{E}F)^2 = f_2/f^2 - 1$. The number of active source nodes at time t is denoted by N_t.

6.4.2 Resource sharing among network nodes

The total transmission capacity of the system is denoted by C and it is shared among the active source nodes and the relay node. If the aggregate rate of traffic flowing

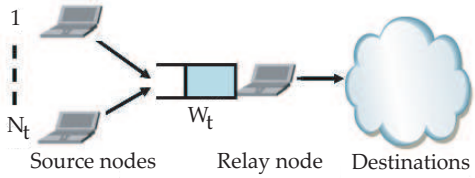

Figure 6.1: *Network model.*

from the source nodes to the relay node exceeds the rate out of the relay node, traffic is stored in a buffer of infinite size which is served by the relay node in a FCFS-fashion. By W_t we denote the workload at the relay node at time t.

The resource-sharing ratio between the relay node and the source nodes is denoted by m, i.e., the relay node obtains capacity $m\mathrm{C}/(N_t + m)$ if $N_t \geq m$ sources are present or if the buffer of the relay node is backlogged, i.e., $W_t > 0$. Otherwise, the relay node and the set of active source nodes each obtain half the capacity ($\mathrm{C}/2$), i.e., the aggregate input rate at the relay node is equal to the output rate. The source nodes always equally share the capacity not used by the relay node. Observe that the entire capacity C is always used if there is work in the system, so that the system is work-conserving. The resource sharing between the source nodes and the relay node is summarized in Table 6.1. The column 'drift' indicates the sign of the *net* input rate into the buffer at the relay node, i.e., it indicates whether the buffer content increases $(+)$, decreases $(-)$ or remains constant (0). Notice from Table 6.1 that the resource sharing at epoch t depends on both N_t and W_t. Figure 6.2 presents a sample-path example of the resource sharing in case $m = 2$, including the performance metrics of Section 6.4.3.

Table 6.1: *Resource sharing between source nodes and relay node.*

Number of	$W_t = 0$			$W_t > 0$		
active sources	source	relay	drift	source	relay	drift
$N_t < m$	$\mathrm{C}/2N_t$	$\mathrm{C}/2$	0	$\mathrm{C}/(m+N_t)$	$m\mathrm{C}/(m+N_t)$	–
$N_t = m$	$\mathrm{C}/2N_t$	$\mathrm{C}/2$	0	$\mathrm{C}/(m+N_t)$	$m\mathrm{C}/(m+N_t)$	0
$N_t > m$	NA	NA	NA	$\mathrm{C}/(m+N_t)$	$m\mathrm{C}/(m+N_t)$	+

Observe that in our model a flow may be present (and receive service) at both the source node and the relay node: at flow initiation the source node immediately starts transmitting fluid to the relay node and parts of the flow may be present at both source and relay node. At some point in time the source node transmits the last (infinitesimally small) 'particle' of the flow to the relay node, and then the source node becomes inactive; this epoch is referred to as the 'arrival of the last particle at the relay node'-epoch, or alternatively as the 'source-departure'-epoch. In case

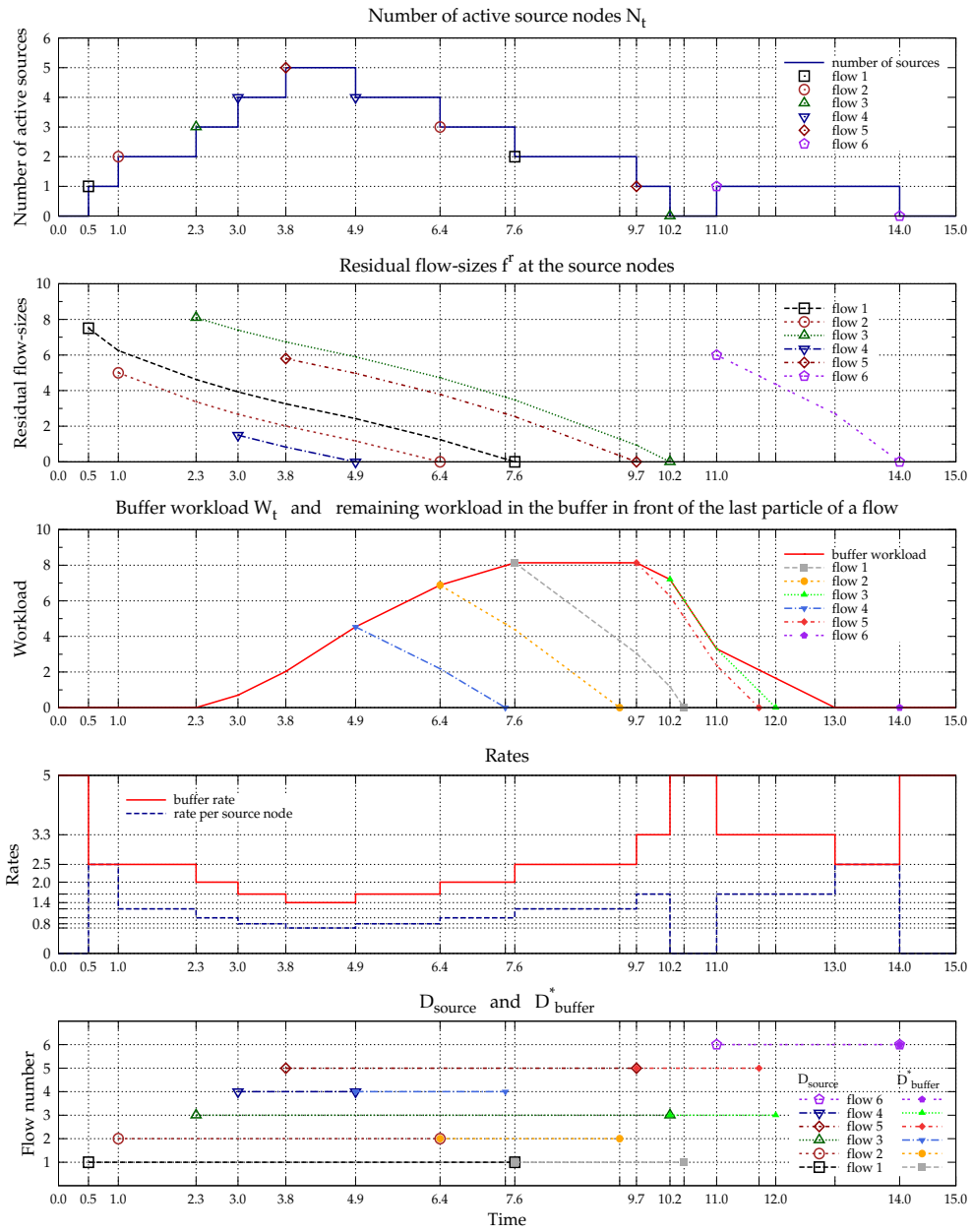

Figure 6.2: *Sample-path example of the resource sharing between source and relay nodes and the resulting performance metrics for resource-sharing ratio* $m = 2$.

the relay node is not backlogged, this last particle will be served instantly and the flow transfer is completed. If the relay node is backlogged, the last particle suffers an additional delay before the flow transfer time from source node to destination is completed.

In the sequel, we often refer to the special case of the resource-sharing ratio $m = 1$ as the 'standard' fluid model; this model captures the resource sharing of the IEEE 802.11 DCF, and was first introduced in [14].

REMARK 6.4.1 (SPECIAL CASE OF THE RESOURCE-SHARING RATIO: $m \in [0, 1]$). *Observe that for $m \in [0, 1]$ the relay node always obtains the full share m of the capacity C when $N_t \geq 1$, i.e., the resource sharing is independent of the workload W_t at the relay node. Consequently, the rates that are granted to the source nodes and the relay node are fully determined by N_t.*

6.4.3 Performance metrics

The main performance metric of interest is the overall flow transfer time D_{overall}, i.e., the time required to completely transfer a flow from source node to destination. The overall flow transfer time is the sum of two other performance metrics: i) the time (D_{source}) a source requires to completely transfer a flow to the relay node, and ii) the delay at the relay node (D^*_{buffer}) experienced by the last particle of the flow (the asterisk indicates that the performance measure relates to the last particle of a flow), i.e.,

$$D_{\text{overall}} = D_{\text{source}} + D^*_{\text{buffer}}.$$

In addition we are interested in the steady-state buffer workload W_{buffer} at an arbitrary moment (which is equal to the steady-state workload at a flow-arrival epoch due to the model assumption of Poisson arrivals), and the buffer workload W^*_{buffer} present at the arrival of a last particle of a flow at the relay node; observe that W^*_{buffer} contributes to the delay D^*_{buffer}.

The notation introduced above is used in Chapters 7 and 8, but in Chapter 9 we use a different, shorter notation. For reasons of conciseness in Chapter 9 we use F, D, and S for respectively D_{source}, D^*_{buffer}, and D_{overall}.

6.5 Characterization of the total workload in a wireless ad-hoc network

This section presents some preliminary analysis, in particular, we relate the total workload W_{total} in our model of the ad-hoc network scenario, described above with

flow-initiation rate λ and flow-size distribution F, to the workload in an M/G/1-queue. Therefore we use the key observations:

1. Each flow is served essentially *twice*; once by the source node to transmit the flow to the relay node, and once by the relay node to forward the flow to the destination node.

2. The overall system is *work-conserving*, i.e., the total service rate of the system is always C if there is work in the system.

We remark that these observations immediately lead to the stability criterion $2\rho < 1$, with $\rho := \lambda f/\mathrm{C}$.

Now it easily seen that the following lemma holds.

LEMMA 6.5.1. *$\mathbb{E}W_{\text{total}}$ corresponds to the virtual waiting-time in an M/G/1-queue with arrival rate λ and service-requirement distribution F with first moment $2f/\mathrm{C}$ and second moment $4f_2/\mathrm{C}^2$. Then,*

$$\mathbb{E}W_{\text{total}} = \left(\frac{2\rho}{1-2\rho}\right) \cdot \frac{1}{2}(\mathrm{C}_F^2 + 1)\frac{2f}{\mathrm{C}} = \left(\frac{2\rho}{1-2\rho}\right) \cdot (\mathrm{C}_F^2 + 1)\frac{f}{\mathrm{C}}. \tag{6.1}$$

Proof The expression for $\mathbb{E}W_{\text{total}}$ follows directly from the Pollaczek-Khintchine mean-value formula. □

Observe that Expression (6.1) is independent of m; this is a direct result of the fact that the total networks is work-conserving, i.e., 'work' always leaves the total system at rate C.

6.6 Validation scenarios

In this section we introduce the validation scenarios and their parameter settings, and we explain the details of the simulation environments used in the numerical studies of Chapters 7 and 8.

6.6.1 Validation scenarios

In the validation experiments in Chapters 7 and 8, we examine the transfer of flows (files) with an average flow size f of 0.12 Mbits, which corresponds to flows of 10 packets of each 1500 Bytes in an actual communication system (cf. [14, 114]). The following flow-size distributions are considered: Deterministic, Erlang-4, Exponential, and Hyper-Exponential (with balanced means, e.g., see p. 359 of [128]) with a Coefficient of Variation (CoV) C_F of $2, 4$ and 16.

The flow-initiation rate is varied such that the load ρ varies from very low loads (0.024) to an almost saturated system (0.48). Recall that, for stability, $\rho < \frac{1}{2}$ is required.

We assume a common radio-link with capacity $\mathrm{C} = 5$ Mbit/s, cf. the motivation in Section 8.3.1. The resource-sharing ratio m is varied over the values $\{0, 1, 2, 3, 5, 10\}$.

6.6.2 Simulation environments

For the numerical evaluations in the following chapters we rely on two simulation tools: a fluid-model simulator and a wireless ad-hoc network simulator. Both simulators are own-built simulation tools in the general-purpose programming language Delphi.

The *fluid-model simulator* exactly captures the behavior of the fluid model that was introduced in Section 6.4; the simulator is used in Chapter 7 to validate the analysis.

The *ad-hoc network simulator* is used in Chapter 8 as a validation of the fluid model. All the details of CSMA/CA contention of the EDCA are included in the simulator, e.g., the back-off mechanism, physical and virtual carrier sensing, and collision handling. The PHY-layer includes propagation- and fading-models and a clear channel assessment (CCA) procedure that results in limited ranges for successfully transmitting and receiving packets and sensing transmissions of other nodes. The PHY-parameters are set according to the IEEE 802.11b standard: RTS-, CTS- and ACK-frames are transmitted at 2 Mbit/s and the data-frames are transmitted at 11 Mbit/s.

The simulation environments are used for the numerical studies in Chapters 7 and 8. In the numerical experiments sufficiently many replications have been simulated in order to obtain small confidence intervals. In all experiments the confidence interval's half-width divided by the estimate is below 5%. We remark that this ratio is only close to 5% for high loads *and* high CoV (i.e., $\mathrm{C}_F = 16$), in the other cases the ratio is usually well below 1%.

Chapter 7

Mean-value analysis of the fluid model

7.1 Introduction

This chapter presents an analysis of the mean values of the performance metrics for the fluid model of Section 6.4. These performance measures include the average overall flow transfer time D_{overall}, i.e., the average time required to transmit a flow from a source node via the relay node to the destination, and the workload W_{buffer} at the relay node. We present analytical expressions for general resource-sharing ratio $m \in [0, \infty)$ and general flow-size distributions.

7.1.1 Contribution

In our analysis we first address the case of exponentially distributed flow-sizes. As the resource sharing is influenced by the workload at the relay node, the number of active source nodes itself does not constitute a Markov chain. However, the model falls within the class of so-called 'fluid queues with feedback' (or: 'feedback fluid-queues'), as introduced in Section 2.1.2. We provide an analysis of the overall flow transfer time.

For the case of generally distributed flow-sizes we cannot rely on the framework of fluid queues with feedback, as the exponentiality assumptions imposed there are crucial. However, we empirically observed the remarkable property that the distribution of the number of active source nodes is *insensitive* to the flow-size distribution (apart from its mean) for *any* general resource-sharing ratio $m \in [0, \infty)$. This insensitivity claim[1] is supported by a sizeable set of simulation experiments, corresponding to a broad range of parameters settings, including combinations of heavy load, a highly variable flow-size distribution, and a large resource-sharing ratio m.

By using the (conjectured) insensitivity as an approximation assumption, and by exploiting the relation between the workload of the total system and the workload of an appropriate M/G/1-queue (cf. Section 6.5), we derive insightful closed-form

[1] We consider the observed insensitivity as a highly surprising fact, since (for $m > 1$) the process describing the number of active source nodes does not fulfill the usual criteria for insensitivity (where we remark that for $m \in [0, 1]$ the insensitivity can be proven in a relatively elementary way, see Remark 7.3.2).

expressions for the mean workload at the relay node. In particular, we show that the mean workload at the relay node for general flow-size distributions is proportional to the mean workload for exponentially distributed flow-sizes; here the associated multiplicative factor is independent of the system parameters (including m), apart from the Coefficient of Variation of the flow-size distribution. In addition, we present an approximation for the overall flow transfer time for general flow-size distributions. The resulting expressions allow for easy numerical evaluation and are thoroughly validated by simulations of the fluid model.

7.1.2 Outline

The outline of this chapter is as follows. The fluid model, as was introduced in Section 6.4, is first analyzed for the special case of exponentially distributed flow-sizes in Section 7.2. In Section 7.3 the analysis is extended to the case of generally distributed flow-sizes; a crucial step in this analysis is played by our claim that the source-node behavior is insensitive to the flow-size distribution apart from its mean for general m. Section 7.4 presents an extensive numerical validation of the analyses of the earlier sections; in particular, we present empirical evidence of the insensitivity claim for a broad set of parameters, including extreme situations. In Section 7.5 we illustrate the benefits of differentiated resource sharing. Finally, Section 7.6 presents concluding remarks.

7.2 Analysis for exponentially distributed flow-sizes

This section presents an analysis of the model for the special case of exponentially distributed flow-sizes, which allows for a detailed analysis of the mean performance metrics. First, in Section 7.2.1 we derive the joint distribution of the number of active source nodes and the workload at the relay node at an arbitrary epoch. Next, this result is used in Section 7.2.2 to determine the workload at the relay node at the arrival epoch of the last particle of a flow. Finally, based on the results obtained in Sections 7.2.1 and 7.2.2, we present in Section 7.2.3 an approximation for the mean delay $\mathbb{E}D^*_{\text{buffer}}$ of the last particle of a flow that added to the mean delay $\mathbb{E}D_{\text{source}}$ yields the mean overall flow transfer time $\mathbb{E}D_{\text{overall}}$.

7.2.1 Joint distribution of the number of source nodes and the buffer workload

The source-node dynamics of the model of Section 6.4 does not constitute a Markov chain; the transition rates depend on both the number of active source nodes N_t and

on the buffer workload W_t, or, more precisely, whether W_t is positive or not. Put differently, there is *feedback* from the buffer content to the source-node behavior, in that information on the buffer content is needed to describe the source-node dynamics. Assuming exponential flow-sizes, the joint distribution of the number of source nodes and the buffer workload can be obtained by using the theory developed in [88] on so-called *feedback fluid-queues*, see e.g. Section 2.1.2. This joint distribution is derived in two steps: first we consider the simplified fluid model *without* feedback (i.e., the source nodes evolve independently of the buffer content), and then the obtained results are translated in terms of the model *with* feedback. An extensive treatment of this derivation is presented in Appendix 7.A; below we restrict ourselves to an excerpt.

First we consider the fluid-queue *without feedback*, i.e., the relay node always obtains share m of the capacity; let $\bar{N}_t$ and $\bar{W}_t$ respectively denote the number of active source nodes and buffer workload at epoch t in this model (in the sequel all quantities with bars '$^-$' relate to the model without feedback). Observe that $\bar{N}_t$ constitutes a Markov chain and has generator matrix $\bar{Q}$ (as given by Expression (7.21) in Appendix 7.A). Further, let $\bar{R}$ be a diagonal matrix where $\bar{R}_n$ is the *net* input rate into the relay node if n sources are active, i.e., $\bar{R}_n := (n-m)\,\mathrm{C}/(n+m)$. We define the stationary distribution of $(\bar{N}_t, \bar{W}_t)$ as

$$F_n(x) := \lim_{t\to\infty} \mathbb{P}\left(\bar{W}_t \le x; \bar{N}_t = n\right) = \mathbb{P}\left(\bar{W} \le x; \bar{N} = n\right).$$

To facilitate the analysis we here assume that a maximum n_{max} is imposed on the number of source nodes that may be simultaneously active; flows that are initiated, if already n_{max} other source nodes are active, are blocked. Observe, however, that if one chooses n_{max} sufficiently large (which we will do in the sequel), this will not have any significant impact on the outcome of the model.

The buffer workload has to satisfy the Kolmogorov forward equations

$$\overrightarrow{F}'(x)\bar{R} = \overrightarrow{F}(x)\bar{Q}'.$$

The spectral expansion of the solution is given by

$$\overrightarrow{F}(x) = \sum_{j=0}^{n_{max}} a_j \overrightarrow{v}_j \exp(z_j x)$$

where $(z_j, \overrightarrow{v}_j)$ is an eigenvalue-eigenvector pair, i.e., a scalar and vector that solve $z_j \overrightarrow{v}_j \bar{R} = \overrightarrow{v}_j \bar{Q}$. For details on how to obtain the coefficients a_j see Appendix 7.A. Clearly, $\mathbb{P}(\bar{W} \le x) = \sum_n F_n(x)$. For stability we assume that the average net input rate is negative, i.e., $\sum_n \omega_n \bar{R}_n < 0$, where ω denotes the distribution of $\bar{N}$, i.e., $\omega_n := F_n(\infty)$.

The joint distribution of the workload and number of active sources for the fluid model *with* feedback follows from the corresponding joint distribution of the fluid

model *without* feedback; this result follows from the crucial observation that both models behave identically *during busy periods*, see [88]. Hence, the joint distribution $G_n(x)$ of (N, W) for the model with feedback is found after some sort of rescaling the distribution $F_n(x)$ of $(\bar{N}, \bar{W})$:

$$G_n(x) = \frac{F_n(x) - \sum_k F_k(0)}{1 - \sum_k F_k(0)} \mathbb{P}(W > 0) + \mathbb{P}(W = 0; N = n). \tag{7.1}$$

For the derivation of the probabilities $\mathbb{P}(W > 0)$ and $\mathbb{P}(W = 0; N = n)$ in (7.1) we refer to Appendix 7.A. The stationary distribution π of the number of active source nodes N follows from (7.1) as

$$\pi_n = G_n(\infty). \tag{7.2}$$

Let π^a (π^d) denote the stationary distribution of the active number of source nodes present at a flow-arrival epoch (left behind at a source-departure epoch, which coincides with the arrival epoch of the last particle of a flow at the relay node). Then we have the following result.

LEMMA 7.2.1. *π, π^a and π^d are identical.*

Proof $\pi = \pi^a$ follows directly from the PASTA-property. $\pi^d = \pi^a$ as N_t is a birth-death process and the number of up-crossings of level n balances the number of down-crossings of level n. □

7.2.2 Flow transfer time and the buffer workload

This subsection first explains how to compute $\mathbb{E}D_{\text{source}}$, and then provides two expressions for $\mathbb{E}W_{\text{buffer}}$, one of them being based on the theory developed in Section 7.2.1. Finally it addresses the expected workload $\mathbb{E}W^*_{\text{buffer}}$ at the relay node at the epoch that a last particle of a flow arrives at the relay node, by considering the sum of the mean workload $\mathbb{E}W_{\text{buffer}}$ present at flow initiation and the mean workload increase $\mathbb{E}\Delta W_{\text{buffer}}$ at the buffer of the relay node during a flow transfer.

Expected flow transfer time $\mathbb{E}D_{\text{source}}$

The expected number of active source nodes is given by

$$\mathbb{E}N = \sum_{n=0}^{\infty} n\pi_n. \tag{7.3}$$

The flow transfer time follows from Little's law:

$$\mathbb{E}D_{\text{source}} = \mathbb{E}N/\lambda. \tag{7.4}$$

Expected buffer workload at an arbitrary moment

The expected buffer workload $\mathbb{E}W_{\text{buffer}}$ at the relay node upon flow arrival can be obtained in two manners which are both presented below: in the first place by relying on the workload distribution, i.e., Expression (7.1), and secondly as a direct application of Lemma 6.5.1.

Buffer workload at an arbitrary moment, using the workload distribution. Observe that $\pi_n^{-1}G_n(x)$ is the workload distribution conditional on n source nodes being active. Then, the expected conditional buffer workload is obtained by

$$\begin{aligned}\mathbb{E}[W_{\text{buffer}}|N=n] &= \pi_n^{-1}\int_0^\infty x\mathrm{d}G_n(x)\\ &= \pi_n^{-1}\frac{1}{1-\sum_i F_i(0)}\mathbb{P}(W>0)\sum_{j=0,j\neq n_+}^{n_{max}}\frac{a_j(\overrightarrow{v}_j)_n}{z_j},\end{aligned}$$

see Appendix 7.A. Hence, the expected unconditional buffer workload is:

$$\mathbb{E}W_{\text{buffer}} = \frac{1}{1-\sum_i F_i(0)}\mathbb{P}(W>0)\sum_{n=0}^{n_{max}}\sum_{j=0,j\neq n_+}^{n_{max}}\frac{a_j(\overrightarrow{v}_j)_n}{z_j}. \tag{7.5}$$

Buffer workload at an arbitrary moment, using the relation between the model and the workload in the corresponding M/G/1 FCFS*-queue.* We define W_{sources} as the aggregate workload at all active source nodes, and W_{total} and W_{buffer} as in respectively Section 6.5. Recall that work at a source node still needs to be served twice, i.e., by the source node and relay node. Then,

$$W_{\text{total}} = W_{\text{sources}} + W_{\text{buffer}}. \tag{7.6}$$

Observe that W_{total} does not depend on the resource-sharing ratio m; in fact, it does not even depend on the service discipline as long as it is work-conserving.

As $\mathbb{E}W_{\text{total}}$ is given by Expression (6.1) with $\mathrm{C}_F^2 = 1$, we are left to derive $\mathbb{E}W_{\text{sources}}$. This follows due to the memoryless property of the exponential distribution of the flow sizes. The expected amount of work at an active source node (i.e., the residual of the flow) equals $2f/\mathrm{C}$. Furthermore the expected number of source nodes simultaneously active is given by Expression (7.3). Hence

$$\mathbb{E}W_{\text{sources}} = \mathbb{E}N\cdot 2f/\mathrm{C},$$

and the expected workload at the buffer of the relay node is:

$$\mathbb{E}W_{\text{buffer}} = \left(\frac{2\rho}{1-2\rho} - \mathbb{E}N\right)\cdot 2f/\mathrm{C}. \tag{7.7}$$

REMARK 7.2.2 (RELATION BETWEEN EXPRESSIONS (7.5) AND (7.7) FOR $\mathbb{E}W_{\text{buffer}}$). *It is interesting to note that Expression (7.5) depends on all the eigenvalue-eigenvectors $(z_j, \overrightarrow{v}_j)$ for $j = \{0, .., n_{max}\}$; on the other hand Expression (7.7) depends just through $\mathbb{E}N$ on the normalized eigenvector that corresponds to the zero eigenvalue, cf. Expressions (7.3) and (7.2). Due to the implicitness of the eigenvalue-eigenvector pairs $(z_j, \overrightarrow{v}_j)$ and corresponding constants a_j, it is not a priori obvious from these expressions that they match.* ♢

Buffer workload at the arrival epoch of the last particle of a flow at the relay node

In this section we derive the expected buffer workload $\mathbb{E}W^*_{\text{buffer}}$ at the epoch that the last particle of a flow arrives at the relay node. Note that the mean buffer workload on flow initiation coincides with the mean workload at an arbitrary epoch (PASTA). Hence, W^*_{buffer} can be obtained using the following relation:

$$\mathbb{E}W^*_{\text{buffer}} = \mathbb{E}W_{\text{buffer}} + \mathbb{E}\Delta W_{\text{buffer}}, \tag{7.8}$$

where ΔW_{buffer} denotes the buffer increase during the transfer time D_{source} of an arbitrary flow.

We are left to derive $\mathbb{E}\Delta W_{\text{buffer}}$. Let ΔW_{total} denote the increase in workload in the total system during the flow transfer time D_{source} by a source node, and let $\Delta W_{\text{sources}}$ denote the increase (which may be negative) of the aggregate workload of all source nodes during D_{source}. By (7.6) we evidently have the following relation:

$$\Delta W_{\text{total}} = \Delta W_{\text{sources}} + \Delta W_{\text{buffer}}. \tag{7.9}$$

LEMMA 7.2.3. $\mathbb{E}\Delta W_{\text{buffer}} = \mathbb{E}\Delta W_{\text{total}}$.

Proof We have to prove that $\mathbb{E}\Delta W_{\text{sources}} = 0$, i.e., the expected amount of work at the source nodes present upon arrival of a flow coincides with the amount present at the corresponding source-departure epoch. This property follows directly from two observations. First, due to Lemma 7.2.1, the expected number of source nodes at the flow-initiation epoch equals the expected number of source nodes at the epoch of the arrival of the last particle of a flow at the relay node. Second, the expected residual flow-sizes at these instances coincide due to the memoryless property of the exponentially distributed flow-sizes. □

Lemma 7.2.3 and Relation (7.9) lead to the following proposition.

PROPOSITION 7.2.4. *The expected increase of the buffer workload $\mathbb{E}\Delta W_{\text{buffer}}$ during the transfer time D_{source} of a flow is given by*

$$\mathbb{E}\Delta W_{\text{buffer}} \quad = \quad (\mathbb{E}N + 1) \cdot 2f/\text{C} - \mathbb{E}N/\lambda. \tag{7.10}$$

Proof Due to Lemma 7.2.3 we are left to compute $\mathbb{E}\Delta W_{\text{total}}$ during the flow transfer time (with mean $\mathbb{E}D_{\text{source}}$) of a tagged flow. The input into the total system is the result of initiations of new flows (including the tagged flow) which arrive at rate λ, each bringing along an amount of work with expected value $2f/\text{C}$ (cf. Lemma 6.5.1). The expected number of arrivals (including the tagged flow) is $\lambda\mathbb{E}D_{\text{source}} + 1$, and consequently the expected input into the total system is $(\lambda\mathbb{E}D_{\text{source}} + 1) \cdot 2f/\text{C}$.

The expected output is $\mathbb{E}D_{\text{source}}\text{C}$, as is readily verified by the following two observations. First, the total system is non-empty during the flow transfer time D_{source} as at least the tagged flow is served during D_{source}. Second, the total system is work-conserving and serves at rate C. Writing the expressions in terms of $\mathbb{E}N$ using (7.4) proves the lemma. □

Notice that the expected workload $\mathbb{E}W^*_{\text{buffer}}$ of Expression (7.8), which is the sum of Expressions (7.7) and (7.10), only depends on the resource-sharing ratio m via $\mathbb{E}N$. Recall that $\mathbb{E}N$ is given by Expression (7.3) that can be determined by Expressions (7.1) and (7.2).

An interesting result follows from rewriting Expressions (7.7) and (7.10) in terms of $\mathbb{E}N$ and considering their ratio. It turns out that, remarkably, the proportionality constant does not depend on m.

COROLLARY 7.2.5. *The expected workload increase $\mathbb{E}\Delta W_{\text{buffer}}$ at the relay node during a flow transfer is proportional to the expected workload $\mathbb{E}W_{\text{buffer}}$ at flow arrival:*

$$\mathbb{E}\Delta W_{\text{buffer}} = \frac{1-2\rho}{2\rho}\mathbb{E}W_{\text{buffer}},$$

and Expression (7.8) can be written as

$$\mathbb{E}W^*_{\text{buffer}} = \frac{1}{2\rho}\mathbb{E}W_{\text{buffer}}.$$

7.2.3 Mean delay $\mathbb{E}D^*_{\text{buffer}}$ of the last particle and mean overall flow transfer time $\mathbb{E}D_{\text{overall}}$

At the moment that the source node has transmitted the full flow into the relay node, the last fluid particle enters the buffer at the relay node, and then the source node becomes inactive. In this subsection we present an approximation for the expected buffer delay $\mathbb{E}D^*_{\text{buffer}}$ of this last particle.

Recall that the last particle does not experience any buffer delay if the buffer is empty. In case the buffer is non-empty, the buffer delay D^*_{buffer} of the last particle is the time required by the relay node to serve the amount of work W^*_{buffer} present upon arrival of that last particle. Recall from Section 6.4 that during D^*_{buffer} the relay node uses the entire resource-sharing ratio m. Hence, during D^*_{buffer} the behavior of the system behaves as the model without feedback presented in Appendix 7.A.1.

Conditional buffer delay of the last particle at the relay node. Let $Y_n(\tau)$ denote the conditional buffer delay, i.e., the time required by the relay node with resource-sharing ratio m to serve an amount τ of fluid if initially n source nodes are active. Here we again assume that there is a maximum n_{max} imposed on the number of source nodes that may be simultaneously active, as in Section 7.2.1. Let $\bar{Q}$ denote the generator matrix without feedback as in (7.21) and $M(s) := -\bar{Q} + sR$ where

$$R := \text{diag}\left\{1, \frac{m}{m+1}, \frac{m}{m+2}, \cdots, \frac{m}{m+n_{max}}\right\}.$$

PROPOSITION 7.2.6. *The expected conditional time required to serve an amount τ of fluid, if initially n source nodes are active, is given by*

$$\mathbb{E}Y_n(\tau) = A_{(n,0)}\tau + \sum_{j=1}^{n_{max}} \frac{A_{(n,j)}}{s_j} e^{s_j\tau} - \sum_{j=1}^{n_{max}} \frac{A_{(n,j)}}{s_j} \tag{7.11}$$

where s_j denote the eigenvalues of $R^{-1}\bar{Q}$. The constants $A_{(n,j)}$ follow from the partial-fraction expansion of

$$s\phi_n(s) = \frac{\det M_{-n}(s)}{\det M(s)}.$$

where $M_{-n}(s)$ is defined as $M(s)$ with the n-th column replaced by $\overrightarrow{1}$.

Proof See Appendix 7.B. □

In Expression (7.11) the eigenvalues s_j for $j \geq 1$ are negative, and hence $\mathbb{E}Y_n(\tau)$ is approximately linear in τ.

Approximation of the buffer delay. We now use Proposition 7.2.6 to approximate the buffer delay experienced by the last particle of the flow. By definition the expected buffer delay $\mathbb{E}Y_n(\tau)$ of the last particle can be expressed as

$$\mathbb{E}D^*_{\text{buffer}} = \sum_{n=0}^{n_{max}} \pi_n^d \int_0^\infty \mathbb{E}Y_n(\tau) w_n^*(\tau) \mathrm{d}\tau,$$

where $w_n^*(\tau)$ is the probability density function of the amount of work at the buffer at a source-departure epoch leaving behind n source nodes, and π^d coincides with π (due to Lemma 7.2.1). Unfortunately, we do not have the density function $w_n^*(\tau)$.

If one assumes, however, that $\mathbb{E}Y_n(\tau)$ is linear in τ, the conditional buffer delay roughly looks like

$$\int_0^\infty \mathbb{E}Y_n(\tau) w_n^*(\tau) \mathrm{d}\tau \approx \mathbb{E}Y_n(\mathbb{E}W^*_{\text{buffer}}). \tag{7.12}$$

Then we obtain the following approximation for the expected delay D^*_{buffer} of the last particle.

APPROXIMATION 7.2.7. *The buffer delay of the last particle can be approximated by*

$$\mathbb{E}D^*_{\text{buffer}} \approx \sum_{n=0}^{n_{max}} \pi_n \mathbb{E}Y_n \left(\mathbb{E}W^*_{\text{buffer}}\right). \tag{7.13}$$

REMARK 7.2.8 (SPECIAL CASE $m = 1$). *As mentioned earlier, the special case m equals 1 was studied in [14]. For exponential flow-sizes and $m = 1$ and $n_{max} \to \infty$ a closed-form expression for the equivalent of (7.11) is available, namely Expression (33) of [29] (also included as Expression (8.3)). It is seen that Expressions (33) of [29] and (7.11) are very similar in nature, in the sense that both expressions consist of a linear term and in addition a term that is exponentially decaying in τ.* ♢

Now we have derived expressions for all the parts of the main performance metric from the user perspective: the mean expected overall flow transfer time $\mathbb{E}D_{\text{overall}}$ is

$$\mathbb{E}D_{\text{overall}} = \mathbb{E}D_{\text{source}} + \mathbb{E}D^*_{\text{buffer}} \tag{7.14}$$

where $\mathbb{E}D_{\text{source}}$ is given by (7.4) and $\mathbb{E}D^*_{\text{buffer}}$ is approximated by (7.13).

7.3 Analysis for generally distributed flow-sizes

This section treats the analysis the performance metrics for *generally* distributed flow-sizes. The analysis presented in this section borrows elements from the approach followed in the previous section for the case of exponentially distributed flow-sizes.

Our analysis relies heavily on knowledge of the stationary distribution of the number of active source nodes, together with the expected residual flow-sizes at the source nodes. In Section 7.3.1 we present an approximation assumption that states that the distribution of the number of active source nodes is insensitive to the flow-size distribution, and we also state two (related) properties concerning the expected residual flow-sizes at the source nodes. (These claims will be thoroughly assessed in Section 7.4.2.) The approximation assumption relates to a general resource-sharing ratio $m \geq 0$, and we use it to derive the mean workload at the relay node in Section 7.3.2. In Section 7.3.3 we then consider the buffer delay of the last particle of a flow, which, together with earlier results, enables us to compute the mean overall flow transfer time. Finally, Section 7.3.4 presents an overview of all the expressions required to evaluate the performance metrics, in the form of a calculation scheme.

7.3.1 Steady-state behavior of the active source nodes

By extensive simulations of the fluid model, for $m \geq 0$, we observed the striking property i) that the source-node behavior (in terms of the distribution of the number

of active source nodes) is *insensitive* to the flow-size distribution, i.e., only the mean flow-size plays a role. In addition, our simulations revealed that the system exhibits two other characteristics that are closely related to insensitive systems, i.e., ii) the residual flow-sizes at the source nodes are very well approximated by the 'usual' excess life distribution as known from renewal theory, and iii) the residual flow-sizes are nearly independent of the number of active source nodes. The numerical evidence for these claims is presented in detail in Section 7.4.2 (see Figures 7.3 and 7.4). We assessed the properties for wide ranges of the parameter settings including (extremely) heavy loads, various flow-size distributions, and high resource-sharing ratios. The simulations indicate that the claim i) is exact, whereas claims ii) and iii) seem to hold as a very accurate approximation; in fact it took a huge number of replications to show that those statements were not exact. This motivates the use of the three properties i)–iii) as *approximation assumptions.*

Let us now formally state the approximation assumptions. As mentioned above, convincing support is provided by the extensive simulation results, to be presented in Section 7.4.2, but it is noted that in Remark 7.3.2 we formally *prove* that for the special case $m \in [0, 1]$ the assumptions holds.

ASSUMPTION 7.3.1.

1. *The stationary distribution of the number of active source nodes is insensitive to the flow-size distribution apart from its mean and is given by Expression (7.2).*

2. *The expected residual flow-size $\mathbb{E}[F^r]$ of a flow at a source node coincides with the expected residual flow-size of a renewal process, i.e.,*

$$\mathbb{E}[F^r] = (1/2)(\mathrm{C}_F^2 + 1) \cdot f = \frac{f_2}{2f}. \tag{7.15}$$

3. *The number of active source nodes N_t and their expected residual flow-sizes $\mathbb{E}[F^r]$ are independent.*

Properties as those mentioned in Assumption 7.3.1 are well-known to hold for stationary symmetric queues, cf., e.g., Kelly [69], Cohen [31], Bonald and Proutière [17], and more recently the work of Zachary [135]. However, the service discipline of our model is *not* symmetric; the requirement that the service rate only depends on N_t is not fulfilled as the service rate also depends on the workload W_t at the buffer of the relay node. We would therefore like to stress that the (empirically observed) insensitivity of Assumption 7.3.1 is a highly remarkable property: to the authors' best knowledge there are no results on other insensitive queueing-systems that do not have a symmetric service-discipline. It is a subject for further research to formally prove this insensitivity.

REMARK 7.3.2 (INSENSITIVITY OF THE SOURCE-NODE BEHAVIOR FOR $m \in [0,1]$). *For $m \in [0,1]$ the relay node obtains a service rate less than or equal to the share that each active source node obtains. Therefore, the relay node always obtains its entire share m if there is work in the system; consequently, the resource sharing only depends on the number of active sources (and no information on the buffer content is needed).*

Hence, the behavior of the source nodes is described by a Processor Sharing queue with state-dependent service rates $r(n) := n\mathrm{C}/(n+m)$ if n source nodes are active. This model is a special case of the so-called Generalized Processor Sharing queue described by Cohen in [31] for which he presented a joint stationary probability/density function of the number of active sources nodes N and their residual service requirements $T := (T(1), ..., T(N))$, cf. formula (7.19) in [31]:

$$\mathbb{P}(N = n, T = \tau) = \frac{\frac{(\lambda\beta)^n}{n!}\varphi(n)}{\sum_{k=0}^{\infty}\frac{(\lambda\beta)^k}{k!}\varphi(k)}\prod_{i=1}^{n}\frac{1-B(\tau(i))}{\beta}, \qquad n = 0, 1, ..., \quad \tau(i) \geq 0, \tag{7.16}$$

where $\varphi(0) := 1$ and $\varphi(n) := (\prod_{i=1}^{n} r(i))^{-1}$, for $n = 1, 2, ...$, and where $B(\cdot)$ denotes the customers' service requirement distribution, β is the mean service requirement and λ the customer arrival rate.

In [31] it was shown that the stationary distribution is insensitive to the flow-size distribution and that it is independent of the residual flow-sizes. Further, [31] establishes that the residual flow-sizes are distributed according to the excess life distribution from renewal theory, see e.g. Expression (7.15), and that the residual flow sizes and the number of active source nodes are independent. A more explicit expression for the distribution of the number of source nodes can be obtained from the local-balance equations:

$$\pi_n = (1-\rho)^{m+1} \cdot \rho^n \prod_{k=1}^{n}\frac{m+k}{k}. \tag{7.17}$$

♢

7.3.2 Mean buffer workload

For the expected buffer workload $\mathbb{E}W_{\text{buffer}}$ we use the relation $\mathbb{E}W_{\text{total}} = \mathbb{E}W_{\text{sources}} + \mathbb{E}W_{\text{buffer}}$. The mean total workload $\mathbb{E}W_{\text{total}}$ is given by (6.1), and $\mathbb{E}W_{\text{sources}}$ follows from Assumption 7.3.1 and equals

$$\mathbb{E}W_{\text{sources}} = \mathbb{E}N \cdot \mathbb{E}[F^r]. \tag{7.18}$$

We obtain the following expression for the expected workload in the buffer.

LEMMA 7.3.3. *Under Assumption 7.3.1,*

$$\mathbb{E}W_{\text{buffer}} = \left(\frac{2\rho}{1-2\rho} - \mathbb{E}N\right) \cdot (\mathrm{C}_F^2 + 1)f/\mathrm{C}. \tag{7.19}$$

Note that Expression (7.19) coincides with Expression (7.7) as $\mathrm{C}_F^2 = 1$ for exponentially distributed flow-sizes. Corollary 7.3.4 results from considering the ratio of Expressions (7.19) and (7.7).

COROLLARY 7.3.4. *Under Assumption 7.3.1, the buffer workload in case of general flow-size distribution relates to the workload in case of exponential flow-size distribution, with the same mean f, in the following manner. In self-evident notation,*

$$\mathbb{E}W_{\text{buffer}} = \frac{(\mathrm{C}_F^2 + 1)}{2}\mathbb{E}W_{\text{buffer}}^{\exp}. \tag{7.20}$$

The important implication of relation (7.20) is that it entails that the expected buffer workload for general flow-size distributions is *proportional to* the expected buffer workload for exponential flow-sizes. It is stressed that the proportionality constant just includes the CoV, and, importantly, that this factor is independent of the resource-sharing ratio m (but recall that, evidently, $\mathbb{E}W_{\text{buffer}}^{\exp}$ *does* depend on m).

Using Expression (7.8), the mean buffer workload $\mathbb{E}W_{\text{buffer}}^*$ at the arrival of the last particle is the sum of $\mathbb{E}W_{\text{buffer}}$ and the mean workload increase $\mathbb{E}\Delta W_{\text{buffer}}$. The latter is derived in Section 7.2.2, and observe that the derivation is independent of the flow-size distribution, i.e., Expression (7.10) holds for general flow-size distributions. As a consequence, imposing Assumption 7.3.1, $\mathbb{E}N$ is still given by (7.3).

7.3.3 Mean buffer delay of the last particle in case of general flow-size distributions

For the mean buffer delay $\mathbb{E}D_{\text{buffer}}^*$ of the last particle we use Approximation (7.13) which is derived in Section 7.2.3, although the approximation is derived assuming exponentially distributed flow-sizes. This procedure is motivated by considering the two ways in which the flow-size distribution has an impact on the buffer delay.

- First, it affects the buffer workload W_{buffer}^* seen by the last particle, but recall that this effect could (under Assumption 7.3.1) be captured, and resulted in Expression (7.10), see the remarks at the end of Section 7.3.2.

- The second effect is on the transient behavior during buffer delay D_{buffer}^* where the resource sharing depends on the number of active source nodes and their residual flow sizes. Recall that during the entire D_{buffer}^* the relay node continuously obtains ratio m, which is, importantly, a symmetric service-discipline (therefore corresponding to an insensitive invariant distribution). Small flows have a small delay anyway, whereas long jobs will see a number of source nodes that is (by approximation) distributed according to this invariant. This suggests that the impact of the distribution of the flow-sizes is only modest.

Table 7.1: *Performance metrics calculation scheme for general flow-size distributions.*

Performance metric	corresponding expression	required expressions	derived by
$\mathbb{E}N$	(7.3)	(7.1)	FQWF, Conjecture 7.3.1
$\mathbb{E}D_{\text{source}}$	(7.4)	(7.3)	FQWF, Conjecture 7.3.1
$\mathbb{E}W_{\text{buffer}}$	(7.20)	(7.3)	FQWF, Conjecture 7.3.1 or Expr. (7.6)
$\mathbb{E}\Delta W_{\text{buffer}}$	(7.10)	(7.3)	Expr. (7.6)
$\mathbb{E}W^*_{\text{buffer}}$	(7.8)	(7.20), (7.10)	Expr. (7.6)
$\mathbb{E}D^*_{\text{buffer}}$	(7.13)	(7.8), (7.11)	Laplace transforms
$\mathbb{E}D_{\text{overall}}$	(7.14)	(7.4), (7.13)	Expr. (7.6)

7.3.4 Calculation scheme of the performance metrics for general flow-sizes

In order to facilitate easy evaluation of all performance metrics involved, we present in Table 7.1 an overview of the expressions required to calculate the performance metrics. For each performance metric we state the equation number of the corresponding expression in this chapter, which other expressions are required to calculate this expression, and how the expression was derived (where FQWF is an abbreviation of 'fluid-queue with feedback').

7.4 Numerical results

This section serves three goals: i) to numerically illustrate the behavior of the system as described in Section 6.4 (or, more specifically, to assess the impact of the ratio m under various loads, and for various flow-size distributions), ii) to provide empirical evidence for Assumption 7.3.1, and iii) to validate the approximations proposed in the previous sections. Recall that the validation scenarios and the parameter settings used in the numerical examinations were already introduced in Section 6.6. Section 7.4.1 presents results for exponentially distributed flow-sizes and general resource-sharing ratios (cf. Section 7.2). Numerical support of Assumption 7.3.1 is provided in Section 7.4.2. Finally, Section 7.4.3 focuses on the performance metrics for general flow-size distributions (cf. Section 7.3). In addition, we refer to Section 8.3.3 which presents, among other things, a numerical evaluation for the special case $m = 1$.

7.4.1 Results for exponentially distributed flow-sizes

This section presents numerical results for exponentially distributed flow-sizes. Figure 7.1 presents the mean flow transfer time $\mathbb{E}D_{\text{source}}$ (left graph) and the buffer workload $\mathbb{E}W_{\text{buffer}}$ (right graph) for different values of resource-sharing ratio m as

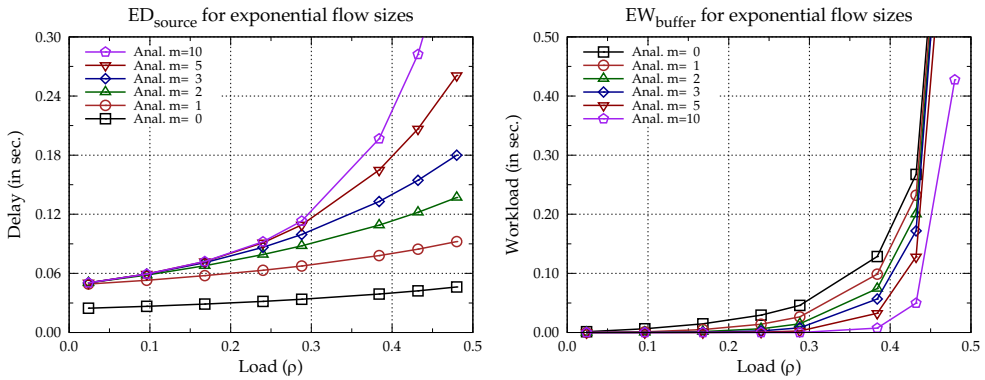

Figure 7.1: *Exponential flow-sizes. Left:* $\mathbb{E}D_{\text{source}}$. *Right:* $\mathbb{E}W_{\text{buffer}}$.

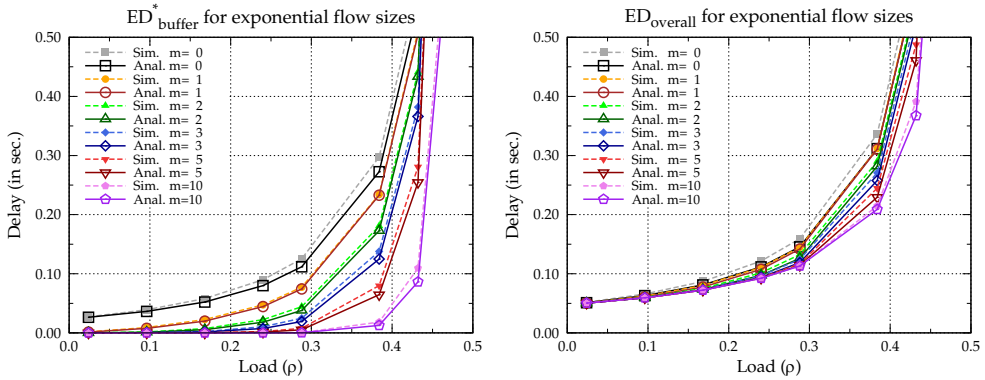

Figure 7.2: *Exponential flow-sizes. Left:* $\mathbb{E}D^*_{\text{buffer}}$. *Right:* $\mathbb{E}D_{\text{overall}}$.

a function of the load. Recall that the results are exact for exponentially distributed flows; therefore we do not compare these results with simulations. The graphs illustrate the influence of resource-sharing ratio m: a small ratio m implies that the source nodes obtain a large share of the capacity resulting in short flow transfer times $\mathbb{E}D_{\text{source}}$ for the source nodes. On the other hand for small ratios m the relay node obtains a small share of the capacity which results in a larger mean buffer workload $\mathbb{E}W_{\text{buffer}}$.

Figure 7.2 presents the approximation of the mean buffer delay $\mathbb{E}D^*_{\text{buffer}}$ (left graph) and the mean overall flow transfer time $\mathbb{E}D_{\text{overall}}$ (right graph). Approximation (7.13) of the buffer delay performs very well: it is close to the simulation results. The right graph shows the overall performance $\mathbb{E}D_{\text{overall}}$. The small error between the analysis and simulation results is solely due to the approximation of the buffer

delay $\mathbb{E}D^*_{\text{buffer}}$, as $\mathbb{E}D_{\text{source}}$ is exact. Also, the graph illustrates that the $\mathbb{E}D_{\text{overall}}$ improves for increasing m. Observe that the curves of $\mathbb{E}D_{\text{overall}}$ for different values of m are close to each other for $\rho < 0.4$; this indicates that the trade-off between $\mathbb{E}D_{\text{source}}$ and $\mathbb{E}D^*_{\text{buffer}}$ is more or less balanced for these regimes.

7.4.2 Numerical evidence for Assumption 7.3.1

Recall that Assumption 7.3.1 consists of three parts, which we will validate separately. For $m \in [0, 1]$ these results can be proven, see Remark 7.3.2; in this section we therefore focus on $m > 1$.

To validate part i) of the assumption (i.e., the distribution of the source nodes is insensitive to the flow-size distribution), we compare the stationary distribution of the number of active source nodes obtained by simulations of the fluid model (for various flow-size distributions) with the exact results for the exponential flow-size distribution. Figure 7.3 shows the results for high loads $\rho \in \{0.38, 0.48\}$ and resource-sharing ratio $m \in \{2, 5, 10\}$. We observe that the distributions from simulation and the analysis coincide for all ranges, i.e., the analytically obtained probability of n active source nodes falls within each of the confidence intervals of the simulated probability, for *all* flow-size distributions. These results offer strong support for the first part of Assumption 7.3.1 for all loads and resource-sharing ratios. Observe also the remarkable shapes of the stationary distributions; in particular, consider the shape for high load and high resource-sharing ratio in the bottom-right panel of Figure 7.3.

Figure 7.4 presents numerical results for the second part of Assumption 7.3.1 (i.e., the expected residual flow-size at a source node coincides with the expected residual excess flow-size from renewal theory). The numerical results from the simulation and the analysis are very close together, although the proposed mean residual flow-size distribution given by Expression (7.15) is not always within the confidence interval of the mean residual flow-sizes at the source nodes as obtained from our simulations. For scenarios with a high load and a flow-size distribution with $\mathrm{c}_F > 1$ we observe that the simulated mean residual flow size is slightly larger than the analytical value. In summary, the curves are very close together, and it seems justified to use the claimed as an approximation assumption.

The third part of Assumption 7.3.1 deals with the independence of the number of active source nodes and their residual flow-sizes. This assumption is mainly required to support Expression (7.18), i.e., $\mathbb{E}W_{\text{sources}} = \mathbb{E}N \cdot \mathbb{E}[F^r]$. Therefore we directly sampled $\mathbb{E}W_{\text{sources}}$, $\mathbb{E}N$, and $\mathbb{E}[F^r]$ from simulations; there we observed that both sides of Expression (7.18) match (up to a high level of precision), and the same applies to the mean residual flow size $\mathbb{E}[F^r]$ (which follows from Expression (7.15)) and the directly sampled value.

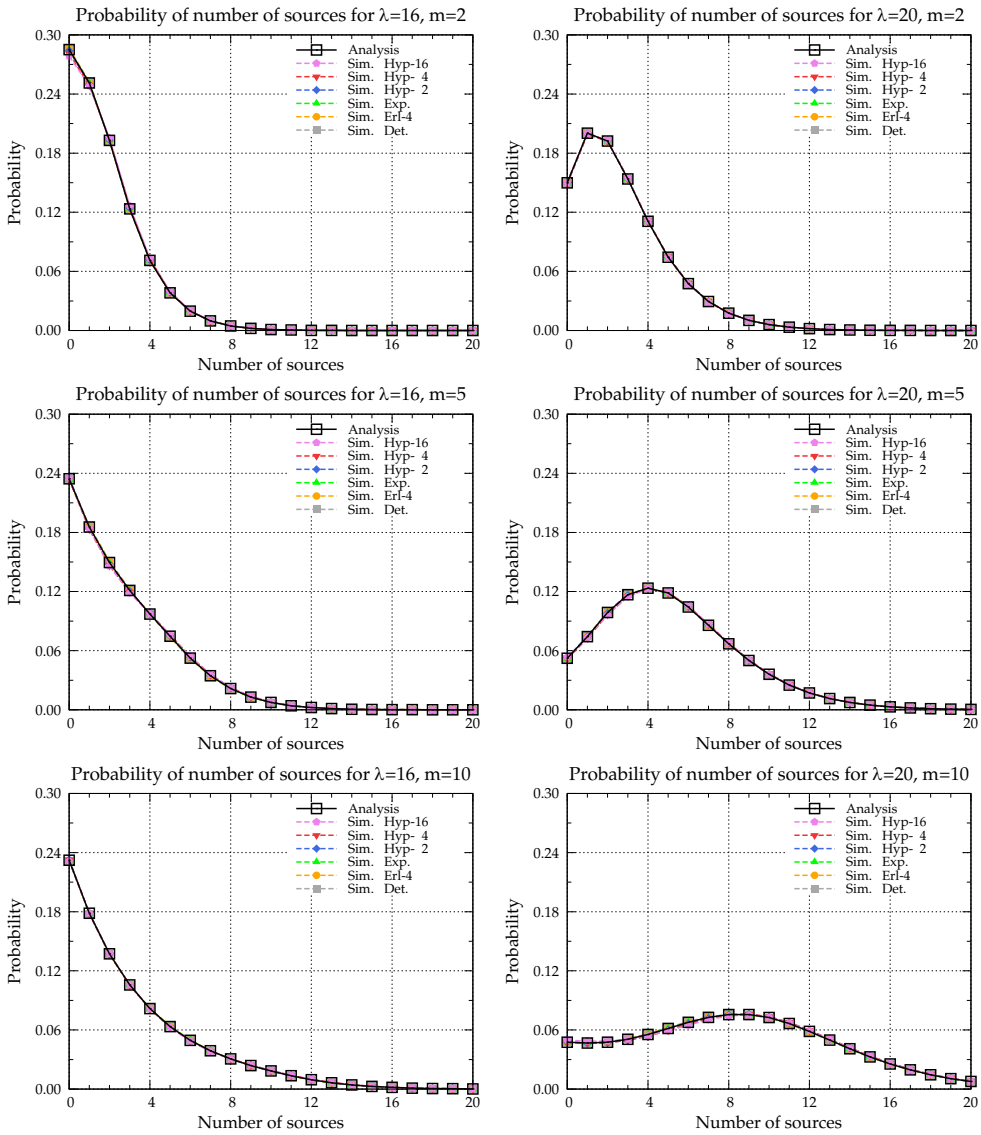

Figure 7.3: *Distributions of the number of source nodes for different flow-size distributions. Left:* $\lambda = 16$, $\rho = 0.38$. *Right:* $\lambda = 20$, $\rho = 0.48$. *Top: ratio* $m = 2$. *Middle: ratio* $m = 5$. *Bottom: ratio* $m = 10$.

7.4.3 Results for general flow-size distributions

This section focuses on the validation of the analysis for general flow-size distributions. We do so by comparing the output of our calculation scheme with esti-

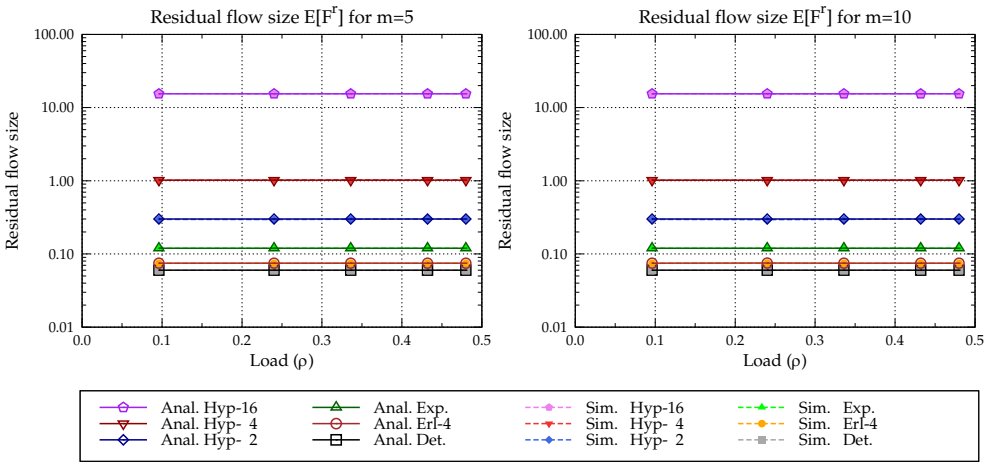

Figure 7.4: *Residual flow-sizes at the source nodes for different flow-size distributions. Left:* $m = 5$. *Right:* $m = 10$.

mates obtained from fluid simulations. We present the main performance metrics for general flow-size distributions and general resource-sharing ratio m. Each graph in Figures 7.5–7.8 shows the performance metrics as a function of the load for various flow-size distributions. In each graph the resource-sharing ratio m is fixed: left $m = 2$ and right $m = 5$. The effects of the load and resource-sharing ratio on the performance metrics are similar to the results for the exponential case of Section 7.4.1 and will not be discussed again.

Figure 7.5 presents the results for the mean flow transfer time $\mathbb{E}D_{\text{source}}$. Note that, due to the (empirically observed) insensitivity of the distribution of the number of active source nodes, the curves coincide for the different flow-size distributions; the deviations between analysis and simulations are less than a percent in general, for $c_F = 16$ less than two percent. The mean buffer workload is shown in Figure 7.6. Recall that, under Assumption 7.3.1, the analysis should be exact for both $\mathbb{E}D_{\text{source}}$ and $\mathbb{E}W_{\text{buffer}}$. At this point we see that, for high CoV's, the analysis overestimates $\mathbb{E}W_{\text{buffer}}$, as is illustrated by the numerical results.

Figure 7.7 shows the approximation of the delay $\mathbb{E}D^*_{\text{buffer}}$ of the last particle. It is observed that the resulting curves are close to the results of the fluid-model simulations. This supports the explanation in Section 7.3.3 that, although the analytical derivation of the expected conditional delay $\mathbb{E}Y_n(\tau)$ relies on exponentially distributed flow-sizes, the flow-size distribution hardly affects the outcome. As a result, the approximation $\mathbb{E}D^*_{\text{buffer}}$ gives a good approximation for general flow-sizes. Finally, Figure 7.8 presents the results for the mean overall flow transfer time $\mathbb{E}D_{\text{overall}}$.

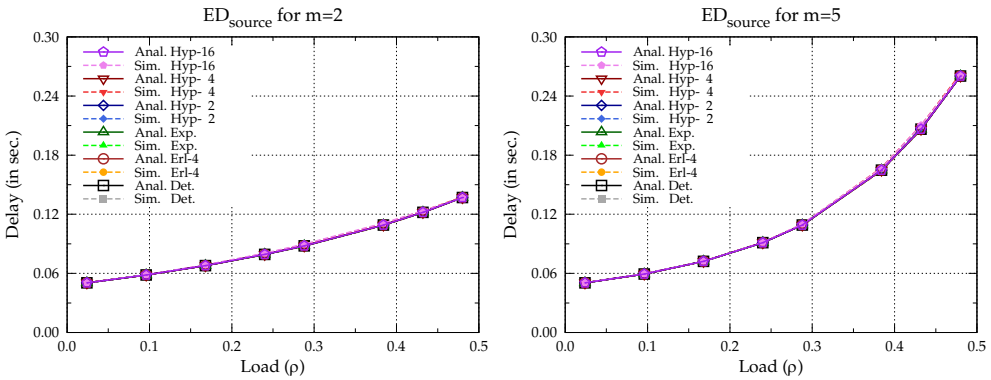

Figure 7.5: *Flow transfer time* $\mathbb{E}D_{\text{source}}$ *for general flow-size distributions. Left:* $m = 2$*. Right:* $m = 5$*.*

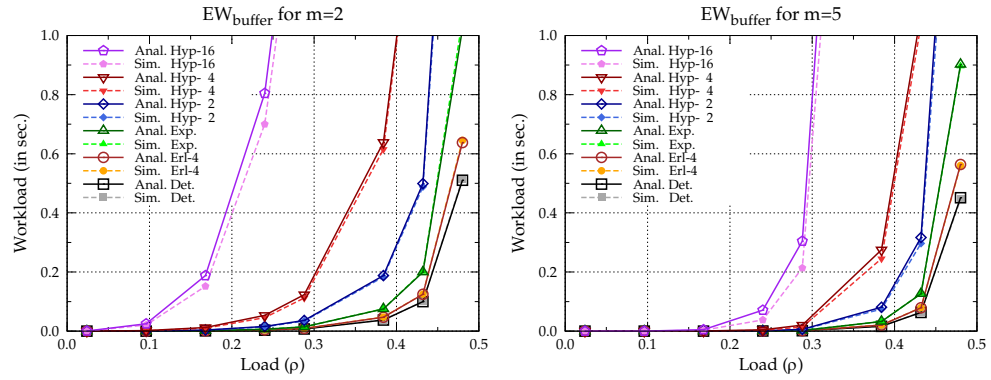

Figure 7.6: *Buffer workload* $\mathbb{E}W_{\text{buffer}}$ *for general flow-size distributions. Left:* $m = 2$*. Right:* $m = 5$*.*

As it is the sum of the exact $\mathbb{E}D_{\text{source}}$ and the (accurately) approximated $\mathbb{E}D^*_{\text{buffer}}$, it implies that the overall flow transfer time has a remarkably good fit.

7.5 Benefits of resource sharing

By varying the resource-sharing ratio m one could try to reduce the *overall* flow transfer time D_{overall}, which is the sum of the delays D_{source} and D^*_{buffer} (cf. Expression (7.14)). Obviously the optimization is a trade-off: by granting a larger share of the capacity to the bottleneck node D^*_{buffer} reduces while D_{source} increases. We investigate the impact of the resource-sharing ratio m on D_{overall}.

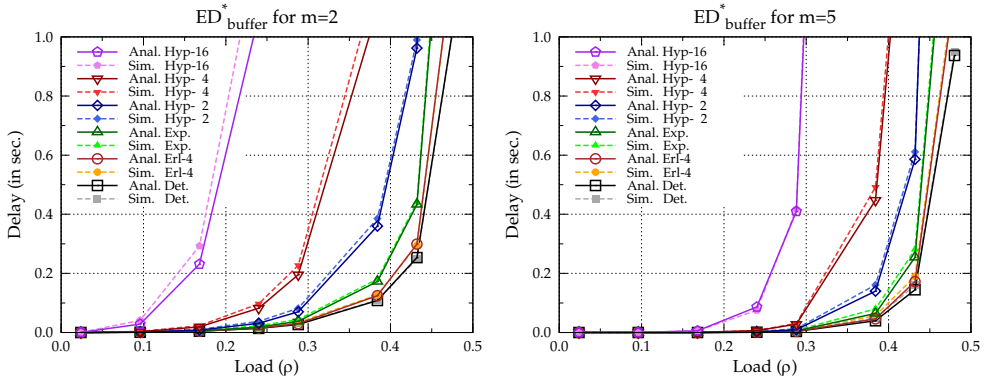

Figure 7.7: *Buffer delay* $\mathbb{E}D^*_{\text{buffer}}$ *for general flow-sizes distributions. Left:* $m = 2$. *Right:* $m = 5$.

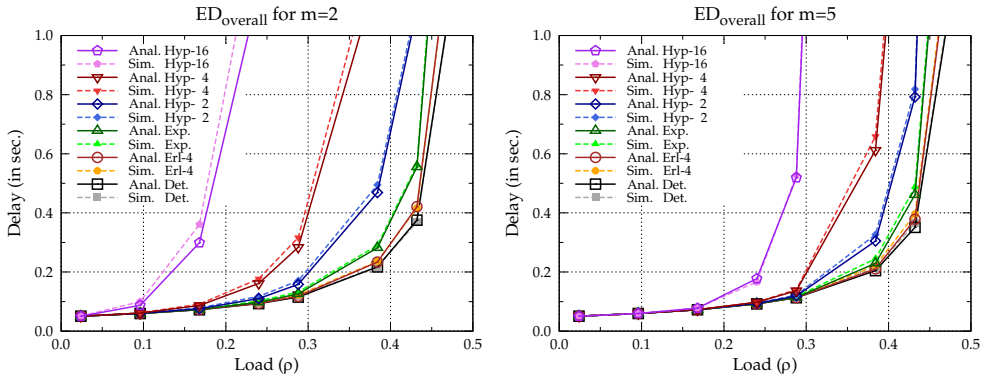

Figure 7.8: *Overall flow transfer time* $\mathbb{E}D_{\text{overall}}$ *for general flow-size distributions. Left:* $m = 2$. *Right:* $m = 5$.

Figure 7.9 illustrates the impact of the resource-sharing ratio m for fixed load $\rho \approx 0.43$ and different flow-size distributions. The left graph presents the mean overall flow transfer time $\mathbb{E}D_{\text{overall}}$, the right graph illustrates the trade-off between $\mathbb{E}D_{\text{source}}$ and $\mathbb{E}D^*_{\text{buffer}}$ for flow-size distributions $\text{C}_F \in \{1, 4, 16\}$. The left graph shows that the $\mathbb{E}D_{\text{overall}}$ is descending in m for all flow-size distributions. In addition, the right graph illustrates, once again, the insensitivity of $\mathbb{E}D_{\text{source}}$ to the flow-size distribution.

Interestingly, $\mathbb{E}D_{\text{overall}}$ coincides for all flow-size distributions at resource-sharing ratio $m = 20$, which was the n_{max} in our numerical experiments. Observe that the buffer remains empty when N_t cannot exceed m, and the relay node will continuously obtain share $\text{C}/2$. As a consequence, $D^*_{\text{buffer}} = 0$ and D_{overall} equals D_{source}. Further recall that $\mathbb{E}D_{\text{source}}$ is insensitive to the flow-size distribution (cf. Assumption

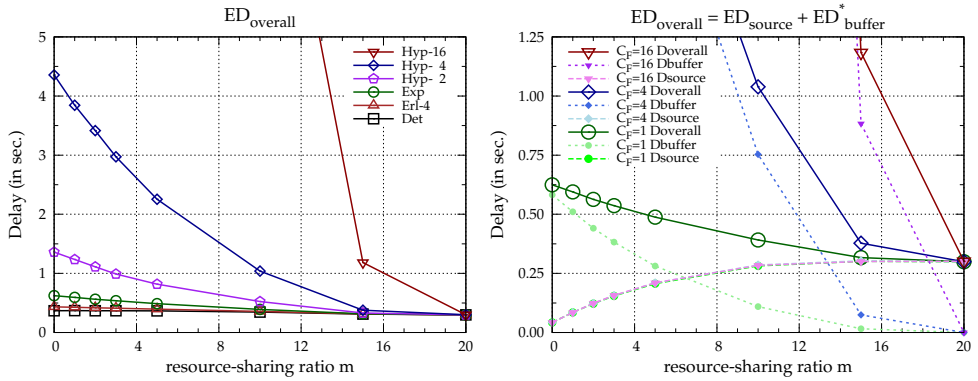

Figure 7.9: *Impact of the resource-sharing ratio m for $\rho = 0.43$. Left: $\mathbb{E}D_{\text{overall}}$ for various flow-size distributions. Right: trade-off between $\mathbb{E}D_{\text{source}}$ and $\mathbb{E}D^*_{\text{buffer}}$ for $\mathrm{C}_F \in \{1, 4, 16\}$.*

7.3.1), but, this result also follows from the observation that the source nodes behave as a Processor Sharing model with service capacity for which it is known that the mean sojourn times are insensitive to the flow-size distribution (apart from its mean). Hence, the mean overall flow transfer time $\mathbb{E}D_{\text{overall}}$ is given by $2f/(\mathrm{C}(1-2\rho))$, and is insensitive to the flow-size distribution.

Observe that, in order to optimize the mean overall flow transfer time, it is best to set the resource-sharing ratio m as large as possible: this results in the shortest mean transfer times.

7.6 Concluding remarks

In this chapter we presented a method to analyze the impact of the resource-sharing policy in a wireless ad-hoc network. We considered a setting where source nodes transmit flows to destinations via a common relay node. We obtained explicit expressions for the means of a number of performance metrics, such as the transfer time of a flow and the workload at the relay node.

The source-node behavior does not constitute a Markov chain (for $m > 1$), but, when assuming exponential flow-sizes, the joint distribution of the number of active source nodes and the workload can be analyzed using feedback fluid queues. We claim the remarkable fact that the obtained stationary distribution of the number of active source nodes is even valid for *generally* distributed flow-sizes, as we argue that the source-node behavior is actually insensitive to the flow-size distribution (apart from its mean); the latter claim is supported through extensive simulation experiments. Under this insensitivity claim we derived a number of expressions (some of

them being exact, others approximations) for the performance metrics under consideration. Again by simulation it was shown that these expressions are highly accurate over a broad set of parameter values.

Topics for further research. Further research includes *service-based* QoS *differentiation* where source nodes can obtain different shares of the capacity based on the priorities of their services; these priorities can even be dynamically adjusted based on the advertized buffer content per node.

Another interesting topic for future research relates to models with multiple hops. This introduces so-called 'hidden nodes', and as a result there is not a single resource C shared by all nodes, but multiple resources shared by non-disjoint subsets of nodes.

7.A Analysis of source-node behavior for exponential flow-sizes

This section presents a more comprehensive analysis of the results presented in Section 7.2.1, i.e., the stationary distribution of the number of active source nodes and the buffer workload.

We assume that flow-sizes are exponentially distributed with mean f. The source-node behavior of the model of Section 6.4 is not an autonomous process; the transition rates depend on both the number of active source nodes N_t and on whether the buffer workload W_t is positive or not. Hence, N_t does not constitute a Markov chain as it requires *feedback* of the workload W_t, e.g., see [88].

We analyze the source-node behavior analogously to [88]. First we analyze the fluid-queue *without feedback*, i.e., the system in which the relay node is always allotted a share $m\mathrm{C}/(n+m)$ of the capacity (when there are n source nodes transmitting); random variables (and other quantities) corresponding to the model without feedback are denoted with a bar ' $^-$ ' on top. Now, $\bar{N}_t$ constitutes a Markov chain and the joint distribution of $(\bar{N}_t, \bar{W}_t)$ is derived in terms of a system of linear differential equations as in the seminal studies on fluid queues [5, 77]. The result without feedback is extended to the case with feedback by the important observation that the behavior *during busy periods* of both models coincide. Finally, the joint distribution (N_t, W_t) of the model with feedback follows from rescaling the distribution $(\bar{N}_t, \bar{W}_t)$ of the model without feedback. In the following we make this procedure precise.

7.A.1 Fluid-queue without feedback

First, we consider the model of Section 6.4 without feedback, i.e., the relay node *always* obtains its entire resource-sharing ratio m. We introduce $\bar{W}_t$ as the buffer workload at time t for a system without feedback. Consequently, $\bar{N}_t$ does not depend on $\bar{W}_t$, and therefore $\bar{N}_t$ constitutes a Markov chain with generator matrix $\bar{Q}$, given through

$$\bar{Q}(i,j) := \begin{cases} \lambda & \text{if } j = i+1, \\ i\mathrm{C}/((m+i)f) & \text{if } j = i-1, \\ 0 & \text{otherwise.} \end{cases} \tag{7.21}$$

where the diagonal elements are such that the rowsums are 0. For technical reasons we first assume that the resource-sharing ratio m is non-integer; the case of integer values of m is explained at the end of this section.

Let $\bar{R}$ be a diagonal matrix where $\bar{R}_n$ is the *net* input rate into the relay node, i.e., if n sources are active then

$$\bar{R}_n = \frac{n-m}{n+m}\mathrm{C}.$$

Denote by D (U) all states with negative (positive) drift, i.e., $\bar{R}_n <$ for all $n \in D$ and $\bar{R}_n > 0$, for all $n \in U$. Let $\bar{Q}_{DD}$, $\bar{Q}_{UU}$ be submatrices obtained by partitioning of $\bar{Q}$ according to the 'up states' and 'down states'.

We define the stationary distribution of $(\bar{N}_t, \bar{W}_t)$ as

$$F_n(x) := \lim_{t\to\infty} P(\bar{W}_t \le x; \bar{N}_t = n) = P(\bar{W} \le x; \bar{N} = n).$$

For the analysis we assume that a maximum n_{max} is imposed on the number of source nodes that may be simultaneously active; flows that are initiated if already n_{max} other source nodes are active are blocked. The buffer workload satisfies the Kolmogorov forward equations $\overrightarrow{F}'(x)\bar{R} = \overrightarrow{F}(x)\bar{Q}'$. The spectral expansion of the solution is given by

$$\overrightarrow{F}(x) = \sum_{j=0}^{n_{max}} a_j \overrightarrow{v}_j \exp(z_j x)$$

where $(z_j, \overrightarrow{v}_j)$ is an eigenvalue-eigenvector pair, i.e., a scalar and vector that solve $z_j \overrightarrow{v}_j \bar{R} = \overrightarrow{v}_j \bar{Q}$. Clearly, $\mathbb{P}(\bar{W} \le x) = \sum_n F_n(x)$. Further, let ω denote the stationary distribution of $\bar{N}_t$ without feedback, hence $\omega_n = F_n(\infty)$. For stability we assume that the average net input rate is negative, that is, $\sum_n \omega_n \bar{R}_n < 0$. As our generator matrix $\bar{Q}$ corresponds to a birth-death chain, all eigenvalues z_j are real [36].

Following Mitra [99], the number of negative eigenvalues n_+ in a stable system (i.e., $\sum_n \omega_n \bar{R}_n < 0$) is equal to the number of states with positive drift, i.e., $n_+ = n_{max} - \lceil m \rceil$; exactly one eigenvalue has value zero and the remaining eigenvalues

are positive. In the remainder we label the eigenvalues z_j such that $z_j < 0$ for $j \in \{0, \ldots, n_+ - 1\}$, $z_{n_+} = 0$, and $z_j > 0$ for $j \in \{n_+ + 1, \ldots, n_{max}\}$.

The coefficients a_j are calculated as follows. When $z_j > 0$, then $a_j = 0$ as the distribution function should be in $[0, 1]$. The other coefficients a_j are computed from $F_i(0) = 0$ for all up states i. Further observe that

$$\omega = \frac{\overrightarrow{v}_{n_+}}{\langle \overrightarrow{v}_{n_+}, \overrightarrow{1} \rangle},$$

where $\langle \cdot, \cdot \rangle$ denotes the (standard) inner product. For computationally efficient numerical schemes, see e.g. [81].

Elwalid and Mitra [39] presented explicit expressions for a number of quantities related to the busy and idle periods of the workload at the relay node. A busy (idle) period is the period during which the workload at the relay node is positive (zero). A busy period starts when the system is empty and N_t becomes larger than m by a flow initiation. A busy period ends when the buffer becomes empty, and then N_t is in a state in D.

Denote by $\overrightarrow{P}$ the distribution of $\bar{N}$ at the end of the busy period. Then, due to Expression (5.9) of [39],

$$\overrightarrow{P} = \frac{1}{\left\langle \overrightarrow{F_D}(0)\bar{Q}_{DD}, \overrightarrow{1} \right\rangle} \overrightarrow{F_D}(0)\bar{Q}_{DD}.$$

Note that the (i, j) entry of $-(\bar{Q}_{DD})^{-1}$ is the mean time spent in state j by N_t, if the process started in state i, before leaving the set D, see e.g. [61]. Then, the mean idle period $\mathbb{E}\bar{I}$ is given by

$$\mathbb{E}\bar{I} = \left\langle -\overrightarrow{P}(\bar{Q}_{DD})^{-1}, \overrightarrow{1} \right\rangle. \tag{7.22}$$

The mean busy period $\mathbb{E}B$ is obtained from $\sum_{i \in D} F_i(0) = \mathbb{E}\bar{I}/(\mathbb{E}\bar{B} + \mathbb{E}\bar{I})$:

$$\mathbb{E}\bar{B} = \mathbb{E}\bar{I} \cdot \frac{1 - \sum_{n \in D} F_n(0)}{\sum_{n \in D} F_n(0)}.$$

REMARK 7.A.1 (INTEGER-VALUED RESOURCE-SHARING RATIO m). *In case of an integer-valued resource-sharing ratio m, state m has zero drift, or, more precisely, $R_m = 0$. Therefore R is singular. In this situation the Kolmogorov forward equations consist of n_{max} differential equations and 1 supplementary algebraic equation. This algebraic equation results from the state with drift zero and hinders obtaining the eigenvalue-eigenvector pairs. Observe that the state with zero drift does not influence the workload distribution and is basically redundant. In Appendix A.1 of [99], Mitra proposes how to reduce the dimension of the system of differential equations by 1 to (obtain a proper system), by eliminating the redundant algebraic equation. Further, it is proven that the eigenvalues of the reduced form coincide with the original form and it is shown how the eigenvectors of the original system are obtained from the reduced system.* ♢

7.A.2 Fluid queue with feedback

Here we consider the model of Section 6.4 which includes feedback of the workload W_t at the relay node. As remarked before, the number of source nodes no longer constitutes a Markov chain. We are interested in the stationary buffer workload denoted by $G_n(x) := \mathbb{P}(W \leq x; N = n)$. Let random variable B (I, respectively) denote a busy (idle) period in the system with feedback, and $\overrightarrow{P}$ the distribution of N at the end of a busy period.

Note that the distributions $\overrightarrow{P}$ and $\overrightarrow{\bar{P}}$ are identical, and also the busy periods B and $\bar{B}$ have the same distribution. Hence,

$$\mathbb{P}(W \leq x; N = n | W > 0) = \mathbb{P}(\bar{W} \leq x; \bar{N} = n | \bar{W} > 0).$$

As a consequence, the stationary distribution $G_n(x)$ of the buffer workload and number of source nodes is

$$\begin{aligned} G_n(x) &= \mathbb{P}(\bar{W} \leq x; N = n | \bar{W} > 0)\mathbb{P}(W > 0) + \mathbb{P}(W = 0; N = n) \\ &= \frac{F_n(x) - \sum_k F_k(0)}{1 - \sum_k F_k(0)} \mathbb{P}(W > 0) + \mathbb{P}(W = 0; N = n). \end{aligned} \tag{7.23}$$

To complete (7.1) we require expressions for $\mathbb{P}(W > 0)$ and $\mathbb{P}(W = 0; N = n)$. Here $\mathbb{P}(W > 0)$ follows from

$$\mathbb{P}(W > 0) = \frac{\mathbb{E}B}{\mathbb{E}I + \mathbb{E}B}.$$

Also

$$\mathbb{E}I = \left\langle -\overrightarrow{P}(Q_{DD})^{-1}, \overrightarrow{1} \right\rangle,$$

cf. Equation (7.22), where Q_{DD} is the square generator matrix of dimension n_+ for the states with downwards drift in case $W_t = 0$, i.e.,

$$Q_{DD}(i,j) := \begin{cases} \lambda & \text{if } j = i+1, \\ \mathrm{C}/2 & \text{if } j = i-1, \\ 0 & \text{otherwise.} \end{cases}$$

Further, $\mathbb{P}(W = 0; N = n)$ corresponds to the n-th element of $-\overrightarrow{P}(Q_{DD})^{-1}$. Finally, the stationary distribution π of the number of active source nodes N_t follows from Equation (7.23), as $\pi_n = G_n(\infty)$.

7.B Proof of Proposition 7.2.6

This proof is essentially along the lines of [19]. Recall that $Y_n(\tau)$ denotes the conditional buffer delay, i.e., the time required by the relay node with resource-sharing

ratio m to serve an amount of work τ if n other jobs are present upon the start. By $\phi_n(s)$ we denote the Laplace transform of $\mathbb{E}Y_n(\tau)$, i.e.,

$$\phi_n(s) = \int_0^\infty e^{-s\tau}\mathbb{E}Y_n(\tau)\mathrm{d}\tau.$$

We obtain an expression for $\phi_n(s)$ by conditioning on the next possible event, namely: flow arrival, source departure of one of the n source nodes, or the relay node completes the service of amount τ. First we define $\nu_n := \lambda_n + \mu_n$, with

$$\lambda_n := \lambda\, 1\{n < n_{max}\}; \quad \mu_n := \frac{\mathsf{C}}{f}\frac{n}{m+n}.$$

Then we obtain the following expression:

$$\begin{aligned}\phi_n(s) \;=\; & \int_0^\infty e^{-s\tau}\int_{\tau\cdot\frac{m+n}{m}}^\infty \nu_n e^{-\nu_n t}\tau\cdot\frac{m+n}{m}\mathrm{d}t\mathrm{d}\tau \;+ \\ & \int_0^\infty e^{-s\tau}\int_0^{\tau\cdot\frac{m+n}{m}} \nu_n e^{-\nu_n t}\left\{t + \frac{\lambda_n}{\nu_n}\mathbb{E}Y_{n+1}\left(\tau - \frac{m}{m+n}t\right)\right. \\ & \qquad \left. + \frac{\mu_n}{\nu_n}\mathbb{E}Y_{n-1}\left(\tau - \frac{m}{m+n}t\right)\right\}\mathrm{d}t\mathrm{d}\tau.\end{aligned}$$

After elementary algebra, this is rewritten as the following system of linear equations:

$$\frac{1}{s} = -\lambda_n\phi_{n+1}(s) + \left(\frac{m}{m+n}s + \nu_n\right)\phi_n(s) - \mu_n\phi_{n-1}(s), \tag{7.24}$$

or, in matrix notation, $\overrightarrow{1} = s\cdot M(s)\overrightarrow{\phi(s)}$ where $M(s) := -\bar{Q} + sR$ with $\bar{Q}$ as in (7.21) and

$$R := \mathrm{diag}\left\{1, \frac{m}{m+1}, \frac{m}{m+2}, \cdots, \frac{m}{m+n_{max}}\right\}.$$

It is readily verified that the equation $\det M(s) = 0$ coincides with $\det(R^{-1}Q - sI) = 0$. In other words: the roots of $\det M(s) = 0$ are the eigenvalues of $R^{-1}Q$. As Q is singular, one of the eigenvalues of $R^{-1}Q$ is 0, say s_0. Further, a straightforward application of 'Geršgorin' yields that all eigenvalues $s_0, \ldots, s_{n_{max}}$ are real, non-positive and unique.

The Laplace transform ϕ_n can be solved from the linear system by applying Cramer's rule to $s\vec{\phi}(s) = (M(s))^{-1}\vec{1}$, i.e.,

$$s\phi_n(s) = \frac{\det M_{-n}(s)}{\det M(s)}. \tag{7.25}$$

where $M_{-n}(s)$ is defined as $M(s)$ with the n-th column replaced by $\vec{1}$. The denominator of the right-hand side of (7.25) is a polynomial of degree $n_{max}+1$ in s. The

above considerations entail

$$s\phi_n(s) = \frac{A_{(n,0)}}{s} + \sum_{j=1}^{n_{max}} \frac{A_{(n,j)}}{s - s_j}$$

where the constants $A_{(n,j)}$ follow from the partial-fraction expansion at $s_0, \ldots, s_{n_{max}}$. Then,

$$\phi_n(s) = \frac{A_{(n,0)}}{s^2} + \sum_{j=1}^{n_{max}} \frac{A_{(n,j)}}{s_j} \frac{1}{s - s_j} - \sum_{j=1}^{n_{max}} \frac{A_{(n,j)}}{s_j} \frac{1}{s}.$$

Finally, inverting the individual parts gives the desired result:

$$\mathbb{E}Y_n(\tau) = A_{(n,0)}\tau + \sum_{j=1}^{n_{max}} \frac{A_{(n,j)}}{s_j} e^{s_j \tau} - \sum_{j=1}^{n_{max}} \frac{A_{(n,j)}}{s_j}.$$

Chapter 8

Validation of the fluid-modeling approach

8.1 Introduction

In this chapter we validate the fluid-modeling approach that was introduced in Chapter 6, and analyzed in Chapter 7. The fluid model (cf. Section 6.4) is validated by simulations of the wireless ad-hoc network that include all the details of the widely used IEEE 802.11 MAC-protocol.

8.1.1 Contribution

The validation of the fluid model is considered separately for the 'standard' fluid model where the relay node and source nodes equally share the capacity (i.e., $m = 1$), and the general case where the relay node can obtain a different share of the capacity (i.e., $m \in [0, \infty)$).

For the standard fluid model with $m = 1$ we validate that it captures the behavior of network nodes that operate according to the IEEE 802.11 MAC-protocol DCF. This model allows for more explicit expressions for the performance metrics than is the case with general resource-sharing ratio $m \in [0, \infty)$, see the analysis in Chapter 7 and the excerpt in Section 8.3.2 (see also [14, 115]). These analytical expressions are numerically evaluated and compared with the results from the wireless ad-hoc network simulator with a detailed implementation of the IEEE 802.11 protocols.

For the case with $m \in [0, \infty)$ we validate that the fluid model captures the resource-sharing behavior of the IEEE 802.11e EDCA. The EDCA provides four QoS-differentiation parameters and we provide a mapping of their values onto the fluid model's parameters, viz. m and C, that model the alternative resource-sharing strategies that can be enforced in real systems. We have to slightly adapt the fluid model of Section 6.4 as the medium capacity C_n and resource-sharing ratio m_n turn out to depend on the number of active source nodes n, albeit only moderately. We discuss a mapping between the parameter settings of IEEE 802.11 and the fluid-flow model, and validate the fluid-flow model and the parameter mapping by means of detailed system simulations.

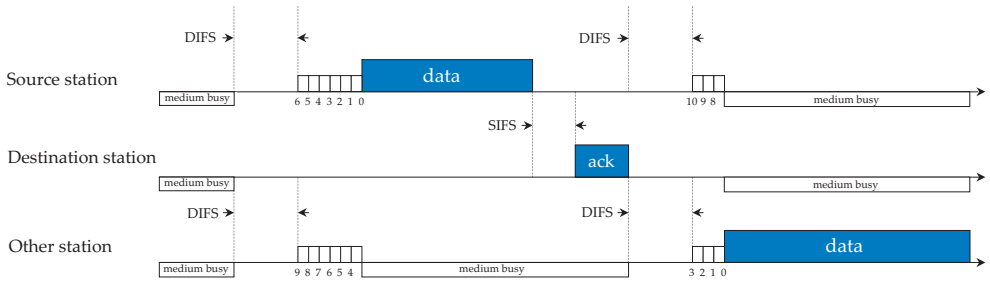

Figure 8.1: *Distributed Coordination Function (*DCF*).*

8.1.2 Outline

This chapter is organized as follows. The IEEE 802.11 DCF and EDCA protocols are explained in Section 8.2. In Section 8.3 we validate the fluid model (of Section 6.4) for the IEEE 802.11 DCF; as mentioned before, this relates to resource-sharing ratio $m = 1$. First we obtain the service capacity used in the fluid model, and we present an excerpt of the analytical results and validate the fluid model by comparing it with ad-hoc network simulations. Section 8.4 validates the fluid model for an IEEE 802.11e EDCA relay node; first we obtain the input parameters of the fluid model and second we validate this model. Finally, Section 8.5 presents concluding remarks and discusses some directions for further research.

8.2 IEEE 802.11 Wireless LAN

In this section we briefly explain the IEEE 802.11 DCF and its enhancements as specified in IEEE 802.11e EDCA in order to support QoS differentiation. The DCF and EDCA were initially standardized in respectively [62] and [65]. In 2007 an update [63] was provided of the original standard [62] which incorporates several amendments, most notably, 802.11a, 802.11b [64], 802.11e [65], and 802.11g.

8.2.1 IEEE 802.11b Distributed Coordination Function (DCF)

Figure 8.1 illustrates the principle of the BASIC access scheme. When a station wants to transmit a data packet, it first senses the medium to determine whether or not the channel is already in use by another station (*physical carrier sensing*). If the channel is sensed idle for a contiguous period of time called DIFS (Distributed InterFrame Space), the considered station transmits its packet. In case the channel is sensed busy, the station must wait until it becomes idle again and subsequently remains idle for

a DIFS period, after which it has to wait another randomly sampled number of time slots before it is permitted to transmit its data packet. This *backoff* period is sampled from a discrete uniform distribution on $\{0, \cdots, \mathrm{CW}_r - 1\}$, with CW_r the contention window after r failed packet transfer attempts (CW_0 is the initial contention window size). The backoff counter is decremented from its initially sampled value until the packet is transferred when the counter reaches zero, unless it is temporarily 'frozen' in case the channel is sensed busy before the backoff counter reaches zero. In the latter case the station continues decrementing its backoff counter once the medium is sensed idle for at least one DIFS period. It is noted that the idea behind the random backoff procedure is to reduce the probability of *collisions*, which occur either when the backoff counters of multiple stations reach zero simultaneously, or in case a so-called hidden station fails to freeze its backoff counter when it cannot sense another station's transmission. In a collision only the strongest signal among multiple concurrent transmissions has a chance of successful *capture* by the intended receiver.

If the destination station successfully captures the transmitted data packet, it responds by sending an ACK (ACKnowledgement message) after a SIFS (Short Inter-Frame Space) time period. A SIFS is shorter than a DIFS in order to give the ACK preference over data packet transmissions by other stations, while it is sufficiently long to allow the stations involved in the considered transfer to switch between transmission and reception mode. If the source station fails to receive the ACK within a predefined time-out period, the contention window size is doubled unless it has reached its maximum window size, upon which the data packet transfer is reattempted. The total number of transmission attempts is limited to $r_{\max}$. Once the data packet is successfully transferred, the contention window size is reset to CW_0 and the entire procedure is repeated to transfer subsequent data packets. If an unfortunate data packet is still not successfully transferred after $r_{\max}$ retransmissions, the MAC layer gives up. It is then up to higher-layer protocols (e.g. UDP (User Datagram Protocol) or TCP) whether the packet is discarded or once again offered to the MAC layer for transmission.

8.2.2 IEEE 802.11e Enhanced Distributed Channel Access (EDCA)

IEEE 802.11e specifies the Enhanced Distributed Coordination Access (EDCA) as the distributed contention mechanism that can provide service differentiation. Whereas an IEEE 802.11b station has only one queue for all traffic, an IEEE 802.11e station (QoS STA) has multiple queues, so-called Access Categories (ACs), and traffic is mapped into one of the ACs according to its service requirements. Each AC contends for the medium using the CSMA/CA mechanism described in Section 8.2.1 using it own set of EDCA parameters values. These EDCA parameters are $\mathrm{CW}_{\min}$, $\mathrm{CW}_{\max}$, AIFS, and the $\mathrm{TXOP}_{\mathrm{limit}}$.

The parameters $\mathrm{CW}_{\min}$ and $\mathrm{CW}_{\max}$ have the same functionality as in the DCF. The

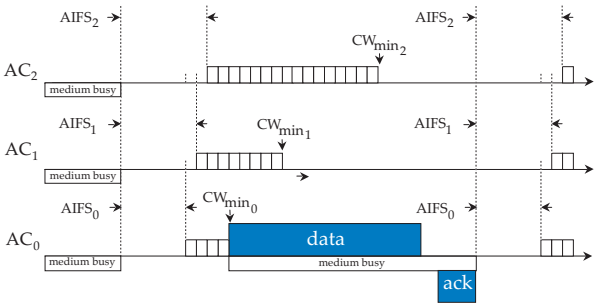

Figure 8.2: QoS STA *with three* Access Categories.

parameter AIFS (Arbitration InterFrame Space) differentiates the time that each AC has to wait before it is allowed to decrement its backoff counter after the medium has become free. In the DCF each station has to wait for a DIFS period while the duration of an AIFS is a SIFS period extended by a discrete number of time slots AIFSN, so AIFS $=$ SIFS $+$ AIFSN $\times$ timeslot (where AIFSN ≥ 2 for QoS STAs and AIFSN ≥ 1 for Quality APs). The $\text{TXOP}_{\text{limit}}$ (Transmission OPportunity (TXOP) limit) is the duration of time that an AC may send after it has won the contention, so it may send multiple packets as long as the last packet is completely transmitted before the $\text{TXOP}_{\text{limit}}$ has passed. Figure 8.2 illustrates the parameters AIFS and CW_{min}.

Obviously, the backoff counters of multiple ACs of one station can reach zero at the same moment, which is called a *virtual collision*. Each QoS STA has an internal scheduler that handles a virtual collision. The AC with the highest priority is given the TXOP and may actually initiate a transmission. The ACs of lower priority are treated as if they experienced a collision, so they have to double their contention window CW and start a new contention for the medium.

8.3 Validation of the fluid model for IEEE 802.11 DCF resource sharing

This section illustrates that the fluid model accurately describes the behavior of source and relay nodes when all are equipped with the IEEE 802.11b DCF. First, we describe how to obtain accurate parameter values for the fluid model. Next, we present a summary of the analytical results that were presented in [14]. Finally, we validate the fluid model by comparing the analytical results with simulation results of the ad-hoc network simulator.

8.3.1 Mapping of IEEE 802.11 DCF parameters

The DCF equally shares the capacity among the contending nodes, hence the resource-sharing ratio m equals 1. Therefore we are left to determine the capacity C.

The capacity C can be obtained by the model of Bianchi [16]. Recall that if at least one source node is active, then the relay node is active as well. As we assume that an active node is continuously contending for a TXOP, the ad-hoc network scenario satisfies the framework of Bianchi's model and C_n corresponds to Bianchi's saturated throughput for $n+1$ nodes. Note that we are interested in the saturated throughput at flow level (i.e., excluding all overhead), although the overheads of all OSI-layers should be taken into account in the calculations.

We consider an IEEE 802.11b wireless ad-hoc network, which entails that all nodes can transmit at a *gross* bit rate of 11 Mbit/s. We assume the use of RTS/CTS-access, then the resulting *net* bit rate is approximately 5.0 Mbit/s independent of the number of active source nodes, cf. the curve 802.11b in the left graph of Figure 8.5.

8.3.2 Analysis of the 'standard' fluid model

The analysis in Chapter 7, obviously, covers the standard case of resource-sharing ratio $m = 1$, but for this special case more explicit (closed-form) expressions can be obtained; see the analysis in paper [14] of which an excerpt is presented next.

The stationary behavior of the active source nodes was already given by Expression (7.17) with $m = 1$, and it is insensitive to the flow-size distribution (cf. Remark 7.3.2). Little's law on the mean number of active source nodes yields

$$\mathbb{E}D_{\text{source}} = \frac{\mathbb{E}N}{\lambda} = 2\frac{f/\mathrm{C}}{1-\rho}. \tag{8.1}$$

The buffer delay D^*_{buffer} is derived from the buffer workload W^*_{buffer} seen by the last particle, which is the sum of the workload W_{buffer} upon flow arrival and the buffer increase ΔW_{buffer} during D_{source}. The amount of work in the buffer at the relay node is the difference between the total amount of work in the system W_{total} (both at the sources and the buffer) and the work remaining at the source W_{sources}, hence

$$\mathbb{E}W_{\text{buffer}} = \mathbb{E}W_{\text{total}} - \mathbb{E}W_{\text{sources}} = \frac{2\rho^2 f_2}{f\mathrm{C}}\frac{1}{(1-2\rho)(1-\rho)}.$$

The expected workload increase during a flow transfer D_{source} is given by

$$\mathbb{E}\Delta W_{\text{buffer}} = \mathbb{E}D_{\text{source}} - 2f/\mathrm{C} = \frac{2f\rho/\mathrm{C}}{1-\rho}.$$

Therefore,

$$\mathbb{E}W^*_{\text{buffer}} = \mathbb{E}W_{\text{buffer}} + \mathbb{E}\Delta W_{\text{buffer}} = \frac{2\rho^2 f_2/f\mathrm{C}}{(1-2\rho)(1-\rho)} + \frac{2f\rho/\mathrm{C}}{1-\rho}.$$

Observe that the buffer delay of the last particle D^*_{buffer} is the time required to serve the amount of work W^*_{buffer} that is present at the buffer upon arrival of the last particle. As the capacity sharing between source nodes and relay node is purely processor sharing, we approximate the buffer delay of the last particle by

$$\mathbb{E}D^*_{\text{buffer}} \approx \sum_{n=0}^{\infty} \pi_n \mathbb{E}X_n(\mathbb{E}W^*_{\text{buffer}}), \tag{8.2}$$

where $\mathbb{E}X_n(\tau)$, the so-called *response time* for jobs in an M/M/1-PS queue presented by Coffman, Muntz, and Trotter (see [29]), is given by

$$\mathbb{E}X_n(\tau) = \tau + \frac{\rho\tau}{1-\rho} + (n(1-\rho)-\rho)\,(f/\mathrm{C})\frac{1-\exp(-(1-\rho)\tau\mathrm{C}/f)}{(1-\rho)^2}. \tag{8.3}$$

For further details about Approximation (8.2) we refer to Approximation 7.2.7 and also to [14].

8.3.3 Numerical results for the fluid modeling of IEEE 802.11 DCF

This section numerically validates the fluid model as an accurate description of the ad-hoc network scenario of Section 6.4.1. The validation using the validation scenario of Section 6.6 and consists of a comparison of:

i) detailed simulations of ad-hoc network,

ii) simulation of the fluid-flow model of Section 6.4, and

iii) the analytical results of Section 8.3.2.

The graphs of Figure 8.3 present the mean buffer workload $\mathbb{E}W_{\text{buffer}}$ at the relay node for an arbitrary packet (left) and last packet of a flow $\mathbb{E}W^*_{\text{buffer}}$ (right). The graphs present three curves: ad-hoc network scenario simulations, fluid-model simulations, and fluid-model analysis. In both graphs it can be seen that the three curves more or less coincide. Only for loads close to the saturation load, the results are less accurate due to the imprecision of the estimated capacity C. Overall the curves indicates that the fluid model accurately describes the ad-hoc network scenario and that the analytical results of Section 8.3.2 are also very good. Further, it can be observed that the buffer occupancy seen by the last particle is only slightly higher than the buffer occupancy upon flow arrival; the relatively short flow transfer time and low number of active source nodes result in a minor increase of the buffer during the flow transfer time.

Figure 8.4 presents the results for the mean buffer delay $\mathbb{E}D^*_{\text{buffer}}$ of the last packet (left) and the mean overall flow transfer time $\mathbb{E}D_{\text{overall}}$ (right). Note that the analytically obtained buffer delay of the last particle in the left graph is based on an approximation (cf. Expression (8.2)). The fluid model captures the behavior of a IEEE

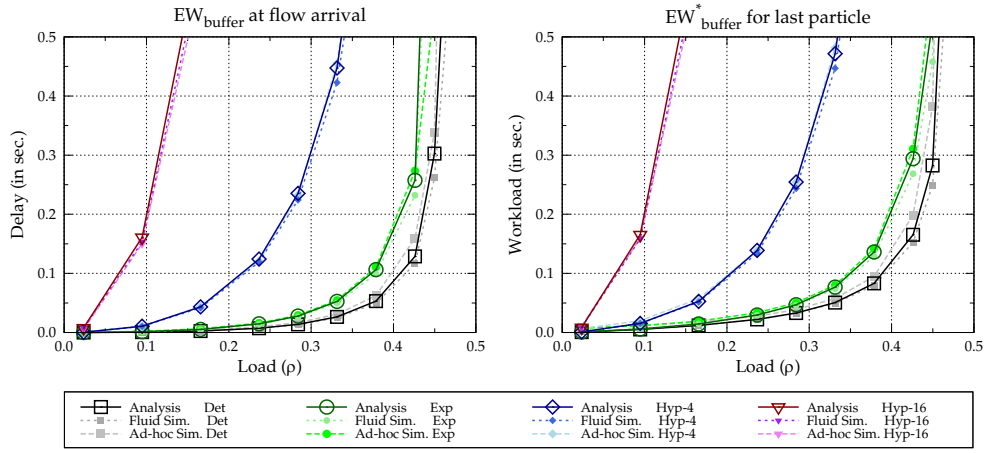

Figure 8.3: *Mean workload. Left: at flow arrival ($\mathbb{E}W_{\text{buffer}}$). Right: last packet ($\mathbb{E}W^*_{\text{buffer}}$).*

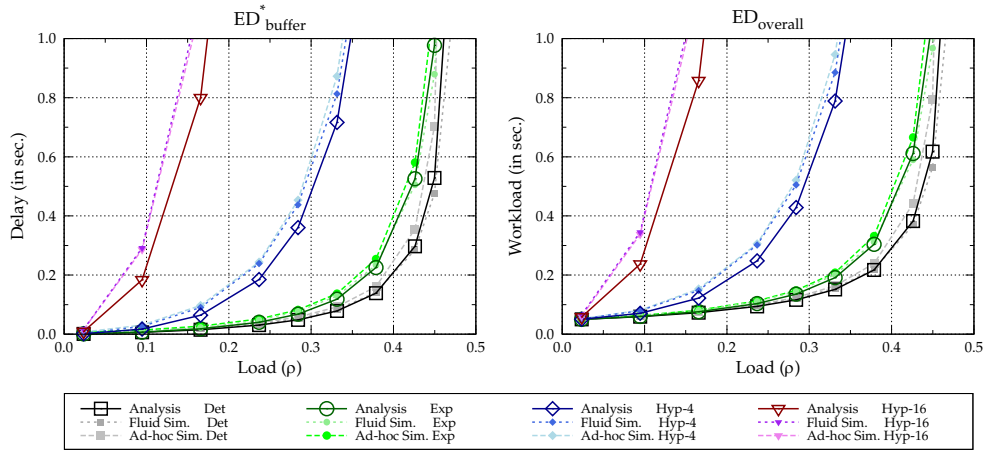

Figure 8.4: *Left: Mean buffer delay of the last packet/particle ($\mathbb{E}D^*_{\text{buffer}}$). Right: Mean overall flow transfer time ($\mathbb{E}D_{\text{overall}}$).*

802.11b nodes very well as the model reflects both the impact of the load and flow-size distribution, except for high loads in which case the results are less accurate. By comparing the graphs it is seen that the mean overall transfer time is almost completely determined by the buffer delay at the relay node.

8.4 Validation of the fluid model for IEEE 802.11e EDCA resource sharing

This section shows that the fluid model accurately captures the behavior of IEEE 802.11e EDCA relay node, and in particular, that it captures the resource sharing between source and relay nodes.

In Section 8.4.1 we introduce a state-dependent variant of the fluid-model of Section 6.4, i.e., the capacity C_n and resource-sharing ratio m_n now dependent on the number of active source nodes n. In Section 8.4.2 the IEEE 802.11e parameters are mapped onto these fluid-model parameters. In Section 8.4.3 we validate our modeling approach by detailed ad-hoc network simulations.

8.4.1 Resource sharing between relay and source nodes

The resource-sharing ratio m_n between the share of the relay node and a source node, and the common capacity C_n depend on the number of active source nodes n, where $m_n \geq 0$ and $\mathrm{C}_n > 0$. Further, the model operates similarly as the fluid model defined in Section 6.4.

The relay node *may* obtain capacity $m_n\mathrm{C}_n/(n+m_n)$, however, only if it can actually use the entire share, viz. the input rate exceeds the output rate (i.e., $N_t \geq m_{N_t}$) or if the buffer is backlogged (i.e., $W_t > 0$). Otherwise the input and output rates are coupled, resulting in capacity share of $\mathrm{C}_{N_t}/2$ for the relay node. The capacity share obtained by the relay node is summarized as follows:

$$\mathrm{C}_{N_t} \times \begin{cases} m_{N_t}/(N_t+m_{N_t}), & \{W_t > 0\} \vee \{N_t \geq m_{N_t}\}, \\ 1/2, & \{W_t = 0\} \wedge \{N_t < m_{N_t}\}, \\ 1, & \{N_t = 0\}. \end{cases}$$

The source nodes always equally share the remaining capacity. The stability condition of this system is $\sum_n \omega_n \mathrm{C}_n(n-m_n)/(n+m_n) < 0$, where ω_n is the steady-state probability of having n active source nodes in the system.

8.4.2 Mapping of IEEE 802.11e parameters

IEEE 802.11e EDCA provides four 'differentiating parameters' (cf. Section 8.2.2), namely $\mathrm{CW}_{\min}$, $\mathrm{CW}_{\max}$, AIFS, and $\mathrm{TXOP}_{\mathrm{limit}}$. Unfortunately, the mapping of the EDCA parameters onto the fluid-model parameters C_n and m_n is not self-evident, see e.g. [113].

In case the relay node is *saturated*, i.e., $\{W_t > 0\} \vee \{N_t \geq m_{N_t}\}$, the fluid model parameters C_n and m_n can be estimated from an extension of the model of Bianchi [16] to two classes with different settings for the differentiating parameters, see e.g. [136]. In particular, the resource sharing in case of n active source nodes can be obtained

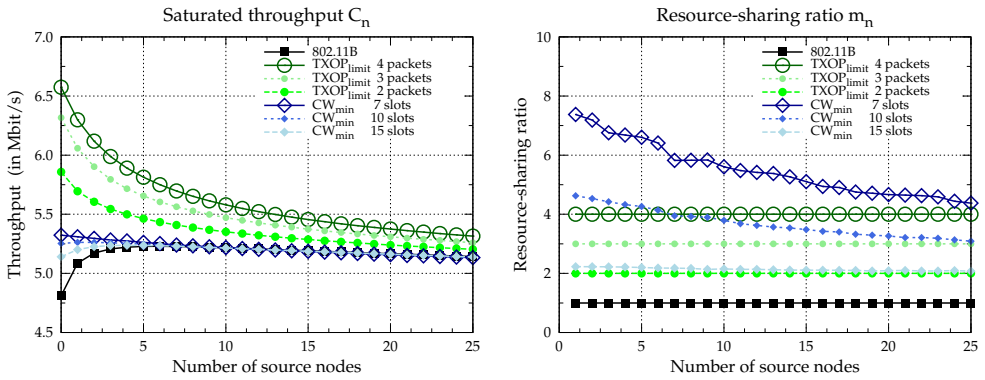

Figure 8.5: *Varying number of source nodes and a relay node with a varying parameter setting. Left: Overall saturated throughput. Right: Resource sharing ratio.*

from the model of [136] with n nodes in one class and a single node (representing our relay node) in the other class. The parameters C_n and m_n for $\{W_t > 0\} \vee \{N_t \geq m_{N_t}\}$ are estimated by respectively the aggregate throughput and the ratio of the per node throughput in the two classes.

In case of a *non-saturated* relay node, i.e., $\{W_t = 0\} \wedge \{N_t < m_{N_t}\}$, the above-mentioned approach would overestimate both the non-saturated capacity C'_n and resource sharing ratio m'_n due to the assumption that all nodes are saturated. Observe that the non-saturated resource sharing ratio m'_n equals n as this resource-sharing ratio couples the input rate into the relay node to the output rate. Next, the non-saturated capacity C'_n is estimated as follows: we consider the same differentiating parameter and its value is set such that it provides for the desired resource-sharing ratio m'_n in the model of [136], the corresponding capacity C_n is used as an estimation of C'_n. For example, when we use differentiating parameter $\mathrm{TXOP}_{\mathrm{limit}} = 3$ and $\{N_t = 2\} \wedge \{W_t = 0\}$, then the resource sharing ratio m'_2 equals 2; therefore C'_2 is estimated by C_2 which is the saturated capacity for $\mathrm{TXOP}_{\mathrm{limit}} = 2$.

Figure 8.5 shows the saturated throughput (left graph) and the resource sharing ratio (right graph) as a function of the number of source nodes. First, we vary the value of $\mathrm{CW}_{\mathrm{min}}$ at the relay node; all other parameters of the relay node and all parameters of the source nodes are set according to the IEEE 802.11b standard. Then we do a similar experiment in which we vary the $\mathrm{TXOP}_{\mathrm{limit}}$ of the relay node, while all other parameters are set according to IEEE 802.11b. The left graph illustrates that higher overall throughputs are obtained by an IEEE 802.11e relay node, especially for parameter $\mathrm{TXOP}_{\mathrm{limit}}$. In the right graph the throughput ratios for parameter $\mathrm{TXOP}_{\mathrm{limit}}$ are trivial, the ratios for parameter $\mathrm{CW}_{\mathrm{min}}$ are examples of non-trivial resource sharing as, intuitively, the throughput ratio for $\mathrm{CW}_{\mathrm{min}}$ is the inverse of the $\mathrm{CW}_{\mathrm{min}}$ parameter-setting ratio. For example, when $\mathrm{CW}_{\mathrm{min}}$ of the relay node

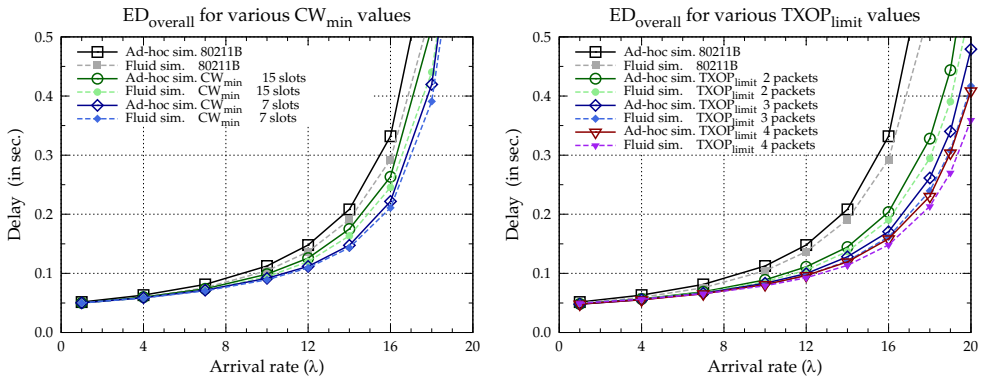

Figure 8.6: *System simulations and fluid-model simulations of the overall flow transfer time. Relay node with varying parameter setting. Left:* $\mathrm{CW_{min}}$. *Right:* $\mathrm{TXOP_{limit}}$.

is set to 7, then the expected ratio is $31/7 \approx 4$, but the realized ratio is larger than 6 for a small number of active source nodes, see e.g. [113] for an explanation on this phenomenon.

8.4.3 Numerical results for the fluid modeling of IEEE 802.11e EDCA

In the present section the fluid model for an IEEE 802.11e relay node is numerically validated by ad-hoc network simulations. The validation scenario were defined in Section 6.6 and the experiments coincide with those of the previous section: one of the parameters $\mathrm{CW_{min}}$ or $\mathrm{TXOP_{limit}}$ of the relay node is varied, all other parameters of the relay node and all parameters of the source nodes are set according to the IEEE 802.11b standard.

For fluid-flow simulations, of which results are presented in Figure 8.6, we use C_n and m_n estimated by the ad-hoc network simulator. The reason is that the fluid model is very sensitive for the used capacity if the offered load is close to the available capacity, cf. Section 8.3.3. Bianchi's model is proven to be accurate and the differences between results of the model and the ad-hoc network scenario simulations are small (just a few percent), but this approach ensures that deviations between the fluid model and the ad-hoc network scenario are solely due to fluid-modeling assumptions.

Figure 8.6 displays a comparison of ad-hoc network scenario simulations and fluid-model simulations. The fluid model simulations slightly underestimate the ad-hoc network scenario simulation results, but the behavior of the differentiating parameters is captured fairly well. The small deviations can be the result of modeling assumptions, e.g., in the fluid-model we assume that C_n and m_n are instantly valid after the number of active source nodes has changed. By slightly modifying the

parameter values, i.e., a minimal reduction of C_n, the results coincide. We conclude that the fluid model of Section 8.4.1 accurately describes the behavior of an IEEE 802.11e relay node in a wireless ad-hoc network.

8.5 Concluding remarks

In this chapter we have shown that the fluid model is an accurate description of multi-hop flows relayed by a performance bottleneck in a wireless ad-hoc network. We have indicated how to map the parameter settings of both IEEE 802.11b and 802.11e relay nodes onto the fluid model, and the validation proves that the fluid modeling is accurate for both types of relay nodes.

Topics for further research. An interesting topic is the implementation of an alternative service disciplines at the relay node. In the above analysis it is assumed that the packet scheduling at the relay node is First Come First Served. Alternative service disciplines, e.g. round robin, may yield considerably smaller mean overall flow transfer times.

Another topic is the investigation of the influence of higher-layer protocols, such as TCP, on the flow transfer time.

Furthermore, it remains to be investigated how to properly implement the resource-sharing policy in wireless ad-hoc networks. Currently, obstacles for implementation are the lack of global knowledge of which nodes currently are bottlenecks and the absence of practical parameter settings to provide the desired resource sharing. A possible implementation is that each node is assigned an infinite $\text{TXOP}_{\text{limit}}$ for all *relay packets*, i.e., a node sends all packets that it has to relay for other nodes in a single TXOP.

Chapter 9

Transforms and tail probabilities for the equal resource-sharing fluid model

9.1 Introduction

In this chapter we consider the 'standard' fluid model of Section 6.4, i.e., the fluid model with $m = 1$ (resulting in so-called coupled input and output rates). In addition, throughout this chapter we assume exponentially distributed flow-sizes. This results in the situation that, when n flows are present, each of these use $\mathrm{C}/(n+1)$ to transmit its traffic into the queue, while the remaining capacity $\mathrm{C}/(n+1)$ is used to drain the queue. It will turn out that we can analyze this model relying on the concept of Markov-modulated fluid queues. We derive the Laplace transforms of all performance metrics and we obtain the tail probabilities by large-deviations analysis. In this chapter we use a different notation than in the previous chapters, as was mentioned earlier in Section 6.4.3.

9.1.1 Markov-modulated fluid queue

Standard Markov fluid queues consist of *traffic sources* feeding into a *queue* that is emptied at a constant rate, say C. The sources are for instance of the exponential ON-OFF type: they alternate between active periods (with a duration that is exponentially distributed with mean μ^{-1} during which traffic is generated at some fixed rate, say p) and silences (which have an exponential distribution with mean λ^{-1}). If there are N of such sources (i.i.d.), and if $Np > \mathrm{C}$, every now and then the buffer of the queue fills. Under the stability condition $Npf < \mathrm{C}$, with $f := \lambda/(\lambda+\mu)$ the fraction of time each source is on, the queue's workload has a steady-state distribution, say $W^\star$. A detailed performance analysis of this workload is available, see e.g. [5].

Standard Markov fluid-queues have been studied extensively. In the seminal studies [5, 77] a system of differential equations (known as Kolmogorov forward equations) is derived for $\mathbb{P}(W^\star \leq x, N^\star = n)$, where $N^\star$ is the number of sources in the on-state in steady-state. Later these results have been extended in many directions. To mention a few: one has considered heterogeneous sources, sources with a more general structure than exponential on-off, see e.g. [99], there have been rather explicit results for the case that the sources have a so-called birth-death structure [36]

or have a countably infinite state-space, see e.g. [131], and also models have been studied in which the source behavior depends on the current workload [88, 120]. In addition there has been considerable interest in so-called large-buffer asymptotics, i.e., expansions of $\mathbb{P}(W^\star > x)$ for large x; these relate nicely to a notion of effective bandwidths [39, 72].

9.1.2 Contribution

The goal of the present chapter is to extend the results for standard Markov fluid queues to our model of a relay node in an ad-hoc network. Interestingly, not even the stability criterion is trivial, as essentially all traffic has to be 'served' twice (it has to be transmitted into the queue, and subsequently it has to be served by the queue); as a result the common stability condition that the mean input rate, say m, be smaller than C does not apply.

The second aim is to characterize the steady-state workload distribution. It is not hard to see that this can be analyzed by setting up a system of Kolmogorov forward equations, but the special structure allows more explicit results. The crucial property of our model with coupled input and output that enables the analysis, is that the queue only drains when there are no flows present. This property entails that our model strongly resembles the classical M/G/1 queueing model, and hence the Laplace transform (LT) of the steady-state workload distribution can be given explicitly.

In standard Markov fluid queues there is a one-to-one mapping between the buffer content that a 'fluid particle' sees upon arrival, and the delay it has: if it sees x units traffic in the queue, it leaves the queue after x/C units of time. As a consequence, for standard Markov fluid queues, the queueing delay distribution follows immediately from the steady-state workload distribution. This is not the case for our model with coupled input and output; more specifically, when considering a tagged fluid particle that arrived at time 0, flows arriving in the future affect the service capacity available to the queue, and hence also the delay of the fluid particle. This makes the analysis of the queueing delay non-standard. We fully characterize its Laplace transform.

Furthermore, we study the flow transfer delay, i.e., the time it takes before the flow has transmitted all its traffic into the queue. This delay is essentially the absorption time of a certain continuous-time Markov chain. Again, the solution is given in terms of Laplace transforms.

The sojourn time of a flow is defined as the flow transfer time of an arbitrary flow increased by the time it takes before the last fluid particle of the flow is served. As these two components are correlated, the Laplace transform of the sojourn time does not immediately follow from the LTs of the buffer delay and the flow transfer delay. We derive the transform of the sojourn time explicitly using the fact that the buffer

content cannot decrease during any flow transfer time.

Having the Laplace transforms of the workload, queueing delay, and flow transfer delay at our disposal, a next question is how the tails of these distributions behave. We show that they decay exponentially, and, relying on large-deviations tools, the decay rates are derived.

9.1.3 Outline

The outline of this chapter is as follows. In Section 9.2 we present a detailed description of the 'standard' fluid model, and we derive its stability condition. In Section 9.3 we relate our model to the classical $\mathrm{M}/\mathrm{G}/1$ queueing model, and we present the Laplace transform (LT) of the steady-state workload distribution. Next, we characterize the Laplace transforms of the queueing delay in Section 9.4, the flow transfer delay in Section 9.5, and the sojourn time in Section 9.6. The tail-probabilities of the workload, queueing delay, and flow transfer delay are presented in Section 9.7. Finally, Section 9.8 concludes and identifies a few challenging subjects for future research. In particular, it discusses to what class of sharing policies (between the flows and the queue) our results can be extended.

9.2 Model and background

In this section, we first give a detailed description of our model, which is a special case of the fluid model presented in Section 6.4. Then we derive the steady-state distribution of the number of flows simultaneously present in the system, allowing us to give a precise stability condition.

9.2.1 Model

Consider a queueing system at which flows arrive according to a Poisson process, transmit traffic into a queue, and leave when ready. When there are n flows active, any flow can transmit its traffic into the queue at rate $\mathrm{C}/(n+1)$, while a rate $\mathrm{C}/(n+1)$ is used to serve the queue; as a consequence, the queue only drains when there are no flows present, while it stays at the same level if exactly one flow is active. Suppose that we impose the admission control policy that the system accommodates maximally $N \in \mathbb{N}$ flows simultaneously; in this way each active flow (as well as the queue) is guaranteed at least a transmission rate $\mathrm{C}/(N+1)$.

We let N_t denote the number of flows present (i.e., feeding traffic into the queue) at time t. It is not hard to see that, under the assumption of exponentially distributed flow sizes (with mean μ^{-1}) and interarrival times with mean λ^{-1}, the process N_t

constitutes a Markov chain on $\{0,\ldots,N\}$, with generator matrix

$$Q := \begin{pmatrix} -\lambda & \lambda & & & & & \\ \mu_1 \mathrm{C} & -\mu_1 \mathrm{C} - \lambda & \lambda & & & & \\ & \mu_2 \mathrm{C} & -\mu_2 \mathrm{C} - \lambda & \lambda & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & & \mu_N \mathrm{C} & -\mu_N \mathrm{C} \end{pmatrix}, \tag{9.1}$$

where $\mu_n := \mu n/(n+1)$. When $N_t = n$, the aggregate traffic rate generated by the flows is $r_{\mathrm{I},n} := \mathrm{C}n/(n+1)$, while the queue's output rate is $r_{\mathrm{O},n} := \mathrm{C}/(n+1)$, such that the net rate of change of the queue is 0 when $Q_t = N_t = 0$, and otherwise, for $n \in \{0,\ldots,N\}$,

$$r_{\mathrm{A},n} := r_{\mathrm{I},n} - r_{\mathrm{O},n} = \mathrm{C}\,\frac{n-1}{n+1}.$$

Define $R_{\mathrm{I}} := \mathrm{diag}\{r_{\mathrm{I}}\}$, $R_{\mathrm{O}} := \mathrm{diag}\{r_{\mathrm{O}}\}$, and $R_{\mathrm{A}} := R_{\mathrm{I}} - R_{\mathrm{O}}$.

Two variants of this model. In a first variant, one lets $N \to \infty$, thus getting a countably infinite state space. This means that there is no admission control imposed on the number of flows.

In a second variant, there are N sources that can be potentially active, and each source has a silence time that is exponentially distributed with mean λ^{-1}. The $q_{n,n+1}$ should be $(N-n)\lambda$ rather than λ (for $n = 0,\ldots,N-1$).

9.2.2 Stability condition

Due to the sharing of the service capacity between the flows and the queue, the stability condition of this model is not standard. Also, a fraction of the flows is rejected because they enter when already N flows are present. In this subsection we find the stability condition and the blocking probability.

To find a condition on λ, μ and C under which the queue is stable, we first determine the equilibrium distribution π of $(N_t)_{t\in\mathbb{R}}$. Trivially, the balance equations are

$$\pi_n \mu_n \mathrm{C} = \pi_{n-1}\lambda, \quad n = 1,\ldots,N.$$

Recursively solving these equations, it is not hard to derive, with ρ defined as $\lambda/(\mu\mathrm{C})$, that

$$\pi_n = \frac{\rho^n (n+1)}{\sum_{k=0}^{N} \rho^k (k+1)}.$$

Standard calculus on the geometric series yields

$$\begin{aligned}\sum_{k=0}^{N}\rho^k(k+1) &= \frac{\mathrm{d}}{\mathrm{d}\rho}\left(\sum_{k=0}^{N}\rho^{k+1}\right)=\frac{\mathrm{d}}{\mathrm{d}\rho}\left(\rho\,\frac{1-\rho^{N+1}}{1-\rho}\right)\\ &= \frac{1-\rho^{N+1}(N+2)+\rho^{N+2}(N+1)}{(1-\rho)^2}.\end{aligned}$$

The equilibrium condition of the fluid model is $\sum_{n=0}^{N}\pi_n r_{\mathrm{A},n}<0$, after considerable algebra translating into

$$\frac{-1+2\rho-\rho^{N+1}N+\rho^{N+2}(N-1)}{1-\rho^{N+1}(N+2)+\rho^{N+2}(N+1)}\cdot \mathrm{C}<0.$$

Due to the PASTA-property, the probability of an arbitrary arriving flow being blocked is

$$\pi_N=\frac{\rho^N(N+1)(1-\rho)^2}{1-\rho^{N+1}(N+2)+\rho^{N+2}(N+1)}.$$

Special case of $N\to\infty$. Interestingly, for $N\to\infty$, the equilibrium probabilities π_n have the form $(1-\rho)^2(n+1)\rho^n$, and the equilibrium condition $(-1+2\rho)\mathrm{C}<0$, or, equivalently, $2\lambda/\mu<\mathrm{C}$ (cf. Section 6.5). The latter condition has an appealing interpretation. In the model with $N\to\infty$, the input process is essentially a Poisson stream (arriving at rate λ) of flows that have mean size μ^{-1}. Every flow has to be processed twice: first it has to be put into the queue, and then it has to be served by the queue. This immediately leads to the stability condition $2\lambda/\mu<\mathrm{C}$.

9.3 Steady-state workload distribution

In this section we study the steady-state workload of the queue. As mentioned in the introduction, one could set up a system of Kolmogorov forward equations as in [5], which, in conjunction with the proper boundary condition, characterize the distribution function (in terms of eigenvalues and eigenvectors of some matrix). Due to the specific structure of our model, however, rather explicit results for the Laplace transform of the steady-state workload can be given. In particular we exploit the property that the buffer content only decreases when no flows are present, and the fact that these periods have an exponential duration, cf. for instance [30, 131]. As a consequence, our model is closely related to the family of M/G/1 systems.

9.3.1 Busy periods

In our analysis of the steady-state workload distribution, we need the notion of busy periods. A busy period B is, in this context, defined as a period that starts at an

epoch at which $(N_t)_{t\in\mathbb{R}}$ jumps from 0 to 1, and ends at a moment that it jumps from 1 to 0. We introduce the auxiliary quantity B_n, for $n = 1, \ldots, N$:

$$B_n := \inf\{t \geq 0 : N_t = n-1 \mid N_0 = n\};$$

evidently $B \stackrel{\mathrm{d}}{=} B_1$. In our analysis we also need the distribution of T, the *net* amount of traffic entering the queue (i.e., the increase of the buffer content) during B. Define $A(s,t) := \int_s^t r_{\mathrm{A},N_u}\mathrm{d}u$. Then $T \stackrel{\mathrm{d}}{=} T_1$, with

$$T_n \stackrel{\mathrm{d}}{=} A(0, B_n).$$

Analysis of the Laplace transform. Using standard arguments, cf. [58, 104, 117], we find the recursion, for $n = 1, \ldots, N-1$,

$$\mathbb{E}e^{-sT_n} = \frac{\lambda}{\lambda + \mu_n \mathrm{C} + r_{\mathrm{A},n}s}\,\mathbb{E}e^{-sT_{n+1}}\mathbb{E}e^{-sT_n} + \frac{\mu_n \mathrm{C}}{\lambda + \mu_n \mathrm{C} + r_{\mathrm{A},n}s}, \tag{9.2}$$

while for $n = N$ the random variable T_n is exponentially distributed with mean $r_N/(\mu_N \mathrm{C})$:

$$\mathbb{E}e^{-sT_N} = \frac{\mu_N \mathrm{C}}{\mu_N \mathrm{C} + r_{\mathrm{A},N}s}. \tag{9.3}$$

The above implies that $\mathbb{E}e^{-sT}$ is the solution of a finite recursion, of which the starting condition is known (namely $\mathbb{E}e^{-sT_N}$). The nature of the formula for $\mathbb{E}e^{-sT}$ is an N-fold iterated fraction.

Mean and second moment. Similarly to the above, we can find a recursion for the mean. It reads

$$\mathbb{E}T_n = \frac{r_{\mathrm{A},n}}{\mu_n \mathrm{C}} + \frac{\lambda}{\mu_n \mathrm{C}}\mathbb{E}T_{n+1} = \ldots = \sum_{i=n}^{N} \frac{\lambda^{i-n} r_{\mathrm{A},i}}{\mu_n \cdots \mu_i \mathrm{C}^{i-n+1}} = \frac{1}{n\mu}\sum_{i=n}^{N} \rho^{i-n}(i-1).$$

In particular,

$$\mathbb{E}T = \frac{1}{\mu}\cdot\frac{\rho}{(1-\rho)^2}\left(1 - \rho^{N-1}N + \rho^N(N-1)\right);$$

for the case $N \to \infty$, this converges to the elegant expression $\rho/(\mu(1-\rho)^2)$. For the second moment we can develop a recursion in the same way, again by distinguishing between the period where the number of flows is n, and the first jump afterwards. We obtain

$$\begin{aligned}\mathbb{E}T_n^2 &= \frac{2r_{\mathrm{A},n}^2}{(\lambda + \mu_n \mathrm{C})^2} + \frac{2r_{\mathrm{A},n}}{\lambda + \mu_n \mathrm{C}}\frac{\lambda}{\lambda + \mu_n \mathrm{C}}(\mathbb{E}T_n + \mathbb{E}T_{n+1}) \\ &\quad + \frac{\lambda}{\lambda + \mu_n \mathrm{C}}(\mathbb{E}T_n^2 + 2\mathbb{E}T_n\mathbb{E}T_{n+1} + \mathbb{E}T_{n+1}^2),\end{aligned}$$

with $\mathbb{E}T_N^2 = 2r_{\mathrm{A},N}^2/(\mu_N \mathrm{C})^2$. The recursion can be restated as $\mathbb{E}T_n^2 = \alpha_n \mathbb{E}T_{n+1}^2 + \beta_n$, with $\alpha_n := \lambda/(\mu_n \mathrm{C})$, and

$$\beta_n := 2\frac{r_{\mathrm{A},n}^2}{\mu_n \mathrm{C}(\lambda + \mu_n \mathrm{C})} + 2\frac{r_{\mathrm{A},n}\lambda}{\mu_n \mathrm{C}(\lambda + \mu_n \mathrm{C})}(\mathbb{E}T_n + \mathbb{E}T_{n+1}) + 2\frac{\lambda}{\mu_n \mathrm{C}}\mathbb{E}T_n \mathbb{E}T_{n+1};$$

notice that β_n, $n = 1, \ldots, N$, are known numbers, in view of the formulae for $\mathbb{E}T_n$ above. The solution of the recursion is, with the 'empty product' defined as 1,

$$\mathbb{E}T_n^2 = \left(\sum_{i=n}^{N-1}\left(\prod_{j=n}^{i-1}\alpha_j\right)\beta_i\right) + \left(\prod_{j=n}^{N-1}\alpha_j\right)\mathbb{E}T_N^2.$$

In particular, by inserting $n = 1$ we derive the second moment of T:

$$\mathbb{E}T^2 = \left(\sum_{i=1}^{N-1}\rho^{i-1} i\beta_i\right) + 2\rho^{N-1}N\frac{r_{\mathrm{A},N}^2}{(\mu_N \mathrm{C})^2}.$$

9.3.2 Steady-state workload

The steady-state workload, say $W^\star$, is, according to Reich's formula [106] (see also Expression (2.1)), distributed as

$$W^\star \stackrel{\mathrm{d}}{=} M := \sup_{t\geq 0} A(-t, 0) \stackrel{\mathrm{d}}{=} \sup_{t\geq 0} A(0, t),$$

where the second equality in distribution is due to the reversibility of $(N_t)_{t\in\mathbb{R}}$. In this subsection, we derive an explicit expression for the LT of M. Define

$$M_i := \sup_{t\geq 0}\{A(0, t) \mid N_0 = i\};$$

clearly $\mathbb{E}e^{-sM} = \sum_{n=0}^{N} \pi_n \mathbb{E}e^{-sM_n}$, hence we have to find expressions for $\mathbb{E}e^{-sM_n}$, for $n = 0, \ldots, N$.

As for $n = 1, \ldots, N$, during periods B_n the queue does not decrease, the random variables T_n are nonnegative almost surely. In fact, $A(0, t)$ attains its maximum either at time 0, or at an epoch at which N_t jumps from 1 to 0. These observations lead to the following equality in distribution:

$$M_n \stackrel{\mathrm{d}}{=} T_n + T_{n-1} + \cdots T_1 + M_0,$$

with $B_n, B_{n-1}, \ldots, B_1, M_0$ independent. This entails that

$$\mathbb{E}e^{-sM_n} = \mathbb{E}e^{-sM_0} \cdot \prod_{i=1}^{n} \mathbb{E}e^{-sT_i},$$

for $n = 0, \ldots, N$ (again defining the empty product to be 1). With a recipe to compute $\mathbb{E}e^{-sT_i}$ given in the previous section, we are left with computing $\mathbb{E}e^{-sM_0}$.

We now introduce an embedding that facilitates easy computation of the LT of M_0. Starting in 0, the maximum of $A(0,t)$ over $t \geq 0$ equals the maximum of $\sum_{j=0}^{i}(X_j - Y_j)$ over $i = 0, 1, \ldots$, with the X_j i.i.d. samples, distributed as T, and the Y_j i.i.d. samples from an exponential distribution with mean C/λ (where also the sequences X_j and Y_j are independent). The LT of the latter maximum is given by the celebrated Pollaczek-Khinchine formula, see for instance [6], so that we arrive at

$$\mathbb{E}e^{-sM_0} = \left(1 - \frac{\lambda \mathbb{E}T}{\mathrm{C}}\right) \frac{s}{s - (\lambda/\mathrm{C})(1 - \mathbb{E}e^{-sT})}.$$

Our final result is stated in the following theorem.

THEOREM 9.3.1. *The* LT *of the steady-state workload is given by,* $s \geq 0$,

$$\mathbb{E}e^{-sW^\star} = \mathbb{E}e^{-sM} = \sum_{n=0}^{N} \pi_n \left(1 - \frac{\lambda \mathbb{E}T}{\mathrm{C}}\right) \frac{s}{s - (\lambda/\mathrm{C})(1 - \mathbb{E}e^{-sT})} \left(\prod_{i=1}^{n} \mathbb{E}e^{-sT_i}\right),$$

where the $\mathbb{E}e^{-sT_i}$ *follow from* (9.2) *and* (9.3).

Moreover, we can also consider the joint distribution of the steady-state workload $W^\star$ and number of flows $N^\star$. It turns out that

$$\mathbb{E}(e^{-sW^\star} 1\{N^\star = n\}) = \pi_n \left(1 - \frac{\lambda \mathbb{E}T}{\mathrm{C}}\right) \frac{s}{s - (\lambda/\mathrm{C})(1 - \mathbb{E}e^{-sT})} \left(\prod_{i=1}^{n} \mathbb{E}e^{-sT_i}\right). \tag{9.4}$$

The above results also enable calculation of the mean steady-state workload:

$$\mathbb{E}W^\star = \left(\frac{1}{2} \frac{\lambda \mathbb{E}T^2}{\mathrm{C} - \lambda \mathbb{E}T}\right) + \left(\sum_{n=0}^{N} \left(\pi_n \sum_{i=1}^{n} \mathbb{E}T_i\right)\right),$$

following the convention that the empty sum is defined as 0.

9.4 Queueing delay distribution

As argued in Section 9.1, it is a nontrivial step to translate the steady-state workload distribution into the queueing delay distribution: for standard Markov fluid queues the buffer content seen by a fluid particle arriving, say at time 0, fully determines the epoch at which it will leave the queue, whereas in our system with coupled input and output the arrivals and departures of flow after 0 has impact. In the first

subsection we analyze the so-called *virtual queueing delay*, i.e., the delay experienced by a fluid particle arriving at a random point in time (i.e., a 'time average'), whereas the second subsection characterizes the queueing delay of an arbitrary fluid particle (i.e., a 'traffic average').

9.4.1 Virtual queueing delay

Let $D^\star$ denote the delay experienced by a fluid particle arriving at the queue in steady state, say for ease at time 0; this type of delay is sometimes referred to as virtual queueing delay. Let $O(0,t)$ denote the amount of output capacity available in the interval $[0,t)$. Then, cf. [74, Section III],

$$\begin{aligned}\mathbb{E}e^{-sD^\star} &= \int_0^\infty e^{-st}\mathbb{P}(D^\star = t)\mathrm{d}t = \int_0^\infty e^{-st}\mathbb{P}(W^\star = O(0,t))\mathrm{d}t \\ &= \sum_{n=0}^{N}\int_0^\infty e^{-st}\mathbb{P}(W^\star = O(0,t), N^\star = n)\mathrm{d}t.\end{aligned}$$

Now define, for $z \geq 0$, the random variable τ_z as the time until z units of service have become available:

$$\tau_z := \inf\{t \geq 0 : O(0,t) = z\} = \inf\left\{t \geq 0 : \int_0^t r_{\mathrm{O},N_s}\mathrm{d}s = z\right\};$$

notice that $O(0,t)$ is increasing in t. Using this notion, we get, with some abuse of notation,

$$\mathbb{E}e^{-sD^\star} = \sum_{n=0}^{N}\int_0^\infty e^{-st}\mathbb{P}(\tau_{W^\star} = t, N^\star = n)\mathrm{d}t,$$

which equals, remarking that $O(0,t)$ depends on $(W^\star, N^\star)$ just through $N^\star$,

$$\sum_{n=0}^{N}\int_0^\infty \int_0^\infty e^{-st}\mathbb{P}(W^\star = z, N^\star = n)\mathbb{P}(\tau_z = t \mid N^\star = n)\mathrm{d}z\mathrm{d}t.$$

Now we interchange the order of integration, to get

$$\sum_{n=0}^{N}\int_0^\infty \mathbb{E}(e^{-s\tau_z} \mid N^\star = n)\mathbb{P}(W^\star = z, N^\star = n)\mathrm{d}z.$$

Hence, to further compute this expression, we are first required to evaluate the expression $\mathbb{E}(e^{-s\tau_z} \mid N^\star = n)$. Fortunately, we have the following proposition at our disposal, cf. [22] and the appendix of [72].

PROPOSITION 9.4.1. *Consider an irreducible, finite-state (with states $0, \ldots, N$), continuous-time Markov chain $(X_t)_{t\in\mathbb{R}}$ with generator Q. Let r be a componentwise positive vector of dimension N, and $R := \mathrm{diag}\{r\}$. Define*

$$\tau_z := \inf\left\{t \geq 0 : \int_0^t r_{X_s}\mathrm{d}s = z\right\},$$

and $\xi_n(s,z) := \mathbb{E}(e^{-s\tau_z} \mid X_0 = n)$. Then, with $\xi(s,z) = (\xi_1(s,z), \ldots, \xi_N(s,z))^{\mathrm{T}}$, and $\mathbf{1}$ an $(N+1)$-dimensional vector with 1's,

$$\xi(s,z) = \exp((R^{-1}Q - sR^{-1})z)\mathbf{1}. \tag{9.5}$$

In addition, the eigenvalues $\delta_0(s), \ldots, \delta_N(s)$ of $R^{-1}Q - sR^{-1}$ are real, negative, and unique $(s > 0)$.

Proof A straightforward conditioning argument yields, with $q_j := -q_{jj}$,

$$\xi_n(s,z) = \sum_{m\neq n} \xi_m(s, z - r_n\,\Delta t)q_{nm}\Delta t + \xi_n(s, z - r_n\,\Delta t)e^{-s\Delta t}(1 - q_n\,\Delta t) \;+\; o(\Delta t).$$

Now writing $e^{-s\Delta t} = 1 - s\Delta t + O((\Delta t)^2)$, subtracting $\xi_n(s, z - r_n\,\Delta t)$ from both sides, dividing the equation by $r_n\Delta t$, and letting $\Delta t \downarrow 0$, we arrive at

$$\frac{\partial}{\partial z}\xi_n(s,z) = \sum_{m=1}^{N} \frac{q_{nm}}{r_n}\xi_n(s,z) - \xi_n(s,z)\frac{s}{r_n}.$$

In matrix-notation, we have that

$$\frac{\partial}{\partial z}\xi(s,z) = (R^{-1}Q - sR^{-1})\xi(s,z),$$

which yields (9.5).

Next we use that Geršgorin's circle theorem, see e.g. [96], implies that each eigenvalue of $M(s) = (m_{ij})_{i,j=0}^{N} := R^{-1}Q - sR^{-1}$ is in at least one of the disks

$$\left\{z \in \mathbb{C} : \left|z - \frac{q_{ii} - s}{r_i}\right| < \sum_{j\neq i}\frac{q_{ij}}{r_i}\right\},$$

and hence all eigenvalues are in the left half plane. Furthermore, the matrix $M(s)$ is real and tridiagonal with $m_{i,i+1}m_{i+1,i} > 0$ for $i = 0, \ldots, N-1$, and hence all its eigenvalues are real and unique, see again [96]. □

Apply Proposition 9.4.1, with continuous-time Markov chain N_t governed by Q as defined by (9.1), and $R := R_{\rm O}$ (which is indeed componentwise positive). Recalling that all eigenvalues $\delta_0(s), \ldots, \delta_N(s)$ of $M(s) := R_{\rm O}^{-1}Q - sR_{\rm O}^{-1}$ are different, so that we can write, for constants γ_{mn} with $m, n = 0, \ldots, N$,

$$\mathbb{E}(e^{-s\tau_z} \mid N^\star = n) = \sum_{m=0}^{N} \gamma_{mn} e^{\delta_m(s)\, z}. \tag{9.6}$$

Then we have found an explicit expression of the LT of the virtual queueing delay.

THEOREM 9.4.2. *For $s > 0$,*

$$\mathbb{E}e^{-sD^\star} = \sum_{n=0}^{N} \sum_{m=0}^{N} \gamma_{mn} \mathbb{E}(e^{\delta_m(s)W^\star} 1\{N^\star = n\}),$$

where the γ_{mn} are as in (9.6)*. The $\delta_n(s)$, for $n = 0, \ldots, N$, are the eigenvalues of $R_{\rm O}^{-1}Q - sR_{\rm O}^{-1}$ (which are negative). An expression for $\mathbb{E}(e^{-sW^\star} 1\{N^\star = n\})$ is available from Theorem 9.3.1.*

9.4.2 'Packet-average' queueing delay

Informally, the previous section gave the LT of the queueing delay 'at an arbitrary point in time'. Clearly, there is a bias between the delay $D^\star$ 'at an arbitrary point in time' and delay $\bar{D}^\star$ 'seen by an arbitrary fluid molecule'. The correction to be made is rather straightforward:

$$\mathbb{E}e^{-s\bar{D}^\star} = \sum_{n=0}^{N} \left(\frac{r_{{\rm I},n}}{\sum_{k=0}^{N} \pi_k r_{{\rm I},k}} \right) \sum_{m=0}^{N} \gamma_{mn} \mathbb{E}(e^{\delta_m(s)W^\star} 1\{N^\star = n\}),$$

cf. Asmussen [6, Proposition 7.2].

9.5 Flow transfer delay distribution

Now we focus on the time F it takes for an arbitrary arriving flow to transmit its traffic. We define the transfer time as the time between arrival and the epoch that its last fluid particle has been transmitted into the queue.

9.5.1 Flow transfer delay

Let the process $(Z_i)_{i\in\mathbb{N}}$ correspond to the number of flows present at (i.e., *just after*) arrival epochs. This process is a Markov chain, with, say, transition matrix $P =$

$(p_{mn})_{m,n=1}^{N}$. It is clear that Z_i can jump only one level up, or in other words, $p_{mn} = 0$ for all $n > m+1$. It can be verified easily that, for $m = 1, \ldots, N$ and $n = 1, \ldots, m+1$,

$$p_{mn} = \left(\prod_{k=n}^{m} \frac{\mu_k \mathsf{C}}{\lambda 1\{k \neq n\} + \mu_k \mathsf{C}} \right) \frac{\lambda}{\lambda + \mu_{n-1} \mathsf{C}}.$$

From this the equilibrium distribution π^Z can be computed efficiently due to the fact that the chain can jump just one level upwards. More directly, however, one can argue that we can use the PASTA-property here, such that

$$\pi_n^Z := \frac{\pi_{n-1}}{\sum_{m=0}^{N-1} \pi_m}. \tag{9.7}$$

We can now compute the LT of the flow transfer delay. Define F as the transfer delay of a tagged flow, that arrives at, say, time 0, when there are $n-1$ flows present (i.e., there are n flows immediately after the arrival of the tagged flow), $n = 1, \ldots, N$. We compute, for $n = 1, \ldots, N$ and $m = 0, \ldots, N-1$,

$$\phi_{nm}(s) := \mathbb{E}(e^{-sF} 1\{N_{F+} = m\} \mid N_0 = n).$$

A standard linear system can be written down, for $n = 1, \ldots, N-1$, cf. the analysis for the finite-capacity processor-sharing queue in [19, Section II]:

$$\phi_{nm}(s) = \frac{1}{\lambda + \mu_n \mathsf{C} + s} \left(\lambda \phi_{n+1,m}(s) + \frac{n-1}{n} \mu_n \mathsf{C}\, \phi_{n-1,m}(s) + \frac{1}{n} \mu_n \mathsf{C}\, 1\{n-1 = m\} \right);$$

here the fraction $1/n$ is the probability that at a departure epoch it is the tagged flow that leaves. We also have

$$\phi_{Nm}(s) = \frac{1}{\mu_N \mathsf{C} + s} \left(\frac{N-1}{N} \mu_N \mathsf{C}\, \phi_{N-1,m}(s) + \frac{1}{N} \mu_N \mathsf{C}\, 1\{N-1 = m\} \right).$$

We have thus derived, for fixed $m = 0, \ldots, N-1$ and s, N linear equations in N unknowns; as in [19] it can be shown that the corresponding matrix is, for any $s > 0$, diagonally dominant and thus non-singular, and hence there is a unique solution. The transform of the flow transfer delay of an arbitrary customer now reads

$$\mathbb{E}e^{-sF} = \sum_{n=1}^{N} \sum_{m=0}^{N-1} \pi_n^Z\, \phi_{nm}(s). \tag{9.8}$$

9.5.2 Representation of flow transfer delay with a phase-type distribution

Alternatively, the flow transfer delay distribution can also be found through a system of Kolmogorov equations. Defining

$$f_{nm}(t) := \mathbb{P}(F > t, N_{F+} = m \mid N_0 = n),$$

it is standard to derive through the usual Δt-argumentation, for $n = 1, \ldots, N$ and $m = 0, \ldots, N-1$,

$$\begin{aligned} f_{nm}(t+\Delta t) &= f_{n+1,m}(t)\,\lambda\Delta t\,1\{n<N\} + f_{n-1,m}(t)\,\mu_n\mathrm{C}\frac{n-1}{n}\Delta t\,1\{n>1\} \\ &\quad + f_{nm}(t)\,(1-(\lambda\,1\{n<N\}+\mu_n\mathrm{C}\,1\{n>1\})\Delta t)\,, \end{aligned}$$

immediately leading to

$$\begin{aligned} f'_{nm}(t) &= \lambda\,1\{n<N\}f_{n+1,m}(t) + \mu_n\mathrm{C}\frac{n-1}{n}\,1\{n>1\}f_{n-1,m}(t) \\ &\quad - (\lambda\,1\{n<N\}+\mu_n\mathrm{C}\,1\{n>1\})f_{nm}(t). \end{aligned}$$

Define the matrix $Q^\star = (q^\star_{mn})_{m,n=1}^N$ through $q^\star_{n,n-1} := q_{n,n-1}\,(n-1)/n$, and $q^\star_{mn} := q_{mn}$ otherwise. Then we have that the vector $f_m(t) := (f_{m1}(t), \ldots, f_{mN}(t))^{\mathrm{T}}$ satisfies $f'_m(t) = Q^\star f_m(t)$. Now also observe that the starting condition $f_{mn}(0)$ (again, fix m) follows from

$$\begin{aligned} &(\lambda\,1\{n<N\}+\mu_n\mathrm{C}\,1\{n>1\})f_{nm}(0) = \\ &\qquad \lambda\,1\{n<N\}f_{n+1,m}(0) + \frac{n-1}{n}\,\mu_n\mathrm{C}f_{n-1,m}(0) + \frac{1}{n}\,\mu_n\mathrm{C}1\{n-1=m\}; \end{aligned}$$

we call the solution $\bar{f}_m := (\bar{f}_{m1}, \ldots, \bar{f}_{mN})^{\mathrm{T}}$. We thus have obtained that

$$f_m(t) = \exp(Q^\star t)\bar{f}_m.$$

As $Q^\star$ is strictly diagonally dominant, it is non-singular. Using Geršgorin's theorem, one can prove that the eigenvalues $\bar{\delta}_1, \ldots, \bar{\delta}_N$ have a negative real part. In addition, as $q^\star_{m,m+1}q^\star_{m+1,m} > 0$ and $Q^\star$ is a real and tridiagonal matrix, all eigenvalues are real and unique [96]. These observations imply that we can find constants $\bar{\gamma}_{nm}$ such that

$$\mathbb{P}(F > t \mid N_0 = n) = \sum_{m=1}^{N} \bar{\gamma}_{nm} e^{\bar{\delta}_m t}. \tag{9.9}$$

Now we can rewrite LT (9.8) as follows. Observe that

$$\mathbb{E}e^{-sU} = 1 - \int_0^\infty \mathbb{P}(U > u) s e^{-su}\mathrm{d}u,$$

for any random variable U on $[0,\infty)$ for which these expectations exist. Hence, we obtain that, using that $\sum_{n=1}^{N} \bar{\gamma}_{mn} = 1$ for all m, and $\sum_{m=1}^{N} \pi_m^Z = 1$,

$$\begin{aligned}\mathbb{E}e^{-sF} &= 1 - \sum_{m=1}^{N} \pi_m^Z \left(\sum_{n=1}^{N} \bar{\gamma}_{nm} \frac{s}{-\bar{\delta}_n + s} \right) = \sum_{m=1}^{N} \pi_m^Z \left(\sum_{n=1}^{N} \bar{\gamma}_{nm} \frac{-\bar{\delta}_n}{-\bar{\delta}_n + s} \right) \\ &= \sum_{n=1}^{N} \bar{\gamma}_n \frac{-\bar{\delta}_n}{-\bar{\delta}_n + s}, \quad \text{with} \quad \bar{\gamma}_n := \sum_{m=1}^{N} \pi_m^Z \bar{\gamma}_{nm}.\end{aligned}$$

We conclude that F has a phase-type distribution, with shape parameters $-\bar{\delta}_1, \ldots, -\bar{\delta}_N$ and weights $\bar{\gamma}_1, \ldots, \bar{\gamma}_N$ (where the latter vector sums to 1).

9.5.3 Mean transfer delay

Consider the mean transfer delay of a flow that finds $n-1$ flows upon arrival ($n = 1, \ldots, N$), i.e.,

$$\mathbb{E}(F \mid N_0 = n) =: \eta_n;$$

at time 0 there are n flows present, including the tagged flow. Clearly, η_n is characterized through the N linear equations

$$\begin{aligned}&(\lambda\, 1\{n < N\} + \mu_n \mathrm{C}\, 1\{n > 1\})\, \eta_n = \\ &\qquad 1 + \lambda\, 1\{n < N\} \eta_{n+1} + \frac{n-1}{n}\, \mu_n \mathrm{C}\, 1\{n > 1\} \eta_{n-1}.\end{aligned}$$

Interestingly, these equations can be solved iteratively, as follows. The first equation gives η_2 in terms of η_1. Then consider the second equation; this gives η_3 in terms of η_1 and η_2, and hence also η_3 in terms of η_1 alone. Continuing in this way, we derive from the j th equation η_{j+1} in terms of η_1. After the $(N-1)$-st equation we have η_1 up to η_N expressed in terms of η_1. Plug these into the N-th equation, and solve η_1, and implicitly also $\eta_2, \ldots, \eta_N$. This procedure, however, does not lead to attractive explicit expressions.

Mean flow transfer delay $\mathbb{E}F$. First consider the limiting case of $N \to \infty$. Then it turns out that the above equations *do* allow a nice explicit solution. Inspired by the results for the processor-sharing queue [121], we try the 'linear solution' $\eta_n = \vartheta_{\mathrm{I}} + \vartheta_{\mathrm{II}}\, n$. Plugging these into our recursion yields the remarkably simple expressions

$$\vartheta_{\mathrm{I}} = \frac{1}{\mu \mathrm{C}} \frac{1}{2-\rho}, \quad \vartheta_{\mathrm{II}} = \frac{1}{\mu \mathrm{C}} \frac{3}{2-\rho},$$

so that

$$\mathbb{E}(F \mid N_0 = n) = \frac{1}{\mu \mathrm{C}} \frac{n+3}{2-\rho}.$$

The unconditioned mean file transfer delay (of an accepted flow) now reads (use PASTA)

$$\begin{aligned}\mathbb{E}F &= \sum_{n=0}^{\infty} \pi_n \mathbb{E}(F \mid N_0 = n+1) = \sum_{n=0}^{\infty} \rho^n (n+1)(1-\rho)^2 \frac{1}{\mu \mathrm{C}} \frac{n+4}{2-\rho} \\ &= \frac{2}{\mu \mathrm{C} - \lambda} = \frac{2}{\mu \mathrm{C}} \frac{1}{1-\rho}.\end{aligned}$$

We remark that the latter quantity can be computed also in a direct way, as follows. The mean number of flows in the system is $\sum_{n=0}^{\infty} n\rho^n(n+1)(1-\rho)^2 = 2\rho/(1-\rho)$, and with 'Little' we get the desired. Note that this result coincides with Expression (8.1) as it should.

'Little' can of course also be used when $N < \infty$; the advantage is that then we do not need explicit expressions for $\mathbb{E}(F \mid N_0 = n)$ to compute $\mathbb{E}F$. It yields

$$\begin{aligned}\mathbb{E}F &= \frac{\sum_{n=0}^{N} n\pi_n}{\lambda(1-\pi_N)} = \frac{\sum_{n=0}^{N} n\rho^n(n+1)(1-\rho)^2}{\lambda(1+\rho^{N+1}N - \rho^N(N+1))} \\ &= \frac{1}{\mu \mathrm{C}} \frac{\sum_{n=0}^{N-1} \rho^n (n+1)(n+2)(1-\rho)^2}{1+\rho^{N+1}N - \rho^N(N+1)};\end{aligned}$$

an explicit (though unattractive) expression for the numerator can be derived by differentiating the finite geometric series $\sum_{n=0}^{N} \rho^n = (1-\rho^{N+1})/(1-\rho)$ twice.

Mean flow transfer delay $\mathbb{E}F(x)$ of a flow of size x. We can also compute the expected flow transfer delay (of an accepted flow) *given* that the flow has size x. It is given by [30]

$$\mathbb{E}F(x) = \frac{x}{\mathrm{C}} \frac{1}{1-\pi_N} \left(\sum_{n=0}^{N-1} \rho^n \frac{c_{n+1}}{n!} \right) \Bigg/ \left(\sum_{n=0}^{N} \rho^n \frac{c_n}{n!} \right),$$

where c_n is the fraction of the service rate C that is dedicated to a single flow, when there are n flows present, i.e., $1/(n+1)$. This formula, which is remarkably enough linear in x, can be simplified to

$$\begin{aligned}\mathbb{E}F(x) &= \frac{x}{\mathrm{C}} \frac{\sum_{n=0}^{N-1} \rho^n (n+1)(n+2)}{\sum_{n=0}^{N-1} \rho^n (n+1)} = \frac{fx}{\mathrm{C}}, \\ &\qquad \text{with } f := \frac{\sum_{n=0}^{N-1} \rho^n (n+1)(n+2)(1-\rho)^2}{1+\rho^{N+1}N - \rho^N(N+1)};\end{aligned}$$

by integrating x out, the above expression for $\mathbb{E}F$ is recovered.

9.6 Sojourn time distribution

In this section we analyze the sojourn time of flows in the system, which is in fact the flow transfer time, increased by the time it takes to serve the last fluid particle of the flow. Notice that these two components are *not* independent, and as a consequence the LT of the sojourn time does not follow immediately from our earlier results.

We first describe the state of the system just after an arrival of an accepted flow. Then we study the transform of the flow transfer time *jointly with* the increase of the buffer during this period. Finally we use these ingredients to find the LT of the sojourn time.

9.6.1 Situation at flow arrival epochs

Here the PASTA-property applies. In other words: the joint distribution of the workload and the number of flows just after an arrival of an accepted flow is given by (9.4). Therefore, associating time 0 with the accepted flow arrival, we write, for $n = 1, \ldots, N$,

$$\begin{aligned} \chi_n(s) &:= \mathbb{E}(e^{-sW_0} 1\{N_0 = n\}) \\ &= \frac{\pi_{n-1}}{\sum_{m=0}^{N-1} \pi_m} \left(1 - \frac{\lambda \mathbb{E}T}{\mathrm{C}}\right) \frac{s}{s - (\lambda/\mathrm{C})(1 - \mathbb{E}e^{sT})} \left(\prod_{i=1}^{n} \mathbb{E}e^{sT_i}\right), \end{aligned} \tag{9.10}$$

cf. also Expression (9.7).

9.6.2 Joint transform of flow transfer delay and workload increment

The goal of this subsection is to compute the transform of the transfer delay F of a job that finds $n-1$ jobs upon arrival ($n = 1, \ldots, N$), jointly with the increment of the workload in this period, say ΔW, and the number of flows present at the end of the transfer (not counting the flow that just left) N_{F+}:

$$\psi_{nm}(\vec{s}) := \mathbb{E}(e^{-s_1 F - s_2 \Delta W} 1\{N_{F+} = m\} \mid N_0 = n),$$

with $\vec{s} \equiv (s_1, s_2)$. Notice that the workload cannot decrease during the flow transfer, and, as a consequence, the distribution of ΔW depends on the past only through N_0 (importantly, the value of W_0 does not play a role).

The $\psi_{nm}(\vec{s})$ satisfy, for $n = 1, \ldots, N-1$, the following system of equations:

$$\begin{aligned} \psi_{nm}(\vec{s}) = &\frac{1}{\lambda + \mu_n \mathrm{C} + s_1 + r_{\mathrm{A},n} s_2} \times \\ &\left(\lambda \psi_{n+1,m}(s) + \frac{n-1}{n} \mu_n \mathrm{C}\, \psi_{n-1,m}(s) + \frac{1}{n} \mu_n \mathrm{C}\, 1\{n-1 = m\}\right). \end{aligned} \tag{9.11}$$

We also have

$$\psi_{Nm}(\vec{s}) = \frac{1}{\mu_N \mathrm{C} + s_1 + r_{\mathrm{A},N} s_2} \times \left(\frac{N-1}{N} \mu_N \mathrm{C}\, \psi_{N-1,m}(s) + \frac{1}{N} \mu_N \mathrm{C}\, 1\{N-1=m\} \right). \tag{9.12}$$

For fixed m and $\vec{s}$, these form a system of linear equations, which is (as earlier) non-singular.

9.6.3 Sojourn time

In our analysis, we use the following decomposition of the sojourn time S: S can be written as the sum of

- the flow transfer delay,
- and the time required to process the last particle of the flow. The buffer content at the end of the flow transfer time can be decomposed into
  - i) the amount of traffic in the buffer at the epoch the flow arrived,
  - ii) the net amount of fluid that entered the buffer during the flow transfer delay.

Above we have seen that the workload at flow arrival (intersected with the event that n flows are present) is characterized through the LT $\chi_n(s)$. On the other hand, the net amount of fluid entering the queue, jointly with the flow transfer delay and intersected with the event that when the tagged flow leaves there are m flows present, given that at the start of the flow transfer n flows were transmitting, is characterized through LT $\psi_{nm}(s)$. Combining these gives, with some abuse of notation, and with τ_z as defined before, the following expression for the LT of S:

$$\begin{aligned} \mathbb{E}e^{-sS} &= \mathbb{E}\exp(-sF - s\tau_{W_0+\Delta W}) \\ &= \int_0^\infty \int_0^\infty \sum_{n=1}^{N} \sum_{m=0}^{N-1} \mathbb{P}(W_0 = x, N_0 = n) \mathbb{E}(e^{-s\tau_{x+y}} \mid N_0 = m) \\ &\qquad \mathbb{E}(e^{-sF} 1\{\Delta W = y, N_{F+} = m\} \mid N_0 = n) \mathrm{d}x\mathrm{d}y. \end{aligned}$$

Now using Proposition 9.4.1, this expression equals

$$\int_0^\infty \int_0^\infty \sum_{n=1}^{N} \sum_{m=0}^{N-1} \mathbb{P}(W_0 = x, N_0 = n) \mathbb{E}(e^{-sF} 1\{\Delta W = y, N_{F+} = m\} \mid N_0 = n) \sum_{k=0}^{N} \gamma_{km} e^{\delta_k(s)\,(x+y)} \mathrm{d}x\mathrm{d}y.$$

We have proven the following result.

THEOREM 9.6.1. *For $s > 0$,*

$$\mathbb{E}e^{-sS} = \sum_{n=1}^{N} \sum_{m=0}^{N-1} \sum_{k=0}^{N} \gamma_{km}\chi_n(-\delta_k(s))\psi_{nm}(s, -\delta_k(s)),$$

where the γ_{mn} are as in (9.6)*, $\chi_n(\cdot)$ as in* (9.10)*, and $\psi(\cdot)$ defined through* (9.11) *and* (9.12)*.*

REMARK 9.6.2. *The above procedure also yields the joint* LT *of the flow transfer time F, and the time $\tau_{W_0+\Delta W}$ it takes to serve the last fluid particle of the flow:*

$$\mathbb{E}\exp(-s_1F - s_2\tau_{W_0+\Delta W}) = \sum_{n=1}^{N} \sum_{m=0}^{N-1} \sum_{k=0}^{N} \gamma_{km}\chi_n(-\delta_k(s_2))\psi_{nm}(s_1, -\delta_k(s_2)).$$

This formula (implicitly) describes the correlation between F and $\tau_{W_0+\Delta W}$. ♢

9.7 Tail probabilities

In this section, we study the tail behavior of $W^\star$, $D^\star$, and F, and S. More specifically, we show that these three random variables decay exponentially, and, in addition, we identify the associated decay rate. We first recall the following collection of results, which were proven in, e.g., [72], relying on the Gärtner-Ellis theorem [34, Theorem 2.3.6]. A key role is played by the asymptotic logarithmic moment generating function (MGF), or cumulant function, and its properties.

PROPOSITION 9.7.1. *Consider an irreducible, finite-state (with states $0, \ldots, N$), continuous-time Markov chain $(X_t)_{t\in\mathbb{R}}$ with generator Q and equilibrium distribution π. Let r be a vector of dimension N such that $m_{\text{A}} := \sum_{n=0}^{N} \pi_n r_n < 0$, and $R := \text{diag}\{r\}$. Define $A(s,t) := \int_s^t r_{X_u}\,\text{d}u$.*

1. *The asymptotic logarithmic* MGF *of $A(0,t)$, i.e.,*

$$\Lambda_{\text{A}}(\theta) := \lim_{t\to\infty} \frac{1}{t} \log \mathbb{E}\exp(\theta A(0,t)),$$

is a convex function, and equals the largest eigenvalue of $Q + \theta R$, irrespective of the value of X_0. With $q_i := \sum_{j\neq i} q_{ij}$, we have that $\Lambda_{\text{A}}(\theta)$ exists for all θ smaller than

$$\min\left\{\frac{q_i}{r_i} : r_i > 0\right\}.$$

2. *For any $x > m_{\text{A}}$,*

$$\lim_{t\to\infty} \frac{1}{t} \log \mathbb{P}\left(\frac{A(0,t)}{t} > x\right) = -I_{\text{A}}(x),$$

with $I_{\rm A}(x) := \sup_\theta(\theta x - \Lambda_{\rm A}(\theta))$*;* $I_{\rm A}(\cdot)$ *is convex,* $I_{\rm A}(m_{\rm A}) = 0$*. Similarly, for* $x < m_{\rm A}$*,*

$$\lim_{t\to\infty} \frac{1}{t} \log \mathbb{P}\left(\frac{A(0,t)}{t} < x\right) = -I_{\rm A}(x).$$

3. *For the steady-state workload* $W^\star$*, which is distributed as* $\sup_{t\geq 0} A(-t,0)$*, it holds that*

$$\lim_{x\to\infty} \frac{1}{x} \log \mathbb{P}(W^\star > x) = -\theta^\star.$$

Here $\theta^\star$ *is the smallest positive eigenvalue solving the eigensystem* $-\theta R x = Q x$*. Alternatively,* $\theta^\star$ *is characterized as the unique positive solution of* $\Lambda_{\rm A}(\theta) = 0$*. Yet a third way of computing the decay rate is*

$$\theta^\star = \inf_{m>0} I_{\rm A}(m)/m. \tag{9.13}$$

REMARK 9.7.2. *An intuitive explanation of the relation (9.13) is the following.* $I_{\rm A}(m)$ *can be interpreted as the cost incurred for the process* $A(0,t)$ *to generate traffic at rate* m*; evidently there is no cost involved when sending at the average rate* $m_{\rm A}$ *(reflected by* $I_{\rm A}(m_{\rm A}) = 0$*), but there is a positive cost for sending at a higher (or lower) rate. Suppose the process generates traffic at rate* $m > 0$*. Then it takes about* x/m *to reach buffer level* x*, and the cost made is* $I_{\rm A}(m)/m$*. There is an evident trade-off between the numerator and the denominator: when choosing* m *small but positive, the cost per unit of time are relatively low, but it takes long to reach* x*, whereas the opposite applies when choosing* m *large. We conclude that the 'most likely speed'* $m^\star$ *is the minimizing argument in*

$$x \left(\inf_{m>0} I_{\rm A}(m)/m\right), \tag{9.14}$$

where (9.14) roughly equals $-\log \mathbb{P}(W^\star > x)$*, for* x *large.* ♢

9.7.1 Decay rate of steady-state workload

The decay rate $\theta^\star$ of $W^\star$ follows immediately from Proposition 9.7.1, with continuous-time Markov chain N_t governed by Q as defined by (9.1), and $R := R_{\rm A}$: $\theta^\star$ is the smallest positive eigenvalue of the system $-\theta R_{\rm A} x = Q x$. In fact, one can prove the stronger statement that $\mathbb{P}(W^\star > x)\exp(\theta^\star x)$ converges to some constant $\kappa > 0$ for $x \to \infty$, and even, for $n = 0, \ldots, N$,

$$\lim_{x\to\infty} \mathbb{P}(W^\star > x, N^\star = n)\exp(\theta^\star x) = \kappa_n, \tag{9.15}$$

for $\kappa_n > 0$, see for instance [77].

Another way to characterize $\theta^\star$ is as follows [89]. Let U_{mn} be the value of $A(0, V_n)$ conditional on $N_0 = m$, where V_n is the epoch of the first entrance of N_t for $t > 0$

to state n. Then $\theta^\star$ can be alternatively characterized as the unique positive solution of $\mathbb{E}e^{\theta U_{mm}} = 1$; remarkably, in [89] it is shown this solution is identical for any $m = 0, \ldots, N$. Now consider $m = 0$. Then U_{00} is distributed as $E+T$, with E exponentially distributed with mean λ^{-1}, T as defined in Section 3, and E and T independent. The equation $\mathbb{E}e^{\theta U_{00}} = 1$ then reduces to

$$\frac{\lambda}{\lambda + \theta \mathrm{C}}\, \mathbb{E}e^{\theta T},$$

or, equivalently, $\theta + (\lambda/\mathrm{C})(1 - \mathbb{E}e^{\theta T}) = 0$. We conclude that the decay rate $\theta^\star$ coincides with (minus) the pole of $\mathbb{E}e^{-sW^\star}$, cf. Theorem 9.3.1.

9.7.2 Decay rate of queueing delay

We next characterize the exponential decay rate of the queueing delay. We here focus on the virtual queueing delay, but it can be verified easily that the same decay rate applies to the 'packet average'.

We first define the cumulant function of the output process, as follows. For $\theta \in \mathbb{R}$,

$$\Lambda_{\mathrm{O}}(\theta) := \lim_{t\to\infty} \frac{1}{t} \log \mathbb{E} \exp(\theta O(0,t)).$$

This function equals the largest eigenvalue of $Q + \theta R_{\mathrm{O}}$, due to Proposition 9.7.1. We also define $I_{\mathrm{O}}(x) := \sup_\theta(\theta x - \Lambda_{\mathrm{O}}(\theta))$, and $m_{\mathrm{O}} := \sum_{n=0} N\pi_n r_{\mathrm{O},n}$. We first observe that, again due to Proposition 9.7.1, irrespective of the number of flows present at time 0,

$$\lim_{u\to\infty} \frac{1}{u} \log \mathbb{P}(O(o,u) \in [i\epsilon u, (i+1)\epsilon u)) =$$

$$\zeta_i(\epsilon) := \begin{cases} -I_{\mathrm{O}}(i\epsilon) & \text{if } m_{\mathrm{O}} < i\epsilon; \\ -I_{\mathrm{O}}((i+1)\epsilon) & \text{if } m_{\mathrm{O}} > (i+1)\epsilon; \\ -I_{\mathrm{O}}(m_{\mathrm{O}}) = 0 & \text{if } i\epsilon < m_{\mathrm{O}} < (i+1)\epsilon, \end{cases}$$

explicitly using the convexity of $I_{\mathrm{O}}(\cdot)$. The following result is [34, Lemma 1.2.15].

LEMMA 9.7.3. *For any finite index set $\mathcal{S}$, and $\omega_i(u) \geq 0$,*

$$\limsup_{u\to\infty} \frac{1}{u} \log \sum_{i\in\mathcal{S}} \omega_i(u) = \max_{i\in\mathcal{S}} \limsup_{u\to\infty} \frac{1}{u} \log \omega_i(u).$$

Now we have collected the prerequisites for the proof of the following result.

THEOREM 9.7.4. *The decay rate of the virtual queueing delay equals*

$$\lim_{t\to\infty} \frac{1}{t} \log \mathbb{P}(D^\star > t) = -\inf_m (I_{\mathrm{O}}(m) + \theta^\star m) = \Lambda_{\mathrm{O}}(-\theta^\star). \tag{9.16}$$

Proof We first prove the first equality in (9.16). We start by establishing the *upper bound*. Conditioning on the value of $O(0,t)$,

$$\begin{aligned}\mathbb{P}(D^\star > t) &= \sum_{n=0}^{N} \mathbb{P}(W^\star > O(0,t), N^\star = n) \\ &\leq \sum_{n=0}^{N}\sum_{i=0}^{\infty} \mathbb{P}(W^\star \geq (i+1)\epsilon t, N^\star = n) \\ &\quad \mathbb{P}(O(0,t) \in [i\epsilon t, (i+1)\epsilon t) \mid N^\star = n). \end{aligned} \tag{9.17}$$

It is clear that for some values of i there is no contribution, due to the fact that the rates in the vector r_{O} are between $\mathrm{C}/(N+1)$ and C. Therefore, we can restrict ourselves to

$$i \in \mathcal{I}_\epsilon, \text{ where } \mathcal{I}_\epsilon := \left\{ i \in \mathbb{N} : \frac{\mathrm{C}}{\epsilon(N+1))} - 1 \leq i \leq \frac{\mathrm{C}}{\epsilon} \right\}.$$

The decay rate of $\mathbb{P}(W^\star \geq (i+1)\epsilon t, N^\star = n)$ is $-\theta^\star(i+1)\epsilon$, independently of n, see (9.15). The decay rate of $\mathbb{P}(O(0,t) \in [i\epsilon t, (i+1)\epsilon t) \mid N^\star = n)$ is $\zeta_i(\epsilon)$, as given above, also independently of n. In view of Lemma 9.7.3, the decay rate of (9.17) is majorized by

$$\max_{i \in \mathcal{I}_\epsilon} \left(-\theta^\star(i+1)\epsilon + \zeta_i(\epsilon) \right).$$

Now let $\epsilon \downarrow 0$; using the continuity of $I_{\mathrm{O}}(\cdot)$, we arrive at

$$\sup_{m \in [\mathrm{C}/(N+1), \mathrm{C}]} (-\theta^\star m - I_{\mathrm{O}}(m)). \tag{9.18}$$

Now we present the *lower bound*, which is established in a similar fashion. Evidently, for any i,

$$\mathbb{P}(D^\star > t) \geq \mathbb{P}(W^\star \geq (i+1)\epsilon t, N^\star = n)\mathbb{P}(O(0,t) \in [i\epsilon t, (i+1)\epsilon t) \mid N^\star = n).$$

The decay rate of the right-hand side of the previous display is $-\theta^\star i\epsilon + \zeta_i(\epsilon)$; as this holds for any i, the supremum over i is still a lower bound. Taking $\epsilon \downarrow 0$, we obtain that the upper bound (9.18) is also lower bound.

We have now proven the first equality in (9.16); the second immediately follows from the duality relation $\Lambda_{\mathrm{O}}(\theta) = \sup_x(x\theta - I_{\mathrm{O}}(x))$, see for instance [38, Theorem VI.4.1]. □

REMARK 9.7.5. *There is an appealing alternative way to characterize this decay rate, cf. Remark 9.7.2. Consider the event that a fluid particle arriving at time 0 has (approximately) virtual delay t. Suppose that, after time 0, the queue drains at rate m, which costs $I_{\mathrm{O}}(m)$ per unit of time. In order to achieve delay t, the workload at time 0 should have been mt.*

Supposing that the queue built up at rate $m' > 0$ before time 0, with cost $I_{\rm A}(m')$ per unit of time, this took $(m/m')t$ time. In other words, we are to minimize

$$\inf_{m,m'>0} \left(I_{\rm A}(m') \, \frac{mt}{m'} + I_{\rm O}(m) \, t \right) = t \left(\inf_{m>0} \left(\theta^\star m + I_{\rm O}(m) \right) \right),$$

where the equality is due to (9.13). ♢

9.7.3 Decay rate of flow transfer delay

The decay rate of the flow transfer delay follows immediately from the phase-type distribution identified in Section 5. Directly from Equation (9.9), we see that

$$\lim_{x\to\infty} \frac{1}{x} \log \mathbb{P}(F > x) = \bar{\delta} := \max_{n=1,\ldots,N} \bar{\delta}_n,$$

i.e., the dominant eigenvalue of $Q^\star$.

9.7.4 Decay rate of sojourn time

We now turn our attention to the tail behavior of the sojourn time. This is a complicated issue, as long sojourn times are due to a combination of i) a high workload when the flow enters, ii) a large flow, iii) a large amount work brought along by flows arriving during the flow transfer time of the tagged flow, iv) a low service speed available to the queue after the flow transmission time (i.e., when the complete flow has been put into the queue). We below sketch how the exponential decay rate can be computed; the arguments can be made precise as in Section 6.2.

Using the representation $S = F + \tau_{W_0+\Delta W}$, we condition on the values of W_0, ΔW, and F. With some abuse of notation,

$$\begin{aligned}
&\mathbb{P}(F + \tau_{W_0+\Delta W} > t) \\
&\quad \approx \sum_{n=1}^{N} \int_0^\infty \mathbb{P}(W_0 = zt \mid N_0 = n) \mathbb{P}(F + \tau_{z+\Delta W} > t, N_0 = n) \mathrm{d}z \\
&\quad \approx \sum_{n=1}^{N} \sum_{m=0}^{N-1} \int_0^\infty \int_0^\infty \int_0^\infty \mathbb{P}(W_0 = zt \mid N_0 = n) \\
&\qquad\qquad \mathbb{P}(F = ft, \Delta W = wt, N_0 = n, N_{F+} = m) \\
&\qquad\qquad \mathbb{P}(\tau_{zt+wt} > t - ft, N_0 = m) \mathrm{d}f \, \mathrm{d}w \, \mathrm{d}z,
\end{aligned}$$

with $f \in (0, 1)$. Now we use the folk theorem that says that the decay rate of an integral equals the decay rate of the maximum of the integrand. We saw earlier that the exponential decay rate (x large) of $\mathbb{P}(W_0 = zt \mid N_0 = n)$ does not depend on n; likewise, the decay rates of the other two probabilities, $\mathbb{P}(\tau_{zt+wt} > t - ft, N_0 = m)$ and $\mathbb{P}(F = ft, \Delta W = wt, N_0 = n, N_{F+} = m)$, do not depend on m and n. They can be computed as follows:

- As before, for $z > 0$,

$$\lim_{t\to\infty} \frac{1}{t} \log \mathbb{P}(W_0 = zt) = -z\theta^\star =: J_1(z).$$

- Similar to the decay rate of F being equal to $\max_{n=1,\ldots,N} \bar{\delta}_n$, i.e., the infimum over all $s < 0$ for which $\mathbb{E}e^{-sF} < \infty$, we have that

$$\lim_{x\to\infty} \frac{1}{t} \log \mathbb{P}(F = ft, \Delta W = wt) = \inf\left\{s_1 f + s_2 w : \mathbb{E}e^{-s_1 F - s_2 \Delta W} < \infty\right\} =: J_2(f, w).$$

  Notice that this decay rate is larger than $-\infty$, as can be seen as follows. Suppose that T is the flow size of the tagged flow. Then, as each flow receives a rate of maximally $\mathrm{C}/2$, we have that $F \geq 2T/\mathrm{C}$. Hence, for $s_1 > -\mu\mathrm{C}/2$,

$$\mathbb{E}e^{-s_1 F - s_2 \Delta W} \geq \frac{\mu}{\mu + 2s_1/\mathrm{C}},$$

  and $\mathbb{E}e^{-s_1 F - s_2 \Delta W} = \infty$ for $s_1 \leq -\mu\mathrm{C}/2$.

- Also, as earlier,

$$\begin{aligned}\lim_{t\to\infty} \frac{1}{t} \log \mathbb{P}(\tau_{zt+wt} > t - ft) &= \lim_{t\to\infty} \frac{1}{t} \log \mathbb{P}(O(0, (1-f)t) < zt + wt) \\ &= -(1-f) I_{\mathrm{O}}\left(\frac{z+w}{1-f}\right) =: J_3(z, f, w),\end{aligned}$$

  with $(z + w)/(1 - f) < m_{\mathrm{O}}$.

Collecting terms, we find that

$$\lim_{x\to\infty} \frac{1}{t} \log \mathbb{P}(S > t) = \sup_{z,f,w} \left(J_1(z) + J_2(f, w) + J_3(z, f, w)\right),$$

where the maximization is over all $z, w > 0$ and $f \in (0, 1)$, such that $(z+w)/(1-f) < m_{\mathrm{O}}$.

9.8 Concluding remarks

An important feature of the model discussed in this chapter is that there is just one state in which the queue drains. It has appeared that this is a key property in our analysis. Importantly, it entails that the dynamics of the number of flows in the system are not affected by the workload process. This enabled the computation of the

LT of the workload, as it brought us into the framework of M/G/1-type of models. Also, it implied that the workload cannot decrease during flow transfer; as a consequence ΔW (as used in Section 9.7) depends on N_0, and not on W_0. We remark that, if the focus is on the *mean* sojourn time, rather than the entire distribution, fairly explicit approximations are possible, see Section 8.3.2.

Topics for further research. An interesting extension would relate to the situation *without* admission control. The complication is that the state-space of $(N_t)_{t\in\mathbb{R}}$ becomes (countably) infinite. The results of Section 9.3 carry over to this situation; still the LT of T can be computed by methods similar to those in [58, 104]. The results of the other sections will change; in any case all matrix-exponentials should be handled with care.

One could also study the situation of multiple relay nodes that are sharing capacity. The complicating factor is that then the dynamics of the flows feeding into one queue will be affected by the workload process in other queues. As a result, this model has the flavor of coupled-processors systems as studied in, e.g., [42], which are notoriously hard to analyze. Other challenging extensions include: i) non-exponential flow-size distribution (for instance regularly varying), ii) heterogeneous flow types, iii) allocation policies that do not depend only on the number of flows present, but also on the buffer content, cf. [120].

References

[1] J. Abate, G. Choudhury, and W. Whitt. *Computational Probability*, chapter An introduction to numerical transform inversion and its application to probability models, pages 257–323. Kluwer, Boston, MA, USA, 1999.

[2] J. Abate and W. Whitt. Numerical inversion of Laplace transforms of probability distributions. *ORSA Journal on Computing*, 7:36–43, 1995.

[3] R. Addie, P. Mannersalo, and I. Norros. Most probable paths and performance formulae for buffers with Gaussian input traffic. *European Transactions on Telecommunications*, 13:183–196, 2002.

[4] R. Addie, T. Neame, and M. Zukerman. Performance evaluation of a queue fed by a Poisson Pareto burst process. *Computer Networks*, 40:377–397, 2002.

[5] D. Anick, D. Mitra, and M. Sondhi. Stochastic theory of data-handling system with multiple sources. *Bell System Technical Journal*, 61:1871–1894, 1982.

[6] S. Asmussen. *Ruin probabilities. Advanced Series on Statistical Science & Applied Probability. Volume 2*. World Scientific, London, UK, 2000.

[7] R. Azencott. *Ecole d'Eté de Probabilités de Saint-Flour VIII-1978*, chapter Grandes déviations et applications, pages 1–176. Lecture Notes in Mathematics, Volume 774. Springer, Berlin, Germany, 1980.

[8] C. Barakat, P. Thiran, G. Iannaccone, C. Diot, and P. Owezarski. Modeling Internet backbone traffic at the flow level. IEEE *Transactions on Signal Processing*, 51:2111–2114, 2003.

[9] R. Bekker and M. Mandjes. A fluid model for a relay node in an ad-hoc network: the case of heavy-tailed input. *Mathematical Methods in Operations Research*, DOI:10.1007/s00186-008-0272-3, CWI-report available at http://ftp.cwi.nl/CWIreports/PNA/PNA-E0703.pdf, 2009.

[10] N. Ben Azzouna, F. Clérot, C. Fricker, and F. Guillemin. A flow-based approach to modeling ADSL traffic on an IP backbone link. *Annals of Telecommunications*, 59:1260–1314, 2004.

[11] S. Ben Fredj, T. Bonald, A. Proutière, G. Régnié, and J. Roberts. Statistical bandwidth sharing: a study of congestion at flow-level. In *Proceedings of Sigcomm*, pages 111–122, 2001.

[12] H. van den Berg, R. Litjens, and F. Roijers. Ad-hoc netwerken robuust en flexibel. *Automatisering Gids*, 34:16–17, 2008.

[13] H. van den Berg, M. Mandjes, R. van de Meent, A. Pras, F. Roijers, and P. Venemans. QoS-aware bandwidth provisioning of IP links. *Computer Networks*, 50:631–647, 2006.

[14] H. van den Berg, M. Mandjes, and F. Roijers. Performance modeling of a bottleneck node in an IEEE 802.11 ad-hoc network. In *Proceedings of AdHoc-Now2006*, pages 321–336, 2006.

[15] T. Bheema Reddy, I. Karthigeyan, B. Manoj, and C. Siva Ram Murthy. Quality-of-service provisioning in ad hoc wireless networks: a survey of issues and solutions. *Ad Hoc Networks*, 4:83–124, 2006.

[16] G. Bianchi. Performance analysis of the IEEE 802.11 Distributed Coordination Function. IEEE *Journal on Selected Areas in Communications*, 18:535–547, 2000.

[17] T. Bonald and A. Proutière. Insensitivity in processor-sharing networks. *Performance Evaluation*, 49:193–209, 2002.

[18] T. Bonald and J. Roberts. Congestion at flow level and the impact of user behaviour. *Computer Networks*, 42:521–536, 2003.

[19] S. Borst, O. Boxma, and N. Hegde. Sojourn times in finite-capacity processor-sharing queues. In *Proceedings of 1st EURO-NGI Conference*, pages 53–60, 2005.

[20] S. Borst, O. Boxma, J. Morrison, and R. Núñez-Queija. The equivalence between processor sharing and service in random order. *Operations Research Letters*, 31:254–262, 2003.

[21] D. Botvich and N. Duffield. Large deviations, the shape of the loss curve, and economies of large scale multiplexers. *Queueing Systems*, 20:293–320, 1995.

[22] A. Brandt and M. Brandt. On the distribution of the number of packets in the fluid flow approximation of packet arrival streams. *Queueing Systems*, 17:275–315, 1994.

[23] J. Bucklew. *Large Deviation Techniques in Decision, Simulation and Estimation*. Wiley, New York, NY, USA, 1990.

[24] J. Bucklew. *Introduction to Rare Event Simulation*. Springer, New York, NY, USA, 2004.

[25] M. Carvalho and J. Garcia-Luna-Aceves. Delay analysis of IEEE 802.11 single-hop networks. In *Proceedings of the 11th* IEEE *International Conference on Network Protocols*, pages 146–155, 2003.

[26] M. Carvalho and J. Garcia-Luna-Aceves. A scalable model for channel access protocols in multihop ad hoc networks. In *Proceedings of MobiCom'04*, pages 330–344, 2004.

[27] S. Cheung, H van den Berg, R. Boucherie, R. Litjens, and F. Roijers. An analytical packet/flow-level modelling approach for Wireless LANs with Quality-of-Service support. In *Proceedings of ITC 19*, pages 1651–1662, 2005.

[28] T. Coenen, H. van den Berg, and R. Boucherie. A flow level model for wireless multihop ad hoc network throughput. In *Proceedings of HET NETs*, pages 34/1–34/10, 2005.

[29] E. Coffman jr., R. Muntz, and H. Trotter. Waiting time distributions for Processor-Sharing systems. *Journal of the ACM*, 17:123–130, 1970.

[30] J. Cohen. Superimposed renewal processes and storage with gradual input. *Stochastic Processes and Applications*, 2:31–58, 1974.

[31] J. Cohen. The multiple phase service network with Generalized Processor Sharing. *Acta Informatica*, 12:245–284, 1979.

[32] M. Crovella and A. Bestavros. Self-similarity in World Wide Web traffic: evidence and possible causes. IEEE/ACM *Transactions on Networking*, 5:835–847, 1997.

[33] K. Dębicki and M. Mandjes. Traffic with an fBm limit: convergence of the workload process. *Queueing Systems*, 46:113–127, 2003.

[34] A. Dembo and O. Zeitouni. *Large deviations techniques and applications, 2nd edition*. Springer, New York, NY, USA, 1998.

[35] J.-D. Deuschel and D. Stroock. *Large Deviations*. Academic Press, Boston, MA, USA, 1989.

[36] E. van Doorn, A. Jagers, and J. de Wit. A fluid reservoir regulated by a birth death-process. *Stochastic Models*, 4:457–472, 1987.

[37] DSL Forum Technical Report TR-126. Triple-play services Quality of Experience (QoE) requirements, 2006.

[38] R. Ellis. *Entropy, large deviations, and statistical mechanics*. Springer, New York, NY, USA, 1985.

[39] A. Elwalid and D. Mitra. Effective bandwidth of general Markovian traffic sources and admission control of high speed networks. IEEE/ACM *Transactions on Networking*, 1:329–343, 1993.

[40] A. Elwalid, D. Mitra, and R. Wentworth. A new approach for allocating buffers and bandwidth to heterogeneous regulated traffic in an ATM node. IEEE *Journal of Selected Areas in Communications*, 13:1115–1127, 1995.

[41] P. Engelstad and O. Østerbø. Non-saturation and saturation analysis of IEEE 802.11e EDCA with starvation prediction. In *Proceedings of MSWiM '05*, pages 224–233, 2005.

[42] G. Fayolle and R. Iasnogorodski. Two coupled processors: the reduction to a Riemann-Hilbert problem. *Zur Wahrscheinlichkeitstheorie verwandte Gebiete*, 47:325–351, 1979.

[43] M. Fiedler and Å. Arvidsson. A resource allocation law to satisfy QoS demands on ATM burst and connection level. *COST257 TD(99)06*, 1999.

[44] L. Flatto. The waiting time distribution for the random order service M/M/1 queue. *Annals of Applied Probability*, 7:382–409, 1997.

[45] C. Foh and M. Zukerman. Performance analysis of the IEEE 802.11 MAC protocol. In *Proceedings of European Wireless '02*, pages 184–190, 2002.

[46] C. Fraleigh, F. Tobagi, and C. Diot. Provisioning IP backbone networks to support latency sensitive traffic. In *Proceedings of* IEEE *Infocom*, pages 375–385, 2003.

[47] Z. Fu, P. Zerfos, H. Luo, S. Lu, L. Zhang, and M. Gerla. The impact of multihop wireless channel on TCP throughput and loss. IEEE *Transactions on Mobile Computing*, 4:209–221, 2005.

[48] A. Ganesh, N. O'Connell, and D. Wischik. *Big Queues*. Lecture Notes in Mathematics, Volume 1838. Springer, Berlin, Germany, 2004.

[49] M. de Graaf, H. van den Berg, R. Boucherie, F. Brouwer, I. de Bruin, H. Elfrink, I. Fernandez-Diaz, S. Heemstra de Groot, R. de Haan, J. de Jongh, R. Núñez Queija, J.-K. van Ommeren, F. Roijers, J. Stemerdink, and E. Tromp. Easy wireless: broadband ad-hoc networking for emergency services. In *Proceedings*

of The Sixth Annual Mediterranean Ad Hoc Networking WorkShop (MEDHOC), pages 32–39, 2007.

[50] M. de Graaf, H. van den Berg, R. Boucherie, H. Elfrink, S. Heemstra de Groot, R. de Haan, A. te Marvelde, J.-K. van Ommeren, F. Roijers, J. Stemerdink, and E. Tromp. Advances in emergency networking. In *Proceedings of Wireless Rural and Emergency Communications Conference*, 2007.

[51] M. Gribaudo and M. Telek. Fluid models in performance analysis. In *SFM*, pages 271–317, 2007.

[52] R. Guérin, H. Ahmadi, and M. Naghshineh. Equivalent capacity and its application to bandwidth allocation in high-speed networks. IEEE *Journal of Selected Areas in Communications*, 9:968–981, 1991.

[53] 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.

[54] F. Guillemin, R. Mazumdar, and A. Simonian. On heavy traffic approximations for transient characteristics of $M/M/\infty$ queues. *Journal of Applied Probability*, 33:490–506, 1996.

[55] F. Guillemin and D. Pinchon. Continued fraction analysis of the duration of an excursion in an $M/M/\infty$ system. *Journal of Applied Probability*, 35:165–183, 1998.

[56] F. Guillemin and D. Pinchon. On the area swept under the occupation process of an M/M/1 queue in a busy period. *Queueing Systems*, 29:383–398, 1998.

[57] F. Guillemin and D. Pinchon. On a random variable associated with excursions in an $M/M/\infty$ system. *Queueing Systems*, 32:305–318, 1999.

[58] F. Guillemin and A. Simonian. Transient characteristics of an $M/M/\infty$ system. *Advances in Applied Probability*, 27:862–888, 1995.

[59] J. He and H. Pung. Fairness of medium access control for multi-hop ad hoc networks. *Computer Networks*, 48:867–890, 2005.

[60] J. He and H. Pung. Performance modelling and evaluation of IEEE 802.11 Distributed Coordination Function in multihop wireless networks. *Computer Communications*, 29:1300–1308, 2006.

[61] R. Howard. *Dynamic Probabilistic Systems, Volume 1*. Wiley, New York, NY, USA, 1971.

[62] IEEE Std 802.11. Part 11: Wireless LAN Medium Access Control (MAC) and physical layer (PHY) specifications, 1997.

[63] IEEE Std 802.11–2007. Part 11: Wireless LAN Medium Access Control (MAC) and physical layer (PHY) specifications. Revision of IEEE std 802.11-1999, 2007.

[64] IEEE Std 802.11b–1999 (R2003). Supplement: higher speed physical layer extension in the 2.4 GHz band. Supplement to IEEE std 802.11, 1999 edition, 1999.

[65] IEEE Std 802.11e–2005. Amendment 8: Medium Access Control (MAC) Quality of Service enhancements. Amendment to IEEE std 802.11–1999 edition (reaff. 2003), 2005.

[66] P. den Iseger. Numerical transform inversion using Gaussian quadrature. *Probability in the Engineering and Informational Science*, 20:1–44, 2006.

[67] ITU-T Recommendation G.114. One-way transmission time, 1996.

[68] J. Keilson. *Markov Chain Models – Rarity and Exponentiality*. Springer-Verlag, New York, NY, USA, 1979.

[69] F. Kelly. *Reversibility and Stochastic Networks*. John Wiley & Sons, New York, NY, USA, 1979.

[70] F. Kelly. Notes on effective bandwidths. In F. P. Kelly, S. Zachary, and I. Ziedins, editors, *Stochastic Networks: Theory and Applications*, pages 141–168. Oxford University Press, 1996.

[71] F. Kelly. Mathematical modelling of the Internet. In Bjorn Engquist and Wilfried Schmid, editors, *Mathematics Unlimited – 2001 and Beyond*, pages 685–702. Springer, 2001.

[72] G. Kesidis, J. Walrand, and C.-S. Chang. Effective bandwidths for multiclass Markov fluids and other ATM sources. IEEE/ACM *Transactions on Networking*, 1:424–428, 1993.

[73] J. Kilpi and I. Norros. Testing the Gaussian approximation of aggregate traffic. In *Proceedings of 2nd Internet Measurement Workshop*, pages 49–61, 2002.

[74] J. Kim and M. Krunz. Fluid analysis of delay and packet discard performance for QoS support in wireless networks. IEEE *Journal on Selected Areas in Communications*, 19:384–395, 2001.

[75] L. Kleinrock. *Queueing Systems. Volume I: Theory*. John Wiley & Sons, New York, NY, USA, 1975.

[76] C. Knessl and Y. Yang. Asymptotic expansions for the congestion period for the $M/M/\infty$-queue. *Queueing Systems*, 39:213–256, 2001.

[77] L. Kosten. Stochastic theory of data-handling systems with groups of multiple sources. In H. Rudin and W. Bux, editors, *Performance of Computer-Communication Systems*, pages 321–331. Elsevier, 1984.

[78] V. Kulkarni and T. Rolski. Fluid model driven by an Ornstein-Uhlenbeck process. *Probability in the Engineering and Informational Sciences*, 8:403–417, 1994.

[79] A. Law and W. Kelton. *Simulation Modeling and Analysis. Third edition*. McGraw-Hill, New York, NY, USA, 2000.

[80] W. Leland, M. Taqqu, W. Willinger, and D. Wilson. On the self-similar nature of Ethernet traffic (extended version). IEEE/ACM *Transactions on Networking*, 2:1–15, 1994.

[81] R. Lenin and P. Parthasarathy. A computational approach for a fluid queue driven by a truncated birth-death process. *Journal Methodology and Computing in Applied Probability*, 2:373–392, 2004.

[82] R. Litjens, F. Roijers, H. van den Berg, R. Boucherie, and M. Fleuren. Analysis of flow transfer times in IEEE 802.11 Wireless LANs. *Annals of Telecommunications*, 59:1407–1432, 2004.

[83] W. Liu and D.-H. Shi. Busy period in $GI^X/G/\infty$. *Journal of Applied Probability*, 33:815–829, 1996.

[84] D. Malone, K. Duffy, and D. Leith. Modeling the 802.11 distributed coordination function in non-saturated heterogeneous conditions. IEEE/ACM *Transactions on Networking*, 15:159–172.

[85] M. Mandjes. *Large Deviations for Gaussian Queues*. Wiley, Chichester, UK, 2007.

[86] M. Mandjes and J.-H. Kim. Large deviations for small buffers: an insensitivity result. *Queueing Systems*, 37:349–362, 2001.

[87] M. Mandjes and P. Mannersalo. Queueing systems fed by many exponential on-off sources: an infinite-intersection approach. *Queueing Systems*, 54:5–20, 2006.

[88] M. Mandjes, D. Mitra, and W. Scheinhardt. Models of network access using feedback fluid queues. *Queueing Systems*, 44:365–398, 2003.

[89] M. Mandjes and A. Ridder. Finding the conjugate of Markov fluid processes. *Probability in the Engineering and Informational Sciences*, 9:297–315.

[90] M. Mandjes and F. Roijers. A fluid system with coupled input and output, and its application to bottlenecks in ad hoc networks. *Queueing Systems*, 56:79–92, 2007.

[91] M. Mandjes and F. Roijers. M/M/∞ transience: tail asymptotics of congestion periods. *Submitted*, 2008.

[92] M. Mandjes, I. Saniee, and A. Stolyar. Load characterization, overload prediction, and load anomaly detection for voice over IP traffic. IEEE *Transactions on Neural Networks*, 16:1019–1028, 2005.

[93] M. Mandjes and W. Scheinhardt. A fluid model for a relay node in an ad hoc network: evaluation of resource sharing policies. *Journal of Applied Mathematics and Stochastic Analysis*, `doi:10.1155/2008/518214`, 2008.

[94] M. Mandjes and M. van Uitert. Sample-path large deviations for tandem queues with Gaussian inputs. In *Proceedings of ITC 18*, pages 521–530, 2003.

[95] M. Mandjes and B. Zwart. Large deviations for sojourn times in processor sharing queues. *Queueing Systems*, 52:237–250, 2006.

[96] M. Marcus and H. Minc. *A survey of matrix theory and matrix inequalities*. Allyn and Bacon, Rockleigh, NJ, USA, 1964.

[97] R. van de Meent, M. Mandjes, and A. Pras. Gaussian traffic everywhere? In *Proceedings of* IEEE *International Conference on Communications*, pages 573–578, 2006.

[98] D. Miorandi, A. Kherani, and E. Altman. A queueing model for HTTP traffic over IEEE 802.11 WLANs. *Computer Networks*, 50:63–79, 2006.

[99] D. Mitra. Stochastic theory of a fluid model of producers and consumers coupled by a buffer. *Advances in Applied Probability*, 20:646–676, 1988.

[100] I. Norros. On the use of fractional Brownian motion in the theory of connectionless networks. IEEE *Journal of Selected Areas in Communications*, 13:953–962, 1995.

[101] T. Oetiker. MRTG: Multi Router Traffic Grapher. Available at: `http://people.ee.ethz.ch/oetiker/webtools/mrtg`, 2006.

[102] M. Parulekar and A. Makowski. M/G/∞ input processes: a versatile class of models for network traffic. In *Proceedings of* IEEE *Infocom*, pages 419–426, 1997.

[103] V. Paxson and S. Floyd. Wide area traffic: the failure of Poisson modeling. IEEE/ACM *Transactions on Networking*, 3:226–244, 1995.

[104] J. Preater. M/M/∞ transience revisited. *Journal of Applied Probability*, 34:1061–1067, 1997.

[105] J. Preater. On the severity of M/M/∞ congested periods. *Journal of Applied Probability*, 39:228–230, 2002.

[106] E. Reich. On the integrodifferential equation of Takács I. *Annals of Mathematical Statistics*, 29:563–570, 1958.

[107] A. Remke, B. Haverkort, and L. Cloth. A versatile infinite-state Markov reward model to study bottlenecks in 2-hop ad hoc networks. In *Proceedings of QEST*, pages 63–72, 2006.

[108] A. Remke, B. Haverkort, G. Heijenk, and L. Cloth. Bottleneck analysis for two-hop IEEE 802.11e ad hoc networks. In *Proceedings of ASMTA*, pages 279–294, 2008.

[109] P. Robert. *Stochastic Networks and Queues*. Springer, New York, NY, USA, 2003.

[110] J. Roberts and L. Massoulié. Bandwidth sharing and admission control for elastic traffic. In *Proceedings of the ITC Specialist Seminar*, pages 185–201.

[111] J. Roberts, U. Mocci, and J. Virtamo. *Broadband Network Teletraffic, Final Report of European Action COST 242*. Springer, Berlin, Germany, 1996.

[112] F. Roijers, H. van den Berg, and X. Fan. Analytical modelling of TCP file transfer times over IEEE 802.11 Wireless LANs. In *Proceedings of ITC 19*, pages 1673–1686, 2005.

[113] F. Roijers, H. van den Berg, X. Fan, and M. Fleuren. A performance study on service integration in IEEE 802.11e Wireless LANs. *Computer Communications*, 29:2621–2633, 2006.

[114] F. Roijers, H. van den Berg, and M. Mandjes. Fluid-flow modeling of a relay node in an IEEE 802.11 wireless ad-hoc network. In *Proceedings of ITC 20*, pages 321–334, 2007.

[115] F. Roijers, H. van den Berg, and M. Mandjes. Performance analysis of differentiated resource-sharing in a wireless ad-hoc network. *Submitted*, 2008.

[116] F. Roijers, M. Mandjes, and H. van den Berg. Analysis of congestion periods of an M/M/∞-queue. *CWI-report*, available at `http://ftp.cwi.nl/CWIreports/PNA/PNA-E0606.pdf`, 2006.

[117] F. Roijers, M. Mandjes, and H. van den Berg. Analysis of congestion periods of an M/M/∞-queue. *Performance Evaluation*, 64:737–754, 2007.

[118] T. Sakurai and S. Hanly. Modelling TCP flows over an 802.11 Wireless LAN. In *Proceedings of European Wireless Conference*, pages 385–391, 2005.

[119] W. Scheinhardt. *Markov-modulated and feedback fluid queues*. PhD thesis, University of Twente, Enschede, The Netrherlands, 1998.

[120] W. Scheinhardt, N. van Foreest, and M. Mandjes. Continuous feedback fluid queues. *Operations Research Letters*, 33:551–559, 2005.

[121] B. Sengupta and D. Jagerman. A conditional response time of the M/M/1 processor-sharing queue. *AT&T Technical Journal*, 2:409–421, 1985.

[122] D. Shanbhag. On infinite server queues with batch arrivals. *Journal of Applied Probability*, 3:274–279, 1966.

[123] A. Shwartz and A. Weiss. *Large Deviations for Performance Analysis. Queues, Communications, and Computing*. Chapman & Hall, London, UK, 1995.

[124] C. Siva Ram Murth and B. Manoj. *Ad hoc wireless networks, Architectures and protocols*. Prentice Hall, New Jersey, USA, 2004.

[125] W. Stadje. Busy period of the queueing system M/G/∞. *Journal of Applied Probability*, 22:697–704, 1985.

[126] L. Takács. *Introduction to the theory of queues*. Oxford University Press, New York, 1962.

[127] M. Taqqu, W. Willinger, and R. Sherman. Proof for a fundamental result in self-similar traffic modeling. *ACM Sigcomm Computer Communication Review*, 27:5–23, 1997.

[128] H. Tijms. *Stochastic models: an algorithmic approach*. Wiley & Sons, 1994.

[129] H. Tijms. *A First Course in Stochastic Models*. Wiley, New York, NY, USA, 2003.

[130] B. Tsybakov. Busy periods in M/M/∞ systems with heterogeneous servers. *Queueing Systems*, 52:153–156, 2006.

[131] J. Virtamo and I. Norros. Fluid queue driven by an M/M/1 queue. *Queueing Systems*, 16:373–386, 1994.

[132] W. Whitt. On the heavy-traffic limit theorem for G/G/∞ queues. *Advances in Applied Probability*, 14:171–190, 1982.

[133] E. Winands, T. Denteneer, J. Resing, and R. Rietman. A finite-source queueing model for the IEEE 802.11 DCF. *European Transactions on Telecommunications*, 16:77–89, 2005.

[134] H. Wu, Y. Peng, K. Long, S. Cheng, and J. Ma. Performance of reliable transport protocol over IEEE 802.11 Wireless LAN: analysis and enhancement. In *Proceedings of* IEEE *Infocom*, pages 599–607, 2002.

[135] S. Zachary. A note on insensitivity in stochastic networks. *Journal of Applied Probability*, 44:238–248, 2007.

[136] J. Zhao, Z. Guo, Q. Zhang, and W. Zhu. Performance study of MAC for service differentiation in IEEE 802.11. In *Proceeding of* IEEE *Globecom*, pages 778–782, 2002.

Acronyms

AC	Access Category
ACK	ACKnowledgement
AIFS	Arbitration IFS
AIFSN	AIFS Number
ADSL	Asymmetric DSL
CoV	Coeffient of Variation
CP	Congestion Period
CSMA/CA	Carrier Sense Multiple Access with Collision Avoidance
CTS	Clear To Send
CW	Contention Window
DCF	Distributed Coordination Function
DIFS	Distributed IFS
DSL	Digital Subscriber Line
DSLAM	DSL Access Multiplexer
EDCA	Enhanced Distributed Channel Access
EIFS	Extended IFS
FCFS	First Come First Served
GPRS	General Packet Radio Service
GPS	Generalized PS
HCCA	HCF Controlled Channel Access
HCF	Hybrid Coordination Function
HSPA	High-Speed Packet Access
IEEE	Institute of Electrical and Electronics Engineers
i.i.d.	independent identically distributed
IFS	InterFrame Space
ICP	InterCongestion Period
iOU	integrated Ornstein-Uhlenbeck

IP	Internet Protocol
IP-TV	IP-Television
IS	Importance Sampling
ISP	Internet Service Provider
ITU	International Telecommunication Union
KPN	Koninklijke PTT Nederland
LAN	Local Area Network
LD	Large Deviations
LDP	Large Deviations Principle
LT	Laplace Transform
MAC	Medium Access Control
MGF	Moment Generating Function
NA	Not Applicable
PASTA	Poisson Arrivals See Time Averages
PHY	PHYsical
PLR	Packet-Loss Ratio
PS	Processor Sharing
QoS	Quality of Service
RTS	Request To Send
SIFS	Short IFS
STA	STAtion
TCP	Transmission Control Protocol
TNO	Nederlandse Organisatie voor Toegepast-Natuurwetenschappelijk Onderzoek
TXOP	transmission (TX) OPportunity
UDP	User Datagram Protocol
UMTS	Universal Mobile Telecommunications System
VoIP	Voice over IP
WLAN	Wireless LAN

Summary

Communication services continuously evolve and become more and more sophisticated. These developments are accompanied by an increasing demand for highly-performing communication networks to properly support these new services. In particular, the so-called Quality of Service (QoS) critically depends on the performance of the underlying network.

This thesis is concerned with two principal performance modeling and analysis issues in communication network environments that are currently attracting substantial attention: i) the dimensioning of IP-network links, and ii) the impact of resource sharing on the performance of wireless ad-hoc networks. These two topics are dealt with in separate parts of this thesis. We use stochastic modeling, and, in particular, queueing theory, to evaluate the performance of the mentioned issues. A central role in our research is played by so-called *fluid models*.

Part I: QoS-aware network dimensioning. In the first part of this thesis we consider the issue of IP-network link dimensioning. It is crucial to find a proper trade-off between providing sufficient capacity in order to achieve the desired QoS, and the related costs.

In Chapter 3 we develop an insightful bandwidth provisioning formula for an IP-network link; it only requires the (envisioned) load of the IP-network link as its input. The provisioning formula takes into account the characteristics of the individual flows and their QoS requirements. The criterion to determine the required capacity is that the probability that the traffic supply exceeds the available bandwidth, over some predefined (small) interval, should be below some small fixed number. The validity of the bandwidth provisioning rule is assessed through extensive measurements performed in several operational network environments.

In Chapters 4 and 5 we consider the $M/M/\infty$ queue which is used as a flow-level model for the occupancy of an IP-network link. We are particularly interested in *congestion periods*, which are defined as periods during which the offered traffic

(number of active users) is continuously above a certain value C. For the congestion periods we are interested in the following performance metrics: the duration D_C, the number of arrivals N_C, and the area A_C which is the amount of offered traffic in excess of the capacity. Knowledge of the characteristics of a congestion period is useful, for instance when dimensioning IP-network links, similar to what we did in Chapter 3. In Chapter 4 we present a procedure to compute all the moments of the quantities of the congestion periods. In Chapter 5 we address the *tail asymptotics* of these quantities. These are particularly interesting if we wish to find a value for C such that the probability that the duration of the congestion period exceeds a given threshold, is kept very small; typically in the order of 10^{-4} to 10^{-6}. We derive the tail asymptotics of the quantities, and we show that these are essentially exponential; the proof techniques stem from large-deviations theory. Additionally, we provide schemes for fast simulations, which yield a substantial speed-up in simulation effort, compared to straightforward simulations.

Part II: Impact of resource sharing on the performance of a wireless ad-hoc network. In the second part of this thesis we focus on wireless ad-hoc networks; these networks can be deployed instantly without a fixed infrastructure or predetermined configuration. An important feature of ad-hoc networks is multi-hop connectivity, i.e., if a node is not directly connected to its destination, it can use other nodes to relay its traffic. The wireless nodes share the radio transmission capacity; therefore, a node which is heavily used by many other nodes, is prone to becoming a performance bottleneck.

In Chapter 6 we introduce our fluid-modeling approach of a varying number of source nodes that transmit flows of data via a common relay node to their destinations. Our main performance metric of interest is the overall flow transfer time, i.e., the time required to entirely transmit a flow from a source node to its destination.

In Chapter 7 we present a mean-value analysis of the fluid model, and, in particular, we investigate the impact on the performance metrics when the relay node may obtain a share of the capacity that is m times as large as the share that each of the source nodes receive. In the analysis we first consider the special case of exponential flow sizes; we analyze the source-node dynamics and the workload at the relay node by our fluid-modeling approach. Then we observe from extensive numerical experimentation over a broad set of parameter values that the distribution of the number of active source nodes is actually *insensitive* to the flow-size distribution. Using this remarkable result as an approximation assumption, we obtain explicit expressions for the mean workload at the relay node and the overall flow transfer time for general flow-size distributions. Finally, we illustrate that the overall flow transfer time of a multi-hop flow improves by granting the relay node a larger share of the medium capacity.

In Chapter 8 the fluid model is validated to accurately describe the behavior of wireless ad-hoc networks based on IEEE 802.11 WLAN technology. We first pro-

pose a mapping between the parameter settings of IEEE 802.11 WLAN and the fluid model, and then we validate the fluid model and parameter mapping by comparison with ad-hoc network simulations that include all the details of the IEEE 802.11 MAC-protocols. The numerical results show that the model accurately captures the behavior of an IEEE 802.11 wireless ad-hoc network.

In Chapter 9 a special case of the fluid model is studied in more detail, i.e., the fluid model with equal sharing of the capacity (viz. $m = 1$) and exponentially distributed flow-sizes. As a result of these assumptions the model under consideration can be regarded as a queueing system with Markov fluid input. We compute the Laplace transforms of the performance measures of our interest. Furthermore, we determine the exponential decay rates of the corresponding tail probabilities, relying on large-deviations theory.

Samenvatting

Communicatiediensten worden steeds functioneler. Een direct gevolg is dat ze steeds veeleisender worden ten aanzien van de communicatienetwerken waarvan ze gebruikmaken. Deze netwerken moeten zorgen voor een snelle en betrouwbare afhandeling van het verkeer dat door de diensten wordt gegenereerd. De kwaliteit van een dienst is dan ook direct afhankelijk van de prestaties van de onderliggende transportnetwerken.

In dit proefschrift beschouwen we de volgende onderzoeksonderwerpen m.b.t. de prestaties van communicatienetwerken: i) het dimensioneren van verbindingen in een IP-netwerk, en ii) het bepalen van de invloed van capaciteitsdeling op de prestaties van draadloze ad-hoc netwerken. Bij het onderzoek naar beide onderwerpen maken we gebruik van stochastische modellering, en in het bijzonder, de wachtrijtheorie. Hierbij staat het gebruik van zogeheten vloeistof-wachtrijmodellen centraal; dit zijn modellen waarbij ervan uitgegaan wordt dat de gebruikers hun verkeer aanbieden als een continue stroom, als ware het 'vloeistof'.

Deel I: kwaliteitsbewust dimensioneren van netwerken. Het eerste deel van dit proefschrift is gewijd aan het dimensioneren van verbindingen in een IP-netwerk. De uitdaging is om enerzijds alle diensten binnen hun gestelde kwaliteitseisen te behandelen, maar anderzijds dit met zo min mogelijk capaciteit te bewerkstelligen vanwege de daaraan gerelateerde kosten.

In Hoofdstuk 3 ontwikkelen we een inzichtelijke dimensioneringsformule voor een IP-netwerk verbinding welke slechts de gemiddelde verkeersbelasting van de verbinding als input nodig heeft. De dimensioneringsformule houdt rekening met de karakteristieken van de individuele verkeersstromen en hun kwaliteitseisen. Het dimensioneringscriterium is dat de kans dat het aangeboden verkeer gedurende een vooraf gedefinieerde kleine tijdsduur groter is dan de capaciteit, kleiner moet zijn dan een bepaalde (kleine) waarde. De validiteit van de dimensioneringsregel is uitgebreid onderzocht door middel van metingen in verschillende operationele

netwerkomgevingen.

In Hoofdstukken 4 en 5 beschouwen we een $\mathrm{M}/\mathrm{M}/\infty$ wachtrijmodel dat kan worden gezien als een vloeistof-wachtrijmodel van de bezetting van een verbinding in een IP-netwerk. We zijn met name geïnteresseerd in *congestieperioden*, dit zijn perioden gedurende welke het aantal klanten continu hoger is dan een bepaald niveau C. Interessante metrieken gerelateerd aan een congestieperiode zijn: de duur D_{C}, het aantal nieuwe klanten N_{C} dat aankomt gedurende de congestieperiode en de hoeveelheid verkeer A_{C} dat in surplus boven C wordt aangeboden. Inzicht in de karakteristieken van congestieperioden is nuttig voor bijvoorbeeld het dimensioneren van verbindingen in een IP-netwerk, zoals in Hoofdstuk 3. In Hoofdstuk 4 presenteren we procedures waarmee alle momenten van de metrieken bepaald kunnen worden. In Hoofstuk 5 bestuderen we de staartkansen van de metrieken van een congestieperiode. Deze zijn interessant voor het dimensioneren, waarbij we de benodigde capaciteit C zo willen bepalen dat de kans zeer klein is dat een congestieperiode langer duurt dan een bepaalde tijd; typisch is deze kans in de orde van 10^{-4} tot 10^{-6}. Gebruikmakend van de theorie van grote afwijkingen tonen we aan dat de staarten van de verdelingen in essentie exponentieel zijn. Daarnaast bepalen we nog het staartgedrag met 'importance sampling', een methode om op efficiënte wijze kleine kansen d.m.v. simulaties te schatten.

Deel II: de invloed van capaciteitsdeling op de prestaties van draadloze ad-hoc netwerken. In het tweede deel beschouwen we draadloze ad-hoc netwerken; dit zijn netwerken die kunnen worden opgezet zonder onderliggende of vooraf gedefinieerde infrastructuur; de gebruikers zijn zelf de knooppunten van het netwerk. Een belangrijk aspect van draadloze ad-hoc netwerken is multi-hop connectiviteit, d.w.z. dat gebruikers die niet direct verbonden zijn met hun eindbestemming, hun verkeer door andere gebruikers kunnen laten 'doorsturen'. Alle gebruikers delen een gezamenlijke radioverbinding die een beperkte capaciteit heeft; men kan zich voorstellen dat gebruikers die vaak verkeer van andere gebruikers moeten doorsturen, een knelpunt kunnen worden.

In Hoofdstuk 6 introduceren we een vloeistof-wachtrijmodel van een gebruiker die door een variërend aantal andere 'brongebruikers' wordt gebruikt als tussenstation om hun uiteindelijke bestemmingen te bereiken. De voornaamste prestatiemaat waarin we geïnteresseerd zijn, is de totale transmissieduur van een file, oftewel de tijd die nodig is om een file geheel van een brongebruiker naar de eindbestemming te versturen.

In Hoofdstuk 7 bepalen we de verwachtingswaarden van de prestatiematen van ons model; in het bijzonder onderzoeken we het geval waarbij het tussenstation een m keer zo groot deel van de gezamenlijke capaciteit mag benutten als ieder van de brongebruikers. In de analyse beschouwen we eerst het speciale geval van exponentieel verdeelde filegroottes; hiervoor onderzoeken we het gedrag van de bronge-

bruikers en de werklast (in de buffer) bij het tussenstation. Aan de hand van een uitgebreide simulatiestudie observeren we vervolgens dat de verdeling van het aantal actieve brongebruikers *ongevoelig* is voor de verdeling van de filegroottes. Dit opmerkelijke resultaat gebruiken we als een benaderingsaanname, waarmee we expliciete uitdrukkingen verkrijgen voor de verwachting van de werklast bij het tussenstation en van de totale transmissietijd voor algemeen verdeelde filegroottes. Tot slot laten we zien dat de totale transmissietijd aanzienlijk verbetert door het tussenstation een groter deel van de gezamenlijke capaciteit toe te kennen.

In Hoofdstuk 8 valideren we dat het vloeistof-wachtrijmodel het gedrag van een draadloos ad-hoc netwerk, gebaseerd op IEEE 802.11 WLAN technologie, accuraat modelleert. Hiervoor leggen we eerst de relatie tussen de parameters van de IEEE 802.11 WLAN protocollen en de instellingen van de parameters van ons vloeistof-wachtrijmodel. Vervolgens vergelijken we het model met ad-hoc netwerk simulaties die alle details van de IEEE 802.11 MAC-protocollen bevatten. De verkregen numerieke resultaten tonen aan dat het vloeistof-wachtijmodel het gedrag van een IEEE 802.11 draadloos ad-hoc netwerk nauwkeurig beschrijft.

In Hoofdstuk 9 wordt een speciaal geval van het model bestudeerd, namelijk het vloeistof-wachtrijmodel met $m = 1$ en exponentieel verdeelde filegroottes. Als gevolg van deze aannames kunnen we de buffer bij het tussenstation beschouwen als een Markov-systeem met vloeistof-input. We bepalen de Laplace getransformeerden voor verschillende prestatiematen. Daarnaast laten we zien, m.b.v. de theorie van grote afwijkingen, dat de staartkansen exponentieel afnemen.

About the author

Frank Roijers was born on July 4, 1975 in Rotterdam and completed grammar school at the Alfrink College in Zoetermeer in 1993. In 1998 he received his master's degree in Bedrijfswiskunde en Informatica (BWI) from the Vrije Universiteit, Amsterdam after a 9 month internship at Shell Research and Technology Centre, Amsterdam (SRTCA). The year 1999 was well-spent on first earning a travel budget and subsequently spending it while backpacking in Southeast Asia.

In 2000 he started his professional career by joining the quantitative optimization department of KPN Research, the research department of KPN which is the largest telecommunication operator in the Netherlands. In 2003 KPN Research was acquired by the Netherlands Organisation for Applied Scientific Research (TNO), and it became the basis of the newly-formed knowledge area TNO Information and Communication Technology.

The more knowledge-oriented approach of TNO offered Frank the opportunity to start working towards a Ph.D, which he began in 2004 under the supervision of prof. dr. Hans van den Berg (University of Twente, TNO) and prof. dr. Michel Mandjes (University of Amsterdam, CWI). This research was conducted in collaboration with the Centrum Wiskunde & Informatica (CWI) and the Korteweg-de Vries Institute (KdVI) for Mathematics of the University of Amsterdam (UvA). From 2006 till 2008 he was also part-time employed at the KdVI. In 2008 he was a visiting researcher at the TMRC (Telecommunication Mathematics Research Center) at Korea university, Seoul, Republic of Korea, for a period of two months. Sofar, his academic research has led to about 15 scientific papers, of which [13, 14, 90, 91, 114, 115, 117] are used as the basis of this thesis.

Currently, the author is fulltime employed as an ICT consultant at TNO Information and Communication Technology.