VRIJE UNIVERSITEIT

RARE EVENT ANALYSIS OF COMMUNICATION NETWORKS

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor aan
de Vrije Universiteit te Amsterdam,
op gezag van de rector magnificus
prof.dr. E. Boeker
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de faculteit der economische wetenschappen en econometrie
op dinsdag 3 december 1996 te 15.45 uur
in het hoofdgebouw van de universiteit,
De Boelelaan 1105

door

Michael Robertus Hendrikus Mandjes

geboren te Zaandam

Promotor:	prof.dr. H.C. Tijms
Copromotor:	dr. A.A.N. Ridder
Referent:	prof.dr. A. Hordijk

Dankwoord

Dit proefschrift doet verslag van het door mij verrichte onderzoek bij de vakgroep Econometrie van de Vrije Universiteit Amsterdam in de periode 1 april 1993 tot 15 september 1996. In de eerste plaats betreft het theoretisch onderzoek naar de probabilistische eigenschappen van zeldzame gebeurtenissen in wachtrijsystemen. Hierbij is echter getracht het belangrijkste veld van toepassingen – namelijk telecommunicatiesystemen – niet uit het oog te verliezen. De wiskundige theorie die ten grondslag ligt aan dit proefschrift staat bekend onder de naam 'large deviations'. Pas gedurende de laatste tien jaar werd het nut van deze theorie met betrekking tot telecommunicatietoepassingen op juiste waarde geschat. Het ziet er naar uit dat large deviations ook in de komende jaren een belangrijke rol zal blijven spelen.

Mijn begeleider tijdens mijn onderzoek was *Ad Ridder*. Hoofdstuk 3 van dit proefschrift berust op gezamenlijk werk, hetgeen niet wegneemt dat Ads aandeel in de andere hoofdstukken ook aanzienlijk is. Vele malen heeft hij tussentijdse versies van mijn artikelen gelezen, waarbij hij vaak met bruikbare kritiek kwam. Daarnaast heb ik ervaren dat Ad buiten het werk om een heel plezierige persoon is om mee om te gaan. Op wat meer afstand werd ik begeleid door mijn promotor *Henk Tijms*. Vooral zijn onuitputtelijke bron van contacten is voor mij van onschatbare waarde geweest. Hoofdstuk 8 heb ik samen met *Richard Boucherie* (Universiteit van Amsterdam) geschreven. Het was interessant om te zien hoe twee gebieden uit de kansrekening (large deviations en productvormen) elkaar kunnen aanvullen.

Gedurende de maanden mei, juni en juli 1994 heb ik gewerkt op het Dr. Neher Laboratorium van KPN Research in Leidschendam. Daar bleek dat mijn theoretisch onderzoek naar efficiënte simulatiemethoden ook bruikbaar was met het oog op het oplossen van praktische kwesties, met name netwerkontwerp. Ik kijk terug op een geslaagde samenwerking met *Eric Smeitink* en *Hans van den Berg*. Samen met die laatste heb ik hoofdstuk 6 geschreven.

Verder ben ik dank verschuldigd aan *Phuoc Tran-Gia*, die me heeft uitgenodigd om gedurende de maanden mei, juni en juli 1996 onderzoek te verrichten aan de Bayerische Julius-Maximilians Universität in Würzburg, Duitsland. Ik heb me hier bezig gehouden met een relatief praktijkgericht project op het gebied van mobiele communicatie. Dat het

behalve een leerzame ook een plezierige tijd was, was mede te danken aan mijn collega's, met name *Notker Gerlich*, *Oliver Rose* en *Kurt Tutschku*.

Ook de leden van de vakgroep Econometrie van de Vrije Universiteit wil ik bedanken. Hierbij wil ik in het bijzonder mijn kamergenoten noemen die ik de afgelopen jaren heb gehad, namelijk *Wilko Bolt*, *Frank Gouweleeuw*, *Kees van der Hoeven*, *Andries Lenstra*, *Harro Merkus* en *Richard Venniker*, maar daarnaast ook *Govert Bijwaard*, *Maarten Cornet*, *Pierre Koning* en *Bert Tieben*. Vooral de vrijdagmiddagen zal ik mij nog lang heugen: dankzij het fenomeen 'bruin café' kon je op voorhand niets zinnigs zeggen over het verdere verloop van de dag. Gelukkig gaven onze 'rondjes Bosbaan' het nodige tegenwicht.

Ik wil ook nog vermelden dat gedeelten van het manuscript gecorrigeerd zijn door *Werner Scheinhardt*, *Hans van den Berg* en *Isaco Meilijson*. *Frank Gouweleeuw* en *Edo Kulkens* hebben veel moeite gedaan om de plaatjes op de juiste plaats te krijgen. *Ruud Koning* is behulpzaam geweest bij het afdrukken van het document.

Tenslotte wil ik mijn paranimfen *Paul Jansen* en *Menno van der Hoorn* bedanken voor zo'n 15 jaar trouwe vriendschap. Verder kan ik het niet nalaten *Gert-Jan Brok*, *Edwin Leuven*, *Dymph van der Maeden*, *João Teixeira* en *Marcus Westra* te noemen. Maar mijn laatste woord van dank gaat uit naar *mijn ouders*, *Francesca*, *Paul* en *Wilbert*. Hun steun was onmisbaar.

Amsterdam, september 1996
Michel Mandjes

Contents

Chapter 1

Rare event analysis of communication networks

The subject of this monograph is the performance analysis of a particular class of communication networks: *multiservice networks*. In this first chapter, we start by giving a brief survey on some technical issues of these networks and explain the design and control problems that have to be solved. The second section shows how to put these matters into a mathematical framework: the problems can be approached by applying probabilistic techniques, e.g., *queueing theory*. Section 3 provides an overview of known results from queueing theory, particularly those of interest in view of design and control of multiservice networks. One important conclusion from Section 3 is that *large deviations* theory is an important tool in performance evaluation, in particular rare event analysis. This theory, to which the greater part of this thesis is devoted, is described in more detail in Section 4. Section 5 outlines this monograph and explains the relationship between the chapters.

1 Introduction: Multiservice networks

The interaction between science and technology on the one hand and society on the other hand can be explained in several ways. Two extreme points of view are the following. The first assumes a mechanism in which technological innovation is basically caused by new scientific and technological developments (*technology push*). Technology is the 'engine of growth'; demand from society follows the supply of new products. The second says the opposite: new technologies are demand-led, i.e., they are developed as a consequence of consumers' demand (*demand pull*).

The growing interest in today's world in high-speed communication networks can be explained from both mechanisms. New technologies (in particular fiber optic technology) enable to transmit and switch data at very high rates. The capacities involved are nowadays typically on the order of gigabits (10^9 bits), per second. On the other hand, there

is a quickly growing demand for new communication systems. Consumers' requirements are shifting both in quantitative and qualitative sense: there is a need for the extension of already existing services (telephony, electronic mail), but also new sophisticated services were recently introduced or will be introduced in the near future (high speed data, video conferencing, home banking, high-definition TV, multimedia, etc.).

1.1 Broadband ISDN and Asynchronous Transfer Mode

The future standard for high-speed networks will be B-ISDN (*Broadband Integrated Services Digital Network*). The two main issues of B-ISDN, in which it differs from 'traditional' communication networks, are 'broadband' and 'integrated services'. 'Broadband' refers to the fact that data is sent through the network at very high speed, 'integrated services' to the fact that B-ISDN supports various kinds of services, sharing common network resources.

There is a number of motives behind the development of B-ISDN. First, B-ISDN's broadband character enables to provide high bandwidths, required to support (future) sophisticated services as high speed data, video, etc. Apart from that, due to its integrated character, it is able to support services in which multiple types of traffic are involved simultaneously, for instance videophone or data transmission accompanied by speech. The resulting network is very flexible: for the introduction of a new service no new network needs to be developed. As a consequence, the development of B-ISDN is attractive for both datacom and telecom industry.

The *Asynchronous Transfer Mode* (ATM) concept is the widely agreed upon concept for the transfer of information across the user-network involved in B-ISDN. It is being standardized by organizations as the International Telecommunications Union (ITU) as well as the ATM Forum. The implementation of ATM can be briefly summarized as follows. Suppose a message has to be sent through the network.

- The data traffic, sent by the user, is packetized into ATM cells, i.e., it is assembled into fixed size elementary transfer entities. These cells consist of a 48 byte information field (reserved for user information), as well as a 5 byte header (containing the 'label', i.e., information required by the network to send the cell from the source to the destination).
- Before the cell transmission can occur, a connection is set-up by a signalling procedure. This procedure allocates the resources that are required by the call. Cells belonging to the same connection follow the same route (the communication is therefore called *connection oriented*).
- The data cells that make use of the same link are merged (or multiplexed) on a first-come-first-serve basis into one single stream. The data units of the individual streams can

be identified by means of their header label. In this respect, it is interesting to compare ATM with another way of data transfer: *Time Division Multiplexing* (TDM). In TDM, the time axis is divided into intervals (or slots) of fixed length. Suppose the network is shared by n connections, then stream i may offer its data during the ith slot in a n-slot frame. Clearly, the individual streams are identified by their position within a time frame, so no header is needed. We say that TDM provides synchronous transfer and ATM asynchronous transfer.

- At arrival at the destination, the ATM cells that belong to a particular connection are reassembled. This is simplified by the fact that all cells of one connection follow the same path, implying that there is no problem of packet sequencing. Of course, in case of a per packet routing this reassembly would be more troublesome.
- After the termination of the call, the occupied network resources are freed.

1.2 Advantages and disadvantages of ATM

The huge advantage of asynchronous transfer is the so-called *statistical multiplexing* effect. This effect can be explained as follows. Suppose a particular link is shared by a number of connections. During a connection, the associate traffic stream is sometimes active, sometimes not (for example telephone traffic). Of course it would be safe to reserve as much bandwidth as is required if all sources are active. This is called peak rate allocation. However, in practice, not all connections on a particular link will be active simultaneously. Therefore, the bandwidth allocation can be done quite efficiently: it suffices to reserve less than peak rate in order to provide a given quality of service (i.e., a very small fraction of cells that are lost, for instance 10^{-9}). Clearly, this is an *efficiency* gain with respect to synchronous transfer, where peak rate allocation is necessary: silences of one source cannot be used by other (active) sources.

In addition, ATM is very *flexible*. It is relatively easy to add a new service to the existing network, as explained above.

There are some disadvantages of integrating several classes of traffic as well. First, the traffic streams offered to the network by sources of different types can be very heterogeneous. Any type of traffic (data, voice, ...) has its own characteristics and service requirements. For data, it is important that the traffic arrives at the destination, the delay is less relevant. For voice the opposite applies: quite a large loss can be tolerated, but the delays may not be very large. In principle, however, ATM handles different classes of traffic (with different burstiness conditions and *Quality of Service* (QoS) criteria) in the same manner. Unless priority mechanisms are introduced, this situation is not optimal for both data and voice traffic. To overcome this problem, also separate buffering of different

traffic classes is proposed.

Also, the asynchronous transfer has some disadvantages. In ATM, a cell header is required, involving some overhead cost, whereas in TDM no header is needed. In fact, the efficiency and flexibility of ATM imply complexity: the traffic streams inside the network are much more difficult to manipulate and control than in TDM.

Finally, the choice of the size of the ATM cell is a compromise. To satisfy the specific requirements of their own clients, telecom and datacom industry had a different preference with respect to the size of the ATM cell. The 'ideal' size of an information packet for voice traffic is relatively small (a 32 byte information field, increased by a 4 byte header), leading to small transfer delays; the 'ideal' size for data traffic $(64 + 5)$ is large, reducing the ratio of overhead. The choice of $48 + 5$ byte is simply an average.

1.3 Network design and traffic management

Problems that arise with respect to the implementation of ATM can be viewed at two levels. The first level is *network design*, i.e., how to develop the network itself, to provide service in the required grade. The solution must be chosen in an economically optimal way, i.e., at lowest cost. The second level is *traffic management* (or traffic control): if the network is in use, what actions have to be performed to maintain the desired service level (satisfying the QoS criteria of the users at lowest cost). This issue requires the development of effective mechanisms for controlling the offered traffic. An important distinction between these levels is that the second kind consists of decisions that have to be taken in *real time*. On the contrary, the calculations needed for network design have less stringent time restrictions.

NETWORK DESIGN can be done by a concept that basically consists of two steps. These steps have to be performed sequentially, but an iterative approach might lead to better results.

- In the first phase the most comprehensive view of the network is considered, i.e., the *topology* of the network has to be determined. It must be decided where to situate the nodes, how to interconnect them, etc. These decisions are based on factors like the equipment to be used, rough estimates for the concentration areas of potential users, and the expected traffic volumes. Subnetworks can be identified, using techniques closely related to combinatorial optimization.
- Given this topology, the network resources must be allocated. This is a *dimensioning* problem: how to choose the buffers of the nodes and the capacities of the links in the network, in order to meet all constraints (with respect to loss and delay), of course, at lowest cost. Clearly, a detailed description of the offered traffic is required to perform

this resource allocation adequately. Here techniques from stochastic analysis, particularly queueing theory, can be applied.

TRAFFIC MANAGEMENT (traffic control) can be viewed as the set of actions, carried out while the network is in use, to make the network behave as desired. We first distinguish different time scales, to which these functions are related.

First, we have *call level* (or connection level): the time scale on which a connection exists, which varies from minutes to several hours. During a call there are different 'states of activity': the source can transmit at its peak rate, at zero rate (a silence during a phone call, for instance), or somewhere in between. This time scale, on which the source alternates between the states, is on the order of microseconds to seconds. A period of being active is called a 'burst', and therefore this level is called *burst level* (or activity level). During bursts, traffic is generated more or less homogeneously in time, which gives rise to approximating the traffic by 'fluid'. Of course, then we abstract from patterns at *cell level*: the level at which the individual ATM cells can be identified. The time scale of this level is typically on the order of microseconds. Having introduced the time scales, we can list some mechanisms to control the communication traffic. These functions protect the network against traffic congestion, in the sense that they enable the network to fulfill the agreed Quality of Service.

Connection acceptance control and call routing are traffic control actions on call level.

- *Connection acceptance control* (or call admission control) controls whether the network resources are sufficient to accept a new connection, without violating the service requirements of all users. When a request for a new connection is initiated (during the 'set-up phase'), it must notify the network its 'parameters'. These parameters consist on the one hand of traffic characteristics (traffic rates, burst lengths, etc.). On the other hand, the user lets the network know its desired quality of service (maximum tolerable loss and delay). Based on this information, and the information of the connections already accepted, it must be decided to accept or reject the call.

- *Call routing* is closely related to connection acceptance control. When the network determines that a connection can be accepted, it must explicitly choose the route from source to destination. Of course, this must be done without violating service requirements of the new, accepted user as well as the users that are already active.

There is also a number of traffic control techniques on burst/cell level, such as user parameter control, shaping, and adaptive rate control.

- *User parameter control.* As said, as a call arrives, it promises to satisfy several traffic characteristics. In return, the network promises to meet the required service level. This can in fact be seen as a contract, negotiated between the network and the user. A typical contract consists of traffic characteristics as for instance the mean rate, peak rate, and mean burst length. However, it cannot be said that the user always obeys its promised activity level: some users attempt to abuse the network or they underestimate their bandwidth requirements. Therefore monitoring of the traffic characteristics is necessary. The User Parameter Control (or 'policing function') is the enforcement that the user is 'confirming', in the sense of transmitting at a larger (mean, peak) rate than mutually agreed upon.

  A way to implement such a protection against 'non-confirming' users, is by a so-called Leaky Bucket mechanism . This mechanism allows transmitting at a larger rate than peak rate, but only temporarily. The cell stream that has to be policed, feeds into a 'pseudo-queue', which is served at the so-called leak rate. When the buffer of this queue runs full, the source has violated its contract, and it is decided that cells of the original traffic stream are discarded. Several algorithms have been proposed to police for instance mean input rate, peak rate or burst length.

- *Traffic shaping.* It is useful to develop methods to decrease the variability of the traffic streams, in order to increase network efficiency (i.e., utilization). For instance, by inserting buffers that are emptied at a constant rate smaller than the peak rate of the incoming traffic, the peak rate of the outcoming traffic is reduced. This kind of procedures is known as traffic shaping. Notice that, in the above-mentioned example, the variability indeed decreases, but delays become larger.

- *Adaptive rate control* is a reactive traffic control action, where user parameter control and shaping are preventive. If congestion occurs, a message is sent to the source, saying that the offered input rate is too high in order to provide the desired QoS. An other reactive mechanism is to discard cells that cause congestion (where the choice between cells is based on information in the header of a cell, that tells the importance of the cell). The last mechanism can be viewed as a kind of *priority* structure.

1.4 Literature

A comprehensive treatment of the ATM concept is provided by Onvural [140] and de Prycker [45]. Coombs, Saviotti, and Walsh [40] comment on the demand pull/technology push debate. Le Boudec [114] is a tutorial, giving some insight into the technical issues of the implementation of ATM. The comparison with TDM can be found in de Vries

[49]. The scheme for network design is based on the report edited by Roberts [155], that also provides an extensive round-up concerning performance evaluation and design of multiservice networks. The three time scales are due to Hui [89]. Furthermore, the list of traffic control measures are adapted from de Vries [49] and Eckberg [56]. The Leaky Bucket algorithm goes back to Turner [179], and is also described in [114] and [155]. Adaptive rate control, in conjunction with the available bit rate service for ATM is explained in Smith, Adams, and Tagg [168]. Stallings [169] summarizes the more technical aspects of ISDN and ATM, in a very general setting. The introduction of Awater [10] gives a brief exposition of recent developments and a look-ahead at the future of multiservice networks, featuring ATM.

2 Queueing network modeling

In the previous section we found that, to achieve the desired grade of service at lowest cost, proper network design and traffic control should be performed. In order to do so, an accurate model has to be developed. Such a model must be simple enough to allow mathematical analysis, or, at least, simulation. On the other hand, it must capture the congestion features that emerge in real communication networks. A proper balance must be chosen between tractability of the resulting mathematical model and the accuracy of the description of the communication system. It appeared that performance analysis of quite large and complex systems is enabled by a *queueing network* modeling.

This modeling basically consists of two elements. The first is *traffic modeling*, aiming to describe the statistical properties of the traffic offered by the network users. In the second place, we have to develop *network modeling*: consisting of topology of the network, buffers, service rates, etc.

2.1 Traffic modeling

Traffic modeling (or source characterization) comprehends the statistical behavior of the traffic offered to the network. Following Hui [89], we can now make a distinction between traffic description on call, burst, and cell scale, as introduced in Section 1.

• CALL LEVEL. Mostly it is assumed that customers of the multiservice network (i.e., requests for a connection) arrive according to some stochastic process, for instance a Poisson process. In general, Poissonian input describes reality reasonably accurate, since there is a large number of potential users, each with a small probability of requesting a connection, see Tijms [173, p. 27-28]. In an ATM network typically multiple types of users are involved. Their call arrival processes can be modeled by introducing multiple

Poisson arrival streams, each with a specific arrival rate.

• BURST LEVEL. We now consider a more detailed time scale: the statistical properties of the traffic (ATM cells) generated by a specific customer, for instance a voice, video, or data source. In 'early' studies (till, say, 1980) the interarrival times of cells were often taken independent and identically distributed (i.i.d., so-called 'renewal input'). However, this assumption is irrealistic for communication applications, e.g. the modeling of ATM traffic. Within a call, traffic generated by a source can be described by a process alternating between busy (or on-) times and idle (or off-) times. During an on-time (*burst*) cells arrive with relatively short interarrival times, during an off-time (*silence*) no traffic arrives. Notice that a correlation structure is introduced: the interarrival times are no longer i.i.d. In this way, we arrive at burst level: the time scale at which sources alternate between the on and off states.

The above mentioned alternation can be modeled by Markov modulation, which was, among others, introduced by Neuts [134], [135]. A Markov modulated Poisson arrival process is a doubly stochastic Poisson process, based on a two-state continuous time Markov chain. If this chain is in the first state, arrivals occur according to a Poisson process with a certain rate; in the other state no arrivals are generated. Implication is that the arrival process is a short-range dependent process: the correlations between the traffic arrived in time intervals that lie k units time apart will fade out, exponentially in k. During bursts, traffic arrives more or less homogeneously in time, giving rise to a (continuous) 'Markov fluid' approximation: the activity level of the source alternates between bursts (generating cells at a constant rate of r per unit time) and silences (transmitting at zero rate), see e.g. Kosten [109]. Notice that, consequently, the detailed behavior of cells is ignored.

However, research showed that some kinds of traffic show burstiness at a large range of time scales. For this type of arrival processes, there is no natural length of a burst, since at a wide range of time scales (ranging from milliseconds to hours) similar traffic patterns can be observed. For that reason, the arrival process is called *self-similar*. A natural way to show self-similarity (or *fractal* behavior) is the following. Plot the arrival rate as a function of time at different time scales. In case of self-similar traffic, after rescaling the vertical axis, it cannot be said which picture belongs to which time scale. This kind of traffic is characterized by long-term dependencies, and, as a consequence, cannot be captured within a Markov modulated framework (which only generates short-term dependency). A lot of qualitative relations derived for arrivals with short-term dependencies (as Markov modulated traffic) do not apply to self-similar traffic. For instance, aggregating traffic streams does not 'smooth' its statistical behavior: instead of a statistical multiplexing effect, the burstiness (i.e., the degree of self-similarity) is intensified [115].

Early references on self-similarity are Fowler and Leland [69], who describe this phenomenon for Local Area Networks, and Leland, Taqqu, Willinger, and Wilson [115], concerning Ethernet traffic. They give a mathematical description of the notion of self-similarity. 'Long-range dependence' means that the autocorrelation between the traffic arrived in slot t and in slot $t+k$ decays hyperbolically in k; in case of short-range dependence this decay is exponentially fast. A measure of long-range dependency is the so-called Hurst-parameter H; $H = 1/2$ is Brownian motion, where $H > 1/2$ shows long-range dependent behavior. Robert [153] mimics self-similar input (on a finite time-scale) by Markov modulated processes, with the intention to use the results for Markov modulated traffic, in order to analyze queues with self-similar input.

• CELL LEVEL. Finally, we arrive at the most detailed time scale: *cell scale.* At the cell scale, we fix the calls that are connected as well as their 'transmission states' (burst or silence). Under these conditions, the ATM cells arrive almost equidistant in time. Most of the time, the aggregate arrival rate is smaller than the service rate. However, due to the asynchronous character of the system, congestion can occur: if several cells arrive 'simultaneously', they cannot be served immediately. The deterministic services make discrete-time models particularly applicable, see Chapter 6 of Roberts [155] and Bruneel and Kim [22] for a survey on these.

COMBINATIONS OF DIFFERENT LEVELS. Models on burst and cell scale can be combined in an integrated model. Examples are Heffes and Lucantoni [84] and other references in Chapter 8 of Roberts [155]. Clearly, for smaller buffers, a few cells can cause a buffer overflow, so the cell-scale effects are dominant. For large buffers, there must have been extremely long bursts and short silences: then burst-scale effects determine the loss behavior. In Hübner and Ritter [87] a model, that covers call as well as burst blocking, is addressed. Hübner and Tran-Gia [88] even consider a three level model: an approximation of the loss probability is given, considering fluctuations on call, burst and cell scale.

BY-PASSING THE MODELING PHASE. Nearly all methods first choose a statistical model, then estimate the parameters, and finally compute or approximate the performance measures as the loss fraction or percentiles of the delay distribution. Several objections can be made to this procedure: there are no strict guidelines to select the statistical model, the number of parameters describing the traffic behavior may be large, the estimation of the parameters can be cumbersome, etc. To overcome these problems, methods have been proposed to skip the modeling and estimation phases. Duffield, Lewis, O'Connell, Russell, and Toomey [54] estimate the so-called entropy of the input traffic, which is directly related to the bandwidth that is required in order to satisfy the QoS requirements. In this way rough estimates for the loss fraction can be derived. Courcoubetis, Kesidis, Ridder,

Walrand, and Weber [42] develop an algorithm covering routing and call admission control without using a model description: every switch of the network constantly observes its spare capacity and extrapolates this to the situation that an additional source were connected.

2.2 Network modeling

After having described the input processes, we now focus on the description of the ATM network itself. The network can be seen as a set of nodes connected by links. At the nodes, connection requests arrive according to some arrival process, see the call level above. Then it is decided whether the call is accepted or not, and a route is determined. Once the source is accommodated, it generates traffic (ATM cells, see burst and cell level) that has to be sent along the nodes on the route towards its destination. In fact all nodes are queues, where the arriving traffic has to be 'processed' ('served'). In other words: at the network nodes, several traffic streams are multiplexed. In ATM networks the service is done at a constant rate (link capacity). If the arrival rate temporarily exceeds the service rate, cells are queued: they are put into a buffer. This buffer has limited size, implying the possibility of cells being lost. So, in fact, each queue is characterized by its service rate and buffer size.

Having described the characteristics of the offered traffic and the ATM network, we have finished our queueing model. Obviously, queueing analysis considerably eases the two tasks formulated in the previous section. If one is able to calculate relevant performance measures (cell loss fractions, waiting time percentiles, blocking probabilities) for queueing systems with given topology, buffers and service rates, one can perform proper *network design.* In either an explicit or an implicit way, the sensitivity of the performance of the system as a function of the design parameters is described. This knowledge helps us to design the network in an optimal way. As indicated in Section 1, combinatorial techniques are very useful as well with respect to network design, especially in order to perform adequate topology design.

On the other hand, the *traffic management* mechanisms can be formulated as queueing phenomena; e.g. de Veciana, Kesidis, and Walrand [48] provide a traffic management framework in terms of queueing analysis. Call acceptance control can simply be performed by comparing performance criteria with and without an additional traffic stream. Also, the Leaky Bucket algorithm can be translated into a queueing model, see the access regulator in Elwalid and Mitra [60]. Congestion control by a priority structure can be modeled by priority queues, see Jaiswal [92], Takács [171]. Shaping is basically inserting an additional queue in order to decrease traffic variability.

3 Approaches for performance evaluation

As we saw in the previous section, in order to perform adequate network design and management in communication networks, insight into the performance of particular queueing systems is useful. Relevant performance indicators are (on burst/cell level) the probabilities of loss due to overflow and extreme delays, and (on call level) the probability of a new arriving call being blocked. These events are usually *rare*, that is, the probability of occurrence of such events is extremely small. An immense literature has been developed to accurately capture these rare events probabilities. Below we shall present a brief summary of methods to evaluate the performance of communication networks, with emphasis on the analysis of rare events. We first consider, on cell/burst level, a single queue in an ATM network. Then, we present a survey on methods used for analyzing call level behavior of ATM networks. We shall discuss, for both of them, the techniques that are used for performance evaluation purposes. These techniques can be divided into five categories, cf. Cohen and Boxma [38].

3.1 Cell/burst level behavior of single ATM links

(i) EXACT ANALYSIS. Classical references on single queues are the textbooks of Kleinrock [106] and Cohen [37]. They provide a thorough analysis, applying continuous-time Markov processes and other techniques. A more recent survey is Chapter 4 of Tijms [173]. A common assumption in these texts is that the input consists of renewal streams.

Neuts [136] considers the class of Quasi Birth-Death processes, in which queues with Markov modulated Poisson arrivals can be fitted. For that reason, this model is more relevant with respect to the analysis of communication systems. Notably, the resulting solutions often have a matrix geometric structure. A relevant extension is the queue with *batch* Markov modulated Poisson input, see Lucantoni [117]; Blondia [14] covers the discrete-time version.

The first successful analysis of a queue fed by (identical, exponential on-off) Markov fluid sources is given in Kosten [109], whereas Anick, Mitra, and Sondhi [5] improved this pioneering work considerably. However, it should be noticed that similar models were analyzed in the context of production systems more or less simultaneously, see e.g. Wijngaard [188]. If a constant service rate is assumed, the entire buffer content distribution can be given explicitly in terms of a so-called *spectral expansion* that emerges from some eigensystem (that contains the generator of the underlying Markov chain as well as the traffic rates). These results were extended by Kosten [111] to a fluid queue fed by multiple classes of general Markov fluid sources, as is the case in an ATM setting. Again the buffer content distribution follows from an eigensystem, but no explicit calculation of

eigenvectors and eigenvalues is possible.

(II) NUMERICAL SOLUTIONS. Unfortunately, exact analysis yields explicit results only for a very restricted class of models: mostly numerical techniques have to be applied to convert the implicit, exact expressions. Typical examples are:

- Suppose a queue can be described by a Markov chain; then the state probabilities follow from the balance equations. This can usually not be done explicitly. However, quite efficient numerical techniques are developed, that solve this system of linear equations, see appendix D of [175].
- Often, solutions remain 'hidden behind the Laplace curtain': only a result in terms of a Laplace transform (or a probability generating function) is available. Therefore, a number of techniques has been developed that numerically invert this kind of transforms, see for instance Abate and Whitt [3].
- Furthermore, numerical techniques are required in case of queues fed by Markov modulated input. Then the performance measures are often given in terms of eigenvectors/eigenvalues, which cannot, in general, be calculated explicitly.

(III) HEURISTICS. If explicit calculation, possibly in conjunction with numerical methods, is not possible or too demanding, one can resort to heuristic methods. A frequently used heuristic is to approximate general distributions by phase-type distributions, as the Erlang distribution or the Coxian distribution. This can be done by fitting a number of moments, for example the mean and second moment. As a consequence, the system allows a Markov chain modeling.

Another accurate heuristic is developed in Tijms [174]. For a considerable class of queueing models, the state probabilities in the infinite-buffer model are far more easy to capture than in the finite-buffer case. For that reason, methods have been developed to heuristically link the solutions of both models.

(IV) ASYMPTOTICAL TECHNIQUES. If exact analysis fails, one might attempt to develop, instead of heuristics, asymptotical expansions. An important class of asymptotics is the heavy traffic approximation by means of diffusion processes, see Borovkov [17] for general models and Knessl and Morrisson [107] for fluid queues. For our applications, however, *large buffer asymptotics* (buffer size tending to infinity) and *large system asymptotics* (number of sources tending to infinity) are more relevant.

Large buffer asymptotics often have an 'asymptotically exponential' form: The probability $\pi(B)$ of a queue of buffer size B being full is (for large values of the buffer size B) of the form

$$\pi(B) \approx \eta \exp[-\theta B], \tag{1.1}$$

where amplitude $\eta > 0$ and decay rate $\theta > 0$ are constants. More formally: $\pi(B)e^{\theta B} \to \eta$, for $B \to \infty$. Of course, this kind of relations is very useful if one is faced with the task of choosing a buffer size, such that the loss fraction is below a certain very small tolerable level ('buffer dimensioning'). For renewal input, we refer to Takahashi [172] and Tijms [175]. For systems with Markov modulated arrivals similar results can be found, e.g., using matrix-geometric methods as [136], and more recently, [2] and [34]. Decay rate θ can be expressed as dominant eigenvalue of some eigensystem.

For Markov fluid input, asymptotic exponentiality has been established as well [5], [111]. Decay rate θ can be found easily. This is done as follows. Suppose the queue is fed by sources $i = 1, \ldots, n$. Kosten [111] found the so-called effective bandwidth functions $C_i(\cdot)$, directly obtainable from the characteristics of source i. These functions determine the decay rate θ: if C denotes the constant service rate, θ is the unique, positive solution of $\sum_i C_i(\theta) = C$. Kesidis, Walrand, and Chang [105] put the effective bandwidth into the context of *large deviations theory* (LD), a collection of techniques that is particularly suitable for the analysis of rare events, explained in more detail in the next section.

As explained, θ can be calculated fairly easily. However amplitude η is difficult to capture. In the case of two-state Markov fluid sources, Stern and Elwalid [170] and Baiocchi and Bléfari-Melazzi [13] found solutions and approximations. In the general case, to get η, an eigensystem with possibly complex eigenvalues has to be solved, which is (particularly for large systems) computationally intractable.

Choudhury, Lucantoni, and Whitt [33], [34] describe how to approach this problem. A common simplification is to replace η by 1: $\pi(B) \approx \exp[-\theta B]$. However, especially in queues fed by a large number of sources, η tends to be typically very small due to the multiplexing effect. However, they also show that for sources (less bursty than Poisson) η can be considerably larger than 1! To overcome this problem, Elwalid, Heyman, Lakshman, Mitra, and Weiss [59] examine an asymptotic approximation of η: $\eta = \eta(n) \approx \eta' \exp[-\theta' n]$, where n denotes the number of sources. This last study applies results from LD.

The counterpart of large buffer asymptotics is *large system asymptotics*. Suppose a queue fed by n (identical) sources, emptied at service rate nC. If n grows, overflows become increasingly rare because of the multiplexing effect, cf. the approximation of $\eta(n)$ in [59]. This kind of rarity is considered in Weiss [184]. If $\pi_n(B)$ denotes the loss probability in the model with n exponential on-off sources, service rate nC and buffer nB, he shows that $n^{-1} \log \pi_n(B)$ tends to a constant (for all C larger than the mean input rate of a single source and all $B \geq 0$). For $B = 0$, this limit can be calculated explicitly, for small and large B accurate approximations are deduced. Again, LD techniques play a crucial role in the analysis. There are also (LD) studies which are asymptotic in the buffer size as well as the number of sources, for instance Botvich and Duffield [18], Tse,

Gallager, and Tsitsiklis [178].

With respect to queueing analysis with long-range dependent input traffic, very few results are known. Duffield and O'Connell [55], [53] (applying LD) and Norros [138], [139] derive large buffer asymptotics. Loosely speaking, they find that loss probability $\pi(B)$ is of 'Weibull-form' $\exp[-\theta B^{2-2H}]$ for some positive constant θ and Hurst-parameter H.

(v) Simulation Techniques. Since the above methods obviously have their deficiencies (especially with respect to accurately capturing η), simulation might be a solution. A problem is that rare events are involved, implying that estimation by direct simulation is very time consuming. To overcome this problem, variance reduction methods have been proposed. One of them is the so-called ReSTART method, see Villen-Altamirano [182]. Suppose the frequency of a buffer overflow must be estimated. Then (i) first estimate the probability of reaching $B/2$, and (ii) next the probability of reaching B, starting in $B/2$. These two events are less rare, and therefore easier to estimate. Glasserman, Heidelberger, Shahabuddin, and Zajic [75] give a more mathematical treatment of this method.

However, importance sampling (in conjunction with large deviations theory) is probably a more powerful simulation technique than ReSTART. It is basically a simulation technique in which is sampled from a probability measure that differs from the actual one, see Glynn and Iglehart [77]. Under this new measure, the rare event under consideration becomes frequent. Unbiasedness is recovered by using the appropriate likelihood ratios: for each sample path during the simulation, the performance measure being estimated is multiplied by a correction factor, expressing the relative likelihood of the observation under the new measure with respect to the old measure. Of course, we are interested in the change of measure providing the largest variance reduction. In a queueing context, this problem can often be tackled by applying large deviations theory.

Large deviations theory enables us to find, for a broad class of queueing systems, the exponential decay rate, above called θ. In fact we derive an asymptotical relation which is weaker than (1.1):

$$\lim_{B\to\infty} \frac{1}{B} \log \pi(B) = -\theta.$$

To get the value of θ, we have to solve a kind of variational problem. Mostly, we have to minimize an entropy function; the optimizing arguments in fact describe the *most probable* trajectory from the equilibrium situation to the rare event. The idea behind importance sampling (in conjunction with large deviations) is to change the probability mechanism such that the average behavior of the model under the new measure coincides with the most probable trajectory to the rare event. This change of measure is endowed with some optimality properties. The first study to combine importance sampling simulations with large deviation techniques was Siegmund [167]. It should be noticed that in a reliability (instead of queueing) context, techniques different from large deviation results must be

used to find an appropriate new measure [78], [85], [133], [164].

3.2 Call level analysis of ATM networks

Effective bandwidths, as introduced above, are also useful in order to perform call acceptance control (CAC) for a single ATM link. Suppose that a loss probability of at most ϵ is allowed; n_r sources of type r are already accepted, $C_r(\cdot)$ is the effective bandwidth of a type r source, and a source of type s must be connected. Approximating η by 1, it has to be checked whether $\sum_r n_r C_r(\theta) + C_s(\theta) \leq C$, for $\theta := -(\log \epsilon)/B$. If this inequality holds the source can be accepted, otherwise it should be rejected. However, notice that, based on the knowledge that (for fluid) η can be considerably smaller than 1, this control mechanism underestimates the remaining available bandwidth.

Notice that the above acceptance control procedure reduces to checking the admission condition in *circuit-switched networks*: C circuits are available; a source i connection requires $C_i(\theta)$ bandwidth in order to guarantee a loss probability below ϵ. To perform CAC in an ATM *network* (instead of a single link), knowledge is required on the bandwidth that is required by a traffic streams of a particular type on *any* link of the network. The studies of de Veciana, Courcoubetis, and Walrand [46] and Chang, Heidelberger, Juneja, and Shahabuddin [28] provide some insight into this matter. Now let A_{jr} denote the bandwidth required by a type r call on link j, let n_r the number of type r calls connected, and let C_j denote the bandwidth available (link capacity) on link j. A call of type s is accepted if and only if $\sum_r n_r A_{jr} + A_{js} \leq C_j$ for all links j.

In this way, CAC in an ATM network reduces to CAC in circuit-switched networks, see Ritter and Tran-Gia [152]. Assuming that type r calls arrive according to a Poisson process with rate λ_r and have holding times with mean μ_r^{-1}, we might try to calculate the probability that a type r call is blocked, given this CAC rule. The model is usually called 'multirate loss model', where multirate refers to the fact that different kinds of customers/services require a different amount of bandwidth. The analysis of this kind of models goes back to Erlang [62]. He considered a telephone network consisting of two switches connected by a given number of circuits.

To analyze Erlang's multirate loss-model, the same methodological subdivision as above applies. (i) *Exact analysis* yields rather explicit results: the steady state distribution of the network occupancy is of *product form*. The analysis of product forms, originated by Jackson [91], has attracted much attention, see for instance Kelly [95] and van Dijk [180].

Although an explicit expression of the blocking probabilities in the multi-class loss model is available, calculation is cumbersome. This is due to the fact that a summation

over a possibly huge amount of states has to be performed, particularly in order to calculate the normalizing constant. Kaufman [93] and Roberts [154] found, in the single-link case, efficient (ii) *numerical solutions* that cope with this problem. The normalizing constant can be calculated, in models of low dimension, by an algorithm that is similar to the one presented in Buzen [25]. Also, an algorithm based on inverting Laplace transforms has recently been proposed by Choudhury, Leung, and Whitt [32].

Another possibility is to use (iii) *heuristics*, for instance those of Lindberger, mentioned in Chapter 4 of [152]. Applying a scaling as in Kelly [96], (iv) *asymptotics* of the blocking probability can be found. Kelly multiplies the arrival rates as well as the link capacities by n, and derives results that are asymptotic in n. Gazdzicki, Lambadaris, and Mazumdar [72] and Chapter 12 of Shwartz and Weiss [166] use LD methods to gain insight into the asymptotical behavior under this scaling. Finally, (v) *simulation techniques* are proposed in order to avoid the summation. Harvey and Hills [83] apply an acceptance-rejection technique, and Ross and Wang [158] develop an importance sampling procedure.

4 Brief introduction to large deviations

In Section 3, we claimed that large deviations (LD) is a powerful technique to give asymptotical rare event analysis of multiservice networks. In this section, we will turn our attention to LD; we first explain the scope of LD, and then we briefly review the theorems that are relevant in the context of this monograph.

4.1 Scope of large deviations theory

Criteria to measure the performance of computer and communication systems can be divided roughly into two classes. On the one hand, several criteria are related to the *average behavior* of the system: what is the mean delay?, what is the mean queue length?, etc. On the other hand, however, the effect of rare calamities can be that huge that it is worthwhile to analyze the so-called *deviant behavior*. As we saw in the previous sections, for network design and traffic control purposes, we must be able to capture performance measures related to the deviant behavior: small probabilities of loss, extreme delays, and call blocking. LD is particularly suited to analyze these rare event probabilities. Within the spectrum of performance analysis techniques presented in Section 3, LD belongs to the asymptotical techniques, and is very useful in order to develop efficient simulation methods.

LD deals with probabilities that tend more or less exponentially to zero in a certain parameter. These situations arise very naturally in many contexts. Think for instance of a random walk with negative drift, and let $\mathrm{P}(B)$ be the probability of ever exceeding

$B > 0$. It can be argued that a reasonable approximation for $\mathrm{P}(2B)$ is $\mathrm{P}(B)^2$. Extending this procedure we find the above-mentioned exponentiality. Similar heuristic arguments can be used to make plausible that the loss fraction of a communication link decays approximately exponentially in the (finite) buffer size.

It should be remarked that LD is not suitable for the analysis of all kinds of rare events. As pointed out in Weiss' tutorial [185], LD addresses rare events that arise as a consequence of a large number of unlikely events happening together. An example is a buffer overflow that is caused by a large number of large bursts. Consequently, LD cannot capture the behavior of a queue in which the tails of the burst length density are heavy (i.e., polynomial), because in that case it is very likely that *one* excessive burst causes overflow, see Anantharam [4] and Asmussen and Klüppelberg [8].

As indicated in Section 3, LD theory has two main features. First, it gives asymptotics (namely the exponential decay rate) of rare event probabilities, and, as a by-product, insight is gained into the question in how the rare event occurs, provided that it occurs: the most probable trajectory. In the second place, LD gives quite useful guidelines how to accelerate rare event simulation.

For instance, consider the probability of a lost cell $\pi(B)$ on a communication link with buffer B. It is interesting to examine how the queue builds up from empty to overflow. Heuristically, this can be examined as follows. It is perhaps reasonable to assume that it would occur with a small positive slope, since this behavior does not deviate much from the average behavior. A disadvantage, however, is that this behavior should be maintained for a very long period of time, in order to reach overflow. On the other hand, if it happens via a trajectory with a very large positive slope, the (very deviant) regime has to be active for a relatively short period. We see that there is some trade-off between these two effects. LD theory enables to quantify a 'cost' of having slope r during one unit of time: $I(r) > 0$, where the *rate function* $I(r)$ increases in $r > 0$. To reach overflow, this slope should be maintained during B/r units time. Clearly, to find the optimal slope, $B/r \cdot I(r)$ has to be minimized. If r^* is the minimizer, $f(t) = r^*t$ (a straight line!) is the optimal trajectory towards overflow. Also $B^{-1} \log \pi(B) \to -\theta$, where $\theta := I(r^*)/r^*$. The knowledge of the optimal path can be used to achieve huge variance reduction in rare event simulations, as explained in Section 3.

4.2 The Large Deviation Principle

A crucial role in LD is played by the large deviations principle (LDP). This principle describes the limiting behavior of a sequence of probability measures $(\mathcal{P}_n)_n$, by introducing a kind of 'cost function for deviant behavior'. We give a brief explanation of this concept

here, using the example of large deviations of sample means. More details on the concept can be found in, e.g., Dembo and Zeitouni [50].

Consider the partial sum of n independent samples from a common distribution, denoted by S_n. According to the (weak) law of large numbers, the sample mean converges (in probability) to the mean μ (supposed to be finite) of the individual random variables. As a consequence, for all positive ϵ,

$$\mathcal{P}\left(\frac{S_n}{n} \in [\mu - \epsilon, \mu + \epsilon]\right) \to 1,$$

as $n \to \infty$. However, what is the probability of the sample mean of the first n samples lying in a set not containing μ? To answer this kind of questions, large deviations theory has been developed.

Suppose a sequence of measures $(\mathcal{P}_n)_n$ that assign a value in $[0, 1]$ to all elements of a certain topological space. Note that in this space open and closed sets are defined. The *rate function* $I(\cdot)$ is a mapping from this space to $[0, \infty]$, satisfying

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathcal{P}_n(F) \leq - \inf_{x\in F} I(x) \text{ and } \liminf_{n\to\infty} \frac{1}{n} \log \mathcal{P}_n(G) \geq - \inf_{x\in G} I(x)$$

for closed F and open G. This setting is called an LDP.

In the case of sample means $\mathcal{P}_n(\cdot) := \mathcal{P}(S_n/n \in \cdot)$, defined on topological space $\mathbb{R}$. Based on the law of large numbers and the large deviations principle, we see that the value of the rate function $I(\cdot)$ in μ must be zero. Also, it can be shown that $I(\cdot)$ is a convex function that is positive elsewhere. In fact, $I(x)$ measures in some sense the discrepancy between x and the theoretical limiting mean μ, or the 'cost' of sample mean x instead of μ. Cramér [43] was the first to identify the rate function $I(\cdot)$ for sample means. Bahadur and Rao [12] and Chernoff [30] gave useful extensions of this LDP.

So far, we assumed the random variables to be independent. However, allowing (short-term) dependencies, Gärtner [71] and Ellis [57], [58] show very sharp generalizations of Cramér's result. Of course, some additional conditions must be imposed in that case.

As we presented an LDP for sample means, LDP for a lot of other stochastic processes can be derived. Consider for instance the so-called empirical distribution of the i.i.d. sequence $(X_i)_i$, defined as follows:

$$L_n(x) := \frac{1}{n} \times \#\{i \in \{1, \ldots, n\} : X_i = x\}.$$

Evidently, this empirical distribution converges (in some sense) to the density of the $(X_i)_i$. As explained above, large deviations provides us the exponential decay rate of rare events, in this case the probability that the empirical distribution L_n based on the first n samples

lies in a set S that does not contain the actual distribution. Sanov [161] found the rate function $J(\cdot)$, which is 0 in the actual density $f_X(\cdot)$ and positive elsewhere. Ellis [58] calls the LDP for sample means 'level 1', the LDP for empirical distributions is 'level 2'. The LDP's of different probabilistic levels are related by contraction principles.

The empirical distribution of a Markov chain satisfies an LDP as well. Clearly, the 'state frequencies' of Markov chains converge to the invariant (if uniquely determined). Miller [128] and Donsker and Varadhan [52] developed large deviations results for this empirical distributions of Markov chains. However, after the establishment of the Gärtner-Ellis theorem, the proofs of these results could be simplified considerably.

Finally, we mention that LDP's can be derived for some classes of stochastic processes, having a kind of 'average path'. Now we are interested in the decay rate of the probability that the path of the stochastic process lies in a set of paths that does not contain this average path. This problem can be solved by minimizing a certain action functional over all paths in the set under consideration. The minimizing argument is the 'most probable trajectory', in case of being in the 'rare set'. Mogulskii [132] found results for random walks, Shwartz and Weiss [166] treat a class of continuous-time, discrete-state processes.

4.3 Literature

In recent years several textbooks in the field of LD appeared. Ellis [58] emphasizes the relationship between large deviations and statistical mechanics. He was the first to identify the several probabilistic levels of LD, see above. Deuschel and Stroock [51] is a detailed but mathematically very demanding treatment of the subject. Bucklew [23] is sometimes more handwaving where tedious mathematical derivations are involved. He pays attention to several engineering applications, e.g. in the field of information theory, detection theory, and quick simulation methods to estimate small loss probabilities in queueing systems. Dembo and Zeitouni [50] is a more technical exposition (on an abstract level) of general principles arising in LD theory.

During the past decade, LD became *the* tool for analyzing rare events in communication systems. For instance, the August 1995 issue of IEEE JSAC – which can be considered as 'state-of-the-art' in the field of ATM networking – consists for about 50% of papers on LD theory or applications. Two surveys on the application of LD in communication are included: Weiss [185] and Chang and Thomas [29]. We also mention the recently published book by Shwartz and Weiss [166]. It partially deals with a rigorous treatment of an important class of large deviations results. Apart from that, a lot of applications in communication networks are provided.

5 Outline

In the course of this thesis, the emphasis gradually shifts from models on very detailed time scale (cell-level) in a very detailed part of the network (one single queue) to models that consider connection level of an ATM network. On cell- (and burst-)scale, we are mostly interested in the asymptotics of the packet loss fraction, on connection level the most important performance criterion is the blocking probability. The chapters are in principle self-contained; the relationship between the chapters is explained in the outline below.

In fact, each chapter contains the rare event analysis of a particular queueing model, i.e., asymptotics of the loss/blocking probability, the most probable way towards this rare event, effective bandwidth results, and importance sampling simulation techniques (mostly with a proof of their optimality with respect to variance reduction). An exception is chapter 6, where the rare event analysis (e.g., the importance sampling program) mainly serves as a tool for resource allocation purposes.

Chapter 2 treats the rare event analysis of the class of single queues, with a renewal batch-input process and general service $\mathrm{GI}^X/\mathrm{G}/1$. This kind of models can be used for cell/burst analysis of a single link in an ATM network, in particular if service times are assumed to be deterministic. Although the arrival streams are of renewal type, they can serve as a first approximation of ATM traffic (usually having some correlation structure), see Kelly [99]. The analysis is to some extent analogous to Sadowsky [160], who considered GI/G/1. The loss probability decays exponentially in the buffer size; the (exponential) decay rate is found. An important condition for these results to hold is the exponential tail of the service time distribution. We also provide a framework to perform call admission control: we derive an expression for the effective bandwidth function, similarly to [99]. Our analysis is closely related to 'change-of-measure' arguments and importance sampling. Importance sampling under a particular change of measure is shown to have some optimality properties. We also comment on importance sampling techniques that alternately use the original and an alternative measure. This chapter is based on Mandjes [124]. A condensed version can be found in Mandjes [119].

Chapter 3 focuses on cell loss on burst level on an ATM link. The input is modeled as a general Markov fluid source, which can be the superposition of a large number of, for instance, on-off sources. The exponential decay rate (in the buffer size) of the loss probability is given. The loss fraction can be estimated by simulation, which can be sped up considerably by applying importance sampling. Explicit results on the optimal change of measure are given: notably, the new measure corresponds to a Markov fluid source, the *conjugate* of the original. It is shown how to efficiently calculate the parameters of this

source. In fact, two techniques are given: the first yields the new source characteristics from an infimization of an entropy function, the second from an eigensystem. It is shown that both approaches are equivalent. The analysis is extended to the ATM link, fed by multiple, non-identical, Markov fluid sources. Applying the notion of effective bandwidth, we find an efficient way to capture the conjugate sources. This chapter has been published as Mandjes and Ridder [125].

In fact, *Chapter 4* considers the same model as dealt with in Chapter 3: burst level performance analysis of an ATM switch. Where Chapter 3 is related to large buffer asymptotics of a single link with Markov fluid input, Chapter 4 concentrates on large system asymptotics of the same model. The exponential decay rate of the overflow probability is examined, as a function of the number of sources n. Analogously to Weiss [184] this decay rate and the optimal trajectory towards overflow are considered for zero, small, and large buffer sizes, where buffer size and service rate are proportional to n. The techniques used here substantially differ from those employed in Chapter 2 and 3: there we used slow random walk large deviations, likelihood ratio techniques, and importance sampling; here we rely on Freidlin-Wentzell-like large deviations [70], calculus of variations, and time reversal. The most significant contribution of this chapter is that we allow the sources to be of *general* Markov fluid type, where in [184] only on-off sources with exponential on and off times are considered. Just as in Chapter 3, the last section addresses the situation in which the queue is fed by multiple classes of sources. This chapter will be published as Mandjes [123].

Chapter 5 is a logical continuation of Chapter 3. Again a large buffer analysis of an ATM system on burst level (Markov fluid input) is given. However, now a tandem system instead of a single link is examined: the output of the first queue serves as input for the second. Due to the constant service rate of the first queue, this queue is basically a *shaper*: the bandwidth required by the output stream of the first queue will be smaller than required by the input. One goal of this chapter is to rigorously deduce an expression for this effective bandwidth. Implicitly, then the decay rate (in the buffer size) of the overflow probability of the second queue is determined. Conjectures from Chang *et al.* [28] on the exponential decay rate and the optimal change of measure (in order to estimate the loss ratio in the second queue) are proven. This chapter is based on Mandjes [120]. A summary of this chapter is given in Ridder and Mandjes [149].

In contrast with the previous chapters, *Chapter 6* is of practical rather than theoretical interest. In the previous chapters, we tried to capture, for fixed resource allocation, the overflow probabilities in ATM networks; here we try to perform resource allocation at lowest possible cost. We find a number of very applicable and intuitive guidelines. The

techniques used are (i) asymptotics for the required bandwidth as a function of the buffer size, for given maximum allowed loss fraction, (ii) heuristics that reduce dimensioning in multi-link systems to dimensioning in single links, and (iii) the fast simulation (importance sampling) program described in the previous chapter. Some comments are made on the question how to optimally (i.e., using the minimal amount of bandwidth) perform resource allocation under delay and loss constraints in tandem and intree networks. This chapter is based on Mandjes and van den Berg [127], where a summary is published in Roberts [156].

Chapter 7 considers the network on connection level. As described earlier, by assigning an effective bandwidth to each connection, the ATM network can, on connection level, be viewed as a (circuit-switched) loss network, as described in Kelly [96]. The state space is an 'integer polyhedron'. The steady-state probabilities appear to be of product form , but explicit calculation of blocking probabilities is impossible because of the huge summations involved. We contribute to the algorithms to capture this blocking probability by developing an importance sampling simulation procedure. The choice of the alternative distribution is closely related to large deviation theory, and it can be calculated by solving a convex programming problem. This chapter will be published as Mandjes [122].

In fact, *Chapter 8* is not in the scope of ATM analysis, but focuses on mobile communications. However, it has a lot in common with the connection-level analysis of ATM networks. Mobile communications are often implemented by a cellular network: the area is divided into cells. In cells that lie far enough apart, the same frequencies can be used without interference. The feasible region of the number of callers active in the cells of the network is again 'integer polyhedral'. If the equilibrium distribution of the cell occupancies is product form, we can use the same techniques to determine blocking probabilities. It appears that the restrictive reversible routing condition of Pallant and Taylor [143] is not required to obtain product form, by introducing a redial mechanism in case of blocking. This chapter is based on Boucherie and Mandjes [19].

Chapter 2

Batch-arrival queues

This chapter deals with the analysis of small overflow probabilities in single-server queues with batch arrivals. First, for the class of $GI^X/G/1$ queues we give analytical expressions for the decay rate (in the buffer size) of these probabilities. In case of Poisson arrivals effective bandwidth results are deduced. Furthermore, we propose a change of measure which enables the execution of importance sampling in an asymptotically optimal way. Simulation results for a specific application show large speedups.

1 Introduction

In queueing theory the analysis of *rare events* has attracted much attention. An important example of these is rejection due to a buffer overflow in a finite capacity queueing system. In order to maintain an adequate level of service, a crucial design issue is how to choose the buffer size to keep the probability of a rejection below a given acceptable level. As this probability can usually not be determined exactly, one might try to approximate it by using an asymptotic expansion for large buffer sizes. For instance, one can show that the rejection probabilities vanish at an exponential rate, under some model assumptions. However, these approximations sometimes only yield rough estimates for buffer sizes in the range of interest. To overcome this problem, one can use simulation methods.

Estimating the statistics of rare events by simulation involves several problems. Since the event of a buffer overflow typically occurs very infrequently, the naive direct *Monte Carlo method* may be infeasible in practice: very long simulation runs are needed to obtain an accurate estimate. To ease the task of estimation by simulation, we can apply variance reduction techniques such as *importance sampling* (Glynn and Iglehart [77]). Simulation data are sampled from a probability measure that differs from the actual measure. In order to maintain unbiasedness, the output must be corrected by a likelihood ratio, expressing the relative likelihood of a realization in the original system with respect to the modified

one. Of course, we want to choose the parameters of the importance sampling model in an optimal way: the speed up should be as large as possible. Cottrell, Fort, and Malgouyres [41] formulated an optimality criterion and found the optimal change of measure in several models. Their analysis is based on *large deviations theory*, in particular the theory of *slow random walks.*

In this chapter we consider batch-arrival queues with a limit on the workload or on the queue length.

- First consider the 'workload model'. Batches of customers arrive at a single-server station according to a renewal process. The interarrival times $(A_n)_n$ have density $a(\cdot)$ and mean $\mathrm{E}A$. The batch sizes $(X_n)_n$ are i.i.d. and independent of the arrival process, and p_k denotes the probability of batch size k. Each customer brings along an amount of work, which is independent of the arrival and batch-size process. These work requirements, $(S_n)_n$, are independent samples from one common distribution with density $s(\cdot)$ and mean $\mathrm{E}S$. The buffer is emptied at a constant rate of, say, 1 per unit time. The buffer contents is restricted to B. We consider several strategies to deal with in a situation of an arriving batch whose work requirement exceeds the remaining buffer capacity. In the case of *complete rejection* a complete batch causing overflow is rejected, while in the case of *partial rejection* only the part of the input in excess of the remaining buffer capacity is lost. We let the offered load to be smaller than 1: $\rho := (\mathrm{E}A)^{-1}\,\mathrm{E}S\,\mathrm{E}X < 1$, assuming the means to be finite.

- In the 'queue-length model' the arrival process is again renewal; interarrival times have density $a(\cdot)$ and mean $\mathrm{E}A$. At an arrival epoch a batch of customers enters the system. The batch-sizes are i.i.d. and distributed as X on $\mathbb{N}$ (with p_k the probability of batch size k), independently of the arrival process. The customers's services are i.i.d. with density $s(\cdot)$. Independence between this service process and the (batch-)arrival process is assumed. The maximum number of customers allowed in the system is B. Again we distinguish between complete rejection ('overflow batch' lost entirely) and partial rejection and assume $\rho < 1$.

Clearly, the queueing processes regenerate themselves at the beginning of a *cycle*, i.e., each time a (batch-)arrival occurs that finds the system empty. We let $\alpha_w(B)$ and $\alpha_q(B)$ be the probabilities of a loss cycle in both models, whereas $\pi_w(B)$ and $\pi_q(B)$ denote the long-run fraction of (completely or partially) rejected customers. In case of single-arrival queues Parekh and Walrand [144] heuristically found the asymptotics of the $\alpha(\cdot)$ and the optimal change of measure to estimate them; Sadowsky [160] gave genuine proofs. Our contribution is to (i) generalize these results to batch-arrival queues and (ii) to treat the simulation of loss fractions $\pi(\cdot)$ instead of $\alpha(\cdot)$.

The organization of our study is as follows. In the first part of this chapter, Section 2, some background on slow random walk theory and importance sampling is provided. We also formulate an optimality criterion for the importance sampling input distributions ('change of measure'), and show how such an optimal change of measure can be constructed.

The second part, Section 3 and 4, consists of new theoretical results on the workload and the queue-length model, respectively. In Section 3 we show that a random walk can be embedded in queues of the workload type so that the theory outlined in Section 2 becomes applicable. It enables us to find the exponential decay rate of $\alpha_w(B)$ and the optimal change of measure to estimate this probability. These topics are addressed in Section 3, where we also extend the effective bandwidth results found by Kelly [99] to the $M^X/G/1$ system and (heuristically) to the $GI^X/G/1$ system.

In Section 4 the queue-length model is examined. It appears that the slow random walk results cannot be applied in this model, only in the particular cases of the $GI^X/M/1$ and $M^X/G/1$ queues. Therefore, other methods are applied to find the decay rate of $\alpha_q(\cdot)$ in the (general) $GI^X/G/1$ model. We find the optimal change of measure to estimate this probability as well.

The last part of the chapter, Section 5, is devoted to practical issues that are related to the importance sampling simulations. We describe how to adapt the simulation approach to make it applicable for estimating the long-run loss fractions $\pi(\cdot)$ instead of $\alpha(\cdot)$. Following Goyal, Shahabuddin, Heidelberger, Nicola, and Glynn [78], this can be done by a procedure in which the importance sampling can be 'turned on and off'. We give some comments on the number of cycles needed to get an estimate whose confidence interval has a length with a specific ratio relative to the estimate (= 'relative error' or 'relative efficiency'). Finally simulation results are given for the loss fractions in a queue fed by multiple (batch-)Poisson sources. The results confirm that we have studied a useful technique to obtain accurate estimates of small overflow probabilities.

2 Importance sampling in conjunction with large deviations: a review

This section is an overview of some basic principles in the field of large deviations that are relevant to this study, e.g., slow random walk theory. Furthermore, we indicate how to apply these results in simulation.

2.1 Slow random walks

The large deviations of sample paths of random walks are treated in Dembo and Zeitouni [50, p. 152-160]. Cottrell *et al.* [41], Parekh and Walrand [144], and Bucklew [23, p. 56-73] consider a more general framework: sample paths with Markov increments. In this section we shall restrict ourselves to i.i.d. increments.

We assume a sequence $\{\xi_i, i \in \mathbb{N}\}$ of i.i.d. random variables, inducing the random walk $\{S_n := \sum_{i=1}^{n} \xi_i, n \in \mathbb{N}\}$. We define the slow random walk $S_n(\cdot)$ as follows:

$$S_n(t) := \frac{1}{n} \sum_{i=1}^{\lfloor nt \rfloor} \xi_i$$

for $t \in [0,T]$, where T is fixed. Just as S_n/n converges to $\mathrm{E}(\xi_1)$ in probability ($\mathrm{E}(\xi_1)$ assumed to be finite), $S_n(\cdot)$ converges (in the supremum metric on the interval $[0,T]$) to $f(\cdot)$ in probability, $f(t)$ given by $\mathrm{E}(\xi_1)\, t$. We assume the random walk to have a drift to $-\infty$, i.e., $\mathrm{E}(\xi_1)$ is negative. Furthermore, we denote the *moment generating function* (mgf) of the increments by $M_\xi(\theta) := \mathrm{E}(e^{\theta \xi_1})$, and the *large deviations rate function* is given by $I_\xi(x) := \sup_\theta \big(\theta x - \log M_\xi(\theta)\big)$, which vanishes at $x = \mathrm{E}(\xi_1)$ and is positive elsewhere.

Mogulskii [132] found that $S_n(\cdot)$ satisfies a large deviations principle, i.e., under some mild conditions on the set A of (differentiable) functions on $[0,T]$ we have

$$\lim_{n\to\infty} \frac{1}{n} \log \mathcal{P}(S_n(\cdot) \in A) = -\inf_{f\in A} \int_0^T I_\xi\big(f'(t)\big)\mathrm{d}t.$$

Let $\alpha(B)$ denote the probability of the random walk $\{S_n, n \in \mathbb{N}\}$ reaching $[B,\infty)$ before hitting $(-\infty, 0]$. We denote by A_T the set of absolutely continuous functions $f(\cdot)$ with $f(0) = 0$, $f(t) \in (0,1)$ for $t \in (0,T)$, and $f(T) = 1$. We find

$$\lim_{B\to\infty} \frac{1}{B} \log \alpha(B) = -\inf_{T>0} \inf_{f\in A_T} \int_0^T I_\xi\big(f'(t)\big)\mathrm{d}t. \tag{2.1}$$

This expression can be simplified [23, page 61-62] to $-\theta^*$, where θ^* is the unique positive solution to the *characteristic equation* $M_\xi(\theta) = 1$. The optimizing $f(\cdot)$ in (2.1) is given by the straight line $M_\xi'(\theta^*)\, t$. From $M_\xi'(0) = \mathrm{E}(\xi_1) < 0$ and the convexity of mgf's, it follows that $M_\xi'(\theta^*)$ is positive, so $M_\xi'(\theta^*)\, t$ is an increasing function. Its slope can be regarded in a sense as the slope of the most probable trajectory of the slow random walk to level 1, or the slope of the optimum path of the random walk to B (for B large).

2.2 Importance sampling

So far we only found a very rough asymptotic expression for $\alpha(B)$ ($\alpha(B) = g(B)e^{-\theta^* B}$, for some function $g(\cdot)$ with $\log g(B)/B \to 0$), and therefore estimation by simulation can

be useful. We might think of simulating the random walk until $(0,B)^c$ is hit, repeat this procedure several times, and estimate $\alpha(B)$ by the proportion of runs the random walk stopped in $[B,\infty)$. More formally: let $\{\xi_i^{(j)}\}$ ($\{S_n^{(j)}\}$, respectively) denote the sequence of increments (partial sums, respectively) in the jth run. T_j is the first epoch in the jth run at which $(0,B)^c$ is hit. Our estimator becomes $(1/n)\sum_{j=1}^n I_j$, where I_j is the indicator function of the event $\{S_{T_j}^{(j)} \geq B\}$. As pointed out in the introduction, this procedure behaves very badly as a result of the rarity of the event involved here.

To overcome this difficulty, we consider the possibility of making use of importance sampling (IS). The original model is governed by some probability measure $\mathcal{P}$, but we generate simulation data under measure $\mathcal{Q}$, with respect to which $\mathcal{P}$ is absolutely continuous. Let the increments have density $f_\xi(\cdot)$ under $\mathcal{P}$ and $g_\xi(\cdot)$ under $\mathcal{Q}$. Define L_j as the likelihood ratio (or Radon-Nikodym derivative, or simply likelihood) of the jth run:

$$L_j := \frac{\mathrm{d}\mathcal{P}}{\mathrm{d}\mathcal{Q}}\left(\xi_1^{(j)},\ldots,\xi_{T_j}^{(j)}\right) = \frac{f_\xi(\xi_1^{(j)})\cdots f_\xi(\xi_{T_j}^{(j)})}{g_\xi(\xi_1^{(j)})\cdots g_\xi(\xi_{T_j}^{(j)})}.$$

With obvious notation, we have that $\alpha(B) \equiv \mathrm{E}^{(\mathcal{P})}(I) = \mathrm{E}^{(\mathcal{Q})}(LI)$. Therefore, weighting the simulation data by the likelihoods L_j yields the unbiased estimator $(1/n)\sum_{i=1}^n L_j I_j$. It can be checked easily that variance reduction is achieved if $I = 1$ implies $L < 1$, see for instance Walrand [183]. As $n \to \infty$, the estimate converges to $\alpha(B)$ a.s., by the law of large numbers.

The number of runs required to obtain a fixed relative efficiency and confidence is approximately proportional to the squared coefficient of variation of LI under the new measure $\mathcal{Q}$: $\mathrm{Var}_B^{(\mathcal{Q})}(LI)/\alpha^2(B)$, see [183, page 335-336]. We are left with the task of finding a measure $\mathcal{Q}$ that minimizes $\mathrm{Var}_B^{(\mathcal{Q})}(LI)$. We use the notion of *asymptotic optimality*, cf. [41], [160], [28]. Since $\mathrm{Var}_B^{(\mathcal{Q})}(LI) = \mathrm{E}_B^{(\mathcal{Q})}(L^2I) - \alpha^2(B) \geq 0$, we have

$$\liminf_{B\to\infty} \frac{1}{B}\log \mathrm{E}_B^{(\mathcal{Q})}(L^2I) \geq \lim_{B\to\infty} \frac{1}{B}\log \alpha^2(B) = -2\theta^*.$$

One calls an IS procedure asymptotically optimal (a.o.) if this lower bound is attained.

We choose measure $\mathcal{Q}$ such that the density of the ξ_i is changed to $g_\xi(x) = f_\xi(x)e^{\theta^* x}$; an increment under this change of measure has mgf $M_\xi(\theta+\theta^*)$. We call $g_\xi(\cdot)$ an *exponentially twisted* version of $f_\xi(\cdot)$. Clearly, the new random walk has a positive drift: $\mathrm{E}^{(\mathcal{Q})}(\xi_1) = M'_\xi(\theta^*) > 0$. We sample from the new distribution $\mathcal{Q}$ (the *conjugate* of the original one, $\mathcal{P}$) until the partial sum attains a value in $(0,B)^c$. Since the conjugate process generates a positive drift, level B is exceeded more frequently. In fact, it is exceeded with probability bounded away from zero, whereas under $\mathcal{P}$ it is hit with exponentially small probability. It can be shown easily that $L = \exp[-\theta^* S_T]$. We found that $L^m I$ is bounded from above by $\exp[-m\theta^* B]$, almost surely, $m \in \mathbb{N}_0$. We conclude that we achieved variance reduction:

$I = 1$ implies $L < 1$. We even have that this change of measure is a.o.:

$$\limsup_{B\to\infty} \frac{1}{B} \log \mathrm{E}_B^{(\mathcal{Q})}(L^2 I) \leq \lim_{B\to\infty} \frac{1}{B} \log e^{-2\theta^* B} = -2\theta^*.$$

Under $\mathcal{P}$ the number of runs that is needed is proportional to $\mathrm{Var}_B^{(\mathcal{P})}(I)/\alpha^2(B) \approx 1/\alpha(B)$, so exponentially increasing in B; under $\mathcal{Q}$ this number turns out to be more or less constant in B, cf. Section 5.

3 Analysis of workload model

As we saw in the previous section, the choice of an ideal importance sampling distribution to estimate the probability of reaching a high level in random walks with a negative drift, comes down to finding the conjugate process. In several queueing processes a random walk can be embedded, which enables us to find the conjugate. This is done in the first part of this section. In the second part we use the $\mathrm{M}^X/\mathrm{G}/1$ results to deduce some ramifications of the effective bandwidth results of Kelly [99]. Heuristically, we give extensions to the $\mathrm{GI}^X/\mathrm{G}/1$ queue.

3.1 Conjugate process of the workload model

Consider the workload model. After removing the boundaries 0 and B, we obtain the so-called free buffer process. In this process a random walk can be embedded easily. Take for S_n the buffer contents just after the nth (batch-)arrival. Then $S_n = \sum_{i=1}^n \xi_i$ with ξ_i i.i.d. and the mgf of the increments is given by

$$M_\xi(\theta) = \int_0^\infty e^{-\theta x} a(x)\mathrm{d}x \times \sum_{k=1}^\infty \Big(\int_0^\infty e^{\theta x} s(x)\mathrm{d}x \Big)^k p_k$$

$$= M_A(-\theta) \sum_{k=1}^\infty \big(M_S(\theta)\big)^k p_k \;=\; M_A(-\theta) M_X\big(\log M_S(\theta)\big),$$

in self evident notation. We assume that the characteristic equation has a unique positive solution, which we denote by θ_w^*. We get, by applying the results of Section 2, $\lim_{B\to\infty}(1/B)\log \alpha_w(B) = -\theta_w^*$.

Note that the existence of θ_w^* implicitly assumes that the moment generating function $M_\xi(\theta)$ is finite for some positive θ. If this is not the case (for instance, if the density of the S_i has a 'heavy tail'), the analysis is far more complicated [8], [7] and no large deviation results can be used.

Now we address the topic of finding the conjugate model, i.e., the optimal IS model with respect to estimating the probability of a loss cycle. As stated in the previous section,

the increments of the 'new random walk' have moment generating function $M_\xi(\theta + \theta_w^*)$. We get that this function equals

$$\frac{M_A(-\theta - \theta_w^*)}{M_A(-\theta_w^*)} \sum_{k=1}^{\infty} \left(\frac{M_S(\theta + \theta_w^*)}{M_S(\theta_w^*)} \right)^k \left(p_k M_A(-\theta_w^*) \big(M_S(\theta_w^*)\big)^k \right). \tag{2.2}$$

From (2.2) we conclude that the a.o. change of measure with respect to estimating $\alpha_w(B)$ is

$$\begin{aligned} a(x) &\to a(x)e^{-\theta_w^* x}/M_A(-\theta_w^*) \\ s(x) &\to s(x)e^{\theta_w^* x}/M_S(\theta_w^*) \\ p_k &\to p_k M_S(\theta_w^*)^k M_A(-\theta_w^*), \end{aligned} \tag{2.3}$$

for $x \geq 0$ and $k \in \mathbb{N}$. Note that (2.3) are indeed densities.

3.2 Effective bandwidth results

In this subsection we consider the situation of multiple distinct sources feeding into a single queue. We suppose N types of traffic sources, which we first assume to be Poisson. In the last part of this subsection we will drop this restriction. There are n_i type i Poisson processes with rate λ_i. A stream of type i brings along batches of customers, where the batch-size is distributed as $X^{(i)}$ with density $p_k^{(i)}$. Each customer provides a random amount of work, distributed as $S^{(i)}$ with density $s_i(\cdot)$, $i = 1, \ldots, N$. The work is put into a buffer that is emptied at a constant rate of 1 per unit time. This setting is motivated by high-speed networks, such as ATM, in which several classes of traffic are multiplexed in a switch. Our purpose is to provide criteria whether or not a source should be admitted, given some service requirement.

Note that we can find an alternative description of this model: there is one Poisson stream with rate $\lambda := \sum_{i=1}^N n_i\lambda_i$ generating batches of customers. The arriving batch is of type i with probability $n_i\lambda_i/\lambda$. Therefore, the offered load is given by

$$\lambda \left\{ \sum_{i=1}^N \frac{n_i\lambda_i}{\lambda} \mathrm{E}S^{(i)}\, \mathrm{E}X^{(i)} \right\} = \sum_{i=1}^N n_i\lambda_i\, \mathrm{E}S^{(i)}\, \mathrm{E}X^{(i)},$$

which we assume to be smaller than 1. As said above, the queueing system under investigation fits in the $\mathrm{M}^X/\mathrm{G}/1$ framework. Therefore, we find θ_w^* via

$$M_\xi(\theta) = \left(\frac{\lambda}{\lambda + \theta} \right) \sum_{i=1}^N \frac{n_i\lambda_i}{\lambda} \left(\sum_{k=1}^{\infty} \big(M_{S^{(i)}}(\theta)\big)^k p_k^{(i)} \right) = 1,$$

which becomes

$$\sum_{i=1}^N n_i \left(\frac{\lambda_i \big(M_{X^{(i)}}(\log M_{S^{(i)}}(\theta)) - 1\big)}{\theta} \right) = 1,$$

or simply $\sum_{i=1}^N n_i C_i(\theta) = 1$. The functions $C_i(\cdot)$ have several nice properties.

- We have that $C_i(0) := \lim_{\theta \downarrow 0} C_i(\theta)$ equals $\lambda_i \, \mathrm{E}S^{(i)} \, \mathrm{E}X^{(i)}$, i.e., the offered load by one source of type i. We can show this by means of Taylor expansions. For θ in a neighborhood of 0, we have $M_{X^{(i)}}\big(\log M_{S^{(i)}}(\theta)\big) = 1 + \theta \, \mathrm{E}S^{(i)} \, \mathrm{E}X^{(i)} + O(\theta^2)$, which implies the statement.

- The function $C_i(\cdot)$ is convex and increasing on $[0, \infty)$. To prove this, we write

$$C_i(\theta) = \lambda_i \sum_{k=1}^{\infty} \left(\frac{\big(M_{S^{(i)}}(\theta)\big)^k - 1}{\theta} \right) p_k^{(i)}.$$

  Note that $\big(M_{S^{(i)}}(\theta)\big)^k$ is an mgf, the mgf of the sum of k drawings from a distribution with density $s_i(\cdot)$. Now letting Y be a non-negative random variable with cumulative distribution function $F_Y(\cdot)$ and mgf $M_Y(\cdot)$, let $\tilde{Y}$ have density $(\mathrm{E}Y)^{-1}\big(1 - F_Y(\cdot)\big)$. Then integration by parts yields

$$M_{\tilde{Y}}(\theta) = \frac{1}{\mathrm{E}Y} \left(\frac{M_Y(\theta) - 1}{\theta} \right).$$

  We find that $C_i(\cdot)$ is a weighted sum (non-negative weights!) of mgf's of random variables having positive mean. Since these are convex and increasing on $[0, \infty)$, so is their sum.

Using the properties above, it is easy to verify that the following relation holds:

$$\sum_{i=1}^{N} n_i C_i(\theta) \leq 1 \iff \lim_{B \to \infty} \frac{1}{B} \log \alpha_w(B) \leq -\theta. \tag{2.4}$$

This result can be viewed as an *effective bandwidth* result, cf. [48], and is a generalization of the results given in Kelly [99], who considered M/G/1 sources. We call $C_i(\cdot)$ the effective bandwidth function of a source of type i, whose interpretation is the following: Suppose that we are faced with a performance criterion that the decay rate of $\alpha_w(B)$ should be smaller than a prespecified value, say $-\theta$. If already n_i $\mathrm{M}^X/\mathrm{G}/1$ sources of type i are accepted ($i = 1, \ldots, N$), and the service requirement is still satisfied, we can determine the remaining capacity by $1 - \sum_{i=1}^{N} n_i C_i(\theta)$. If the bandwidth of the source to be accommodated fits in this space, it can be accepted without violating the performance criterion. In other words: the effective bandwidths can be used to perform call acceptance control. We note that for $\theta = 0$ the effective bandwidth constraint simplifies to the stability constraint. A general treatment of the effective bandwidth concept can be found in Whitt [186].

We can heuristically extend the above relations to a multi-class $\mathrm{GI}^X/\mathrm{G}/1$ setting in the following way. We let the interarrival times of type i traffic be i.i.d. and distributed

as $A^{(i)}$, with density $a_i(\cdot)$. We first notice that the effective bandwidth function $C_i(\cdot)$ in the $M^X/G/1$ case can be regarded as follows: Suppose only one traffic stream of type i is feeding into the buffer, then $C_i(\theta)$ represents that service rate, such that the decay rate of $\alpha_w(B)$ is $-\theta$. In other words: it can be checked that $C_i(\theta)$ in the multi-class $M^X/G/1$ system is the solution for C in

$$M_{A^{(i)}}(-C\theta)M_{X^{(i)}}\big(\log M_{S^{(i)}}(\theta)\big) = 1, \text{ so } C_i(\theta) = -\frac{1}{\theta}M^{-1}_{A^{(i)}}\left(\frac{1}{M_{X^{(i)}}\big(\log M_{S^{(i)}}(\theta)\big)}\right). \quad (2.5)$$

Of course, this function $C_i(\cdot)$ can be calculated for the multi-class $GI^X/G/1$ case (instead of $M^X/G/1$) as well! However, this effective bandwidth function remains heuristic, because we cannot formulate an analogue of (2.4). This is due to the fact that cycles (as defined in the introduction) are not regenerative anymore in the multi-class $GI^X/G/1$ queue, so $\alpha_w(B)$ is not defined.

A nice feature of the function $C_i(\theta)$, as defined implicitly by relation (2.5), is that $\lim_{\theta\downarrow 0} C_i(\theta) = (\mathrm{E}A^{(i)})^{-1}\,\mathrm{E}S^{(i)}\,\mathrm{E}X^{(i)}$ remains valid, just as in the multi-class $M^X/G/1$ case.

It is a matter of straightforward calculus to determine the conjugate of this multi-class queueing model:

$$\begin{aligned} a_i(x) &\to a_i(x)e^{-\theta^*_w C_i(\theta^*_w)x}/M_{A^{(i)}}(-\theta^*_w C_i(\theta^*_w)) \\ s_i(x) &\to s_i(x)e^{\theta^*_w x}/M_{S^{(i)}}(\theta^*_w) \\ p^{(i)}_k &\to p^{(i)}_k M_{S^{(i)}}(\theta^*_w)^k M_{A^{(i)}}(-\theta^*_w C_i(\theta^*_w)), \end{aligned} \quad (2.6)$$

θ^*_w solving $\sum_{i=1}^N n_i C_i(\theta) = 1$, $x \geq 0$ and $k \in \mathbb{N}$.

4 Analysis of queue-length model

For the second model – the queue-length model – we know how to embed a random walk only in some special cases, such as the $GI^X/M/1$ and $M^X/G/1$ queues (as a result of memoryless properties). Consequently, we can use the results of Section 2 to find the decay rate of $\alpha_q(\cdot)$. We start this section by examining these two models. Notice that we denote by θ_q the decay rate of $\alpha_q(B)$, i.e. $\lim_{B\to\infty}(1/B)\log\alpha_q(B)$, assuming to exist.

In the (general) $GI^X/G/1$ model, no random walk can be embedded, and therefore no slow random walk theory is applicable. Therefore, we have to develop other methods to characterize the decay rate of $\alpha_q(\cdot)$ in this case. That is done in Theorem 4.1. We end this section by giving an a.o. change of measure.

- *The* $GI^X/M/1$ *case.* Let the service times be exponentially distributed with mean μ^{-1}. Let the ξ_i now represent the increments of the queue-length process between two

consecutive batch-arrivals. After removing boundaries 0 and B, these are i.i.d. and their mgf is given by

$$M_\xi(\theta) = M_X(\theta) \int_0^\infty \sum_{k=0}^\infty \frac{e^{-\mu x}(\mu x)^k}{k!} e^{-\theta k} a(x)\mathrm{d}x = M_X(\theta) M_A(\mu(e^{-\theta} - 1)).$$

Equating this expression to 1 yields θ_q^*. It can be checked that the decay rates θ_w^* and θ_q^* are related by $\theta_q^* = \log M_S(\theta_w^*)$. Just as in Section 3, one can determine the conjugate process: p_k becomes under the new measure $p_k e^{\theta_q^* k}/M_X(\theta_q^*)$, μ is replaced by $\mu e^{-\theta_q^*}$ and $a(x)$ by $a(x)\exp\big(\mu(e^{-\theta_q^*} - 1)x\big)M_X(\theta_q^*)$ (which is a density). Using the relations between θ_w^* and θ_q^* and (2.3), we conclude that the optimal changes of $a(\cdot)$, $s(\cdot)$ and the distribution of X in both models coincide in the $\mathrm{GI}^X/\mathrm{M}/1$ case.

- *The* $\mathrm{M}^X/\mathrm{G}/1$ *case.* Again we are able to embed a random walk (in the free buffer process). We define the ξ_i to be the net increase of the number of customers between two consecutive service completions. Its mgf is given by, λ denoting the rate of the Poisson arrival process,

$$e^{-\theta} \int_0^\infty \sum_{k=0}^\infty \frac{e^{-\lambda x}(\lambda x)^k}{k!} \big(M_X(\theta)\big)^k s(x)\mathrm{d}x = e^{-\theta} M_S\left(\lambda(M_X(\theta) - 1)\right).$$

Equating to 1 yields θ_q^*. We again find that the relation between θ_w^* and θ_q^* is given by $\theta_q^* = \log M_S(\theta_w^*)$ (or $\theta_w^* = \lambda(M_X(\theta_q^*) - 1)$). Under the a.o. $\mathcal{Q}$ we have that p_k changes to $p_k e^{\theta_q^* k}/M_X(\theta_q^*)$, λ becomes $\lambda M_X(\theta_q^*)$ and $s(x)$ is replaced by $s(x)\exp\big(\lambda(M_X(\theta_q^*)-1)x\big)e^{-\theta_q^*}$ (density!). Again we can deduce that this change of measure is equal to (2.3).

In Section 3 we found that the change of measure (2.3) is a.o. in the workload model with respect to estimation of $\alpha_w(B)$. The above calculations show that (2.3) is, in the $\mathrm{GI}^X/\mathrm{M}/1$ and $\mathrm{M}^X/\mathrm{G}/1$ model, a.o. in order to estimate $\alpha_q(B)$ as well. This gives rise to the idea that it might also be a.o. in the $\mathrm{GI}^X/\mathrm{G}/1$ case.

Furthermore, in both models examined above, we found the decay rate of $\alpha_q(B)$ to equal $\log M_S(\theta_w^*)$. This suggests that this is also valid in the $\mathrm{GI}^X/\mathrm{G}/1$ case.

We prove these two properties in this section, by following the arguments of Sadowsky [160] to some extent. We first deduce the relation between the two decay rates; this relation enables us to conclude asymptotical optimality of the change of measure (2.3).

THEOREM 4.1. Calling the decay rates of the probability of a loss cycle in both models θ_w^* and θ_q^*, respectively, they are related as follows:

$$\theta_q^* = \log M_S(\theta_w^*).$$

PROOF. We let A_k be the interarrival time between batch k and batch $k+1$; S_k is the service time of customer k; X_k denotes the batch-size of batch k, where $k \in \mathbb{N}$. Define

stopping times

$$T := \inf\left\{k > 0 : \sum_{i=1}^{k} A_i > \sum_{i=1}^{X_1+\cdots+X_k} S_i\right\}$$

$$T(B) := \inf\left\{k > 0 : X_1 + \cdots + X_k > B \text{ and } \sum_{i=1}^{k-1} A_i < \sum_{i=1}^{X_1+\cdots+X_k-B} S_i\right\}.$$

Note that T does not depend on buffer size B. Clearly, $\alpha_q(B)$ can be characterized as $\mathcal{P}(T(B) < T)$.

UPPER BOUND. Suppose we execute importance sampling, making use of the new densities defined in (2.3), which we will call $\mathcal{Q}$ or the θ_w^*-twisted version of the original measure. Start a cycle and continue simulating until an overflow occurs or the cycle ends. Let I be the indicator function of the event $\{T(B) < T\}$ and L be the likelihood. Denote by $(A_1, A_2, \ldots)$ the interarrival times, $(S_1, S_2, \ldots)$ the service times, and $(X_1, X_2, \ldots)$ the batch-sizes during this realization. Clearly, we have that $\alpha_q(B) = \mathrm{E}_B^{(\mathcal{Q})}(LI)$, so $\alpha_q(B)$ is bounded from above by the expected value of L (under measure $\mathcal{Q}$) conditioned on an overflow during that cycle $\mathrm{E}_B^{(\mathcal{Q})}(L \mid I = 1)$.

Therefore, suppose $I = 1$. Let C denote the number of customers whose service has been started before the overflow. In the IS simulation, we have scheduled $T(B) - 1$ interarrival times. It can be verified that the likelihood becomes

$$L = \prod_{i=1}^{T(B)-1} \left(e^{\theta_w^* A_i} M_A(-\theta_w^*)\right) \prod_{i=1}^{C} \left(e^{-\theta_w^* S_i} M_S(\theta_w^*)\right) \prod_{i=1}^{T(B)} \left(\frac{1}{M_S(\theta_w^*)^{X_i} M_A(-\theta_w^*)}\right), \tag{2.7}$$

which can be simplified to

$$M_S(\theta_w^*)^{C-\sum_{i=1}^{T(B)} X_i} \exp\left(\sum_{i=1}^{T(B)-1} \theta_w^* A_i - \sum_{i=1}^{C} \theta_w^* S_i\right) \Big/ M_A(-\theta_w^*).$$

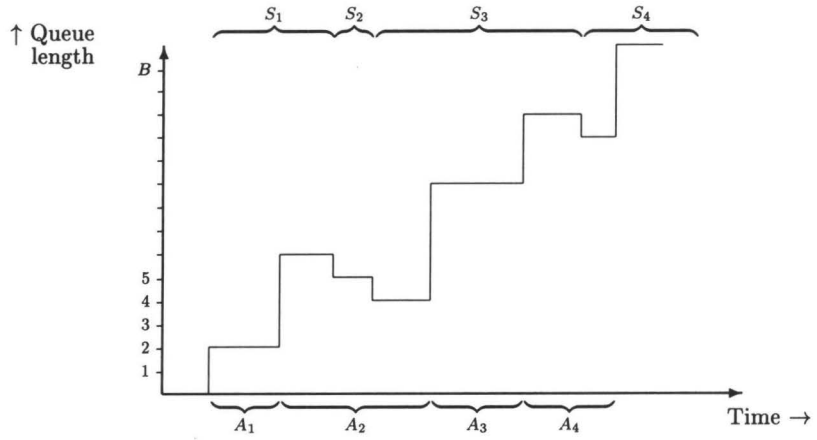
We use the convention that sum (products) defined over an empty range are defined 0 (1, respectively). The difference between the number of customers that entered the system and those who have been served (the queue backlog) exceeds (at an overflow) B, so clearly C is not larger than $X_1 + \cdots + X_k - B$. Apart from that, $\sum_{i=1}^{T(B)-1} A_i < \sum_{i=1}^{C} S_i$. We conclude that an upper bound for $\alpha_q(B)$ is given by $M_S(\theta_w^*)^{-B}/M_A(-\theta_w^*)$. It follows that

$$\limsup_{B\to\infty} \frac{1}{B} \log \alpha_q(B) \leq -\log M_S(\theta_w^*),$$

i.e., we established the upper bound.

LOWER BOUND. Again notice that $\alpha_q(B) = \mathrm{E}_B^{(\mathcal{Q})}(L|I = 1)\mathrm{E}_B^{(\mathcal{Q})}(I)$. First consider $\mathrm{E}_B^{(\mathcal{Q})}(I) = \mathcal{Q}(T(B) < T)$. Note that (since $\mathcal{Q}$ generates a positive drift) we have

Figure 1: Overflow cycle. $B = 14$.
Batch-sizes are $X_1 = 2$, $X_2 = 4$, $X_3 = 5$, $X_4 = 3$, *and* $X_5 = 4$.
$T(B) = 5$, $C = 4$, *and* $C' = 3$.



(i) $\mathcal{Q}(T(B) < \infty) = 1$ for all B and (ii) $\mathcal{Q}(T = \infty) =: q > 0$. Some calculation yields that for all B

$$\left| \mathcal{Q}(T(B) < T) - \mathcal{Q}(T = \infty) \right| = (1 - q) \times \mathcal{Q}(T(B) < T \mid T < \infty),$$

which tends to 0. We find

$$\lim_{B \to \infty} \mathcal{Q}(T(B) < T) = \mathcal{Q}\left(\forall k > 0 : \sum_{i=1}^{k} A_i \leq \sum_{i=1}^{X_1 + \cdots + X_k} S_i \right) = q > 0. \tag{2.8}$$

Consequently it remains to prove the lower bound for $\mathrm{E}_B^{(\mathcal{Q})}(L \mid I = 1)$.

Assume $I = 1$. We let C' denote the number of services that have been started at the arrival of batch $T(B) - 1$. We have $C' > X_1 + \cdots + X_{T(B)-1} - B$ (no overflow!) and (by definition of C') $\sum_{i=1}^{T(B)-2} A_i > \sum_{i=1}^{C'-1} S_i$. It follows that $C - C' < X_{T(B)}$. Simple algebra yields, cf. equation (2.7),

$$L = \prod_{i=1}^{T(B)-2} \left(e^{\theta_w^* A_i} M_A(-\theta_w^*) \right) \prod_{i=1}^{C'-1} \left(e^{-\theta_w^* S_i} M_S(\theta_w^*) \right) \prod_{i=1}^{T(B)-1} \left(\frac{1}{M_S(\theta_w^*)^{X_i} M_A(-\theta_w^*)} \right)$$

$$\times \left(e^{\theta_w^* A_{T(B)-1}} M_A(-\theta_w^*) \right) \prod_{i=C'}^{C} \left(e^{-\theta_w^* S_i} M_S(\theta_w^*) \right) \left(\frac{1}{M_S(\theta_w^*)^{X_{T(B)}} M_A(-\theta_w^*)} \right),$$

for $T(B) > 2$. For $T(B) = 1, 2$ similar expressions hold. Now suppose that the batch sizes (service times) are a.s. bounded by m_X (m_S, respectively). We can check that, given $I = 1$,

$$L \geq \frac{M_S(\theta_w^*)^{-B}}{M_A(-\theta_w^*)} \left(\frac{e^{-\theta_w^* m_S}}{M_S(\theta_w^*)} \right)^{m_X},$$

with probability 1. Thus, under the assumptions $m_X < \infty$ and $m_S < \infty$, we have

$$\liminf_{B\to\infty} \frac{1}{B} \log \alpha_q(B) \geq -\log M_S(\theta_w^*).$$

It is a technical matter (see Ney and Nummelin [137], Sadowsky [160]) to generalize this lower bound to unrestricted distributions (using a truncation argument). We give here a variation of that proof. Let $E(m)$ be the event that for all batches 1 to T the batch sizes are not larger than m_X and the customers do not require more than m_S service time. Choose $m := (m_X, m_S)$ large enough so that on $E(m)$ we have with positive probability that $S_1 + \cdots S_{X_1} \geq A_1$. Let $\alpha_q(B, E(m))$ be the probability on the intersection of $E(m)$ and a loss cycle and $\alpha_q(B \mid E(m))$ the probability of a loss cycle conditional on $E(m)$. Obviously

$$\alpha_q(B) \geq \alpha_q(B, E(m)) = \alpha_q(B \mid E(m))\, \mathcal{P}(E(m)).$$

We notice that $\mathcal{P}(E(m))$ is constant in B. We define the following conditional mgf's, the mgf's of batch size and service time under the condition $E(m)$:

$$M_X(\theta \mid m_X) := \frac{\sum_{k=0}^{m_X} e^{\theta k} p_k}{\sum_{k=0}^{m_X} p_k}, \qquad M_S(\theta \mid m_S) := \frac{\int_0^{m_S} e^{\theta x} s(x)\mathrm{d}x}{\int_0^{m_S} s(x)\mathrm{d}x}.$$

It is easy to verify that for positive θ, $M_X(\theta \mid m_X)$ increases to $M_X(\theta)$ and $M_S(\theta \mid m_S)$ to $M_S(\theta)$ as m_X and m_S tend to ∞. Let $\theta_w^*(m)$ be the positive solution of the characteristic equation

$$M_A(-\theta) M_X\big(\log M_S(\theta \mid m_S) \mid m_X\big) = 1,$$

parametrized by m. Let $\mathcal{Q}_m$ denote the $\theta_w^*(m)$-twisting of the original conditional distributions. We get that $\alpha_q(B \mid E(m))$ equals the mean of LI under the measure $\mathcal{Q}_m$. Using the 'bounded batch-size and service-time case', we find that $\alpha_q(B \mid E(m))$ dominates

$$\frac{M_S(\theta_w^*(m) \mid m_S)^{-B}}{M_A(-\theta_w^*(m))} \left(\frac{e^{-\theta_w^*(m) m_S}}{M_S(\theta_w^*(m) \mid m_S)} \right)^{m_X} \times \mathrm{E}_B^{(\mathcal{Q}_m)}(I). \tag{2.9}$$

The last factor of the right hand side of (2.9) converges ($B \to \infty$) to the (positive) probability (2.8) under $\mathcal{Q}_m$, which is a constant (in B). We get for arbitrary m_S, m_X the following lower bound for the decay rate of $\alpha_q(B)$:

$$\liminf_{B\to\infty} \frac{1}{B} \log \alpha_q(B) \geq \liminf_{B\to\infty} \frac{1}{B} \log \alpha_q(B \mid E(m))\, \mathcal{P}(E(m)) \geq -\log M_S(\theta_w^*(m) \mid m_S).$$

Now we let m approach (∞, ∞), in such a way that m_X as well as m_S are increasing. $M_X(\cdot \mid \cdot)$ is increasing in both arguments, so (for fixed positive θ) if $m \to (\infty, \infty)$

$$M_A(-\theta)M_X\big(\log M_S(\theta \mid m_S) \mid m_X\big) \uparrow M_A(-\theta)M_X(\log M_S(\theta)).$$

It follows that $\theta_w^*(m) \downarrow \theta_w^*$, and therefore $M_A(-\theta_w^*(m))$ increases to $M_A(-\theta_w^*)$. As a consequence, we have for their reciprocals

$$M_X\big(\log M_S(\theta_w^*(m) \mid m_S) \mid m_X\big) \downarrow M_X(\log M_S(\theta_w^*)). \tag{2.10}$$

Again noting that $M_X(\theta \mid m_X)$ is increasing in θ (θ positive) as well as m_X, it is straightforward to obtain that $\log M_S(\theta_w^*(m) \mid m_S)$ cannot increase anywhere along the m-path to (∞, ∞), because this would violate the monotonicity in (2.10). We conclude that, as m approaches (∞, ∞), $-\log M_S(\theta_w^*(m) \mid m_S)$ increases monotonically to $-\log M_S(\theta_w^*)$. We have found a lower bound for the decay rate of $\alpha_q(B)$:

$$\liminf_{B\to\infty} \frac{1}{B} \log \alpha_q(B) \geq -\log M_S(\theta_w^*),$$

as required. ■

Similarly to the slow random walk results of Section 2, we are able to prove optimality properties of the IS technique. We have on the one hand that $\mathrm{E}_B^{(\mathcal{Q})}(L^2 I) \geq \alpha_q^2(B)$, implying

$$\liminf_{B\to\infty} \frac{1}{B} \log \mathrm{E}_B^{(\mathcal{Q})}(L^2 I) \geq \lim_{B\to\infty} \frac{1}{B} \log \alpha_q^2(B) = -2 \log M_S(\theta_w^*).$$

But our choice of IS makes sure that this lower bound is reached:

$$\limsup_{B\to\infty} \frac{1}{B} \log \mathrm{E}_B^{(\mathcal{Q})}(L^2 I) \leq \lim_{B\to\infty} \frac{1}{B} \log \frac{M_S(\theta_w^*)^{-2B}}{M_A(-\theta_w^*)^2} = -2 \log M_S(\theta_w^*).$$

Consequently, this simulation procedure is a.o.

5 Importance sampling of loss fractions

We have determined for the queueing model under consideration the optimal change of measure of the original model, in order to estimate the probability of a loss cycle $\alpha(\cdot)$. However, in practice we are often interested in more detailed quantities like the long-run fraction of customers who are (completely or partially) rejected. Regenerative analysis yields

$$\pi(B) = \frac{\mathrm{E}_B^{(\mathcal{P})}(N)}{\mathrm{E}_B^{(\mathcal{P})}(D)}, \tag{2.11}$$

where $\pi(B)$ can be read as $\pi_w(B)$ or $\pi_q(B)$. Here N denotes the number of rejected customers during one cycle, D is the number of customers arriving during a cycle, and the index B represents the buffer size.

5.1 Description of the simulation procedure

The denominator of (2.11) can be estimated in a straightforward way using the original model. However, the numerator includes the rare event of a buffer overflow. Using the original distributions (say $\mathcal{P}$) the estimation becomes troublesome, as explained in the introduction. An alternative is sampling from the IS distributions ($\mathcal{Q}$) that are asymptotically optimal with respect to estimating $\alpha_w(B)$ and $\alpha_q(B)$. Indeed, due to the 'positive drift' overflows are not rare, but notice that the length of a cycle may attain very large values. To circumvent this practical difficulty, we use a new measure $\mathcal{R}$, composed from $\mathcal{P}$ and $\mathcal{Q}$, in the following way.

(i) Start from an empty system and simulate the process under $\mathcal{Q}$. If the cycle ends before a customer is rejected (despite the 'positive drift' under $\mathcal{Q}$), we put $N := 0$ and the simulation of this cycle is finished. If on the other hand overflow is reached, we denote the likelihood of the sample path by L and go to step (ii).

(ii) We finish the cycle, by simulating under $\mathcal{P}$. We denote by N the number of customers lost during this cycle.

Obviously we have that $\mathrm{E}_B^{(\mathcal{P})}(N) = \mathrm{E}_B^{(\mathcal{R})}(LN)$. Note that we add in superscript the underlying probability model. The technique described above is called 'measure specific dynamic importance sampling', see [78], [28]: specific measures are used to estimate both numerator and denominator. 'Dynamic' refers to the fact that the IS is 'turned on' until an overflow and 'turned off' thereafter.

Repeat the above recipe n times; the obtained values are $(LN)_1, \ldots, (LN)_n$. We also simulate n cycles (under $\mathcal{P}$) to obtain $D_1, \ldots, D_n$. An unbiased estimate for the numerator is $(\overline{LN})_n := \sum_{i=1}^n (LN)_i/n$, for the denominator $\overline{D}_n := \sum_{i=1}^n D_i/n$, each converging a.s. to the corresponding means. Therefore, we have the following efficient estimator for $\pi(B)$:

$$(\hat{\pi}(B))_n := \frac{(\overline{LN})_n}{\overline{D}_n}.$$

This estimator is asymptotically normally distributed with mean $\pi(B)$ and variance

$$\sigma_n^2 := \frac{\frac{1}{n}\big(\mathrm{Var}_B^{(\mathcal{R})}(LN) + \pi^2(B)\mathrm{Var}_B^{(\mathcal{P})}(D)\big)}{\big(\mathrm{E}_B^{(\mathcal{P})}(D)\big)^2}.$$

This variance expression is the familiar variance of a ratio estimator, with zero covariance-term, since the observations of the numerator and denominator are observed from different cycles. An (approximate) confidence interval is given by $(\hat{\pi}(B))_n \pm q_{1-\alpha/2}\sqrt{(\hat{\sigma}_n^2)}$, where

q_α denotes the $\alpha-$quantile of the normal distribution. Here $\hat{\sigma}_n^2$ estimates σ_n^2, taking for $\mathrm{Var}_B^{(\mathcal{R})}(LN)$ the sample variance

$$\frac{1}{n-1}\sum_{i=1}^{n}\left((LN)_i - (\overline{LN})_n\right)^2,$$

for $\mathrm{Var}_B^{(\mathcal{P})}(D)$ an analogous expression, for $\pi(B)$ the estimate $(\hat{\pi}(B))_n$, and for $\mathrm{E}_B^{(\mathcal{P})}(D)$ the sample mean $\overline{D}_n$.

Very similar to proving that $\mathcal{Q}$ is a.o. with respect to estimating $\alpha(\cdot)$, it is straightforward to show that $\mathcal{R}$ is a.o. with respect to estimating $\pi(\cdot)$. This property is shown in [121]. A key result in this proof is that $\alpha(\cdot)$ and $\pi(\cdot)$ have identical decay rates.

5.2 Simulation results

This subsection treats the fast simulation of overflow probabilities in a multiple source $\mathrm{M}^X/\mathrm{G}/1$ queue as described in Subsection 3.2. This example is of the 'workload-type'; the simulation of queues with restricted queue length can be done similarly. The model under the original probability measure $\mathcal{P}$ is characterized by 6 types of sources. Note that the batch-sizes can attain value $0 \notin \mathbb{N}$, but it can be verified that all results derived earlier remain valid.

Table 1: Model under original measure

Type	Number	Arrival rate	Batch size	Job size	Offered load
1	10	0.010	$\sim$ Geom(0.600)	$\sim$ Erl(3, 3.000)	0.083
2	10	0.020	$\sim$ Bin(3, 0.333)	$\sim$ Erl(2, 8.000)	0.050
3	5	0.100	$\sim$ Pois(1.000)	$\sim$ Erl(2, 10.000)	0.100
4	20	0.010	$\sim$ Bin(2, 0.333)	$\sim$ SE(0.500, 3.000)	0.111
5	2	0.020	$\sim$ Det(1)	$\sim$ SE(1.000, 4.000)	0.050
6	2	0.100	$\sim$ Pois(0.500)	$\sim$ Det(1.000)	0.100

Here $\mathrm{Erl}(n, \lambda)$ denotes the Erlang(n) distribution with scale parameter λ. $\mathrm{SE}(a, \lambda)$ is defined as the shifted exponential distribution with location parameter a and scale parameter λ : this distribution corresponds to the sum of an $\mathrm{Exp}(\lambda)$ rv and a positive constant a.

Following the lines of Subsection 3.2, we can find the decay rate $-\theta_w^*$ of $\alpha_w(B)$ and $\pi_w(B)$: -0.9487. Using the transformations (2.6), it is easy to check that the exponentially twisted version of an $\mathrm{Erl}(n, \lambda)$ rv is again $\mathrm{Erl}(n)$, but with scale parameter $\lambda - \theta_w^*$. Analogously: $\mathrm{SE}(a, \lambda)$ becomes $\mathrm{SE}(a, \lambda - \theta_w^*)$. It is a matter of elementary calculus to obtain that the $\mathrm{Bin}(n, p)$ distributions must be replaced by

$$\mathrm{Bin}\left(n, \frac{M_{S^{(i)}}(\theta_w^*)p}{M_{S^{(i)}}(\theta_w^*)p + 1 - p}\right)$$

distributions; using $\mathcal{Q}$ instead of $\mathcal{P}$, Geom(p) becomes Geom$\big(1-(1-p)M_{S^{(i)}}(\theta_w^*)\big)$, and Pois($\lambda$) is replaced by Pois$\big(\lambda M_{S^{(i)}}(\theta_w^*)\big)$. Finally, we multiply λ_i by $M_{X^{(i)}}\big(\log M_{S^{(i)}}(\theta_w^*)\big)$. We get for $\mathcal{Q}$

Table 2: Model under importance sampling measure

Type	Number	Arrival rate	Batch size	Job size	Offered load
1	10	0.031	$\sim$ Geom(0.330)	$\sim$ Erl(3, 5.051)	0.550
2	10	0.026	$\sim$ Bin(3, 0.392)	$\sim$ Erl(2, 7.051)	0.088
3	5	0.125	$\sim$ Pois(1.221)	$\sim$ Erl(2, 9.051)	0.168
4	20	0.021	$\sim$ Bin(2, 0.540)	$\sim$ SE(0.500, 2.051)	0.449
5	2	0.068	$\sim$ Det(1)	$\sim$ SE(1.000, 3.051)	0.180
6	2	0.221	$\sim$ Pois(1.291)	$\sim$ Det(1.000)	0.570

We changed the stable queueing system into an unstable one: the offered load is 2.005 instead of 0.494. As explained in the previous section we can combine measure $\mathcal{P}$ and its conjugate $\mathcal{Q}$ to a new measure $\mathcal{R}$. As pointed out in the introduction, several rejection strategies can be used. we will consider:

1. Complete 'overflow batch' lost.

2. Jobs of the 'overflow batch' are put in random order. Those jobs that fit in the buffer entirely are accepted, the remainder is rejected (the 'overflow-job' is rejected completely).

3. Jobs in the 'overflow batch' are put in random order. The part of them that exceeds level B is lost. Consequently, the 'overflow-job' is processed partially.

4. Jobs in the 'overflow batch' are put in increasing order. The fobs fitting in the buffer entirely are processed, the others are lost. We might hope to decrease the long run fraction of customers lost, compared with strategy 2.

We estimated the following performance measures by simulation: (i) the probability of an overflow $\alpha_w(B)$, (ii) the long-run fraction of customers whose job is not processed completely $\pi_{w,i}(B)$, and (iii) the fraction of work lost $\eta_{w,i}(B)$, $i = 1, \ldots, 4$. We conclude with a comparison between the usual direct simulation method (regenerative, as explained in [113]) and the quick simulation method described above. We are interested in performance indicators like the number of cycles and simulation time required to obtain an estimate with a certain fixed confidence and relative efficiency.

5.3 Evaluation of the simulation results

For our results we used confidence of 95%, and relative efficiency (i.e., the ratio of the confidence interval half-length to the estimated value) 10%. In the tables we list estimates, the number of runs needed (divided by 1000) and the computer time, where m indicates minutes and s seconds. The simulation jobs are executed on a 486 personal computer. Because of large simulation times we omit results for direct simulation and larger values of B. We notice that $\eta_{w,i}(B)$ shows the same asymptotical behavior as $\alpha_w(B)$ and $\pi_{w,i}(B)$ (for $i = 1, \ldots, 4$):

$$\lim_{B\to\infty} \frac{1}{B} \log \alpha_w(B) = \lim_{B\to\infty} \frac{1}{B} \log \pi_{w,i}(B) = \lim_{B\to\infty} \frac{1}{B} \log \eta_{w,i}(B) = -\theta_w^*. \tag{2.12}$$

As shown theoretically in [121], we find that the number of cycles $C_{\mathrm{DS}}(B)$ needed to get some fixed level of confidence in the direct simulation, blows up exponentially (in buffer size B) with rate θ_w^*. In fact, in the tables we find that $C_{\mathrm{DS}}(B)$ is more or less inversely proportional to $\pi_w(B)$. The simulation time grows approximately proportionally to $C_{\mathrm{DS}}(B)$.

Table 3: Direct simulation

B	Discipline 1 $\hat{\pi}_{w,1}(B)$	Discipline 2 $\hat{\pi}_{w,2}(B)$	Discipline 3 $\hat{\pi}_{w,3}(B)$	Discipline 4 $\hat{\pi}_{w,4}(B)$	$\hat{\alpha}_w(B)$
4	$1.42{\cdot}10^{-2}$	$9.07{\cdot}10^{-3}$	$1.31{\cdot}10^{-2}$	$8.08{\cdot}10^{-3}$	$8.72{\cdot}10^{-3}$
	182	189	199	177	41
	1 m	1 m	1 m	1 m	13 s
6	$2.04{\cdot}10^{-3}$	$1.32{\cdot}10^{-3}$	$1.87{\cdot}10^{-3}$	$1.18{\cdot}10^{-3}$	$1.38{\cdot}10^{-3}$
	1437	1429	1464	1355	278
	10 m	10 m	10 m	9 m	2 m
8	$3.09{\cdot}10^{-4}$	$1.94{\cdot}10^{-4}$	$2.87{\cdot}10^{-4}$	$1.80{\cdot}10^{-4}$	$1.92{\cdot}10^{-4}$
	10451	9812	10802	9122	2006
	69 m	64 m	72 m	62 m	12 m

Table 4: Direct simulation

B	Discipline 1 $\hat{\eta}_{w,1}(B)$	Discipline 2 $\hat{\eta}_{w,2}(B)$	Discipline 3 $\hat{\eta}_{w,3}(B)$	Discipline 4 $\hat{\eta}_{w,4}(B)$
4	$2.65{\cdot}10^{-2}$	$1.91{\cdot}10^{-2}$	$1.26{\cdot}10^{-2}$	$1.80{\cdot}10^{-2}$
	32	32	45	33
	13 s	13 s	18 s	13 s
6	$4.28{\cdot}10^{-3}$	$2.63{\cdot}10^{-3}$	$1.67{\cdot}10^{-3}$	$3.14{\cdot}10^{-3}$
	222	246	360	217
	2 m	2 m	2 m	2 m
8	$5.89{\cdot}10^{-4}$	$3.88{\cdot}10^{-4}$	$2.68{\cdot}10^{-4}$	$4.15{\cdot}10^{-4}$
	1752	1780	1830	1778
	12 m	12 m	13 m	12 m

Table 5: Importance Sampling

B	Discipline 1 $\hat{\pi}_{w,1}(B)$	Discipline 2 $\hat{\pi}_{w,2}(B)$	Discipline 3 $\hat{\pi}_{w,3}(B)$	Discipline 4 $\hat{\pi}_{w,4}(B)$	$\hat{\alpha}_w(B)$
4	$1.41 \cdot 10^{-2}$	$9.27 \cdot 10^{-3}$	$1.26 \cdot 10^{-2}$	$8.38 \cdot 10^{-3}$	$8.64 \cdot 10^{-3}$
	7, 12 s	7, 13 s	7, 16 s	8, 14 s	1, 1 s
6	$2.10 \cdot 10^{-3}$	$1.27 \cdot 10^{-3}$	$1.88 \cdot 10^{-3}$	$1.17 \cdot 10^{-3}$	$1.28 \cdot 10^{-3}$
	8, 20 s	8, 21 s	9, 25 s	8, 22 s	1, 1 s
8	$3.14 \cdot 10^{-4}$	$2.03 \cdot 10^{-4}$	$2.74 \cdot 10^{-4}$	$1.69 \cdot 10^{-4}$	$1.99 \cdot 10^{-4}$
	8, 26 s	8, 29 s	9, 32 s	9, 32 s	1, 1 s
10	$4.98 \cdot 10^{-5}$	$3.04 \cdot 10^{-5}$	$4.26 \cdot 10^{-5}$	$2.69 \cdot 10^{-5}$	$3.00 \cdot 10^{-5}$
	8, 31 s	8, 35 s	9, 41 s	8, 36 s	1, 2 s
15	$4.31 \cdot 10^{-7}$	$2.62 \cdot 10^{-7}$	$3.63 \cdot 10^{-7}$	$2.37 \cdot 10^{-7}$	$2.68 \cdot 10^{-7}$
	8, 50 s	8, 52 s	9, 59 s	8, 55 s	1, 2 s
20	$3.88 \cdot 10^{-9}$	$2.29 \cdot 10^{-9}$	$3.07 \cdot 10^{-9}$	$2.03 \cdot 10^{-9}$	$2.09 \cdot 10^{-9}$
	8, 1 m	8, 1 m	9, 1 m	9, 1 m	1, 3 s

Table 6: Importance Sampling

B	Discipline 1 $\hat{\eta}_{w,1}(B)$	Discipline 2 $\hat{\eta}_{w,2}(B)$	Discipline 3 $\hat{\eta}_{w,3}(B)$	Discipline 4 $\hat{\eta}_{w,4}(B)$
4	$2.74 \cdot 10^{-2}$	$1.79 \cdot 10^{-2}$	$1.21 \cdot 10^{-2}$	$1.79 \cdot 10^{-2}$
	2, 3 s	2, 4 s	2, 4 s	2, 4 s
6	$4.08 \cdot 10^{-3}$	$3.03 \cdot 10^{-3}$	$1.60 \cdot 10^{-3}$	$3.04 \cdot 10^{-3}$
	2, 5 s	2, 5 s	2, 6 s	2, 6 s
8	$5.77 \cdot 10^{-4}$	$4.25 \cdot 10^{-4}$	$2.55 \cdot 10^{-4}$	$4.41 \cdot 10^{-4}$
	2, 7 s	2, 7 s	2, 7 s	2, 8 s
10	$9.05 \cdot 10^{-5}$	$6.52 \cdot 10^{-5}$	$4.05 \cdot 10^{-5}$	$5.77 \cdot 10^{-5}$
	2, 9 s	2, 9 s	2, 9 s	2, 9 s
15	$7.61 \cdot 10^{-7}$	$5.57 \cdot 10^{-7}$	$3.57 \cdot 10^{-7}$	$5.52 \cdot 10^{-7}$
	3, 19 s	2, 13 s	2, 13 s	2, 13 s
20	$6.99 \cdot 10^{-9}$	$4.60 \cdot 10^{-9}$	$3.08 \cdot 10^{-9}$	$4.40 \cdot 10^{-9}$
	2, 18 s	2, 17 s	2, 17 s	2, 17 s

Defining $C_{\text{IS}}(B)$ analogously, it can be shown that $\lim_{B\to\infty}(1/B)\log C_{\text{IS}}(B) = 0$. In other words: the number of cycles required by the IS technique is subexponential in B. From our simulation results we even find that $C_{\text{IS}}(B)$ is in fact almost constant in B! The simulation time grows more or less linear in the buffer size. This is because the length of a cycle under $\mathcal{R}$ is approximately proportional to B: the contents increases from 0 to B approximately linearly (under conjugate measure $\mathcal{Q}$) and then decreases from B back to 0 approximately linearly as well (under original measure $\mathcal{P}$).

In the simulation example we find that strategy 4 indeed improves the long-run average number of rejected customers, compared with 2. Finally, we conjecture that probably even stronger properties than (2.12) hold: from the experiments it appears that the probability

to be estimated multiplied by $e^{\theta_w^* B}$ tends to a constant.

6 Conclusions

We have deduced large buffer asymptotics of the loss probability in batch-arrival queues. We saw that the slow random walk is a powerful tool, provided that a random walk can be embedded. If this is not possible, as in the $GI^X/G/1$ queue-length model, other techniques must be used, e.g., the likelihood ratio techniques of Theorem 4.1.

We also conclude that the IS technique permits us to execute fast tests on overflow probabilities, varying the buffer size. If the arrival or service process is changed as well, then a new alternative probability measure has to be calculated in order to perform the fast simulation.

Chapter 3

Markov fluid queues with large buffers

This chapter addresses characteristics of finite-buffer queues with Markov modulated fluid input, particularly those related to their deviant behavior. Our aim in this chapter is to find rough asymptotics for the probability of a loss cycle. Apart from that, we derive some properties of the fluid process in case of the buffer contents reaching a high level (which process we call the conjugate of the original process). Our main goal is to obtain methods to find the rate matrix of this conjugate process. For this purpose we use large deviations techniques, but we consider the governing eigensystem as well, and we discuss the relation between these two approaches. We extend the analysis to the multiple source case. Finally, we use the obtained results in simulation. We examine variance reduction by importance sampling in a multiple source example. The new statistical law of the fluid process is based on the conjugate rate matrices.

1 Introduction

Nowadays in engineering, probability and operations research much attention is paid to *Markov modulated fluid queues*. Their interest lies in the possibility to model probabilistically the buffer behavior in ATM switches. Several research issues and mathematical analyses have been initiated by these fluid models. Before we introduce our model and notation, we first give a brief overview of these.

The fluid model can be considered as a queueing system with finite buffer that is filled (continuously) by input streams and emptied at a constant rate. The buffer is used by one or more customers (sources), generating Markov modulated input streams. The main idea behind the model is that we assume that the input of queueing systems may be highly correlated in time: the arrival pattern consists of bursty moments (peaks) alternating with periods with more quiet input. This means that the fluid model is a more natural representation of reality than models with 'homogeneous' input, e.g. Poisson processes.

If the input consists of identical independent on-off sources, we may consider the modulating process as a birth-death Markov chain. This model (with uniform birth and death rates) is analyzed by Anick, Mitra, and Sondhi [5]; later, the model with general birth and death rates was solved by van Doorn, Jagers, and de Wit [181]. A system of linear differential equations, describing the evolution of the buffer contents in time, is derived. Equilibrium probabilities of the buffer contents and other performance parameters are derived from the spectral expansion of the solution of the associated eigensystem. Therefore, this approach is referred to as the *governing eigensystem* method. Later, Kosten [110], Mitra [130], Stern and Elwalid [170], and Elwalid and Mitra [60] found further generalizations. This approach allows us to examine the rare event of a buffer overflow: the deviant behavior appears to be dominated by an eigenvalue of the eigensystem mentioned above.

Recently, the *large deviations* approach of analyzing fluid systems has received much attention. An asymptotic approach based on large deviations is given by Weiss [184] and leads to analyzing the occurrence of rare events in models with a large number of (identical) sources. On the other hand, a lot of other authors derive asymptotical expressions in the buffer size by using large deviations techniques. Examples are Ridder and Walrand [150], who consider one source, and de Veciana, Olivier, and Walrand [47], examining the particular case of a birth-death source.

The approaches mentioned above mostly assume the traffic intensity to be smaller than 1. In case of heavy traffic, *diffusion approximations* are discussed by Knessl and Morrison [107] for the model of Anick *et al.* [5] and by Kobayashi and Ren [108] for extensions to multiple types of traffic. Expressions for the steady-state distribution of the buffer contents are derived in these studies.

An important issue in the study of ATM networking models is the *effective bandwidth* problem. Suppose that there are several types of sources. It is very relevant to know how many of each of them can be accepted in order to maintain a reasonable service level, e.g., keeping loss probabilities below some prescribed small value. Commonly one writes the constraint $\sum_{k=1}^{K} n_k C_k \leq C$, where n_k is the number of type k sources each having an effective bandwidth C_k and C is the deterministic service rate. Kelly [99] analyzes this problem in a 'traditional' queueing context ('renewal input' streams), Gibbens and Hunt [74] consider correlated input: Markov fluid on-off sources. Kesidis, Walrand, and Chang [105] and Elwalid and Mitra [61] link the effective bandwidth concept to the eigensystem method mentioned above and the large deviations concept. They find that the effective bandwidth can be expressed as a Perron-Frobenius eigenvalue of an eigensystem. The same approach for discrete time ATM models has been pursued by Chang, Heidelberger, Juneja, and Shahabuddin [28]. Whitt [186] gives a detailed summary on the effective

bandwidth concept.

Since the techniques mentioned above mostly yield only *asymptotics* of the probabilities of interest, simulation can be quite useful. However, using ordinary Monte Carlo estimators, long simulation runs are required for good relative efficiency of the estimates, particularly when the loss probabilities are small (say of the order 10^{-9}). Importance sampling may lead to variance reductions and hence 'fast simulations' are possible, see e.g. Siegmund [167] and Glynn and Iglehart [77]. Cottrell, Malgouyres, and Fort [41] find with large deviations techniques an exponential change of measure that is in some sense optimal; Parekh and Walrand [144] apply this result in a queueing context. To execute the 'fast simulation' we have to determine the statistical characteristics of the *conjugate model,* i.e., the model under the new probability measure. Feller [68, Ch. XI and XII], Asmussen [6], and Anantharam [4] concentrate on this subject in a general setting, Ridder [148] finds sufficient conditions for the parameters of the conjugate model in case of a Markov fluid process.

The modulating input process can be considered as a continuous-time Markov chain $X(\cdot)$, on a finite state space $E = \{1, \ldots, d\}$, with infinitesimal generator $\Lambda := (\lambda_{ij})_{i,j=1}^{d}$. We recall the convention $\lambda_{ii} := -\sum_{j \neq i} \lambda_{ij} =: -\lambda_i$. We assume that the Markov chain is irreducible, and π denotes it unique steady state distribution. While the Markov chain visits state i, the buffer is filled continuously at a constant rate of $r_i > 0$ units fluid per unit time, representing e.g. ATM cells or communication packets. The input pair (Λ, r) represents the characteristics of the source that uses the buffer. The buffer is emptied by a continuous output flow with rate $C > 0$ cells per unit time, independently of the current state of the modulating process. The buffer has 'large' but finite size B. We may regard (B, C) as the system characteristics. To guarantee the stability of this system, we assume throughout $\langle \pi, r \rangle < C$, where $\langle \cdot\,, \cdot \rangle$ denotes the inner product. To avoid trivialities, we suppose that there is at least one state i with a rate r_i that is larger than C.

Defining $S(t)$ as the amount of fluid in the buffer at time t (where $t \geq 0$), the development of the buffer contents is given by the following differential equation:

$$S(t + \mathrm{d}t) = S(t) + (r_{X(t)} - C)\,\mathrm{d}t,$$

provided that the right-hand side is feasible: $0 \leq S(t) + (r_{X(t)} - C)\,\mathrm{d}t \leq B$. Otherwise the buffer contents remains on equal level (i.e., 0 or B): $S(t + \mathrm{d}t) = S(t)$.

In a way described in Ridder and Walrand [150] and Ridder [148] we can identify cycles in the evolution of $S(\cdot)$, consisting of a busy part followed by an idle part. In this chapter we give rough estimates for the probability of an arbitrary cycle being an overflow cycle. In case of buffer size B, we denote this probability by $\alpha(B)$ throughout. Notice that these cycles are not strictly regenerative, since the busy period cannot start in particular states

of the underlying Markov chains, e.g. those with $r_i < C$. How to overcome this difficulty is explained in Ridder [148] and Kesidis and Walrand [103].

Furthermore, it is clear that during an overflow the underlying Markov chain does not behave according to the transition rates of the modulating Markov chain: $S(\cdot)$ hits $[B, \infty)$ although we assume the offered load to be smaller than 1. In case of a cycle reaching overflow (B large), the Markov chain behaves according to the conjugate (or dual) rate matrix $M := (\mu_{ij})_{i,j}^{d}$. The contribution of this present chapter is to discuss two ways to find this conjugate model. The first uses large deviations theory, in particular *slow random walk* theory, in order to find the decay rate of $\alpha(B)$ and the dual rate matrix. The problem boils down to an eigenvalue problem, and we propose a straightforward, efficient algorithm to solve this. The second method gives the decay rate and dual matrix as the results of a variational problem: the minimization of an entropy function. We also prove equivalence of both methods.

The organization of this chapter is as follows. Section 2 addresses the slow random walk approach, to find decay rate and conjugate model In Section 3, we come to the same decay rate and dual matrix, but they emerge as the result of a variational problem: the minimization of an entropy function. Section 4 uses the previous results to execute importance sampling in an asymptotically optimal way. A multiple source case is considered. Conclusions can be found in the final section.

2 Slow random walk approach

Tackling the problem of finding asymptotics for $\alpha(B)$ we may investigate the free buffer process, just like in Ridder and Walrand [150]. They consider a process $\bar{S}(.)$ denoting the amount of fluid in the system after removing the boundaries 0 and B. They fix an initial state, say 1, and let ξ_j be the increment of fluid between the $(j-1)$-th and jth return to this state. These increments are independently and identically distributed, and therefore slow random walk theory becomes applicable [23, Ch. 4]. Defining the large deviations rate function or *Legendre-Fenchel transform* by

$$I_1(x) := \sup_{\theta} \big(\theta x - \log \mathrm{E} \exp(\theta \xi_1)\big),$$

it appeared that the following rough asymptotic relation is valid:

$$\lim_{B\to\infty} \frac{1}{B} \log \alpha(B) = \inf_{T>0} T\, I_1\left(\frac{1}{T}\right) = -\theta^*, \tag{3.1}$$

where θ^* denotes a positive solution of the characteristic equation $\mathrm{E} \exp(\theta \xi_1) = 1$. The renewal reward theorem says that $\langle \pi, r \rangle - C$ equals the ratio of $\mathrm{E}\xi_1$ and $\mathrm{E}Y_1$, where Y_1

denotes the time elapsed between two consecutive returns to state 1. We see that ξ_1 obviously has a negative mean. Now it is easy to prove that the characteristic equation has a unique positive root. However, how to find this root?

For the moment generating function $\mathrm{E}\exp(\theta\xi_1)$ it is in general not possible to find a simple closed expression. Conditioning on the jump from state 1, one may easily verify that it can be written as follows

$$\begin{aligned}\mathrm{E}\exp(\theta\xi_1) &= \left(\frac{\lambda_{12}}{\lambda_1}\times\frac{\lambda_1}{\lambda_1-(r_1-C)\theta}\times x_2\right)+\cdots+\left(\frac{\lambda_{1d}}{\lambda_1}\times\frac{\lambda_1}{\lambda_1-(r_1-C)\theta}\times x_d\right)\\ &= \frac{\lambda_{12}x_2+\cdots+\lambda_{1d}x_d}{\lambda_1-(r_1-C)\theta},\end{aligned}$$

where the variables x_i (where $i=2,\ldots,d$) denote the moment generating functions of the net amount of fluid generated by the free buffer model starting in state i until absorption in state 1. They are implicitly defined by

$$\begin{aligned}x_i &= \left(\frac{\lambda_{i1}}{\lambda_i-(r_i-C)\theta}\right)+\sum_{j=2,j\neq i}^{d}\left(\frac{\lambda_{ij}}{\lambda_i}\times\frac{\lambda_i}{\lambda_i-(r_i-C)\theta}\times x_j\right)\\ &= \frac{\lambda_{i1}+\lambda_{i2}x_2+\cdots+\lambda_{i,i-1}x_{i-1}+\lambda_{i,i+1}x_{i+1}+\cdots+\lambda_{id}x_d}{\lambda_i-(r_i-C)\theta}.\end{aligned}$$

So for a fixed θ, the moment generating function in this argument can be found by solving a system of $d-1$ linear equations in $d-1$ unknowns, which can be done easily using standard Gauss-Jordan procedures. Applying for instance a bisection method we can find the value of θ^*. The domain for θ^* is

$$D(\theta) := \{\theta : \lambda_i-(r_i-C)\theta > 0,\ i=1,\ldots,d\}.$$

In this domain the equations have two solutions: 0 and θ^*. Defining $x_1\equiv 1$ and recalling that $\lambda_i=-\lambda_{ii}$, we get for $i=1,\ldots d$:

$$-(r_i-C)\theta^* x_i = \sum_{j=1}^{d}\lambda_{ij}x_j.$$

We define the diagonal matrix R by $\mathrm{diag}\{r_1,\ldots,r_d\}$. As treated in Stern and Elwalid [170] we may ignore the case of $r_i=C$ for any $i\in E$. Then the system above reduces to $-\theta^* x=(R-CI)^{-1}\Lambda x$. In other words: $-\theta^*$ is eigenvalue of $(R-CI)^{-1}\Lambda$, with associated right eigenvector $x=(1,x_2,\ldots,x_d)^T$. Since the x_i represent moment generating functions, they are all strictly positive. We also see that the choice of the initial state (in the system above: state 1) does not affect the value of θ^* nor the associated eigenvector (besides the normalization: in case of initial state $i\in E$ we have $x_i\equiv 1$). We conclude that we have found the connection between the slow random walk approach and the governing

eigensystems, cf. related work of Anick *et al.* [5], Mitra [130], Stern and Elwalid [170], and Elwalid and Mitra [61].

As pointed out in Kesidis, Walrand, and Chang [105] the *effective bandwidth* of a Markov fluid source (for fixed θ) can be regarded as the output rate $C(\theta)$, such that the decay rate of $\alpha(B)$ equals $-\theta$. Again this problem can be converted into a system of equations. For fixed θ and C, the moment generating function can be calculated by solving the set of linear equations; effective bandwidth $C(\theta)$ is that value of the output rate, such that $\mathrm{E}\exp(\theta\xi_1)$ equals 1.

The effective bandwidth concept eases the task of deciding whether a source should be accepted or not, in case of multiple sources sharing a single buffer that is emptied at a rate C (call acceptance control). Suppose that a service level θ should be maintained, and source $1,\ldots,K$ are connected, having bandwidths $C_1(\theta),\ldots,C_K(\theta)$, such that $\sum_{i=1}^{K} C_i(\theta) < C$. If the effective bandwidth of the source to be accommodated fits in the 'remaining space' $C - \sum_{i=1}^{K} C_i(\theta)$ it can be accepted, otherwise it should be rejected. As said in Elwalid and Mitra [61], the effective bandwidth can also be seen as the maximal eigenvalue of $R + \Lambda/\theta$.

It is well-known that (Asmussen and Rubinstein [9]), in case of an overflow, the increments ξ_i have an *exponentially twisted distribution.* This means the following: If the increments of the free buffer process between two visits to state 1 have cumulative distribution function $F(\cdot)$, the increments along the path to level B (for large B) approximately have distribution function

$$\int_{-\infty}^{u} e^{\theta^* x}\mathrm{d}F(x) \quad \text{instead of} \quad \int_{-\infty}^{u} \mathrm{d}F(x) = F(u). \tag{3.2}$$

We call the increments of the twisted process ξ_i^*. It is easy to check that (3.2) gives the following 'translation property' holds for the moment generating function of the 'old' and 'new' increments:

$$\mathrm{E}\exp\left((\theta+\theta^*)\xi_1\right) = \mathrm{E}\exp\left(\theta\xi_1^*\right). \tag{3.3}$$

The new increments induce the conjugate process, that can be used for variance reduction objectives in simulation, as mentioned in the introduction. To perform these simulations, it is important to find the rate matrix $M = (\mu_{ij})_{i,j=1}^{d}$ of a modulating Markov chain that goes with ξ_1^*. However, since in general no closed expression for the moment generating function can be found, it is unclear at this stage how to find these rates.

In a very direct way, we can find sufficient conditions to satisfy (3.3). Consider an arbitrary cycle from state 1 to state 1, visiting $1 = i_0 \to i_1 \to \cdots \to i_n \to i_0 = 1$, satisfying $i_j \neq i_{j+1}$, $j = 0,\ldots,n-1$, $i_0 \neq i_n$. The chain stays a time t_k in state i_k. The

probability of this realization to occur is

$$\left(\frac{\lambda_{i_0 i_1}}{\lambda_{i_0}}\right)\cdots\left(\frac{\lambda_{i_n i_0}}{\lambda_{i_n}}\right) \times \big(\lambda_{i_0}\exp(-\lambda_{i_0}t_0)\big)\cdots\big(\lambda_{i_n}\exp(-\lambda_{i_n}t_n)\big)\mathrm{d}t_0\cdots\mathrm{d}t_n. \tag{3.4}$$

So the moment generating function of ξ_1 in $\theta+\theta^*$ equals

$$\sum_{\text{all cycles}}\int_{-\infty}^{\infty}\int_{t_j:\sum(r_{i_j}-C)t_j=x} e^{(\theta+\theta^*)x}\big(\lambda_{i_0 i_1}\cdots\lambda_{i_n i_0}\big) \times \big(e^{-\lambda_{i_0}t_0}\cdots e^{-\lambda_{i_n}t_n}\big)\mathrm{d}t_0\cdots\mathrm{d}t_n\mathrm{d}x,$$

where the summation is over all cycles starting and ending in 1 and the inner integration over all t_j satisfying $\sum_{j=0}^{n} r(i_j)t_j = x$. This expression equals $\mathrm{E}\exp(\theta\xi_1^*) =$

$$\sum_{\text{all cycles}}\int_{-\infty}^{\infty}\int_{t_j:\sum(r_{i_j}-C)t_j=x} e^{\theta x}\big(\mu_{i_0 i_1}\cdots\mu_{i_n i_0}\big) \times \big(e^{-\mu_{i_0}t_0}\cdots e^{-\mu_{i_n}t_n}\big)\mathrm{d}t_0\cdots\mathrm{d}t_n\mathrm{d}x,$$

cf. Ridder [148]. Clearly, if the following conditions are met, we have an equality:

- for any subset $\{i_0, i_1, \ldots, i_n\}$ of the state space E (with $n \in \mathbb{N}$ and satisfying $i_j \neq i_{j+1}$, $j = 0, \ldots, n-1$, $i_0 \neq i_n$)

$$\lambda_{i_0 i_1}\cdots\lambda_{i_n i_0} = \mu_{i_0 i_1}\cdots\mu_{i_n i_0}. \tag{3.5}$$

  In words: the products of subsequent transition rates in cycles of the chain with rate matrix M are equal to these products of the original chain with transition matrix Λ.

- for any $i \in E$, we have

$$\mu_i = \lambda_i - \theta^*(r_i - C). \tag{3.6}$$

  This is a simple relation between the rowsums of the original transition matrix Λ and those of M. We see that in case of $r_i < C$ for some state i, we have that $\mu_i > \lambda_i$. This means that average time the chain spends per visit in state i is under M smaller than under Λ. Conversely, if $r_i > 0$, the mean time of a visit to state i is shorter under Λ than under M. This is intuitively reasonable: If the fluid process reaches a high level the visits to states with a positive rate must be longer and those to states with a negative rate shorter than in the original chain.

Suppose we use the following transformation of the original transition rates ($i \neq j$)

$$\mu_{ij} = \lambda_{ij}\frac{x_j}{x_i}, \tag{3.7}$$

where $x = (1, x_2, \ldots, x_d)$ is the right eigenvector of the matrix $(R-CI)^{-1}\Lambda$ corresponding to eigenvalue $-\theta^*$. Clearly, condition (3.5) is satisfied. Moreover, we have

$$\mu_i = \sum_{j\neq i}\mu_{ij} = \lambda_i + \sum_{j=1}^{d}\lambda_{ij}\frac{x_j}{x_i} = \lambda_i - (r_i - C)\theta^*,$$

yielding that (3.6) is met as well.

We conclude that it is sufficient to choose the μ_{ij} $(i \neq j)$ according to (3.7), $\mu_i := \sum_{j \neq i} \mu_{ij}$. However, this transformation appears to be necessary as well. This can be derived as follows. Until now, we found sufficient conditions for the dual transition rates, and we showed that they can be found by solving an eigensystem. Because $\mathrm{E}\exp(\theta\xi_1) = 1$ has only roots θ^* and 0 (both with multiplicity 1, since the moment generating function has a positive slope in θ^* and a negative in 0), the other $d-2$ eigenvalues must either lie outside the domain or be complex. Therefore, the conjugate process is determined uniquely, implying that transformation (3.7) is necessary as well.

3 Equivalence with entropy functions

The change of measure established in the previous section can also be found by examining entropy functions. The new transition matrix appears to be the optimizing argument in a variational problem, as we will see in this section.

As treated in Ridder and Walrand [150], we can find the most likely equilibrium distribution ρ of the underlying Markov chain, given that the buffer contents reaches a high level. This is done as follows. Instead of examining the roots of $\mathrm{E}\exp(\theta\xi_1) = 1$, we can find θ^* as a result of the optimization program $\inf T\, I_2(\tilde{\rho})$. Here T ranges over $(0, \infty)$ and $\tilde{\rho}$ over all probability measures on E such that the inner product of $\tilde{\rho}$ and r is $1/T$. The level 2 large deviation rate function for finite state continuous time Markov chains $I_2(\cdot)$ is given in Donsker and Varadhan [52]. It can be shown that the result of this program equals expression (3.1), see [150]. The optimizing argument can be regarded as the most likely equilibrium distribution of the modulating chain in case of overflow.

This method has several drawbacks. Firstly, in general the function $I_2(\cdot)$ cannot be evaluated easily. Only in very few special cases a simple closed-form expression for $I_2(\cdot)$ can be calculated, for instance in case of birth death Markov chains. Secondly, it is true that we derive an important characteristic of the conjugate process, namely the long-run distribution of the modulating Markov process ρ, but this does not determine uniquely the complete probabilistic behavior of this chain. For this reason, we are not able to perform importance sampling with an 'optimal' change of measure. To do that, we have to find the dual rate matrix M.

To cope with the second drawback mentioned in the previous paragraph, we must not only consider the fraction of time the process spends in each state (resulting in ρ), but also the fraction of jumps which are from $i \in E$ to $j \in E$, where $i \neq j$. To explain the main ideas of this concept we first consider an irreducible discrete time Markov chain $\{X_n, n \in \mathbb{N}_0\}$ on a finite state space, say $E := \{1, \ldots, d\}$. Let $P = (p_{ij})_{i,j=1}^d$ be the matrix

containing the transition probabilities; π_P is its invariant. Let $R_n(i,j)$ be the *empirical pair measure*:

$$R_n(i,j) := \frac{1}{n}\sum_{k=1}^{n} 1_{\{i\}}(X_{k-1})1_{\{j\}}(X_k).$$

As we know, $R_n(i,j)$ converges to $\pi_P(i)p_{ij}$, almost surely $(n \to \infty)$. Note that for large n it holds that both marginals of $R_n(i,j)$ are identical:

$$\lim_{n\to\infty}\sum_{j=1}^{d} R_n(i,j) = \lim_{n\to\infty}\sum_{j=1}^{d} R_n(j,i)$$

with probability 1. It is clear that, in case of the large deviations of $R_n(\,\cdot\,,\,\cdot\,)$, we only need to consider matrices $R = (r_{ij})_{i,j=1}^{d}$ with all entries non-negative, adding up to 1 and $\sum_{j=1}^{d} r_{ij} = \sum_{j=1}^{d} r_{ji}$. As can be found in Dembo and Zeitouni [50], $R_n(\,\cdot\,,\,\cdot\,)$ obeys a large deviations principle with rate function

$$I_3(R) := \sum_{i=1}^{d}\sum_{j=1}^{d} r_{ij}\log\left(\frac{r_{ij}}{p_{ij}}\right) - \sum_{i=1}^{d} r_{i\cdot}\log r_{i\cdot},$$

with $r_{i\cdot} := \sum_{j=1}^{d} r_{ij}$. Here $0\log 0$ and $\log(0/0)$ are defined to be 0. The entropy function $I_3(\cdot)$ can be rewritten as a *relative entropy* between two transition matrices, see Mandjes [118]:

$$I_3(Q \mid P) := \sum_{i=1}^{d}\pi_Q(i)\sum_{j=1}^{d} q_{ij}\log\left(\frac{q_{ij}}{p_{ij}}\right).$$

The notion 'relative entropy' enables us to examine the probability that a Markov chain with actual transition matrix P acts as if obeying different transition probabilities (the entries of Q) during a substantial period. In fact, $I_3(Q \mid P)$ provides us the exponent of the probability of observing a chain that behaves according to Q instead of P. By discretization (cf. Donsker and Varadhan [52]) we can derive also a relative entropy of a rate matrix M (with invariant ρ) with respect to a transition matrix $\Lambda := (\lambda_{ij})_{i,j=1}^{d}$ of a continuous time Markov chain $X(\cdot)$:

$$I_3(M \mid \Lambda) := \sum_{i=1}^{d}\rho_i\sum_{j=1,j\neq i}^{d}\mu_{ij}\log\left(\frac{\mu_{ij}}{\lambda_{ij}}\right) + \sum_{i=1}^{d}\rho_i\big(\mu_{ii}-\lambda_{ii}\big),$$

see Kesidis and Walrand [104]. A heuristic derivation of this entropy is sketched in Mandjes [118]. One can show that the following contraction principle holds:

$$I_2(\rho) = \inf_{M:\rho M=0} I_3(M \mid \Lambda),$$

cf. de Veciana *et al.* [47]. Therefore, the following two optimization programs are equivalent, and equal to θ^*:

$$\left(\inf_{T>0,\langle\widetilde{\rho},r\rangle=1/T+C} TI_2(\widetilde{\rho}) = \inf_{\widetilde{\rho}:\langle\widetilde{\rho},r\rangle>C}\frac{I_2(\widetilde{\rho})}{\langle\widetilde{\rho},r\rangle - C}\right) \quad\text{and}\quad \inf_{\widetilde{M}:\ \langle\widetilde{\rho},r\rangle>C}\frac{I_3(\widetilde{M}\mid\Lambda)}{\langle\widetilde{\rho},r\rangle - C}, \tag{3.8}$$

where $\widetilde{\rho}$ denotes the invariant of $\widetilde{M}$: $\widetilde{\rho}\widetilde{M} = 0$ and $\sum_{i=1}^{d} \rho_i = 1$. We conclude that the drawbacks mentioned earlier in this section are removed: the entropy function is a simple closed form expression and the minimization yields an entire rate matrix. We claim that the latter expression in the previous display is optimized by the rates found in Section 2.

In Section 2 it was explained how to find an conjugate random variable ξ_1^* of ξ_1 satisfying the translation relation (3.3). Clearly we have $\mathrm{E}\exp(-\theta^*\xi_1^*) = 1$, yielding that $\theta^{**} = -\theta^*$. Again applying (3.3), we get that the conjugate of this conjugate is again the original random variable: $\xi_1^{**} = \xi_1$. Moreover, we find easily

$$\lim_{B\to\infty} \frac{1}{B} \log \mathcal{P}\left(\left\{ \sum_{k=1}^{j} \xi_k, j \in \mathbb{N} \right\} \text{ hits } [B,\infty) \text{ during a cycle} \right) =$$

$$\lim_{B\to\infty} \frac{1}{B} \log \mathcal{P}\left(\left\{ -\sum_{k=1}^{j} \xi_k^*, j \in \mathbb{N} \right\} \text{ hits } [B,\infty) \text{ during a cycle} \right) = -\theta^*.$$

Therefore, combining the above results, we found the following duality relation, $\widetilde{\pi}$ invariant of $\widetilde{\Lambda}$:

$$\frac{I_3(M \mid \Lambda)}{\langle \rho, r\rangle - C} = \inf_{\widetilde{M}:\langle\widetilde{\rho},r\rangle > C} \frac{I_3(\widetilde{M} \mid \Lambda)}{\langle \widetilde{\rho}, r\rangle - C} = \theta^* = \inf_{\widetilde{\Lambda}:\langle\widetilde{\pi},r\rangle < C} \frac{I_3(\widetilde{\Lambda} \mid M)}{-\langle \widetilde{\pi}, r\rangle + C} = \frac{I_3(\Lambda \mid M)}{-\langle \pi, r\rangle + C}.$$

From the definition of the relative entropy $I_3(\cdot \mid \cdot)$, we conclude that the fact that θ^* is finite implies that $\lambda_{ij} = 0 \Longleftrightarrow \mu_{ij} = 0$ (for $i, j \in E, i \neq j$). We derived that the Markov chains equipped with transition matrices Λ and M, respectively, have the same chain structure. We conclude that the chain governed by M is irreducible as well; given M, ρ is uniquely determined and consists of merely positive components.

In order to solve the minimization in the right hand side of (3.8), we introduce the *Lagrangian* which has to be minimized over the transition matrices $\widetilde{M}$ with $\langle \widetilde{\rho}, r\rangle > C$ and which contains already the condition $\widetilde{\rho}\widetilde{M} = 0$:

$$L(\widetilde{M}, \widetilde{\rho}, \widetilde{K}) := \frac{I_3(\widetilde{M} \mid \Lambda)}{\langle \widetilde{\rho}, r\rangle - C} - \sum_{i=1}^{d} \widetilde{K}_i \left(\sum_{j\neq i} \widetilde{\mu}_{ij}\widetilde{\rho}_i - \sum_{j\neq i} \widetilde{\mu}_{ji}\widetilde{\rho}_j \right)$$

$$= \frac{\sum_{i=1}^{d} \left(\widetilde{\rho}_i \sum_{j=1, j\neq i}^{d} f(i,j) \right)}{\langle \widetilde{\rho}, r - C\rangle} - \sum_{i=1}^{d} \widetilde{K}_i\, g(i),$$

where $f(\cdot,\cdot)$ and $g(\cdot)$ are defined for fixed $\widetilde{M}$ and $\widetilde{\rho}$:

$$\begin{aligned} f(i,j) &:= \widetilde{\mu}_{ij} \log(\widetilde{\mu}_{ij}/\lambda_{ij}) + \lambda_{ij} - \widetilde{\mu}_{ij} \text{ for distinct } i \text{ and } j \in E \\ g(i) &:= \sum_{j\neq i} \widetilde{\mu}_{ij}\widetilde{\rho}_i - \sum_{j\neq i} \widetilde{\mu}_{ji}\widetilde{\rho}_j \text{ for } i \in E. \end{aligned}$$

We make a couple of remarks:

- Since $I_3(\cdot \mid \Lambda)$ as well as $\langle \cdot, r - C \rangle$ are proportional in $\tilde{\rho}$, we may omit the normalization equation $\sum_{i=1}^{d} \tilde{\rho}_i = 1$.
- For notational convenience, we did not remove the equation $g(d) = 0$, although redundant.

It is a well-known result that a necessary condition for the minimum is that all partial derivatives equal 0. Equating the partial derivatives of the $\tilde{\mu}_{ij}$ (where $i \neq j$), $\tilde{\rho}_i$ and $\widetilde{K}_i$ to 0 yields subsequently:

$$\frac{\partial L}{\partial \tilde{\mu}_{ij}} = \tilde{\rho}_i \left(\frac{\log(\tilde{\mu}_{ij}/\lambda_{ij})}{\langle \tilde{\rho}, r \rangle - C} - \widetilde{K}_i + \widetilde{K}_j \right) = 0 \,; \tag{3.9}$$

$$\begin{aligned} \frac{\partial L}{\partial \tilde{\rho}_i} &= \frac{\sum_{j \neq i} f(i,j)(\langle \tilde{\rho}, r \rangle - C) - \left(\sum_{k=1}^{d} \tilde{\rho}_k \sum_{j \neq k} f(k,j) \right)(r_i - C)}{(\langle \tilde{\rho}, r \rangle - C)^2} - \widetilde{K}_i \sum_{j \neq i} \tilde{\mu}_{ij} + \sum_{j \neq i} \widetilde{K}_j \tilde{\mu}_{ij} \\ &= \frac{\sum_{j \neq i} f(i,j)}{\langle \tilde{\rho}, r \rangle - C} - \frac{\left(\sum_{k=1}^{d} \tilde{\rho}_k \sum_{j \neq k} f(k,j) \right)(r_i - C)}{(\langle \tilde{\rho}, r \rangle - C)^2} + \sum_{j \neq i} (\widetilde{K}_j - \widetilde{K}_i) \tilde{\mu}_{ij} = 0 \,; \end{aligned} \tag{3.10}$$

$$\frac{\partial L}{\partial \widetilde{K}_i} = -\sum_{j \neq i} \tilde{\mu}_{ij} \tilde{\rho}_i + \sum_{j \neq i} \tilde{\mu}_{ji} \tilde{\rho}_j = 0 \ \left(\text{or } g(i) = 0 \right). \tag{3.11}$$

From now on we examine relations (3.9), (3.10) and (3.11) a bit more precisely. Suppose that the infimum is attained in (M, ρ, K). Since the vector ρ consists of only positive entries, equation (3.9) yields $\log(\mu_{ij}/\lambda_{ij}) = (\langle \rho, r \rangle - C)(K_i - K_j)$, where $i, j \in E$ and $i \neq j$. Now consider an arbitrary subset $\{i_0, i_1, \ldots, i_n\}$ of E, with $n \in \mathbb{N}$ and satisfying the following constraints: $i_j \neq i_{j+1}$, $j = 0, \ldots, n-1$, $i_0 \neq i_n$. It follows immediately that (defining $i_{n+1} := i_0$)

$$\log\left(\frac{\mu_{i_0 i_1}}{\lambda_{i_0 i_1}} \right) + \cdots + \log\left(\frac{\mu_{i_n i_0}}{\lambda_{i_n i_0}} \right) = (\langle \rho, r \rangle - C) \times \left(\sum_{j=1}^{n+1} (K_{i_{j-1}} - K_{i_j}) \right).$$

Clearly, the right-hand side of the previous display equals 0. We see that we obtain that condition (3.5) is necessary. Some elementary calculus yields for all $i \in E$

$$\frac{L(\widetilde{M}, \tilde{\rho}, \widetilde{K})(r_i - C)}{\langle \tilde{\rho}, r \rangle - C} = \left(\sum_{j \neq i} \frac{f(i,j)}{\langle \tilde{\rho}, r \rangle - C} + \sum_{j \neq i} (\widetilde{K}_j - \widetilde{K}_i) \tilde{\mu}_{ij} \right) - \frac{\partial L}{\partial \tilde{\rho}_i} - \frac{r_i - C}{\langle \tilde{\rho}, r \rangle - C} \sum_{j=1}^{d} \widetilde{K}_j \, g(j).$$

Under the necessary conditions (3.10) and (3.11), the first and the last term of the right-hand side equal 0 in (M, ρ, K). Since θ^* is the value of the Lagrangian in a stationary point, plugging in condition (3.9), $\theta^*(r_i - C)/(\langle \rho, r \rangle - C)$ can be rewritten as follows:

$$\frac{\theta^*(r_i - C)}{\langle \rho, r \rangle - C} = \left(\sum_{j \neq i} \frac{f(i,j)}{\langle \rho, r \rangle - C} + \sum_{j \neq i} (K_j - K_i) \mu_{ij} \right)$$

$$= \left(\sum_{j \neq i} \frac{f(i,j)}{\langle \rho, r \rangle - C} - \sum_{j \neq i} \frac{\mu_{ij} \log(\mu_{ij}/\lambda_{ij})}{\langle \rho, r \rangle - C} \right) = \sum_{j \neq i} \frac{(\lambda_{ij} - \mu_{ij})}{\langle \rho, r \rangle - C}.$$

We found the second necessary condition (3.6).

Note that the condition $\langle \rho, r \rangle > C$ excludes the trivial solution $M = \Lambda$ and $\theta^* = 0$. We saw that it necessarily holds that (for distinct i and j) $\lambda_{ij} = 0 \Longrightarrow \mu_{ij} = 0$ and otherwise

$$\frac{\mu_{ij}}{\lambda_{ij}} = \frac{\exp(\,(\langle \rho, r \rangle - C)K_i\,)}{\exp(\,(\langle \rho, r \rangle - C)K_j\,)},$$

for solution K. These relations show the existence of some vector y (coordinatewise positive) such that

$$\mu_{ij} = \lambda_{ij} \frac{y_j}{y_i}.$$

Suppose that we found such a vector y, we have because of necessary condition (3.6) that

$$\sum_{j \neq i} \lambda_{ij} \frac{y_j}{y_i} = \sum_{j \neq i} \lambda_{ij} - \theta^*(r_i - C),$$

which can be read as $\sum_{j=1}^{d} \lambda_{ij} y_j = -(r_i - C)\theta^* y_i$. We see that y equals the eigenvector x mentioned before, except possibly a normalizing constant which does not influence relation (3.7). We conclude that the dual rate matrices found by both methods coincide.

We remark that the entries of the optimizing rate matrix in (3.8) are of the form (3.7). Therefore the minimization program can be considerably simplified: $d - 1$ variables are involved rather than $d^2 - d$. We conclude that we have encountered an important reduction property. Using some minimization code, the solution of this variational problem can be derived relatively easily now.

De Veciana *et al.* [47] find the decay rate as well as the dual transition matrix in case of a modulating Markov process of birth-death type. To find M and θ^*, they propose in fact solving (3.5) and (3.6). Our analysis, on the other hand, covers the general case and, apart from that, simplifies the search for the dual matrix.

4 Fast simulation by importance sampling

This section deals with 'fast simulation' methods to estimate overflow probabilities. In the first subsection we show that our choice of the importance sampling distribution is within a certain class optimal. The second subsection gives a simulation example, in which multiple Markov fluid sources are involved, sharing one buffer.

4.1 Asymptotic optimality

In finite buffer models the choice of buffer size B is often an important issue. We may choose B for instance such that the probability of a loss cycle, $\alpha(B)$, is very small, typically

in the order of 10^{-9}. For these design purposes it is useful to have accurate estimates of $\alpha(B)$.

From the preceding sections we found the decay rate of $\alpha(B)$, but not an approximation of the probability itself. Therefore, we may try to estimate it by means of *Monte Carlo simulation*: generate cycles of the fluid process and estimate $\alpha(B)$ by the ratio of the number of overflow cycles to the total number of cycles. For large levels B the number of samples to draw must be large in order to obtain good relative efficiency of the estimate. For instance, is $\alpha(B)$ of the order 10^{-9}, then the number of cycles required is of the order 10^{11} if a confidence of 95% and a relative efficiency of 10% is required. This imposes strong demands on the random generator. Apart from that, this Monte Carlo method would obviously be very time consuming.

To cope with the problems mentioned in the previous paragraph, we may tackle the problem using importance sampling, as introduced in Chapter 2. In Ridder [148] some details on the implementation are treated. We assume some underlying probability triple $(\Omega, \mathcal{F}, \mathcal{P})$, under which the original Monte Carlo simulation would be executed. We start the simulation of the process in an empty system and finish a run either at a buffer overflow or a return on level 0. Each $\omega \in \Omega$ is of the form

$$\omega = \big((i_0, t_0),\ (i_1, t_1),\ \ldots,\ (i_n, t_n)\big),$$

where i_k represents the state of the chain after the kth jump during this sample cycle and t_k measures the amount of time spent in state i_k. Clearly a cycle is a loss cycle if $\sum_{k=0}^{n}(r(i_k) - C)t_k \geq B$, whereas it represents a return to level 0 if $\sum_{k=0}^{n}(r_{i_k} - C)t_k \leq 0$.

Consider the possibility of executing the simulation based on another probability $\mathcal{Q}$, such that $\mathcal{P}$ is absolutely continuous relative to $\mathcal{Q}$. Since the chain behaves, in case of buffer overflows, as under M instead of Λ it seems reasonable to associate $\mathcal{Q}$ with the fluid model governed by an underlying Markov chain with rates M. Then the likelihood of a cycle satisfies

$$L(\omega) = \frac{\mathrm{d}\mathcal{P}}{\mathrm{d}\mathcal{Q}}(\omega) = \frac{\lambda_{i_0 i_1} \cdots \lambda_{i_{n-1} i_n} \lambda_{i_n}}{\mu_{i_0 i_1} \cdots \mu_{i_{n-1} i_n} \mu_{i_n}} \exp\left(-\sum_{k=0}^{n}(\lambda_{i_k} - \mu_{i_k})t_k\right),$$

cf. density (3.4). According to (3.6), the exponent in the previous likelihood equals $-\theta^* \sum_{k=0}^{n}(r_{i_k} - C)t_k$, which is smaller than $-\theta^* B$ in an overflow cycle. Bearing in mind the parametrization (3.7)

$$\frac{\lambda_{i_0 i_1} \cdots \lambda_{i_{n-1} i_n} \lambda_{i_n}}{\mu_{i_0 i_1} \cdots \mu_{i_{n-1} i_n} \mu_{i_n}} = \frac{x_{i_0}}{x_{i_n}} \frac{\lambda_{i_n}}{\mu_{i_n}} \leq \max_{i,j \in E} \frac{x_i}{x_j} \frac{\lambda_j}{\mu_j} =: K.$$

As explained in Chapter 2, $\alpha(B) = \mathrm{E}_B^{(\mathcal{Q})}(LI)$, L denoting the likelihood, I being 1 in case of an overflow cycle and 0 otherwise. The sample mean $\sum_{i=1}^{n} L_i I_i / n$ (simulation

performed under measure $\mathcal{Q}$) is therefore an unbiased estimator. As in Chapter 2, to gain insight into its variance performance, we again use the notion of asymptotic optimality. Since variances are non-negative, we have $\mathrm{E}_B^{(\mathcal{Q})}(L^2 I) \geq \alpha^2(B)$. Applying (3.1), we get that for all $\mathcal{Q}$

$$\liminf_{B\to\infty} \frac{1}{B} \log \mathrm{E}_B^{(\mathcal{Q})}(L^2 I) \geq -2\theta^*.$$

But, with our specific choice of the new law of the fluid process we have that $\mathrm{E}_B^{(\mathcal{Q})}(L^2 I)$ is dominated by $K^2\, e^{-2\theta^* B}$. Therefore our choice is asymptotically optimal:

$$\limsup_{B\to\infty} \frac{1}{B} \log \mathrm{E}_B^{(\mathcal{Q})}(L^2 I) \leq \lim_{B\to\infty} \frac{2}{B} \log K - 2\theta^* = -2\theta^*.$$

It is a well-known result that in case of an overflow the buffer content builds up essentially linear with slope $\langle \rho, r\rangle - C > 0$ (cf. Anantharam [4]), exactly the drift of the fluid process under M. We find that the average behavior of the fluid system under M looks like the most probable trajectory under Λ under condition of an overflow. The drift of the new process becomes positive instead of $\langle \pi, r\rangle - C < 0$.

4.2 A simulation example

Following the procedures explained above, we now work out the concepts of 'fast simulation'. It appears that the 'fast simulation' method provides us a possibility to speed up simulation runs substantially. The rates of the dual transition matrices can be calculated from the optimization program in the righthand-side of display (3.8) in conjunction with the reduction property mentioned in the previous section, or the system of equations proposed in Section 2. These methods enable us to find the conjugate process even for systems of considerable proportions.

As an example we now treat the analysis of a multiplexing system fed by *multiple independent Markov fluid sources.* In Section 1, we associated a customer loading packets into a buffer with a continuous time Markov chain $X(\cdot)$ and traffic rate function. In other words: a customer is characterized by (Λ, r), whereas (B, C) reflects the system characteristics. In this example, there are several customers (Λ_i, r_i), $i = 1, \ldots, K$, sharing the same buffer (B, C). The sources behave statistically independent in time.

Without loss of generality, we first consider two Markov fluid sources. Customer i is recorded by Markov chain $X_i(\cdot)$ on $\{1, \ldots, d_i\}$, equipped with rate matrix Λ_i and traffic rate function r_i, $i = 1, 2$. The two sources can be combined to one aggregate source on $\{1, \ldots, d_1\} \times \{1, \ldots, d_2\}$. The input per unit time is $r_{1,i} + r_{2,j}$ if the first modulating Markov chain is in state i and the second in j. In matrix notation, one may equivalently rewrite the traffic rate matrix R as the Kronecker sum of R_1 and R_2, i.e., $R_1 \oplus R_2$, see

Elwalid and Mitra [61]. Analogously, the generator of the aggregate chain can be given by $\Lambda := \Lambda_1 \oplus \Lambda_2$. The following conclusions can be drawn:

- As we mentioned before, the dual rates are provided by the governing eigensystem $-\theta^* x = (R - CI)^{-1}\Lambda x$, or equivalently $Cx = (R + (1/\theta^*)\Lambda)x$. This system can be rewritten as follows:

$$Cx = ((R_1 \oplus R_2) + (1/\theta^*)(\Lambda_1 \oplus \Lambda_2))\, x = ((R_1 + (1/\theta^*)\Lambda_1) \oplus (R_2 + (1/\theta^*)\Lambda_2))\, x.$$

  Analogously to Elwalid and Mitra [61], we obtain that the depletion rate C can be decomposed as $C_1 + C_2$, where C_i is eigenvalue of the matrix $\Lambda_i + (1/\theta^*)R_i$, with accompanying eigenvector x_i. The C_i can be regarded as the depletion rate of a single source to make the decay rate of the loss probability of that source $-\theta^*$, i.e., the effective bandwidth.

- Every non-diagonal entry of Λ_1 (Λ_2) is d_2 (d_1) times present in Λ. We may ask ourselves whether this structure carries over, i.e., do the corresponding entries in its conjugate transition matrix coincide as well? In other words: does it hold that $M = M_1 \oplus M_2$, for rate matrices M_1 and M_2? This is indeed the case and a direct implication of the fact that the eigenvector x is a Kronecker product: $x = x_1 \otimes x_2$. Defining $X := \mathrm{diag}\{x\}$, we get the new rate matrix by calculating $M := X^{-1}\Lambda X - \theta^*(R - CI_{d_1 d_2})$; here I_n stands for the identity matrix of order n. It is easily verified that

$$R - CI_{d_1 d_2} = (R_1 - C_1 I_{d_1}) \oplus (R_2 - C_2 I_{d_2}),$$

  where C_1 and C_2 denote the effective bandwidths of both sources. On the other hand, using some elementary properties from Kronecker algebra:

$$\begin{aligned} X^{-1}\Lambda X &= (X_1^{-1} \otimes X_2^{-1})(\Lambda_1 \oplus \Lambda_2)(X_1 \otimes X_2) \\ &= (X_1^{-1} \otimes X_2^{-1})(\Lambda_1 \otimes I_{d_2})(X_1 \otimes X_2) + (X_1^{-1} \otimes X_2^{-1})(I_{d_1} \otimes \Lambda_2)(X_1 \otimes X_2) \\ &= (X_1^{-1}\Lambda_1 X_1) \otimes (X_2^{-1} I_{d_2} X_2) + (X_1^{-1} I_{d_1} X_1) \otimes (X_2^{-1}\Lambda_2 X_2) \\ &= (X_1^{-1}\Lambda_1 X_1) \otimes I_{d_2} + I_{d_1} \otimes (X_2^{-1}\Lambda_2 X_2) = (X_1^{-1}\Lambda_1 X_1) \oplus (X_2^{-1}\Lambda_2 X_2). \end{aligned}$$

  We get $M = M_1 \oplus M_2$, where $M_i = (X_i^{-1}\Lambda_i X_i) - \theta^*(R_i - C_i I_{d_i})$, $i = 1, 2$. Of course, this decoupling result can be extended (inductively) to higher dimensions.

We conclude with a numerical example, concerning the simulation of a multiclass traffic system. We distinguish between three types of customers loading data packets into a finite buffer. All sources have peak rates 5 Mbit/s, while their mean rates are 512 kbit/s. We have 40 sources of the on-off type, where the average burst length is 50 msec. Furthermore, we have 5 sources of both type 2 and type 3, which are modeled by 3-state and 4-state

modulating Markov chains, respectively. They consist, apart from the bursty state and the off-state, of states with moderate traffic rates. The rate matrices (in s^{-1}) are:

$$\Lambda_1 = \begin{pmatrix} * & 20 \\ 2.28 & * \end{pmatrix}, \Lambda_2 = \begin{pmatrix} * & 1 & 20 \\ 2 & * & 10 \\ 1 & 3.83 & * \end{pmatrix}, \Lambda_3 = \begin{pmatrix} * & 1 & 2 & 20 \\ 2 & * & 3 & 10 \\ 1 & 1 & * & 10 \\ 1 & 4.60 & 1 & * \end{pmatrix}$$

The traffic rates are $r_1 = (5,0)^T$, $r_2 = (5,1,0)^T$, and $r_3 = (5,1,0.5,0)^T$, in Mbits/s. The depletion rate C amounts to 35 Mbit/s, and therefore the offered load is 0.73. Making use of the effective bandwidths, we can find the dual rates. We get $\theta^* = 1.5018$, the bandwidths of the three types are 0.7192, 0.6358, and 0.6106, respectively. The dual rate matrices are

$$M_1 = \begin{pmatrix} * & 13.57 \\ 3.36 & * \end{pmatrix}, M_2 = \begin{pmatrix} * & 0.77 & 13.67 \\ 2.59 & * & 8.86 \\ 1.46 & 4.32 & * \end{pmatrix}, M_3 = \begin{pmatrix} * & 0.78 & 1.46 & 14.17 \\ 2.56 & * & 2.80 & 9.06 \\ 1.37 & 1.07 & * & 9.72 \\ 1.41 & 5.08 & 1.03 & * \end{pmatrix}.$$

We already remarked that the cycles are not iid, and therefore we use a batch means approach, see Kesidis and Walrand [103]. In the table we give 95% confidence estimates, where we simulated until the ratio of the confidence interval half-length to the estimate (i.e., the relative efficiency) was below 10%. It is well known that, under the original measure, the number of cycles to get a fixed level of confidence is proportional to the reciprocal of the probability to be estimated. Therefore, it increases exponentially fast (at a rate of θ^*) in the buffer size. However, under the asymptotically optimal importance sampling measure $\mathcal{Q}$, this number is more or less a constant in B. Finally we notice that the simulation procedure can be extended in a straightforward way to a technique which enables us to estimate the long-run fraction of time spent at level B, or the long-run fraction of data packets lost, see Chang *et al.* [28].

In the Table 1 the estimates are given using the direct method as well as the optimal importance sampling measure. We also gave the number of cycles required to get this estimate. Because of large simulation times, we omitted direct estimates for large buffer sizes.

From the table we empirically find that the probability of an overflow cycle is asymptotically exponential with decay rate θ^* and an unknown amplitude η :

$$\alpha(B)\exp(\theta^* B) \to \eta, \text{ where } B \to \infty.$$

We conclude that fast simulation is an important tool in buffer dimensioning. It enables us also to find the value of the depletion rate C such that the overflow probability is below

Table 1: Simulation results.

	Fast simulation	
B ($\times 10^6$ bit)	$\hat{\alpha}(B)$	# runs ($\times 10^2$)
1	$3.76 \cdot 10^{-2}$	8
2	$7.28 \cdot 10^{-3}$	8
3	$1.41 \cdot 10^{-3}$	8
4	$3.34 \cdot 10^{-4}$	8
6	$1.50 \cdot 10^{-5}$	9
10	$3.91 \cdot 10^{-8}$	8
15	$2.23 \cdot 10^{-11}$	9

	Direct simulation	
B ($\times 10^6$ bit)	$\hat{\alpha}(B)$	# runs ($\times 10^4$)
1	$3.51 \cdot 10^{-2}$	1
2	$6.91 \cdot 10^{-3}$	6
3	$1.53 \cdot 10^{-3}$	26

a certain level. However, then we have to execute simulations with different values of C, and that means that different optimal importance sampling parameters are involved as well.

5 Conclusions

We have considered a Markov fluid model, in which customers use a buffer of finite size B. The overflow probability emerging in this model, $\alpha(B)$, can be approached on specific probabilistic levels, with different advantages and disadvantages. Based on slow random walk results, we may find the decay rate of $\alpha(B)$ and the optimal change of the underlying probabilistic model. The root of $\mathrm{E}\exp(\theta\xi_1) = 1$ must be found, where each evaluation of the moment generating function is equivalent to solving a linear system of $d-1$ equations in $d-1$ unknowns. We also treated an alternative method to find θ^* and the dual rate matrix M, using entropy functions. In this case the minimum of a function of $d-1$ variables has to be determined, while a system of d linear equations in the same number of unknowns (i.e., the invariant has to be determined!) must be solved in each function evaluation. It is not clear yet whether the procedure proposed in Section 2 or the minimization of Section 3 is computationally less demanding.

Then we have focused on applying these expressions for variance reduction objectives. We wanted to execute Monte Carlo simulations in order to estimate the overflow probability. We changed the underlying probability model such that the negative drift of the fluid system was replaced by a positive one with the slope of the optimal trajectory that causes overflows. The new estimators have variance properties obviously superior to the direct Monte Carlo estimators: the required simulation time is persistently orders of magnitude smaller. Therefore, 'fast simulation' permits us to perform tests on buffer sizes and service rates to gain insight into the consequences for overflow probabilities.

Chapter 4

Markov fluid queues with many sources

This chapter is concerned with overflows in queues fed by Markov fluid input. The results are asymptotic in the number of sources, i.e., we let the number of users grow large. The main objectives of this study are to characterize overflow probability as well as 'most probable way' in which overflow occurs. Applying large deviations techniques, known results [184] for exponential on-off sources are extended to general Markov fluid input. Successively, zero, small, and large buffers are treated. Finally, results for multiclass input are given.

1 Introduction

In view of the implementation of ATM (asynchronous transfer mode) based high-speed communication networks, the performance evaluation of particular queueing systems attracts much attention. These systems can be described by queueing models, in which a large number of on-off traffic sources feed into a buffer that is emptied at constant rate. *Large deviations* (LD) theory is particularly appropriate to analyze this kind of large traffic systems. In the first place, LD yields rough asymptotics of the probability of a rare event, in this case cell loss due to buffer overflow. Usually, this is done by solving an associate variational problem. However, as a by-product insight is gained into the way the underlying stochastic process reaches the rare event under consideration. In fact, the 'most probable' trajectory to overflow can be found from the optimizing argument of the variational problem. Freidlin and Wentzell [70] provide a comprehensive treatment of this subject.

To simplify the analysis, the (discrete) cell streams are approximated by (continuous) *fluid*. Shwartz and Weiss [166, Ch. 13], [184] successfully analyzed queues fed by a large number, say n, of on-off fluid sources with exponentially distributed on and off periods. They developed rough approximations of the overflow probability, asymptotically in n.

More precisely, for zero, small, and large buffers, a number $I > 0$ was found such that the overflow probability is (asymptotically) equal to $\exp[-nI]$.

However, several studies showed that in some cases the assumption of exponential on and off times is not realistic. For instance, O'Reilly and Ghani [141] propose sources with exponential on times and hyperexponential silences, not fitting into the framework of [184]. Therefore, it is interesting to analyze whether the results of Shwartz and Weiss can be extended to more general sources. The contribution of this paper is that this is indeed possible: nearly all the results carry over to queues fed by *general Markov fluid* sources. This class of sources is very rich: it allows for instance on-off sources with phase-type (Coxian, hyperexponential, Erlang, etc.) on and off times. In the Markov fluid framework even on-off sources with general on and off times can be embedded, because each distribution on the positive half-axis can be approximated arbitrarily closely by a mixture of Erlang distributions with common 'shape-parameter' [162].

In the present chapter, we let the number of sources n grow large. Also, we take the deterministic service rate nC and buffer size nB. Under the assumption that the mean input rate of a single source is smaller than C, the event of a buffer overflow (for any $B > 0$) is rare for large n: we will show that the overflow probability decays exponentially in n to 0. We will derive the associate decay rate and other overflow characteristics that are asymptotical in the number of sources n.

In a lot of previous articles, however, one derived *large buffer asymptotics* of overflow probabilities: overflow is rare because of large buffers instead of large numbers of input sources. Chapter 3 of this monograph can be regarded as one of these. With this respect, we mention, among many other studies, the initiating article of Anick, Mitra, and Sondhi [5] (identical exponential on-off sources) and Kosten [111] (multiple types of traffic, in accordance with the multiclass input of ATM-networks). Elwalid and Mitra [61] and Kesidis, Walrand, and Chang [105] relate the large buffer asymptotics to the *effective bandwidth* concept. Kesidis and Walrand [103] and Mandjes and Ridder [125] focused on estimating cell loss due to overflow by applying importance sampling (simulation) techniques, which are based on LD theory and in particular the solution of the above-mentioned variational problems.

The organization of this chapter is as follows. Section 2 briefly reviews some main results of large deviations theory that are used throughout this paper. Section 3 addresses overflows in queues without buffers. Surprisingly, the concept of time reversal appears to be a useful tool in order to characterize the most probable trajectory to overflow. For positive buffer B, the associate variational problem cannot be solved explicitly. However for the limiting regimes $B \downarrow 0$ and $B \to \infty$ accurate approximations are deduced, as

reported in Sections 4 and 5. Finally, Section 6 addresses the topic of a multiclass model.

2 Preliminaries

The purpose of this section is to review some basic results from LD. We pay attention to theorems concerning a large number of independent and identically distributed (i.i.d.) random variables, and a large number of i.i.d. Markov processes.

2.1 Large deviations of i.i.d. random variables

Consider an i.i.d. sequence $(X_i)_{i\in\mathbf{N}}$ with partial sums $(S_n)_{n\in\mathbf{N}}$: $S_n := \sum_{i=1}^n X_i$. The strong law of large numbers says $S_n/n \to \mathrm{E}X_1$ a.s., provided that $\mathrm{E}|X_1| < \infty$. However, in our applications we are more interested in large deviations from this limiting value, e.g. the probability $\mathcal{P}(S_n/n \geq a)$ for some $a > \mathrm{E}X_1$ and n large. LD [23], [50], [58] describes the asymptotics of this kind of rare event probabilities. We recall some basic results that are relevant for this study.

Let $M(\cdot)$ be the moment generating function (mgf) of a single increment, i.e., $M(\theta) := \mathrm{E}\exp(\theta X_1)$, assumed to be finite in a neighborhood of 0. $I(\cdot)$ is the large deviations rate function or convex conjugate of $\log M(\cdot)$: $I(u) := \sup_\theta(\theta u - \log M(\theta))$. Under weak conditions on the set U, Cramér's theorem says

$$\lim_{n\to\infty} \frac{1}{n} \log \mathcal{P}(S_n/n \in U) = -\inf_{u\in U} I(u). \tag{4.1}$$

A sufficient condition on the set U is [58, p. 35]: $\inf_{u\in \mathrm{int}\, U} I(u) = \inf_{u\in \mathrm{cl}\, U} I(u)$, $\mathrm{int}\, U$ and $\mathrm{cl}\, U$ denoting the interior and closure, respectively, of U. The function $I(\cdot)$ is convex, with minimal value 0 at $\mathrm{E}X_1$. Consequently, if U contains $\mathrm{E}X_1$, the decay rate of the probability under consideration is 0, in agreement with the laws of large numbers. Also, for all $a > \mathrm{E}X_1$ the decay rate of $\mathcal{P}(S_n/n \geq a)$ equals $I(a)$. More accurate asymptotics are the so-called Bahadur-Rao extensions [50, p. 95-98] of Cramér's theorem (4.1):

$$\mathcal{P}\left(\frac{S_n}{n} \geq a\right) e^{nI(a)}\sqrt{n} \to K(a) \quad \text{and} \quad \mathcal{P}\left(\frac{S_n}{n} \geq a + \frac{b}{n}\right) e^{nI(a)}\sqrt{n} \to K(a)e^{-\theta(a)b}, \tag{4.2}$$

as $n \to \infty$. Here $K(\cdot)$ is a positive function, whose precise form is irrelevant here; $\theta(a)$ solves $(\log M)'(\theta) = a$.

Analogously to the LD of sample means, the asymptotics of the empirical distribution can be considered. Define the empirical distribution in x after n samples by $L_n(x) := n^{-1} \cdot \#\{i = 1, \ldots, n : X_i = x\}$. $L_n(\cdot)$ converges weakly [58, p. 32] to $\mathcal{P}_X(\cdot)$, i.e. the

distribution of the X_i. Sanov's theorem says that, under mild conditions,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathcal{P}(L_n \in U) = - \inf_{\mu \in U} \int_{-\infty}^{\infty} \log \left(\frac{\mathrm{d}\mu(x)}{\mathrm{dP}_X(x)} \right) \mathrm{d}\mu(x), \tag{4.3}$$

where $\mathrm{d}\mu/\mathrm{dP}_X = \infty$ if μ is not absolutely continuous w.r.t. P_X. If U contains P_X, the right hand side indeed yields 0, in accordance with the convergence of L_n to P_X. Loosely speaking, an optimizing μ in the right hand side can be interpreted as the most probable distribution of the X_i in case of $L_n \in U$ (for n large).

2.2 Large deviations of i.i.d. Markov processes

Consider n independent continuous-time Markov chains, having common rate matrix $\Lambda := (\lambda_{ij})_{i,j=1}^d$ of finite dimension d. We assume the Markov chains to be irreducible, and therefore endowed with unique invariant π (satisfying $\pi\Lambda = 0$ and $\sum_{i=1}^d \pi_i = 1$). Let $Z_n^{(i)}(t)$ be the fraction of all n processes (or the state frequency) in state i at time t. As explained in [166], the limiting process of $Z_n(\cdot) = (Z_n^{(1)}(\cdot), \ldots, Z_n^{(d)}(\cdot))^{\mathrm{T}}$ (for $n \to \infty$) is the deterministic process $Z_\infty(\cdot)$, given by the system of linear differential equations (cf. Kolmogorov's forward differential equations)

$$Z_\infty^{(i)\prime}(t) = \sum_{j\neq i} Z_\infty^{(j)}(t)\lambda_{ji} - \sum_{j\neq i} Z_\infty^{(i)}(t)\lambda_{ij} = \sum_{j=1}^{d} Z_\infty^{(j)}(t)\lambda_{ji}, \quad \text{or} \quad Z_\infty'(t) = \Lambda^{\mathrm{T}} Z_\infty(t).$$

Methods to solve this kind of systems of differential equations are well-known [86]. Explicitly, if $Z_\infty(0)$ is given, we can write $Z_\infty(t) = \exp(\Lambda^{\mathrm{T}} t) Z_\infty(0)$. For a fixed t, $Z_\infty(t)$ can be calculated using an uniformization technique, see Tijms [175, p. 154-156]. Shwartz and Weiss [166] present the LD of $Z_n(\cdot)$, i.e., the probability that $Z_n(\cdot)$ lies far away from $Z_\infty(\cdot)$, for large n. One of their results, of particular interest for this paper, can be phrased as follows. Define

$$\log M_x(\theta) := \sum_{i=1}^{d} \sum_{j\neq i} x_i \lambda_{ij} \left(e^{\theta_j} e^{-\theta_i} - 1 \right), \tag{4.4}$$

in fact reflecting the joint log mgf of the number of jumps in all directions, if the state frequencies are x_i. Here $\theta \in \mathbb{R}^d$, $x \in \mathbb{P}^d := \{x \in \mathbb{R}^d : x_i \in [0,1], \sum_{i=1}^d x_i = 1\}$. Furthermore, we define the multidimensional convex conjugate of $\log M_x(\cdot)$: for u such that $\sum_{i=1}^d u_i = 0$, $I_x(u) := \sup_\theta(\langle \theta, u\rangle - \log M_x(\theta))$, where $\langle \theta, u\rangle := \sum_{i=1}^d \theta_i u_i$.

Let U be a set of (differentiable almost everywhere) functions from $[0,T]$ (T positive, possibly ∞) to $\mathbb{P}^d$, representing the paths of the state frequencies of the Markov chain in time. Again, under mild conditions on U, formally treated in [166],

$$\lim_{n\to\infty} \frac{1}{n} \log \mathcal{P}(Z_n(\cdot) \in U \text{ on } [0,T]) = - \inf_{f\in U} \int_0^T I_{f(t)}(f'(t))\mathrm{d}t. \tag{4.5}$$

We call $\int I_{f(t)}(f'(t))\mathrm{d}t$ the entropy or action functional, evaluated in path f. Inserting 'average path' $f(\cdot) = Z_\infty(\cdot)$ yields $\int_0^\infty I_{f(t)}(f'(t))\mathrm{d}t = 0$: if U contains the limiting behavior, the decay rate equals 0, cf. laws of large numbers. In most cases, the variational problem in the right hand side of (4.5) is difficult to solve. However, for special classes of sets U, the infimization can be done explicitly applying calculus of variations [73].

The optimizing f is called the 'optimum path' and reflects the most probable trajectory of $Z_n(\cdot)$ under condition of $Z_n(\cdot) \in U$ on $[0,T]$, n large. In fact, it can be said that if the event under consideration happens, it happens via this optimum trajectory (cf. the interpretation of the optimizing probability measure in Sanov's theorem).

3 Zero buffers

Consider a queueing system, fed by a large number (n) of independent and statistically identical Markov fluid sources. Each source is associated with a continuous-time, finite-dimensional Markov chain with infinitesimal generator $\Lambda = (\lambda_{ij})_{i,j=1}^d$. This rate matrix is assumed to be irreducible, implying the existence of a unique invariant π. If a Markov chain is in state i, fluid is generated at a constant rate $r_i \geq 0$. This kind of sources are of the so-called Markov fluid type. Notice that the parameters can be chosen such that they represent on-off sources with phase-type on and off distributions.

The fluid generated by the sources is led into a queue, which is emptied at a constant rate nC. To avoid trivialities, we assume C smaller than the peak rate of a source. For the sake of stability, C is larger than the mean rate of a single source, $\langle \pi, r \rangle$. Denote by H the elements x of $\mathbb{P}^d$ such that $\langle x, r \rangle \geq C$. Obviously, an overflow in the zero-buffer queue occurs if $Z_n(t) \in H$ for some t, $Z_n(\cdot)$ as defined in Section 2. We aim to characterize the frequency of this event, asymptotically in n. Clearly the event represents a large deviation, since its probability tends to 0 as n grows large. We prove that this decay is exponential, and derive an expression for the corresponding decay rate. We also find the most probable path of $Z_n(\cdot)$ from the equilibrium situation towards a distribution in H.

3.1 Decay rate of the overflow probability

Let A_j be the input rate generated by source j ($j = 1, \ldots, n$) at a point in time at which the system is in equilibrium: $A_j = r_i$ with probability π_i. Based on Cramér (4.1), we have the following rough asymptotics of the input rate being larger than the output rate:

$$\lim_{n\to\infty} \frac{1}{n} \log \mathcal{P}\left(\sum_{j=1}^{n} A_j \geq nC \right) = -I(C), \tag{4.6}$$

where $I(\cdot)$ is the convex conjugate $\log M(\cdot)$, $M(\cdot)$ being the mgf of A_1. The value of θ solving $(\log M)'(\theta) = u$, called $\theta(u)$, is the optimizing argument in the definition of $I(u)$. Consequently, $I(u) = \theta(u)u - \log M(\theta(u))$ and $I'(u) = \theta(u)$.

LD allows us to find the most probable state frequencies, conditional on overflow. Let $L_n(i)$ denote the fraction of Markov chains (in equilibrium) that is in state i. Then $\{\sum_{j=1}^n A_j \geq nC\} = \{L_n \in H\}$. It follows from Sanov (4.3), that

$$\lim_{n\to\infty} \frac{1}{n} \log \mathcal{P}\left(\sum_{j=1}^{n} A_j \geq nC\right) = -\inf_{\alpha\in H} \sum_{i=i}^{d} \alpha_i \log\left(\frac{\alpha_i}{\pi_i}\right).$$

The infimum is attained for $\alpha_i(0) := \pi_i \exp(\theta(C)r_i)/M(\theta(C))$, $i = 1, \ldots, d$ (use Lagrangian optimization) [1]. This statement can be somewhat rigorized in the following way. Invoking Bahadur-Rao (4.2), we find for large n,

$$\lim_{n\to\infty} \mathcal{P}\left(A_j = r_i \,\middle|\, \sum_{k=1}^{n} A_k \geq nC\right) = \lim_{n\to\infty} \pi_i \left(\frac{\mathcal{P}\left(\sum_{k=1}^{n-1} A_k \geq nC - r_i\right)}{\mathcal{P}\left(\sum_{k=1}^{n} A_k \geq nC\right)}\right)$$

$$= \lim_{n\to\infty} \pi_i \left(\frac{\mathcal{P}\left((n-1)^{-1}\sum_{k=1}^{n-1} A_k \geq C + (n-1)^{-1}(C - r_i)\right)}{\mathcal{P}\left(n^{-1}\sum_{k=1}^{n} A_k \geq C\right)}\right) = \pi_i e^{I(C)-\theta(C)(C-r_i)}.$$

Using the relations between $I(\cdot)$, $\theta(\cdot)$ and $M(\cdot)$, the above limit indeed equals $\alpha_i(0)$.

3.2 Optimal path towards overflow

Recapitulating, we found the decay rate (in n) of the probability of an overflow $I(C)$, and we derived the state frequencies α when hitting the hyperplane $\partial H := \{x \in \mathbb{P}^d | \langle \pi, r\rangle = C\}$. However, it is also possible to find the optimal path f of the $Z_n(\cdot)$ (in $\mathbb{P}^d$) from the equilibrium π to $\alpha(0)$. In order to derive an explicit expression for this path, we shall use the large deviations for i.i.d. Markov chains. Let U denote the set of paths, consisting of elements $f : \mathbb{R} \to \mathbb{P}^d$ with $f(0) = \pi$ and reaching ∂H for some $T(f) > 0$. Then, according to (4.5), the decay rate of the overflow probability reads

$$\inf_{f\in U} \int_0^{T(f)} I_{f(t)}(f'(t))\mathrm{d}t.$$

As explained in theorem 2.1 of [59] and [166, Ch. 2], this expression must equal $I(C)$, where $I(C)$ follows from (4.6). We shift the time axis such that overflow is reached at time 0. In other words, we get

$$I(C) = \inf_{f\in U'} \int_{-T(f)}^{0} I_{f(t)}(f'(t))\mathrm{d}t, \tag{4.7}$$

[1] Here $\alpha(\cdot)$ has argument 0, since it reflects the most probable distribution to hit ∂H in the *zero*-buffer case. A proper definition of the function $\alpha(\cdot)$ is given in the next section.

where U' consists of $f : \mathbb{R} \to \mathbb{P}^d$ that equal π in $-T(f)$ and are in ∂H at time 0.

We will now find a path f that solves (4.7) and is, for that reason, an optimum path. Consider a rate matrix with entries $\widetilde{\lambda}_{ij} := \lambda_{ji}\pi_j/\pi_i$, being the transition rates of the time-reversed of the modulating Markov process [95]. Let $f(\cdot)$ be the solution of the differential equation $f'(t) = -\widetilde{\Lambda}^{\mathrm{T}} f(t)$, value π for $t \to -\infty$ and α for $t = 0$. Consequently, for $t \leq 0$,

$$f_i'(t) = \sum_{j \neq i} \frac{\pi_j}{\pi_i} \lambda_{ji} f_i(t) - \sum_{j \neq i} \frac{\pi_i}{\pi_j} \lambda_{ij} f_j(t) \tag{4.8}$$

Our 'guess' for an optimum trajectory from π to α (in time interval $(-\infty, 0]$) is the path that solves this system of linear differential equations. We can show that this is true. Suppose that f satisfies the differential equations. Then, by inserting (4.8),

$$\begin{aligned}\int_{-\infty}^{0} I_{f(t)}(f'(t))\mathrm{d}t &= \int_{-\infty}^{0} \sup_{\theta} \left(\sum_{i=1}^{d} \theta_i f_i'(t) - \sum_{i=1}^{d} \sum_{j \neq i} f_i(t) \lambda_{ij} \left(e^{\theta_j} e^{-\theta_i} - 1 \right) \right) \mathrm{d}t \\ &= \int_{-\infty}^{0} \sup_{\theta} \left(-\sum_{i=1}^{d} \sum_{j=1}^{d} \theta_i \frac{\pi_i}{\pi_j} \lambda_{ij} f_j(t) - \sum_{i=1}^{d} \sum_{j \neq i} f_i(t) \lambda_{ij} \left(e^{\theta_j} e^{-\theta_i} - 1 \right) \right) \mathrm{d}t\end{aligned}$$

To get the θ that optimizes the integrand, given the values of $f(t)$ and $f'(t)$, we differentiate the integrand with respect to θ_i and equate it to 0:

$$\sum_{j \neq i} \lambda_{ij} \left(f_i(t) \frac{e^{\theta_j}}{e^{\theta_i}} - f_j(t) \frac{\pi_i}{\pi_j} \right) - \sum_{j \neq i} \lambda_{ji} \left(f_j(t) \frac{e^{\theta_i}}{e^{\theta_j}} - f_i(t) \frac{\pi_j}{\pi_i} \right) = 0.$$

Inserting $\theta_i(t) := \log(f_i(t)/\pi_i)$ indeed yields 0. This choice of $\theta_i(t)$ gives, cf. equation (4.4), that $\log M_{f(t)}(\theta_1(t), \ldots, \theta_d(t)) = 0$:

$$\sum_{i=1}^{d} \sum_{j \neq i} f_i(t) \lambda_{ij} e^{\theta_j(t)} e^{-\theta_i(t)} = \sum_{i=1}^{d} \sum_{j \neq i} f_j(t) \lambda_{ij} \frac{\pi_i}{\pi_j} \overset{\text{(i)}}{=} \sum_{i=1}^{d} \sum_{j \neq i} f_i(t) \lambda_{ji} \frac{\pi_j}{\pi_i} \overset{\text{(ii)}}{=} \sum_{i=1}^{d} \sum_{j \neq i} f_i(t) \lambda_{ij}.$$

Here (i) is justified by reversing the order of summation, and (ii) by π being invariant of Λ. We now conclude that the minimum achievable value in (4.7) is attained:

$$\begin{aligned}\int_{-\infty}^{0} I_{f(t)}(f'(t))\mathrm{d}t &= \int_{-\infty}^{0} \left(\sum_{i=1}^{d} \log \left(\frac{f_i(t)}{\pi_i} \right) f_i'(t) \right) \mathrm{d}t \\ &= \sum_{i=1}^{d} \int_{\pi_i}^{\alpha_i(0)} \log \left(\frac{u}{\pi_i} \right) \mathrm{d}u = \sum_{i=1}^{d} \alpha_i(0) \log \left(\frac{\alpha_i(0)}{\pi_i} \right) = I(C).\end{aligned}$$

We conclude that we proved an appealing result, namely that the most probable path from π to $\alpha(0)$ of the original process is the path from $\alpha(0)$ to π of the time-reversed process (having generator $\widetilde{\Lambda}$), with a reversed time axis. In other words: f satisfies $f'(t) = -\widetilde{\Lambda}^{\mathrm{T}} f(t)$ on $(-\infty, 0]$, with $f(0) = \alpha(0)$. In this way, large deviations theory and time reversal coincide very nicely, as reported in earlier studies (see for instance Shwartz and Weiss [165] on the M/M/1 queue).

4 Small buffers

The previous section dealt with queues with zero buffers. It is noteworthy that queues with no (or small) buffers are particularly suited to process traffic with stringent delay constraints (voice). However, for other types of traffic (data), positive (or even large) buffers might be attractive. This section deals with the decay rate of the overflow probability in case of small buffers; the next section addresses large buffers. We take the buffer, if n sources are connected, nB. We call the decay rate of the loss probability in this model $I^*(B)$. Note that $I^*(0) = I(C)$, as considered in Section 3. In [59] bounds for $I^*(B)$ are found in case of reversible sources.

4.1 Basic results

We will first review some of basic results that we need in the remainder of this paper. These are extensively treated in Shwartz and Weiss [166]. From (4.5), decay rate $I^*(B)$ is the minimization of the action functional over U, where U consists of (differentiable a.e.) functions $f : \mathbb{R} \to \mathbb{P}^d$ such that $f(-\infty) = \pi$, $f(0) \in \partial H$, yielding overflow at time T_B:

$$\inf_{f \in U} \int_{-\infty}^{T_B} I_{f(t)}(f'(t))\mathrm{d}t = \inf_{\alpha,\beta \in \partial H} \left(\inf_{f \in U_{\alpha,\beta}} \int_{-\infty}^{T_B} I_{f(t)}(f'(t))\mathrm{d}t \right). \tag{4.9}$$

The equality is basically 'conditioning' on the state frequencies at time 0 and T_B; $U_{\alpha,\beta}$ is defined as the functions $f \in U$ such that $f(0) = \alpha$ and $f(T_B) = \beta$. The requirement $\beta \in \partial H$ is based on the fact that for an f that minimizes the left hand side of (4.9), it holds that $f(T_B) \in \partial H$, as explained in [166, Exercise 13.7].

We decompose the infimization between brackets into two parts:

$$I^*(B) = \inf_{\alpha,\beta \in \partial H} \left(\sum_{i=1}^{d} \alpha_i \log\left(\frac{\alpha_i}{\pi_i}\right) + \inf_{f \in U_{\alpha,\beta}} \int_0^{T_B} I_{f(t)}\left(f'(t)\right) \mathrm{d}t \right). \tag{4.10}$$

The first term represents time interval $(-\infty, 0]$: it is the entropy of hitting ∂H in a distribution α, starting in π (use (4.3), cf. Section 3). The second part is the entropy of the trajectory to overflow, starting in α and reaching overflow while being in β. The minimizing α and β, say $\alpha(B)$ and $\beta(B)$, are in fact the most probable distributions to enter and leave H in order to cause a buffer overflow.

We refer to (4.10) as the *unconstrained problem*, whereas the infimum between brackets (α and β fixed) is the *constrained problem*. Since for an optimizing path f it holds that $f \in \mathrm{int}H$ for all $t \in (0, T_B)$, see [166], we get that $\int_0^{T_B} \langle f(t), r - C \rangle \mathrm{d}t = B$. Consequently, the Lagrangian representation of the constrained problem is (with multiplier K)

$$\inf_{f \in U_{\alpha,\beta}} \int_0^{T_B} h(f(t), f'(t))\mathrm{d}t \text{ , where } h(a,b) := I_a(b) - K\langle a, r - C \rangle, \tag{4.11}$$

From the calculus of variations we find the first order necessary conditions for an optimal f, known as the Euler equations [73]:

$$\frac{\partial}{\partial a}h(a,b)\bigg|_{a=f(t),b=f'(t)} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial}{\partial b}h(a,b)\bigg|_{a=f(t),b=f'(t)}\right). \tag{4.12}$$

These equations will be used in the section on large buffers.

4.2 Approximations for small buffers

Weiss [184] considered two-state (on-off) Markov fluid sources. In that case, clearly the fraction of sources that is on at time 0 and T_B is determined uniquely. Thus α and β are known, and we are left with the task of determining the solution of the constrained problem. The constrained problem can be solved explicitly: the fraction of sources in the on-state (between 0 and T_B) along the optimal path turns out to be a hyperbolic cosine. However, for general sources we did not succeed in finding similar results. Therefore, our aim is to find accurate estimates for the decay rate of the overflow probability and the optimal path.

In the remainder of this section we heuristically construct an approximation f for the optimal path (in the interval $[0, T_B]$) and the decay rate in case of small B.

- In the previous section we found $\alpha(0)$: the most probable state frequencies in order to reach H. It is plausible that the most probable tuple in which H is reached in order to build up a small buffer contents B lies close to this distribution. Consequently, we choose $f(0) = \alpha(0)$. Analogously, and recalling the system of differential equations (4.8), the left derivative of the optimum path in 0 can be approximated by $\lim_{\epsilon \downarrow 0} f_i'(\epsilon) = -\sum_{j=1}^{d} \pi_i/\pi_j \lambda_{ij} f_j(0) =: \alpha_i'(0)$. The 'principle of smooth fit' [166, page 374] says that the first derivative of the optimum path is continuous, and therefore we let the left and right derivative of the f_i at 0 coincide.

- H is left (at time T_B) where the time-reversed process would reach H to build up a small contents B, i.e., close to where the time-reversed process would reach H in the zero buffer case. Using the fact that Λ and $\tilde{\Lambda}$ have the same invariant, we choose $f(T_B) = f(0) = \alpha(0)$. Using arguments similar to those above, we approximate the derivative of the optimum path at T_B by $f_i'(T_B) = \sum_{j=1}^{d} \lambda_{ji} f_j(0) =: \beta_i'(0)$.

For $t \in [0, T_B]$ we let $f_i(t)$ equal

$$\alpha_i(0) + \alpha_i'(0)t - \left(\frac{\beta_i'(0) + 2\alpha_i'(0)}{T_B}\right)t^2 + \left(\frac{\alpha_i'(0) + \beta_i'(0)}{T_B^2}\right)t^3,$$

meeting all the obtained $4d$ constraints. Because $\int_0^{T_B} \langle f(t), r - C \rangle \mathrm{d}t = B$ it follows that

$$T_B := \sqrt{\frac{12B}{\langle \alpha'(0) - \beta'(0), r \rangle}}.$$

Returning to (4.10) the left term of $I^*(B)$ reduces, by inserting $\alpha = \alpha(0)$, to $I(C)$. The value of $\int_0^{T_B} I_{f(t)}\left(f'(t)\right) \mathrm{d}t$ still has to be calculated. We notice that $f'(\cdot)$ changes rapidly in $[0, T_B]$, whereas $f(\cdot)$ is more or less constant. Therefore, similar to [184], we approximate the integral by

$$\int_0^{T_B} I_{f(0)}(f'(t))\mathrm{d}t = \int_0^{T_B} I_{\alpha(0)} \left(\alpha'(0) - \frac{2t}{T_B}\left(\beta'(0) + 2\alpha'(0)\right) + \frac{3t^2}{T_B^2}\left(\alpha'(0) + \beta'(0)\right) \right) \mathrm{d}t$$

$$= T_B \int_0^1 I_{\alpha(0)} \left(\alpha'(0) - 2u\left(\beta'(0) + 2\alpha'(0)\right) + 3u^2\left(\alpha'(0) + \beta'(0)\right) \right) \mathrm{d}u,$$

which is proportional to $\sqrt{B}$. We get that $\lim_{B \downarrow 0}(I^*(B) - I(C))/\sqrt{B}$ is a positive constant.

EXAMPLE. As said, in case of exponential on-off sources, exact results are calculable, but tedious numerical computations are required. To avoid these, Weiss [184] developed an approximation as well. However, we emphasize that both the exact method and Weiss' approximation do not apply to general Markov fluid sources, as opposed to our approximation technique established above.

In this example we will compare the exact results with both Weiss' and our approximation. Consider sources with exponential on and off times. State 1 is the busy state: fluid is generated at rate 1; state 2 is the idle state. Define $\lambda := \lambda_{21}$ and $\mu := \lambda_{12}$. Furthermore, $M := \mu C$ and $L := \lambda(1 - C)$. The following approximations were derived in [184] for the the maximal input rate during the optimum path, the time to overflow, and the decay rate

$$\begin{aligned} f_{\max} &\approx C + (\sqrt{M} - \sqrt{L})^2 \frac{\sqrt{B}}{\gamma}, \quad T_B \approx \log\left(\frac{M}{L}\right) \frac{\sqrt{B}}{\gamma}, \\ I^*(B) &\approx I(C) + 2\gamma\sqrt{B}, \quad \text{where } \gamma := \sqrt{(L+M)\log(M/L) - 2(M-L)}. \end{aligned}$$

Our approach yields after quite a lot of calculus

$$\begin{aligned} f_{\max} &\approx C + \frac{1}{4}\sqrt{6(M-L)}\sqrt{B}, \quad T_B \approx \sqrt{\frac{6B}{M-L}}, \\ I^*(B) &\approx I(C) + T_B \left\{ \frac{L+M}{4} + \frac{1}{2}\frac{LM}{M-L}\log\left(\frac{L}{M}\right) + \frac{L-M}{4}\log\left(\frac{L}{M}\right) \right\}. \end{aligned}$$

Suppose $\lambda = 1$, $\mu = 2$, $C = 0.5$. We notice that, with respect to the value of the decay rate $I^*(B)$, our approach and Weiss' approximation coincide to a large number of digits. In the next table we list, for several values of the buffer size, the exact value, Weiss' result, and our result, respectively.

Table 1: Comparison of results

B	$f_{\max}$			T_B		
0.005	0.530	0.530	0.531	0.247	0.246	0.245
0.01	0.542	0.543	0.543	0.351	0.348	0.346
0.02	0.559	0.561	0.561	0.500	0.492	0.490
0.05	0.590	0.596	0.597	0.812	0.778	0.775

5 Large buffers

In the previous section we deduced an approximation for the decay rate and the optimal path for small buffers, using the exact results for $B = 0$. Here, we will find the solution for infinitely large B, yielding an approximation for large buffers.

The constrained problem

In case of large buffer sizes and a *fixed* number of sources, we know the most probable distribution of the sources, given that overflow occurs. This distribution, say ρ, is evidently different from π and can be calculated as follows, see Chapter 3 and [125]. Solve the eigensystem

$$\Lambda x = -\theta(R - CI)x, \tag{4.13}$$

with $R := \mathrm{diag}\{r\}$ and I the d-dimensional identity matrix. Let θ^* be the smallest positive eigenvalue, and x a corresponding eigenvector (which can be chosen componentwise positive). Then, conditional on overflow, the rates λ_{ij} seem to be replaced by $\mu_{ij} = \lambda_{ij}x_j/x_i$ for $i \neq j$ and $\mu_{ii} = \lambda_{ii} + \theta^*(r_i - C)$. The invariant of these μ's is ρ. $M = (\mu_{ij})_{i,j=1}^d$ is called the dual generator of Λ.

Based on these observations, our 'guess' for the optimal path (to build up an infinitely large buffer) in the constrained problem is the following. Fix an $\alpha, \beta \in \partial H$. Then the optimal path consists of two regimes:

A. First, the path goes from α to ρ, via the equilibrium path of the μ's. This path, satisfying $f'(t) = M^{\mathrm{T}} f(t)$, is defined on $[0, \infty)$, with $f(0) = \alpha$.

B. Then the optimal trajectory goes from ρ to β via path g. $\widetilde{M}$ denoting the dual generator of $\widetilde{\Lambda}$, g is given by $g'(t) = -\widetilde{M}^{\mathrm{T}} g(t)$ on $(-\infty, 0]$ with $g(0) = \beta$. (Notice the shift of the time axis!) To show that $g(-\infty)$ indeed equals ρ, it must be shown that $\widetilde{M}$ has invariant ρ as well, as is done below.

We now determine the dual rates $\tilde{\mu}_{ij}$ of the time reversed process and show that indeed ρ is invariant of these rates. We must solve $\widetilde{\Lambda} y = -\theta(R - CI)y$. It can be shown easily

that this eigensystem is solved for $\theta = \theta^*$ and y such that $y_i = \rho_i/\pi_i x_i$:

$$\begin{aligned}\sum_{j=1}^{d} \tilde{\lambda}_{ij} y_j &= \sum_{j=1}^{d} \lambda_{ji} \frac{\pi_j}{\pi_i} y_j = \sum_{j=1}^{d} \lambda_{ji} \frac{\pi_j}{\pi_i} \frac{\rho_j}{\pi_j x_j} = \left(\sum_{j=1}^{d} \lambda_{ji} \frac{x_i}{x_j} \right) \frac{\rho_j}{\pi_i x_i} \\ &= \lambda_{ii} \frac{\rho_i}{\pi_i x_i} + \left(\sum_{j \neq i} \mu_{ji} \rho_j \right) \frac{1}{\pi_i x_i} = \lambda_{ii} \frac{\rho_i}{\pi_i x_i} + \left(\sum_{j \neq i} \mu_{ij} \rho_i \right) \frac{1}{\pi_i x_i} \\ &= \left(\lambda_{ii} + \sum_{j \neq i} \mu_{ij} \right) y_i = -\theta^* (r_i - C) y_i,\end{aligned}$$

as desired. Now elementary algebra yields $\rho \widetilde{M} = 0$, as desired:

$$\tilde{\mu}_{ij} = \tilde{\lambda}_{ij} \frac{y_j}{y_i} = \lambda_{ji} \frac{\rho_j / x_j}{\rho_i / x_i} = \mu_{ji} \frac{\rho_j}{\rho_i} \implies \sum_{j \neq i} \tilde{\mu}_{ij} \rho_i = \sum_{j \neq i} \mu_{ji} \rho_j = \sum_{j \neq i} \mu_{ij} \rho_i = \sum_{j \neq i} \tilde{\mu}_{ji} \rho_j.$$

As an aside we mention that we derived a commutational property: the dual of the time reversed process is the time reversed of the dual process.

Having done this preliminary work, we can now calculate the action functional explicitly for the trajectory of our 'guess'. Differentiating the supremand in the definition of $I_{f(t)}(f'(t))$ with respect to θ_i yields the first order conditions

$$f_i'(t) + \sum_{j \neq i} f_i(t) \lambda_{ij} e^{\theta_j} e^{-\theta_i} - \sum_{j \neq i} f_j(t) \lambda_{ji} e^{\theta_i} e^{-\theta_j} = 0. \tag{4.14}$$

During the trajectory from α to ρ, the path of our 'guess' satisfies

$$f_i'(t) = \sum_{j=1}^{d} \mu_{ji} f_j(t) = \sum_{j \neq i} \mu_{ji} f_j(t) - \sum_{j \neq i} \mu_{ij} f_i(t) = \sum_{j \neq i} \lambda_{ji} \frac{x_i}{x_j} f_j(t) - \sum_{j \neq i} \lambda_{ij} \frac{x_j}{x_i} f_i(t). \tag{4.15}$$

Substituting (4.15) in (4.14) yields $\theta_i(t) = \log x_i$ for all $t \in [0, \infty)$. We find

$$\begin{aligned}I_{f(t)}(f'(t)) &= \sum_{i=1}^{d} (\log x_i) f_i'(t) - \sum_{i=1}^{d} \sum_{j \neq i} f_i(t) \lambda_{ij} \left(\frac{x_j}{x_i} - 1 \right) \\ &= \sum_{i=1}^{d} (\log x_i) f_i'(t) + \sum_{i=1}^{d} f_i(t) \theta^* (r_i - C).\end{aligned} \tag{4.16}$$

In the same way we find that on the trajectory from ρ to β the optimizing $\theta_i(t)$ is equal to $\log(g_i(t) x_i / \rho_i)$, yielding for $t \in (-\infty, 0]$

$$\begin{aligned}I_{g(t)}(g'(t)) &= \sum_{i=1}^{d} \log \left(\frac{x_i g_i(t)}{\rho_i} \right) g_i'(t) - \sum_{i=1}^{d} \sum_{j \neq i} \left(\lambda_{ij} g_j(t) \frac{x_j / \rho_j}{x_i / \rho_i} - \lambda_{ij} g_i(t) \right) \\ &= \sum_{i=1}^{d} \log \left(\frac{x_i g_i(t)}{\rho_i} \right) g_i'(t) + \sum_{i=1}^{d} g_i(t) \theta^* (r_i - C).\end{aligned} \tag{4.17}$$

We can prove the optimality of our 'guess'. On the path from α to ρ we check Euler's equations (4.12), recalling the definition of h from (4.11), as follows:

$$\begin{aligned} h(a,b) &= \sum_{i=1}^{d}(\log x_i)b_i - \sum_{i=1}^{d}\sum_{j\neq i} a_i\lambda_{ij}\left(\frac{x_j}{x_i}-1\right) - K\sum_{i=1}^{d}(r_i-C)a_i, \\ \frac{\partial}{\partial a_i}h(a,b)\Big|_{a=f(t),b=f'(t)} &= \sum_{j\neq i}\lambda_{ij}\left(\frac{x_j}{x_i}-1\right) - K(r_i-C) = (-\theta^*-K)(r_i-C), \\ \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial}{\partial b_i}h(a,b)\Big|_{a=f(t),b=f'(t)}\right) &= \frac{\mathrm{d}}{\mathrm{d}t}\log x_i = 0. \end{aligned}$$

Conclude that the first order conditions are satisfied with $K = -\theta^*$. A similar procedure can be executed for the '$\rho \to \beta$-path', yielding that $g(\cdot)$ is optimal, with $K = -\theta^*$ as well.

The unconstrained problem

The optimum trajectory in the unconstrained problem consists of four regimes, all inducing an entropy. In the following list, these entropies are functions of α and β. The most probable distributions to enter and leave H still have to be determined by minimizing the total entropy over α and β.

A'. The first step is a path from π to α via f given by $f'(t) = -\tilde{\Lambda}^{\mathrm{T}}f(t)$, for t in $(-\infty, 0]$, with $f(0) = \alpha$. According to Sanov's theorem (4.3), the entropy is given by $\sum_{i=1}^{d}\alpha_i\log(\alpha_i/\pi_i)$, cf. (4.10).

A. Then from α to ρ via f as explained in the subsection on the constrained problem. The corresponding entropy can be found by inserting (4.16):

$$\int_0^\infty I_{f(t)}(f'(t))\mathrm{d}t = \sum_{i=1}^{d}(\rho_i-\alpha_i)\log x_i + \theta^*\left(\int_0^\infty \langle f(t), r-C\rangle \mathrm{d}t\right).$$

B. The third step leads from ρ to β as above. Using (4.17), the entropy is given by

$$\int_{-\infty}^0 I_{g(t)}(g'(t))\mathrm{d}t = \sum_{i=1}^{d}\beta_i\log\left(\frac{x_i\beta_i}{\rho_i}\right) - \sum_{i=1}^{d}\rho_i\log x_i + \theta^*\left(\int_{-\infty}^0 \langle g(t), r-C\rangle \mathrm{d}t\right).$$

B'. Finally, the path goes from β back to π via g given by $g'(t) = \Lambda^{\mathrm{T}}g(t)$ for nonnegative t, with $g(0) = \beta$. The corresponding entropy is 0, since this is the limiting behavior $Z_\infty(\cdot)$, as noticed in Section 2.

The entropy functions together are equal to

$$\sum_{i=1}^{d}\beta_i\log\left(\frac{\beta_i}{\pi_i y_i}\right) + \sum_{i=1}^{d}\alpha_i\log\left(\frac{\alpha_i}{\pi_i x_i}\right) + \theta^*\left(\int_0^\infty \langle f(t), r-C\rangle \mathrm{d}t + \int_{-\infty}^0 \langle g(t), r-C\rangle \mathrm{d}t\right). \tag{4.18}$$

The two integrals are together exactly the (infinite) buffer contents that had to be built up. To find the most probable α and β, we are therefore left with the task of minimizing the first two terms. Of course, the minimization is subject to the constraints $\alpha, \beta \in \partial H$. Consider the Lagrangian

$$\sum_{i=1}^{d} \beta_i \log\left(\frac{\beta_i}{\pi_i y_i}\right) + \sum_{i=1}^{d} \alpha_i \log\left(\frac{\alpha_i}{\pi_i x_i}\right) + K_1\left(\sum_{i=1}^{d} \alpha_i - 1\right) + K_2\left(\sum_{i=1}^{d} \beta_i - 1\right). \tag{4.19}$$

The minimum of (4.19) is attained for $\alpha_i(\infty) = \pi_i x_i / \langle \pi, x \rangle$ and $\beta_i(\infty) = \pi_i y_i / \langle \pi, y \rangle$. We did not take into account the requirements $\alpha, \beta \in \partial H$, but these are met automatically:

$$-\theta^* \langle \alpha(\infty), r - C \rangle \;=\; -\sum_{i=1}^{d} \theta^* \frac{\pi_i x_i}{\langle \pi, x \rangle}(r_i - C) \overset{\text{(i)}}{=} \sum_{i=1}^{d} \frac{\pi_i}{\langle \pi, x \rangle} \sum_{j=1}^{d} \lambda_{ij} x_j \overset{\text{(ii)}}{=} 0,$$

where (i) is because of (4.13) and (ii) due to $\sum_{i=1}^{d} \pi_i \lambda_{ij} = 0$. Analogously $\beta(\infty) \in \partial H$.

Finally, we pay some attention to a remarkable phenomenon. Suppose, the hyperplane is hit in the optimal α derived above. Coming from π the hyperplane ∂H is hit in $\alpha(\infty)$ with slope

$$\lim_{\epsilon \downarrow 0} f_i'(-\epsilon) = \sum_{j \neq i} \lambda_{ji} \frac{\pi_j}{\pi_i} \alpha_i(\infty) - \sum_{j \neq i} \lambda_{ij} \frac{\pi_i}{\pi_j} \alpha_j(\infty) = \sum_{j \neq i} \lambda_{ji} \frac{\pi_j x_i}{\langle \pi, x \rangle} - \sum_{j \neq i} \lambda_{ij} \frac{\pi_i x_j}{\langle \pi, x \rangle}.$$

Also, the optimal path leaves ∂H from $\alpha(\infty)$ towards ρ with slope

$$\lim_{\epsilon \downarrow 0} f_i'(\epsilon) = \sum_{j \neq i} \mu_{ji} \alpha_j(\infty) - \sum_{j \neq i} \mu_{ij} \alpha_i(\infty) = \sum_{j \neq i} \lambda_{ji} \frac{\pi_j x_i}{\langle \pi, x \rangle} - \sum_{j \neq i} \lambda_{ij} \frac{\pi_i x_j}{\langle \pi, x \rangle}.$$

The same equality holds in $\beta(\infty)$. So the optimal path is smooth (continuous and the derivative is continuous) in the most probable distributions $\alpha(\infty)$ and $\beta(\infty)$ in which the hyperplane is hit and left, coinciding with the 'principle of smooth fit' [166].

Inserting $\alpha(\infty)$ and $\beta(\infty)$ into (4.18), we find the following approximation of $I^*(B)$ for large B:

$$-\log\left(\sum_k \pi_k x_k \;\times\; \sum_k \pi_k y_k\right) + \theta^* B.$$

6 Multiple types of sources

ATM traffic can be characterized by multiple types of traffic (data, voice, video, ...) having widely varying burstiness properties. For this reason, the situation of identical (statistically independent) sources, as presented in the previous sections is not very realistic. Following Kosten [110], we consider multiple classes of sources, for convenience we take two classes. Suppose n sources, of which $n\gamma_1 \in \mathbb{N}$ are of type 1 and $n\gamma_2$

of the second type, where $\gamma_1, \gamma_2 \in [0,1]$ and $\gamma_1 + \gamma_2 = 1$. Let (Λ_k, π_k, r_k) be the infinitesimal generator, its invariant, and the traffic rates, respectively, of a type k source $(k = 1, 2)$. Furthermore, we define the moment generating function of the generated traffic by $M_k(\theta) := \sum_i \pi_{k,i} \exp(\theta r_{k,i})$ and $I_k(\cdot)$ as its convex conjugate. Also, $\theta_k(u) := I_k'(u)$.

ZERO BUFFERS. First we investigate the decay rate of the probability of an overflow in a zero buffer queue

$$\mathcal{P}\left(\sum_{k=1}^{2}\sum_{i=1}^{n\gamma_k} A_{k,i} \geq nC\right),$$

$A_{k,i}$ representing the traffic generated by the ith type k source (in equilibrium). To characterize the decay rate of the above probability, we will use a number of insightful, heuristic arguments. The results can be rigorized in a manner similar to that presented in Section 3.

A simple conditioning argument yields

$$\mathcal{P}\left(\sum_{k=1}^{2}\sum_{i=1}^{n\gamma_k} A_{k,i} \approx nC\right) \approx \int_0^C \mathcal{P}\left(\sum_{i=1}^{n\gamma_1} A_{1,i} \approx n(C-C')\right)\mathcal{P}\left(\sum_{i=1}^{n\gamma_2} A_{2,i} \approx nC'\right) \mathrm{d}C'.$$

Laplace's principle [166, p. 12] says that the decay rate of the above integral equals the decay rate of the maximal value of the integrand. These observations yield for the decay rate of the probability of a filling buffer

$$\lim_{n\to\infty} \frac{1}{n} \log\left\{\sup_{C'\in[0,C]} \mathcal{P}\left(\sum_{i=1}^{n\gamma_1} A_{1,i} \approx n(C-C')\right)\mathcal{P}\left(\sum_{i=1}^{n\gamma_2} A_{2,i} \approx nC'\right)\right\} =$$

$$-\inf_{C'\in[0,C]}\left\{\gamma_1 I_1\left(\frac{C-C'}{\gamma_1}\right) + \gamma_2 I_2\left(\frac{C'}{\gamma_2}\right)\right\} = -\inf\sum_{k=1}^{2}\gamma_k I_k(C_k),$$

where C_1 and C_2 in the final expression must be chosen such that $\sum_{k=1}^2 \gamma_k C_k = C$. Call the minimizers in this formula simply C_1 and C_2, and (re)define H as the set of distributions (x_1, x_2) such that $\sum_{k=1}^2 \gamma_k \langle x_k, r_k\rangle \geq C$. Similarly to Section 3, the most probable distributions to enter H are given by

$$\alpha_{k,i}(0) = \frac{\pi_{k,i}\exp(\theta_k(C_k) r_{k,i})}{M_k(\theta_k(C_k))}, \quad k = 1, 2. \tag{4.20}$$

The paths towards $(\alpha_1(0), \alpha_2(0))$ from the equilibrium distribution (π_1, π_2) are given by $\exp(-\tilde{\Lambda}_k^{\mathrm{T}} t)\alpha_k(0)$, $k = 1, 2$ and $t \in (-\infty, 0]$.

SMALL BUFFERS. The results for small buffers of Section 4 can be extended easily to the heterogeneous case. The hyperspace H is hit by the optimal path close to the distributions given by (4.20), just as in Section 4. Also an approximation of the slope at this moment can be found: $(\alpha_1'(0), \alpha_2'(0)) = (-\tilde{\Lambda}_1^{\mathrm{T}}\alpha_1(0), -\tilde{\Lambda}_2^{\mathrm{T}}\alpha_2(0))$. On the other hand, using the

time-reversed process, we can approximate the distribution at which the hyperspace is left and its derivative there: H is left near $(\beta_1(0), \beta_2(0)) = (\alpha_1(0), \alpha_2(0))$ with slope $(\beta_1'(0), \beta_2'(0)) = (\Lambda_1^{\mathrm{T}}\alpha_1(0), \Lambda_2^{\mathrm{T}}\alpha_2(0))$. The optimal path on $[0, T_B]$ can be approximated by a polynomial of degree 3, leading to the approximation for $I^*(B)$

$$\sum_{k=1}^{2} \gamma_k \left(I_k(C_k) + T_B \int_0^1 I_{\alpha_k(0)} \left(\alpha_k'(0) + 2u(-\beta_k'(0) - 2\alpha_k'(0)) + 3u^2(\alpha_k'(0) + \beta_k'(0)) \right) \mathrm{d}u \right),$$

where the epoch of overflow is given by

$$T_B = \sqrt{\frac{12B}{\sum_{k=1}^{2} \gamma_k \langle \alpha_k'(0) - \beta_k'(0), r_k \rangle}}.$$

LARGE BUFFERS. We finish this subsection by treating the large buffer case. First define the so called effective bandwidth [61] of a source of type k, $C_k(\theta)$, by the largest real value of C that solves the eigensystem (with eigenvector x_k)

$$\Lambda_k x_k = -\theta(R_k - CI)x_k$$

where $R_k := \mathrm{diag}\{r_k\}$, $k = 1, 2$. We define θ^* as the solution of $\sum_{k=1}^{2} \gamma_k C_k(\theta) = C$. Furthermore, call the eigenvectors from the definition of $C_k(\theta^*)$ simply x_k, $k = 1, 2$.

It can be checked easily that the effective bandwidth of a source with the original transition rates Λ_k, i.e. the function $C_k(\cdot)$, and the effective bandwidth based on the time-reversed rates $\tilde{\Lambda}_k$, i.e. $\tilde{C}_k(\cdot)$, coincide. Consequently, θ^* also solves $\sum_{k=1}^{2} \gamma_k \tilde{C}_k(\theta) = C$, yielding eigenvectors y_1 and y_2. Similarly to Section 5, we deduce the following asymptotics for the decay rate of the loss probability $I^*(B)$ for large B:

$$-\sum_{k=1}^{2} \gamma_k \log\left(\langle \pi_k, x_k \rangle \langle \pi_k, y_k \rangle \right) + \theta^* B.$$

7 Conclusions

We investigated a communication link fed by a large number of general Markov fluid sources. For on-off sources with exponential on and off times this model has been investigated thoroughly [166], [184]. This chapter shows the possibility of extending the results to general Markov fluid sources. In fact, the case of zero buffers and infinite buffers can be analyzed exactly and lead to accurate approximations for small and large buffers, respectively. Also, the extension to multiple types of traffic is established.

Decay rate $I^*(B)$ provides only rough insight into the asymptotic behavior of the overflow probability. For practical purposes (for instance dimensioning and traffic control) more detailed information is required. For that reason, fast simulation techniques can be

developed to estimate the probability itself, instead of its decay rate. Another interesting subject of future research is the analysis of $I^*(B)$ for moderate values of B, instead of zero, small, and large buffers.

Chapter 5

Markov fluid tandem queues

The model considered is a communication network with a two-node tandem structure. The input consists of a number of Markov modulated fluid sources and feeds into a first queue. The output serves as input for a second queue. We roughly characterize the tail probability of the second queue. Chang *et al.* [28] proposed an importance sampling technique to estimate this probability; we show an optimality property of this method. Finally, simulation results are given, showing a large speed-up.

1 Introduction

An important design issue in high-speed digital network architectures (as ATM) is the allocation of switching and transmission resources, in such a way that certain quality of service (QoS) criteria are met (with respect to loss, delay). Large numbers of traffic sources with widely varying burstiness conditions are integrated at the entrance of the network, but obviously the characteristics of the traffic streams change when passing network switches. It is clear that it is necessary to gain insight into this change of burstiness in order to perform adequate resource allocation for all links of the network.

This chapter focuses on the loss constraint: buffers and service rates must be chosen such that the cell loss ratio is kept below a given acceptable level. To ease the task of resource allocation, one attempted to approximate the loss ratio for given buffers and link capacities. Particularly, for a broad class of single queues, it was shown that the loss ratio has an exponentially decreasing tail: for large B, an accurate approximation is $\eta \exp[-\theta B]$, for positive amplitude η and decay rate θ that do not depend on B. Consequently, for given service rate, this asymptotic relation enables to find the right buffer size.

In order to determine an appropriate service rate, an important notion is the *effective bandwidth*, see Hui [89], Guérin, Ahmadi, and Naghshineh [80]. Suppose a queue is fed by

a number of traffic streams and emptied at a constant rate. Then it is possible to assign to each source a bandwidth, viz. the service rate that should be offered to this single source in order to achieve some service requirement. Mostly, the bandwidth required by the superposition of the sources to guarantee the QoS, is approximated by the sum of the individual bandwidths. Notably, if the service criterion under consideration requires that the decay rate of the loss fraction is θ, this additivity property is *exact*. Then the resulting bandwidths, as a function of parameter θ, are mostly called 'effective bandwidth' functions. An extensive study on effective bandwidths of classes of Markovian traffic sources can be found in Elwalid and Mitra [61], while more general input processes are considered by Whitt [186], Chang [27], and de Veciana and Walrand [48]. In many studies one has already succeeded in finding analytical expressions of the effective bandwidth functions of several ATM arrival processes, for instance Markov fluid sources and Markov modulated Poisson sources, see Kesidis, Walrand, and Chang [105].

In recent years, attention was also paid on assigning effective bandwidths to departure processes, see Chang, Heidelberger, Juneja, and Shahabuddin [28] and de Veciana, Courcoubetis, and Walrand [46]. They attempted to extend the analysis to an important class of queueing networks: *intree networks*, coming much closer to 'ATM-reality' than single queues. Their analysis had two aims. Their first purpose was to determine the decay rate of the overflow probability of any queue in the network. Clearly, a decay rate by itself only provides a rough impression of the tail of the distribution. For that reason, their second goal was to develop fast simulation methods (based on *importance sampling*) that provide numerical values. However, a number of questions remained unanswered.

(i) Chang *et al.* [28] developed a conjecture of the decay rate of the level crossing probability of any particular queue in the intree network. They indeed found that the decay rate was bounded from above by the conjectured expression, but they could not prove the lower bound.

(ii) In [28] an efficient simulation procedure was proposed. This procedure is based on importance sampling, and the alternative probability measure was meant to be such that its variance performance were optimal. However, this optimality property was not shown.

The contribution of this chapter is the solution of these problems, for the special case of a two-node tandem model with Markov modulated fluid input.

We begin this study by introducing the model, reviewing some basic results on importance sampling and effective bandwidths, and explaining the simulation technique proposed by [28]. Section 3 deals with rough asymptotics of the overflow probability in

the second queue. Also the optimality of the importance sampling technique, that was proposed in Chang *et al.* [28], is proven. In Section 4 we comment on the estimation of the long-run fraction of fluid that is lost (the fluid loss ratio), and give simulation results. A summary of results and conclusions is found in the final section.

2 Model description - Importance sampling

The performance of a queueing system with respect to loss can be measured by means of several criteria, for instance the cell loss ratio or the mean time to overflow (starting with an empty system). However, as pointed out in the introduction, in most cases the mathematical analysis yields only rough, asymptotic characteristics of these performance measures. Then, simulation can be used for obtaining numerical values. But the performance criteria include the rare event of a buffer overflow, and therefore (using direct simulation) a huge effort is required to get an accurate estimate. This motivates the research on applying variance reduction techniques, e.g., importance sampling.

Importance sampling uses some alternative probability model, under which the rare event under consideration occurs more frequently. The system is simulated under this measure; the simulation output is translated back to the original model to obtain unbiased estimates. Obviously, we are left with the task of finding the alternative probability measure providing the largest variance reduction.

In the first part of this section, we describe our model formally. Then we give a brief outline of importance sampling and we relate our study to earlier work. The choice of an appropriate alternative probability model is closely related to the concept of effective bandwidths, of which we will review the main ideas. We conclude this section by Chang's conjecture of the optimal change of measure.

2.1 Model description

The model we consider is a so-called two-link tandem model. The first queue is fed by a Markov fluid source. A fluid source is characterized by (i) an infinitesimal generator $\Lambda = (\lambda_{ij})_{i,j}$ of a finite-state, irreducible continuous-time Markov chain with invariant π and (ii) a traffic rate vector r. Fluid is generated at constant rate r_i while the modulating Markov chain is in state i. In practice, several (independent) sources will feed into the system, but these can be superimposed to one source [61].

The effective bandwidth, as mentioned in the introduction, arises from an eigensystem [110], [61]. Let R be $\text{diag}\{r\}$, then the effective bandwidth $C(\theta)$ of a (Λ, r) source is given by the largest (real) eigenvalue of $R + \Lambda/\theta$. This function increases from the mean rate $\mu := \sum_i \pi_i r_i$ (for $\theta \downarrow 0$) to the peak rate $r_p := \max_i r_i$ (for $\theta \to \infty$). Kesidis *et al.* [105]

found the following alternative characterization of $C(\cdot)$. Let $A(t)$ denote the amount of fluid generated by the Markov fluid source during $[0,t]$, and define for real θ the asymptotic log moment generating function

$$M(\theta) := \lim_{t\to\infty} \frac{1}{t} \log \mathrm{E} \exp(\theta A(t)).$$

It was shown that $C(\theta) = M(\theta)/\theta$. Also, $M(\cdot)$ is convex, with $M'(0) = \mu$.

The arrival process obeys a certain limiting regime, which is described by the following lemma, frequently used in Section 3.

LEMMA 2.1. There exist positive $J(a)$ and $H(a)$ such that for all $t \geq 0$,

$$\mathcal{P}\left(\frac{A(t)}{t} \geq a\right) \leq H(a)e^{-J(a)t} \qquad (a > \mu)$$

$$\mathcal{P}\left(\frac{A(t)}{t} \leq a\right) \leq H(a)e^{-J(a)t} \qquad (a < \mu)$$

As a consequence, $A(t)/t$ converges to $M'(0) = \mu$ a.s.

PROOF. Analogously to Theorem 2.5 of [28], the upper bounds follows from Chebychev's inequality. The almost sure convergence is due to the Borel-Cantelli lemma. ■

The first queue is emptied at a constant rate $C_1 > \mu$. The output of the first queue is led into a second queue, serviced by a channel of constant capacity $C_2 \in (\mu, C_1)$. Notice that this last assumption is no restriction, because the other cases are either trivial ($C_2 \geq C_1$ means that overflow in queue 2 is impossible) or not of interest ($C_2 \leq \mu$ means that overflow is not rare). Also, we assume that $r_p > C_1$, because otherwise the tandem system is essentially a single queue. The buffers are assumed to be infinitely large.

2.2 Importance sampling

Importance sampling is a variance reduction technique, whose main idea is to simulate under a probability measure (say $\mathcal{Q}$) that differs from the actual one (say $\mathcal{P}$). Unbiasedness is maintained by weighing each observation by a likelihood. Suppose, for some random variable X, we want to estimate $\mathrm{E}^{(\mathcal{P})}(X)$ (in self-evident notation), of course assuming $\mathrm{E}^{(\mathcal{P})}|X| < \infty$. In case of direct simulation, we would draw a sample $X_1, \ldots, X_n$ according to $\mathcal{P}$, yielding unbiased estimator $\sum_{i=1}^n X_i/n$. However, consider the possibility of generating a sample under $\mathcal{Q}$. If $\mathcal{P}$ is absolutely continuous with respect to $\mathcal{Q}$, the Radon-Nikodym theorem implies the existence of a likelihood ratio L such that $\mathrm{E}^{(\mathcal{P})}(X) = \mathrm{E}^{(\mathcal{Q})}(LX)$. Simulating the process under $\mathcal{Q}$, we find that an unbiased estimator is $\sum_{i=1}^n L_i X_i/n$, where L_i denotes the likelihood of the observation in run i. Now it remains to point out how to capture this likelihood. We let every X be determined by a

sequence (of random length) of independent random variables $Y_1, \ldots, Y_T$, all having densities under both $\mathcal{P}$ and $\mathcal{Q}$. Then the likelihood is simply the ratio of the (joint) density under $\mathcal{P}$ in $(Y_1, \ldots, Y_T)$ and the density under $\mathcal{Q}$ in $(Y_1, \ldots, Y_T)$. For details, see [77].

Returning to our setting of the tandem queue network, X is for instance the indicator function of an overflow in the second queue, which is rare under $\mathcal{P}$. This suggests that, in the importance sampling, the parameters should be changed such that the second queue becomes unstable. Of course, an interesting issue is the choice of $\mathcal{Q}$ such that the variance reduction is maximal. Sadowsky [160] showed for GI/G/m queues that an exponential twist of the densities of the original model is optimal within a broad class of importance sampling distributions, Parekh and Walrand [144] heuristically motivate a similar twist for networks of GI/G/1 queues. Glasserman and Kou [76] give conditions for an alternative measure to be optimal in order to estimate rare event probabilities in tandem Jackson networks. In contrast with our study, the authors of [76] consider the probability of the network population reaching a large value, instead of the individual buffer contents. An extensive survey on fast simulation methods is [85].

Kesidis and Walrand [103] propose a change of measure for single queues with a Markov fluid arrival process, but they do not treat the variance performance of the estimator. Chang *et al.* [28] show for single queues (in discrete time) with Markov modulated input that the optimal change of measure is again exponential and closely related to the effective bandwidth concept. Their reasoning can be extended easily to (continuous-time) Markov fluid sources. The next subsection deals with the computation of the exponential twist of a Markov fluid source.

2.3 Exponential twisting - Effective bandwidths

The source (Λ, r) can be (exponentially) θ-twisted in the following way. Consider a real θ. Find the largest real eigenvalue of $R + \Lambda/\theta$. The corresponding right eigenvector x is componentwise positive and uniquely determined to constant multiples [61], and we recall that the eigenvalue is the effective bandwidth $C(\theta)$. Let the new rates $\lambda_{ij}(\theta)$ be $\lambda_{ij}x_j/x_i$ for $i \neq j$ and $\lambda_{ii}(\theta) := \lambda_{ii} + \theta(r_i - C(\theta))$. The traffic rates r remain unchanged. If the structure of the source is more complicated than simply on-off, the eigensystem can still be solved numerically in an efficient way [125].

LEMMA 2.2. Let $(\Lambda(\sigma), r)$ be the σ-twisted version of (Λ, r). Let $\mathrm{E}^{(\Lambda)}$ ($\mathrm{E}^{(\Lambda(\sigma))}$, respectively) denote expectation under rate matrix Λ ($\Lambda(\sigma)$). Then for all real θ,

$$\lim_{t\to\infty} \frac{1}{t} \log \mathrm{E}^{(\Lambda(\sigma))} \exp(\theta A(t)) = \lim_{t\to\infty} \frac{1}{t} \log \frac{\mathrm{E}^{(\Lambda)} \exp((\sigma+\theta)A(t))}{\mathrm{E}^{(\Lambda)} \exp(\sigma A(t))}.$$

PROOF. Let $C(\cdot)$ and $D(\cdot)$ be the effective bandwidth functions of the (Λ, r) and $(\Lambda(\sigma), r)$

source, respectively. Equivalently, it has to be shown that $(\sigma+\theta)C(\sigma+\theta)$ is equal to $\sigma C(\sigma)+\theta D(\theta)$. Our proof consists of two steps. First we show that $f := (\sigma C(\sigma)+\theta D(\theta))(\theta+\sigma)^{-1}$ is eigenvalue of $R+\Lambda/(\sigma+\theta)$. Then we prove that there does not exist any larger eigenvalue.

(i). Let x and y be the right eigenvectors from the definitions of $C(\sigma)$ and $D(\theta)$:

$$\sigma(C(\sigma)-r_i)=\sum_j \lambda_{ij}\frac{x_j}{x_i} \quad \text{and} \quad \theta(D(\theta)-r_i)=\sum_j \lambda_{ij}(\sigma)\frac{y_j}{y_i}.$$

Then z, with $z_i = x_i y_i$, is right eigenvector of $R+\Lambda/(\sigma+\theta)$, with eigenvalue f:

$$(\sigma C(\sigma)+\theta D(\theta))z_i = \left(\sigma r_i + \sum_j \lambda_{ij}\frac{x_j}{x_i} + \theta r_i + \sum_j \lambda_{ij}(\sigma)\frac{y_j}{y_i}\right) z_i =$$

$$(\sigma+\theta)r_i z_i + \left(\sum_{j\neq i}\lambda_{ij}(\sigma) + \lambda_{ii} + \sum_{j\neq i}\lambda_{ij}\frac{z_j}{z_i} + \lambda_{ii}(\sigma)\right) z_i = (\sigma+\theta)r_i z_i + \left(\sum_j \lambda_{ij} z_j\right).$$

(ii). Now suppose there exists an eigenvalue g of $R+\Lambda/(\sigma+\theta)$ that is larger than f, with accompanying (componentwise positive) right eigenvector w:

$$(\sigma+\theta)(g-r_i)=\sum_j \lambda_{ij}\frac{w_j}{w_i}.$$

Let $C(\sigma)$ be the largest eigenvalue of $R+\Lambda/\sigma$, with right eigenvector x. Then

$$\left(\frac{(\sigma+\theta)g-\sigma C(\sigma)}{\theta}\right)\frac{w_i}{x_i} = \frac{1}{\theta}\left((\sigma+\theta)r_i + \sum_j \lambda_{ij}\frac{w_j}{w_i} - \sigma r_i - \sum_j \lambda_{ij}\frac{x_j}{x_i}\right)\frac{w_i}{x_i}. \tag{5.1}$$

Now notice that

$$\sum_j \lambda_{ij}\frac{w_j}{w_i} = \sum_{j\neq i}\lambda_{ij}(\sigma)\frac{x_i w_j}{x_j w_i} + \lambda_{ii} \quad \text{and} \quad \sum_j \lambda_{ij}\frac{x_j}{x_i} = \sum_{j\neq i}\lambda_{ij}(\sigma) + \lambda_{ii} = -\lambda_{ii}(\sigma) + \lambda_{ii}.$$

As a consequence, the right hand side of (5.1) reads

$$r_i\frac{w_i}{x_i} + \frac{1}{\theta}\left(\sum_j \lambda_{ij}(\sigma)\frac{w_j}{x_j}\right).$$

Thus, $R+\Lambda(\sigma)/\theta$ has eigenvalue $(\sigma+\theta)g\theta^{-1} - \sigma C(\sigma)\theta^{-1} > D(\theta)$. Contradiction. ■

The following twisting is of particular importance. Let $\tilde{\theta}$ solve $M'(\theta)=C_1$. Due to the previous lemma, if the arrival process is $\tilde{\theta}$-twisted, its asymptotic log-moment generating function is given by $N(\theta) := M(\theta+\tilde{\theta}) - M(\tilde{\theta})$. Applying Lemma 2.1, the mean arrival rate converges (a.s.) to $N'(0) = M'(\tilde{\theta}) = C_1$. We conclude that the $\tilde{\theta}$-twisting provides a load 1 queue.

2.4 The change of measure

We define a busy period of a queue as the interval until the queue returns empty, started from empty. (Notice that in case of fluid queues the output rate can be positive, although the queue is empty!) Chang *et al.* [28] investigated the change of measure in order to estimate the probability of overflow in queue 2, during a busy period of this queue. They proposed the following change of measure. Let θ^* be the (positive, unique) solution of

$$C_D(\theta) = C_2, \text{ where } C_D(\theta) := \begin{cases} C(\theta) & \text{if } \theta \leq \tilde{\theta}; \\ C_1 - \frac{\tilde{\theta}}{\theta}\left(C_1 - C(\tilde{\theta})\right) & \text{if } \theta > \tilde{\theta}. \end{cases} \tag{5.2}$$

The arrival process should be twisted by the smallest of $\tilde{\theta}$ and θ^*. A consequence is that (under this new measure) the load of the first queue is not larger than 1, where the second queue is unstable (load larger than 1). This can be seen as follows, applying the previous lemmas. Call the mean arrival rate under the above change of measure ν. (i) If $\tilde{\theta} \leq \theta^*$, $\nu = C_1$. Since $C_1 > C_2$, the second queue has load larger than 1. (ii) Suppose $\tilde{\theta} > \theta^*$. From the (strict) convexity of $M(\cdot)$, $\nu = M'(\theta^*) < M'(\tilde{\theta}) = C_1$. Furthermore,

$$\nu = M'(\theta^*) = C(\theta^*) + \theta^* C'(\theta^*) > C(\theta^*) = C_2,$$

since θ^* is positive and $C(\cdot)$ increases. Consequently, $\tilde{\theta} > \theta^*$ yields $\nu \in (C_2, C_1)$, implying that the first queue is stable where the second is not.

The intuition behind $\nu \in (C_2, C_1]$ is the following: In 'optimal importance sampling', we let typical behavior under the new measure coincide with deviant behavior under the old one. Loosely speaking, we choose the parameters of the new measure $\mathcal{Q}$ such that its 'average trajectory' equals the 'most likely trajectory' to overflow under $\mathcal{P}$ (cf. Anantharam [4]). Now distinguish between the following two cases:

- First suppose $\nu \leq C_2$. Then the mean output rate of the first queue is ν, since $\nu \leq C_2 < C_1$. Consequently, overflow in the second queue remains rare. So, ν must be chosen larger than C_2.

- Suppose on the other hand $\nu > C_1$. Then the mean output rate of queue 1 will be C_1. But this is also the case if $\nu = C_1$! However, it is more likely that the source transmits at rate C_1 than at a larger rate (more precisely: transmitting at rate C_1 is less deviant from the 'actual' rate μ than transmitting at a larger rate). Consequently, the optimal arrival rate in order to cause overflow will not be larger than C_1.

The notation $C_D(\cdot)$ is used because Chang *et al.* [28] conjectured it to be the effective bandwidth of the departure process, cf. also Corollary 3.1 in [46].

3 Analysis of the level crossing probability

This section deals with the asymptotics of the level crossing probability $\alpha(B)$, i.e., the probability that the buffer contents of the second queue exceeds, during a busy period, level B. This probability is very useful in determining the mean time to overflow [160].

Observe that at the start of a busy period of the second queue the first queue is empty as well, as a consequence of $C_1 > C_2$. We let β denote the steady state distribution of the state of the modulating Markov chain at the beginning of the busy period of the second queue. The main result of this section deals with the asymptotics of $\alpha(B)$:

THEOREM 3.1. Let θ^* solve $C_D(\theta) = C_2$. Then $B^{-1} \log \alpha(B) \to -\theta^*$ as $B \to \infty$.

In Subsection 3.1 we show upper bound $\limsup B^{-1} \log \alpha(B) \leq -\theta^*$, Subsection 3.2 shows that $\liminf B^{-1} \log \alpha(B) \geq -\theta^*$. Although the upper bound was already in Chang *et al.* [28], we give our own proof. This is because of the fact that the relations derived in order to prove this upper bound can be used in the lower bound.

Before starting, we first give the following definitions. $D(t)$ ($O(t)$, respectively) is the amount of fluid left from the first (second) queue in $[0, t]$. $X(t)$ is the state of the modulating Markov chain at time t. $Q_i(t)$ is the buffer contents of queue i at time t, $i = 1, 2$.

3.1 Upper bound

We call $\Lambda(\theta)$ the exponential twisted version of Λ, according to the change of measure $\mathcal{Q}$ proposed by Chang *et al.* [28], which was explained in the previous section. Here θ is the smallest of θ^* and $\tilde{\theta}$. We call $\mathcal{T}_B$ the first epoch at which the second queue exceeds B, and $\mathcal{T}$ the first epoch at which this queue is empty again:

$$\mathcal{T}_B := \inf\{t > 0 : D(t) - O(t) = B\} \quad \text{and} \quad \mathcal{T} := \inf\{t > 0 : D(t) - O(t) = 0\}.$$

Call $\mathcal{I}_B$ the indicator function of $\{\mathcal{T}_B < \mathcal{T}\}$, we have $\alpha(B) = \mathrm{E}^{(\mathcal{P})}(\mathcal{I}_B) = \mathrm{E}^{(\mathcal{Q})}(L\mathcal{I}_B)$, L denoting the likelihood ratio of the sample path. The experiment is started (at time 0) from a situation in which both queues are empty, and the modulating chain is distributed according to β. We let the sequence $((I_0, T_0), (I_1, T_1), \ldots, (I_N, T_N))$ denote the states of the modulating chain and the times spent in those states, during the experiment.

LEMMA 3.2.

$$\alpha(B) = \mathrm{E}^{(\mathcal{Q})}\left(\frac{x_{I_0}}{x_{I_N}} \exp\left[-\theta A(\mathcal{T}_B) + \theta C(\theta)\mathcal{T}_B\right] \mathcal{I}_B\right). \tag{5.3}$$

PROOF. As said before, $\alpha(B) = \mathrm{E}^{(\mathcal{Q})}(L\mathcal{I}_B)$. Define $\lambda_i := -\lambda_{ii}$ and $\lambda_i(\theta)_i := -\lambda_{ii}(\theta)$. We

write L as the ratio of the joint densities of the sample path:

$$\lambda_{I_0} \exp\left[-\lambda_{I_0} T_0\right] \left(\frac{\lambda_{I_0 I_1}}{\lambda_{I_0}}\right) \cdots \left(\frac{\lambda_{I_{N-1} I_N}}{\lambda_{I_{N-1}}}\right) \lambda_{I_N} \exp\left[-\lambda_{I_N} T_N\right]$$

divided by a similar expression, where the λ's are replaced by $\lambda(\theta)$'s. Now L reads, using the relations $\lambda_{ij}(\theta) = \lambda_{ij} x_j / x_i$ and $\lambda_{ii}(\theta) = \lambda_{ii} + \theta(r_i - C(\theta))$,

$$\frac{\lambda_{I_N}}{\lambda_{I_N}(\theta)} \frac{x_{I_0}}{x_{I_N}} \exp\left[\sum_{k=0}^{N} \left(\lambda_{I_k}(\theta) - \lambda_{I_k}\right) T_k\right] = \frac{\lambda_{I_N}}{\lambda_{I_N}(\theta)} \frac{x_{I_0}}{x_{I_N}} \exp\left[-\sum_{k=0}^{N} \theta \left(r_{I_k} - C(\theta)\right) T_k\right]. \quad (5.4)$$

The exponent in (5.4) equals $\exp\left[-\theta A\left(T^{(N)}\right) + \theta C(\theta) T^{(N)}\right]$, where $T^{(N)} := \sum_{k=0}^{N} T_k$. Now suppose $I_N = i$. Then, on $\{\mathcal{I}_B = 1\}$, we have the following equality in distribution (due to the memoryless property of the exponential distribution)

$$A\left(T^{(N)}\right) - C(\theta) T^{(N)} = A(\mathcal{T}_B) - C(\theta)\mathcal{T}_B + (r_i - C(\theta))X,$$

where $(A(\mathcal{T}_B) - C(\theta)\mathcal{T}_B)$ and X are independent and X is exponential with mean $(\lambda_i(\theta))^{-1}$. Now the stated follows from

$$\mathrm{E}^{(\mathcal{Q})} \exp\left[-\theta(r_i - C(\theta))X\right] = \frac{\lambda_i(\theta)}{\lambda_i(\theta) + \theta(r_i - C(\theta))} = \frac{\lambda_i(\theta)}{\lambda_i}. \qquad \blacksquare$$

Invoking Lemma 3.2, we proof the upper bound as follows. Notice that the first factor of the right hand side of (5.3) can be simply bounded (uniformly in B) by the maximum over all possible indices: $\max_{i,j} x_i / x_j$. Therefore, it suffices to prove the upper bound for the exponential part of (5.3). On $\{\mathcal{I}_B = 1\}$ we have (i) $D(\mathcal{T}_B) - C_2\mathcal{T}_B = B$. Furthermore, for all $t \geq 0$ (ii) $D(t) \leq A(t)$ and (iii) $D(t) \leq C_1 t$. Suppose $\tilde{\theta} \leq \theta^*$. Then the source is twisted by $\tilde{\theta}$. It can be checked that $C_D(\theta^*) = C_2$ is equivalent to $\tilde{\theta} C(\tilde{\theta}) = (C_2 - C_1)\theta^* + \tilde{\theta} C_1$. The exponent of (5.3) can be bounded as follows:

$$-\tilde{\theta} A(\mathcal{T}_B) + \tilde{\theta} C(\tilde{\theta}) \mathcal{T}_B = (\theta^* - \tilde{\theta})(D(\mathcal{T}_B) - C_1\mathcal{T}_B) + \theta^*(C_2\mathcal{T}_B - D(\mathcal{T}_B)) + \tilde{\theta}(D(\mathcal{T}_B) - A(\mathcal{T}_B))$$

which is smaller than $-\theta^* B$. If $\tilde{\theta} > \theta^*$, we have $C(\theta^*) = C_2$. We find the same upper bound via $-\theta^*(A(\mathcal{T}_B) - C_2\mathcal{T}_B) \leq -\theta^*(D(\mathcal{T}_B) - C_2\mathcal{T}_B) = -\theta^* B$. This proves the upper bound of Theorem 3.1.

3.2 Lower bound

We start the proof of the lower bound by establishing a few useful lemmas that describe the behavior of $\mathcal{T}_B$. Lemma 3.4 is proven in the appendix.

LEMMA 3.3. $\mathrm{E}^{(\mathcal{Q})} A(\mathcal{T}_B) / \mathrm{E}^{(\mathcal{Q})} \mathcal{T}_B \to \nu$ as $B \to \infty$, with ν as defined in Subsection 2.4.

PROOF. Define $\mathcal{T}_B'$ as the first epoch after $\mathcal{T}_B$ at which the modulating chain is in its starting state $X(0) = i^*$: $\mathcal{T}_B' := \inf\{t \geq \mathcal{T}_B : X(t) = i^*\}$. The arrival process regenerates after a return to state i^*. Let N_B be the number of regenerations in $[0, \mathcal{T}_B']$; A_i and τ_i are the fluid generated in the ith regeneration cycle and the duration of the ith cycle, respectively; write A and τ for the generic variables. Clearly, the event $\{N_B = n\}$ does not depend on $\{A_{n+1}, A_{n+2}, \ldots\}$ nor $\{\tau_{n+1}, \tau_{n+2}, \ldots\}$. Notice that $\mathrm{E}^{(\mathcal{Q})}\tau < \infty$ since an irreducible, finite-state Markov chain is involved; also $\mathrm{E}^{(\mathcal{Q})}A \leq r_p \mathrm{E}^{(\mathcal{Q})}\tau < \infty$. Consequently, we may apply 'Wald':

$$\mathrm{E}^{(\mathcal{Q})}A(\mathcal{T}_B') = \mathrm{E}^{(\mathcal{Q})}N_B \cdot \mathrm{E}^{(\mathcal{Q})}A \quad \text{and} \quad \mathrm{E}^{(\mathcal{Q})}\mathcal{T}_B' = \mathrm{E}^{(\mathcal{Q})}N_B \cdot \mathrm{E}^{(\mathcal{Q})}\tau.$$

Then by the renewal reward theorem and Lemma 2.1,

$$\mathrm{E}^{(\mathcal{Q})}A(\mathcal{T}_B') = \frac{\mathrm{E}^{(\mathcal{Q})}A}{\mathrm{E}^{(\mathcal{Q})}\tau}\,\mathrm{E}^{(\mathcal{Q})}\mathcal{T}_B' \overset{\text{a.s.}}{=} \lim_{t\to\infty}\frac{A(t)}{t}\,\mathrm{E}^{(\mathcal{Q})}\mathcal{T}_B' \overset{\text{a.s.}}{=} \nu\,\mathrm{E}^{(\mathcal{Q})}\mathcal{T}_B'.$$

It follows that

$$\frac{\mathrm{E}^{(\mathcal{Q})}A(\mathcal{T}_B)}{\mathrm{E}^{(\mathcal{Q})}\mathcal{T}_B} = \nu + \frac{\nu(\mathrm{E}^{(\mathcal{Q})}\mathcal{T}_B' - \mathrm{E}^{(\mathcal{Q})}\mathcal{T}_B) + (\mathrm{E}^{(\mathcal{Q})}A(\mathcal{T}_B) - \mathrm{E}^{(\mathcal{Q})}A(\mathcal{T}_B'))}{\mathrm{E}^{(\mathcal{Q})}\mathcal{T}_B}. \tag{5.5}$$

Now we examine the interval $[\mathcal{T}_B, \mathcal{T}_B']$. Define U_{ij} as the time it takes for the Markov chain to get in j, starting in i. It is seen easily that $\mathrm{E}^{(\mathcal{Q})}(\mathcal{T}_B' - \mathcal{T}_B)$ is bounded from above by $u := \max_{i,j}\mathrm{E}^{(\mathcal{Q})}U_{ij}$, which is finite because of the irreducibility of the modulating Markov chain. Also, $\mathrm{E}^{(\mathcal{Q})}(A(\mathcal{T}_B') - A(\mathcal{T}_B)) \leq r_p u$. Noticing that $\mathcal{T}_B \geq B/(C_1 - C_2)$, the second term of the right hand side of (5.5) converges to 0 as $B \to \infty$. ■

LEMMA 3.4. $\limsup_{B\to\infty}\mathrm{E}^{(\mathcal{Q})}\mathcal{T}_B/B \leq (\nu - C_2)^{-1}$.

LEMMA 3.5. $\mathrm{E}^{(\mathcal{Q})}D(\mathcal{T}_B)/\mathrm{E}^{(\mathcal{Q})}\mathcal{T}_B \to \nu$ as $B \to \infty$.

PROOF. Since $D(t) \leq A(t)$ it suffices to show that $\liminf \mathrm{E}^{(\mathcal{Q})}D(\mathcal{T}_B)/\mathrm{E}^{(\mathcal{Q})}\mathcal{T}_B \geq \nu$ (use Lemma 3.3). Clearly,

$$D(\mathcal{T}_B) = B + O(\mathcal{T}_B) = B + C_2\mathcal{T}_B - rS_B, \tag{5.6}$$

for an $r \in [0, C_2)$ and S_B the total duration of time intervals (till $\mathcal{T}_B$) on which $Q_2(\cdot) = 0$. But if $Q_2(t) = 0$, then also $Q_1(t) = 0$ or equivalently $A(t) = D(t)$. It follows that

$$\{Q_2(t) = 0\} = \{D(t) - O(t) = 0\} = \{A(t) - O(t) = 0\} \subset \{A(t) - C_2 t \leq 0\}, \tag{5.7}$$

and therefore by Lemma 2.1, $\mathrm{E}^{(\mathcal{Q})}S_B$ is bounded by a constant (uniformly in B):

$$\mathrm{E}^{(\mathcal{Q})}S_B \leq \mathrm{E}^{(\mathcal{Q})}S_\infty = \int_0^\infty \mathcal{Q}(Q_2(t) = 0)\mathrm{d}t \leq \int_0^\infty \mathcal{Q}(A(t) \leq C_2 t)\mathrm{d}t \leq \int_0^\infty H(C_2)e^{-J(C_2)t}\mathrm{d}t,$$

which is finite. Since $\mathcal{T}_B \geq B/(C_1 - C_2)$, $\mathrm{E}^{(\mathcal{Q})}S_B/\mathrm{E}^{(\mathcal{Q})}\mathcal{T}_B \to 0$. By invoking Lemma 3.4, the stated follows immediately from (5.6). ■

Now we can start the actual proof of the lower bound. We first notice that Lemma 3.2 implies

$$\alpha(B) = \mathrm{E}^{(\mathcal{Q})}\left(\frac{x_{I_0}}{x_{I_N}} \exp\left[-\theta A(\mathcal{T}_B) + \theta C(\theta)\mathcal{T}_B\right] | \mathcal{I}_B = 1\right) \mathcal{Q}(\mathcal{I}_B = 1). \tag{5.8}$$

Clearly $\mathcal{Q}(I_B = 1) \geq \mathcal{Q}(\forall t > 0 : Q_2(t) > 0)$, which is positive due to the unstability of the second queue, cf. part (iii) of Theorem 2.5 in [28]. So we only have to find the lower bound for the conditional expectation. A lower bound (uniformly in B) for the first factor is trivial: $\min_{i,j} x_i/x_j$. For the exponential part, first consider the case $\tilde{\theta} \leq \theta^*$; applying (5.2)

$$\begin{aligned}\tilde{\theta}A(\mathcal{T}_B) - \tilde{\theta}C(\tilde{\theta})\mathcal{T}_B &= \tilde{\theta}A(\mathcal{T}_B) - \theta^*(C_2 - C_1)\mathcal{T}_B + \tilde{\theta}C_1\mathcal{T}_B \\ &= \theta^*(D(\mathcal{T}_B) - C_2\mathcal{T}_B) + \theta^*(C_1\mathcal{T}_B - D(\mathcal{T}_B)) + \tilde{\theta}(A(\mathcal{T}_B) - C_1\mathcal{T}_B).\end{aligned}$$

Noting that $D(\mathcal{T}_B) - C_2\mathcal{T}_B = B$ on $\{\mathcal{I}_B = 1\}$, the decay rate of $\alpha(B)$ is larger than

$$\begin{aligned}&-\theta^* + \liminf_{B\to\infty} \frac{1}{B} \log \mathrm{E}^{(\mathcal{Q})}\left(\exp\left[\theta^*(D(\mathcal{T}_B) - C_1\mathcal{T}_B) + \tilde{\theta}(C_1\mathcal{T}_B - A(\mathcal{T}_B))\right] | \mathcal{I}_B = 1\right) \geq \\ &-\theta^* + \liminf_{B\to\infty} \frac{1}{B}\mathrm{E}^{(\mathcal{Q})}\left(\left[\theta^*(D(\mathcal{T}_B) - C_1\mathcal{T}_B) + \tilde{\theta}(C_1\mathcal{T}_B - A(\mathcal{T}_B))\right] | \mathcal{I}_B = 1\right)\end{aligned}$$

applying Jensen's inequality. With the same line of reasoning, we get, for the case $\tilde{\theta} > \theta^*$, that $\liminf B^{-1} \log \alpha(B)$ majorizes

$$-\theta^* + \liminf_{B\to\infty} \frac{1}{B}\mathrm{E}^{(\mathcal{Q})}(\theta^*(D(\mathcal{T}_B) - A(\mathcal{T}_B)) | \mathcal{I}_B = 1)\,.$$

Again applying that, uniformly in B, $\mathcal{Q}(\mathcal{I}_B = 1)$ lies between two positive constants, and denoting $\mathrm{E}^{(\mathcal{Q})}(X; F) := \int_F x \mathrm{d}\mathcal{Q}(X = x)$, the lower bound is an immediate consequence of the following lemma.

LEMMA 3.6.

$$\lim_{B\to\infty} \frac{\mathrm{E}^{(\mathcal{Q})}(A(\mathcal{T}_B) - \nu\mathcal{T}_B; \mathcal{T}_B < \mathcal{T})}{B} = \lim_{B\to\infty} \frac{\mathrm{E}^{(\mathcal{Q})}(D(\mathcal{T}_B) - \nu\mathcal{T}_B; \mathcal{T}_B < \mathcal{T})}{B} = 0. \tag{5.9}$$

PROOF. Consider the first limit of (5.9); the second limit is analogous. Trivially,

$$\frac{\mathrm{E}^{(\mathcal{Q})}(A(\mathcal{T}_B) - \nu\mathcal{T}_B; \mathcal{T}_B < \mathcal{T})}{B} = \frac{\mathrm{E}^{(\mathcal{Q})}(A(\mathcal{T}_B) - \nu\mathcal{T}_B)}{B} - \frac{\mathrm{E}^{(\mathcal{Q})}(A(\mathcal{T}_B) - \nu\mathcal{T}_B; \mathcal{T}_B > \mathcal{T})}{B}. \tag{5.10}$$

The first term of the right hand side (rhs) vanishes due to Lemmas 3.3 and 3.4:

$$\lim_{B\to\infty}\left|\frac{\mathrm{E}^{(\mathcal{Q})}(A(\mathcal{T}_B)-\nu\mathcal{T}_B)}{B}\right| \le \limsup_{B\to\infty}\frac{\mathrm{E}^{(\mathcal{Q})}\mathcal{T}_B}{B}\cdot\lim_{B\to\infty}\left|\frac{\mathrm{E}^{(\mathcal{Q})}A(\mathcal{T}_B)}{\mathrm{E}^{(\mathcal{Q})}\mathcal{T}_B}-\nu\right| = 0. \tag{5.11}$$

The second term of the rhs of (5.10) can be treated as follows. If $\mathcal{T}_B > \mathcal{T}$, the second queue starts a new busy period before reaching overflow. Let S denote the epoch of the start of this busy period. Then the interval $[0,\mathcal{T}_B]$ can be divided into $[0,S)$ and $[S,\mathcal{T}_B] = [S, S+\mathcal{T}_B']$, where $\mathcal{T}_B'$ denotes the length of the time interval between S and $\mathcal{T}_B$. Notice that the situation of the queueing system at S equals the situation at time 0, except possibly the state of the modulating Markov chain. It follows that

$$\frac{\mathrm{E}^{(\mathcal{Q})}(A(\mathcal{T}_B)-\nu\mathcal{T}_B;\mathcal{T}_B>\mathcal{T})}{B} = \frac{\mathrm{E}^{(\mathcal{Q})}(A(S)-\nu S;\mathcal{T}_B>\mathcal{T})}{B} + \frac{\mathrm{E}^{(\mathcal{Q})}(A(\mathcal{T}_B')-\nu\mathcal{T}_B')}{B}. \tag{5.12}$$

Now notice that Lemmas 3.3 and 3.4 did not depend on the state of the chain at the start of the busy period of queue 2. Therefore the second term in the rhs of (5.12) converges to 0, analogously to (5.11). Showing that the the numerator of the first term in the rhs of (5.12) has a finite upper bound, uniformly in B, we are done. This is done as follows.

Due to the triangle inequality and $\{\mathcal{T}_B > \mathcal{T}\} \subset \{S < \infty\}$,

$$|\,\mathrm{E}^{(\mathcal{Q})}(A(S)-\nu S;\mathcal{T}_B>\mathcal{T})\,| \le (r_p+\nu)\mathrm{E}^{(\mathcal{Q})}(S;S<\infty).$$

Let $\lceil a\rceil$ be the smallest integer larger than or equal to a. Then $\mathrm{E}^{(\mathcal{Q})}(S;S<\infty)$ is dominated by $\mathrm{E}^{(\mathcal{Q})}(\lceil S\rceil;S<\infty)$, which is smaller than

$$\sum_{n=1}^{\infty} n\mathcal{Q}(\exists t\in(n-1,n]: Q_2(t)=0) \le \sum_{n=1}^{\infty} n\mathcal{Q}(\exists t\in(n-1,n]: A(t)\le C_2 t),$$

using inclusion (5.7). But if $A(t)\le C_2 t$ for a $t\in(n-1,n]$, then also $A(n)\le C_2 n + r_p - C_2$. Now choose $\epsilon < \nu - C_2$. Then for $n \ge N := \lceil (r_p - C_2)/\epsilon\rceil$,

$$\{\exists t\in(n-1,n]: Q_2(t)=0\} \subset \left\{\frac{A(n)}{n}\le C_2+\epsilon\right\}.$$

Noticing that $C_2+\epsilon<\nu$, we invoke Lemma 2.1 and get the following upper bound for $\mathrm{E}^{(\mathcal{Q})}(S;S<\infty)$:

$$\sum_{n=1}^{N} n + \sum_{n=N+1}^{\infty} n\mathcal{Q}\left(\frac{A(n)}{n}\le C_2+\epsilon\right) \le \sum_{n=1}^{N} n + \sum_{n=N+1}^{\infty} nH(C_2+\epsilon)e^{-J(C_2+\epsilon)n},$$

which is finite. ■

3.3 Optimality of the importance sampling procedure

In case of estimating a probability by a Monte Carlo procedure, the number of samples to be drawn to get a fixed relative efficiency is proportional to the reciprocal of the probability to be estimated. Therefore, in order to estimate $\alpha(B)$, the number of runs blows up more or less exponentially in the buffer size. Executing the simulation as described in the previous section, we empirically find that the number of runs required is polynomial in the buffer size, sometimes even more or less constant. In this section we will prove that, within some class of distributions, there is no alternative importance sampling measure that results in a better variance performance.

The quality of the proposed procedure is determined by the variance of the observations $L\mathcal{I}_B$ under the alternative measure $\mathcal{Q}$: $\mathrm{Var}^{(\mathcal{Q})}(L\mathcal{I}_B) = \mathrm{E}^{(\mathcal{Q})}(L^2\mathcal{I}_B^2) - (\alpha(B))^2$. This variance is non-negative, and consequently for all $\mathcal{Q}$ we have

$$\liminf_{B\to\infty} \frac{\log \mathrm{E}^{(\mathcal{Q})}(L^2\mathcal{I}_B^2)}{\log \mathrm{E}^{(\mathcal{Q})}(L\mathcal{I}_B)} \geq 2.$$

Following the terminology of Chang *et al.* [28], Sadowsky [160], we call a procedure asymptotically optimal if for some measure $\mathcal{Q}$ this lower bound is attained. By Theorem 3.1, we are left to prove

$$\limsup_{B\to\infty} \frac{1}{B} \log \mathrm{E}^{(\mathcal{Q})}(L^2\mathcal{I}_B^2) \leq -2\theta^*.$$

But, for the alternative measure proposed in [28] this property holds, which can be seen easily as follows.

We take L to be $x_{I_0}/x_{I_N} \exp[-\theta A(\mathcal{T}_B) + \theta C(\theta)\mathcal{T}_B]$, which is allowed based on Lemma 3.2. Then the upper bound of the first moment of $L\mathcal{I}_B$ (under $\mathcal{Q}$) carries over to the second moment: $L^2\mathcal{I}_B^2 \leq Ce^{-2\theta^* B}$ a.s. for some positive constant C. This proves the optimality.

4 A simulation example

So far we concentrated on the probability of overflow during a busy period. However, in practice one is more interested in the fraction of arriving cells (fluid) that is lost (in a finite-buffer setting). It can be shown that this performance measure can be estimated similarly, by turning on and off the importance sampling. This procedure is originally due to Goyal, Shahabuddin, Heidelberger, Nicola, and Glynn [78].

As an example we consider a two-node tandem model fed by two groups of on-off sources, cf. Kosten [110]. The first (second) group consists of 25 (50) sources with

exponential on-time with mean 2/3 (4/3), exponential off-time with mean 5/2 (5/3) and on rate 2 (1). The mean input rate is 32.75, $C_1 = 43$ and $C_2 = 36$. The buffer size of the first queue is large (100). For different values of the buffer size B of the second queue, we estimate the fluid loss fraction $\zeta(B)$, with 95% confidence and 15% relative efficiency. The value of θ^* is 0.1716; the simulations are performed on a 486 personal computer, 433 DXS.

We conclude from table 1 that the number of busy periods of queue 2 needed is more or less constant in B, whereas the required simulation time grows approximately linearly. The loss fraction is asymptotically exponential, as shown by the results: $\hat{\zeta}(B)\exp[\theta^* B]$ tends to a constant, say η. Since $-\log\eta$ is rather large, $-(1/B)\log\hat{\zeta}(B)$ converges slowly to θ^*. For $B = 5, 10, 15, 20$ we also obtained direct estimates, but the required time as well as the number of busy periods exponentially grow in B. For $B = 20$, already a simulation time of half an hour was involved. Extrapolating this rate of growth, a direct estimate of $\zeta(50)$ would take about 80 hours.

Table 1: Quick simulation results

B	$\hat{\zeta}(B)$	$-(1/B)\log\hat{\zeta}(B)$	$\hat{\zeta}(B)\exp[\theta^* B]$	time	busy periods
5	$8.39 \cdot 10^{-3}$	0.956	$1.98 \cdot 10^{-2}$	1:03	$9 \cdot 10^2$
10	$2.66 \cdot 10^{-3}$	0.593	$1.48 \cdot 10^{-2}$	1:50	$9 \cdot 10^2$
15	$1.17 \cdot 10^{-3}$	0.450	$1.54 \cdot 10^{-2}$	2:20	$10 \cdot 10^2$
20	$4.91 \cdot 10^{-4}$	0.381	$1.52 \cdot 10^{-2}$	2:41	$9 \cdot 10^2$
25	$2.25 \cdot 10^{-4}$	0.336	$1.64 \cdot 10^{-2}$	3:02	$9 \cdot 10^2$
30	$8.91 \cdot 10^{-5}$	0.311	$1.54 \cdot 10^{-2}$	3:20	$10 \cdot 10^2$
35	$3.88 \cdot 10^{-5}$	0.290	$1.57 \cdot 10^{-2}$	3:54	$10 \cdot 10^2$
40	$1.62 \cdot 10^{-5}$	0.276	$1.55 \cdot 10^{-2}$	4:35	$11 \cdot 10^2$
45	$6.43 \cdot 10^{-6}$	0.266	$1.45 \cdot 10^{-2}$	5:15	$11 \cdot 10^2$
50	$3.01 \cdot 10^{-6}$	0.254	$1.60 \cdot 10^{-2}$	5:57	$10 \cdot 10^2$

5 Conclusions and further research

We have discussed the analysis of loss probability $\alpha(B)$ in a tandem communication network. It is explained how to find the exponential decay rate θ^* of $\alpha(B)$. Consequently $\alpha(B)$ is equal to $\eta(B)\exp[-\theta^* B]$, for some unknown function $\eta(\cdot)$ with $\log\eta(B) = o(B)$. The approximation $\exp[-\theta^* B]$ behaves very poorly. Therefore, we have to resort to simulation to find a numerical value for $\alpha(B)$. Chang *et al.* [28] developed an importance sampling technique: simulate the model under another measure than the original one. They found a change of measure providing a substantial speed up. We gave a proof of optimality of this simulation procedure. The simulation results show that this quick simulation is a very powerful technique that enables us to describe the performance of a communication system as a function of the switching and transmission resources.

In spite of all theoretical results derived in this chapter, there is still a need for more practically applicable methods to dimension the buffers and capacities in the network. Also, in this chapter we only considered the loss constraint, while in most practical situations there is a delay constraint as well. In Mandjes and van den Berg [127], which can be regarded as complementary to this study, these problems are tackled.

An interesting extension of this study would be a rigorous proof of Chang's conjecture for intree networks with fluid input. Consider for instance an intree network with depth 2: the output of a number (> 1) of first phase queues feeds into a second phase queue. However, at the beginning of a busy period of the second phase queue, *the first phase queues need not to be empty,* in contrast to the tandem model. This property makes the analysis much more complicated. Insight is required into the steady state buffer contents of the first phase queues at the start of a busy period of the second phase queue, in order to prove Chang's conjectures.

Another interesting subject of future research is the possible extension to more general networks, such as the following: Let two sources feed into a queue. After being served, the source 1 cells leave the system, whereas the source 2 cells are led into a second queue. In order to analyze this model, one must be able to characterize the individual traffic streams leaving from the first queue, instead of the aggregate traffic stream. De Veciana *et al.* [46] derived conditions under which the effective bandwidths of the individual input and output processes coincide, but no general expression for the effective bandwidth function has been found so far.

6 Appendix

In this appendix Lemma 3.4 is proven: $\limsup_{B\to\infty} \mathrm{E}^{(\mathcal{Q})}\mathcal{T}_B/B \leq (\nu - C_2)^{-1}$.

PROOF. First suppose $C_1 > \nu$. Consider the set of epochs

$$V := \{t : Q_1(t) = 0, Q_1(t+) > 0, X(t) = i^*\}.$$

The departure (and arrival!) process of the first queue regenerate at epochs in V. Define

$$\mathcal{T}'_B := \inf\{t \in V \mid D(t) - C_2 t \geq B\}.$$

Clearly, $\mathcal{T}_B \leq \mathcal{T}'_B$ a.s., so it suffices to show the stated with $\mathcal{T}_B$ replaced by $\mathcal{T}'_B$. Let N_B denote the number of regenerations up to $\mathcal{T}'_B$, and A_i (D_i) the fluid arrived at (left from) the first queue in the ith regeneration cycle and τ_i its duration; write A, D, and τ for their generic variables. Clearly, $A_i = D_i$, since the first queue is empty at epochs in V.

Suppose $\mathrm{E}^{(\mathcal{Q})}\tau$ is finite, implying that $\mathrm{E}^{(\mathcal{Q})}A \leq r_p \mathrm{E}^{(\mathcal{Q})}\tau < \infty$. Applying 'Wald', renewal

reward, and Lemma 2.1,

$$\frac{\mathrm{E}^{(\mathcal{Q})}(D(\mathcal{T}'_B) - C_2\mathcal{T}'_B)}{\mathrm{E}^{(\mathcal{Q})}\mathcal{T}'_B} = \frac{\mathrm{E}^{(\mathcal{Q})}N_B \cdot \mathrm{E}^{(\mathcal{Q})}(D - C_2\tau)}{\mathrm{E}^{(\mathcal{Q})}N_B \cdot \mathrm{E}^{(\mathcal{Q})}\tau} = \frac{\mathrm{E}^{(\mathcal{Q})}(A - C_2\tau)}{\mathrm{E}^{(\mathcal{Q})}\tau} \overset{\text{a.s.}}{=} \lim_{t\to\infty}\frac{A(t)}{t} - C_2,$$

which is equal to $\nu - C_2$ a.s. Consider the random walk with positive drift $(\sum_{i=1}^{n}(D_i - C_2\tau_i))_n$. From the definition of $\mathcal{T}'_B$, $\mathrm{E}^{(\mathcal{Q})}(D(\mathcal{T}'_B) - C_2\mathcal{T}'_B) = B + \mathrm{E}^{(\mathcal{Q})}_B(W)$, where $\mathrm{E}^{(\mathcal{Q})}_B(W)$ is the expected overshoot of the random walk over level B. Now suppose the increments of the random walk have a finite second moment, then as a result Theorem 10.5 of [82] is applicable, stating that $\mathrm{E}^{(\mathcal{Q})}_B(W)$ tends to a constant as $B \to \infty$. We get the desired.

Above we applied 'Wald' and renewal reward. However, to use these results we need to prove $\mathrm{E}^{(\mathcal{Q})}\tau < \infty$ and $\mathrm{E}^{(\mathcal{Q})}\tau^2 < \infty$. This is done as follows. Consider the set $V' := \{t : Q_1(t) = 0, Q_1(t+) > 0\}$. These epochs are no regenerations, since the state of the modulating Markov chain may vary. Therefore, the periods between two consecutive epochs in V' are called a-cycles [28] rather than cycles. Let N_j be the number of a-cycles that start off when the modulating chain is in state j $(j = 1, \ldots, d)$ during a regeneration cycle; τ_j is the length of such an a-cycle. Notice that $\{N_j = n\}$ does not depend on the τ_j's during the next regeneration cycle $\{\tau_{j,n+1}, \tau_{j,n+2}, \ldots\}$. Therefore, by 'Wald':

$$\mathrm{E}^{(\mathcal{Q})}\tau = \mathrm{E}^{(\mathcal{Q})}\left(\sum_{j=1}^{d}\sum_{i=1}^{N_j}\tau_{j,i}\right) = \sum_{j=1}^{d}\mathrm{E}^{(\mathcal{Q})}\left(\sum_{i=1}^{N_j}\tau_{j,i}\right) = \sum_{j=1}^{d}\mathrm{E}^{(\mathcal{Q})}N_j\mathrm{E}^{(\mathcal{Q})}\tau_j.$$

The state of the modulating chain at the start of the jth a-cycle is called Y_j. One sees easily that $(Y_j)_j$ is a (discrete-time) irreducible Markov chain. For all j, $\mathrm{E}^{(\mathcal{Q})}N_j$ is smaller than the mean number of a-cycles during a regeneration cycle, i.e. the mean passage time (of $(Y_j)_j$) of state i^*, starting in i^*, which is finite according to Theorem 4.4.2 in [102].

So we are left to prove that the duration of an a-cycle has (under $\mathcal{Q}$) a finite mean: for all j, $\mathrm{E}^{(\mathcal{Q})}\tau_j < \infty$. This can be done as follows. We decompose τ_j into a part in which $Q_1(\cdot)$ is positive (the busy period, $\tau_j^{(B)}$) and a part in which $Q_1(\cdot)$ is zero (the idle period $\tau_j^{(I)}$). Denoting by $\lceil a \rceil$ the smallest integer larger than or equal to a, we have by Lemma 2.1, uniformly in j,

$$\mathrm{E}^{(\mathcal{Q})}\tau_j^{(B)} \le \mathrm{E}^{(\mathcal{Q})}\lceil\tau_j^{(B)}\rceil = \sum_{n=1}^{\infty} n\mathcal{Q}\left(\lceil\tau_j^{(B)}\rceil = n\right) \le \sum_{n=1}^{\infty} n\mathcal{Q}\left(\bigcap_{0\le t\le n-1}\{Q_1(t) > 0\}\right)$$

$$= \sum_{n=1}^{\infty} n\mathcal{Q}\left(\bigcap_{0\le t\le n-1}\{A(t) > C_1 t\}\right) \le \sum_{n=1}^{\infty} n\mathcal{Q}\left(\frac{A(n-1)}{n-1} > C_1\right) \le \sum_{n=1}^{\infty} nH(C_1)e^{-J(C_1)(n-1)},$$

being finite. Also, $\mathrm{E}^{(\mathcal{Q})}\tau_j^{(I)} \le u$, as defined in the proof of Lemma 3.3. We conclude that $\mathrm{E}^{(\mathcal{Q})}\tau < \infty$. In an analogous manner, the finiteness of the second moments (under $\mathcal{Q}$) of τ and A can be shown.

Now suppose $C_1 = \nu$. Then $\mathrm{E}^{(\mathcal{Q})}\tau = \infty$, so the above argument does not apply. Now define $\mathcal{T}'_{B,\epsilon}$ as $\mathcal{T}'_B$, where the model is changed such that all input rates r_i are multiplied by $1 - \epsilon$. But

then the first queue is stable. With $\mathrm{E}^{(\mathcal{Q})}\mathcal{T}'_B \leq \mathrm{E}^{(\mathcal{Q})}\mathcal{T}'_{B,\epsilon}$ and the above result for a stable first queue, we get for small enough ϵ,

$$\limsup_{B\to\infty} \frac{\mathrm{E}^{(\mathcal{Q})}\mathcal{T}'_B}{B} \leq \limsup_{B\to\infty} \frac{\mathrm{E}^{(\mathcal{Q})}\mathcal{T}'_{B,\epsilon}}{B} \leq \frac{1}{\nu(1-\epsilon) - C_2}.$$

Now let $\epsilon \downarrow 0$. ■

Chapter 6

Buffer and bandwidth allocation in ATM systems

This chapter studies the optimal allocation of bandwidth and buffer space in single and multi-link ATM systems, subject to loss and delay constraints. First, a useful approximate relation between buffer size and required bandwidth is derived for the single-link model with Markov modulated fluid input. This relation is shown to be asymptotically exact, for the case that the buffer size becomes large. Second, buffer and bandwidth allocation for multi-link models, e.g. tandem and intree networks, is studied. For that purpose we use importance sampling simulations. Particularly, it is found from the numerical results that the asymptotic relation for the single-link case is also useful for the determination of the required bandwidth for traffic streams in multi-link systems. In conjunction with some heuristic ideas, we propose manageable and reasonably accurate guidelines for resource allocation. In particular, under certain delay and loss constraints, it is shown how to optimally allocate buffer space and bandwidth in tandem and intree networks.

1 Introduction

In ATM-based multiservice networks heterogeneous traffic streams (voice, data, video) share common network resources (bandwidth, buffer space). Due to the widely varying traffic characteristics and service requirements of the sources, it is a challenging task to perform ATM resource allocation and traffic control. A basic problem arising in ATM network dimensioning and traffic control is to determine the *required bandwidth* for a set of traffic streams offered to a buffered link, i.e., the minimal bandwidth that is needed in order to achieve a predefined service level. Insight into the relation between the characteristics of the offered traffic, buffer capacity, desired service level, and the required

bandwidth is needed to efficiently allocate ATM network resources and to adequately perform call acceptance control.

SINGLE LINKS. Many authors have studied the required bandwidth for a set of traffic streams multiplexed on a single link. Essentially, the approach is mostly the following. Suppose, the link is emptied at a constant rate C. A well-known result is that the overflow probability is of the form $\eta(C)\exp[-\theta(C)B]$ for $B \to \infty$. Since amplitude $\eta(C)$ is difficult to calculate, one commonly approximates it by 1, see e.g. Guérin *et al.* [80]. Usually, it is easy to calculate $\theta(C)$ numerically, and therefore it is not difficult to dimension C, i.e., to find the smallest C such that $\exp[-\theta(C)B]$ is below a predefined level, say ϵ.

A problem is that the approximation $\eta(C) \approx 1$ is in general not very accurate. Suppose for simplicity the case of n identical sources feeding into a buffer emptied at rate nC. Then it can be shown that the approximation $\exp[-\theta'(C)n]\exp[-\theta(C)B]$ is substantially better, where $\theta(C)$ is the same as above, see e.g. Weiss [184], Choudhury *et al.* [34], Botvich and Duffield [18]. Usually, $\theta'(C) > 0$, so the amplitude $\eta(C)$ decreases asymptotically in n; this is the mathematical description of the so-called multiplexing effect.

As said before, $\theta(C)$ and $\theta'(C)$ can be calculated rather easily, but mostly there are no explicit formulas available; usually, they are given as complicated variational problems, coming from large deviations expressions [184], [18]. Therefore, it would be useful to have formulas with a clearer physical interpretation. Apart from that, it would be very interesting to know under what circumstances the extremely simple formula $\exp[-\theta(C)B]$ is reasonably accurate.

MULTI-LINKS. Only few papers are dedicated to assigning bandwidths in multi-link systems, e.g. tandem links. An important question here is how the required bandwidth changes when a traffic stream passes several links, i.e., what is the smoothing effect of link buffering on the offered traffic? Chang *et al.* [28] and De Veciana *et al.* [46] give some first answers on this question, but these are not very explicit. Another important issue is the optimal allocation of buffers to the different links in order to minimize the total required bandwidth.

The contribution of this chapter is twofold. (i) We derive for the *single link* case large buffer asymptotics of the bandwidth required to achieve a predefined loss probability. The asymptotic formula depends on the input traffic characteristics only through the mean and variance. Apart from that, we mathematically show when the approximation $\eta(C) \approx 1$ is justified. (ii) We consider *multi-link networks*, e.g., *tandem and intree networks*. In order to study these models and to validate our theoretical insights, we have developed a simulation program which estimates small cell loss ratios quickly by using importance sampling with the 'optimal change of measure' [41], [144]. We use the methods described

in [28], [120], [125] and Chapter 5 of this monograph. Using this tool a large number of scenarios can be checked in a reasonable amount of time: a simulation takes typically a few minutes (instead of hours or days). Simulation results show that asymptotics similar to the one deduced for the single link can be used. Together with some new heuristic ideas and the approximations of Guérin *et al.* [80], we propose manageable and reasonably accurate guidelines for resource allocation. In particular it is shown how to optimally allocate buffer space and bandwidth, subject to loss and delay constraints.

This chapter is organized as follows. Section 2 deals with derivation, corollaries, implications, and examples of the asymptotic formula for the required bandwidth of a set of traffic streams offered to a single link. In Section 3, required bandwidths in multi-link systems (tandem and intree networks) are studied. In particular, we study the optimal allocation of bandwidth and buffer space to the links. theoretical results for resource allocation in ATM based networks. Finally, a summary of the results and the conclusions are given in Section 4, together with some topics identified for further research.

2 Single link model, theoretical aspects

In our analysis, it is assumed that the traffic sources generate Markov modulated fluid. This means that each source is characterized by the infinitesimal generator Λ of an irreducible continuous-time Markov chain (of dimension d) and a traffic rate vector r: if the modulating chain is in state i, cells are generated at rate r_i. Arriving traffic is put into a buffer which is emptied at constant rate C. Without loss of generality we can assume one source, since individual sources can be merged to one aggregate Markov fluid source [61].

In this section, assuming an infinite buffer, we will focus on determining the minimal required service rate C_B such that the probability (i.e., the long-run fraction of time) that the buffer occupancy exceeds some level B, P_B, is below ϵ. As said in the introduction, P_B is of the form $\eta(C)\exp[-\theta(C)B]$, where $\theta(C)$ can be calculated easily, but $\eta(C)$ is difficult to capture, cf. Kosten [111]. Approximating $\eta(C)$ by 1, C can be dimensioned by $\theta^{-1}(-\log\epsilon/B)$. However, $\eta(C)$ can be considerably smaller than 1, implying that this approximation behaves poorly. In this section it is shown that this is justified in the case of a high load; then $\eta(C) \approx 1$. In addition, we derive a insightful asymptotic ($B \to \infty$) formula for C_B. In literature, the function $\theta^{-1}(\cdot)$ is called the *effective bandwidth*, in the sequel denoted by $C(\cdot)$. $C(\theta)$ can be interpreted as the service rate that yields a overflow probability with decay rate θ.

2.1 Analysis

It is clear that the required service rate C_B decreases to the mean input rate m as the buffer size B tends to ∞. In this section, we will show that C_B decays asymptotically hyperbolically to m:

$$\lim_{B\to\infty}(C_B - m)B \to -\log\epsilon\left(\frac{v}{2}\right), \tag{6.1}$$

where v is the *asymptotic variance* of the arrival process defined as follows. $A(t)$ denoting the amount of fluid arrived up to time $t \geq 0$, then

$$v := \lim_{t\to\infty}\frac{1}{t}\left(\mathrm{E}A(t)^2 - (\mathrm{E}A(t))^2\right). \tag{6.2}$$

The proof of this result runs as follows. First we summarize basic theory concerning the asymptotic behavior of the level exceedance probability P_B for large B. As a second step, we derive an asymptotic expression for the required level $B = B_C$ to achieve $P_B = \epsilon$, when $C \downarrow m$, which is in some sense the dual problem of finding C_B. Finally, from the result for B_C we come to the desired result (6.1).

Let us recall the analysis of the probabilistic behavior of a fluid flow queueing system, see also [110], [111], [60], [61]. The buffer occupancy distribution can be captured by solving an eigensystem, that arises from the governing differential equations. In this way, the problem is reduced to finding all eigenvalue/eigenvector pairs (θ, x) satisfying $-\theta x = x\Lambda(R - CI)^{-1}$, R denoting $\mathrm{diag}\{r\}$. Clearly, Λ is singular (all rows add up to 0), so there is a zero eigenvalue, say θ_1. Notice that a corresponding left eigenvector is the invariant π of Λ: $\pi\Lambda = 0$. If C is larger than the mean input rate $m = \sum_{i=1}^{d}\pi_i r_i$, it can be shown that there exists at least one positive eigenvalue. Let θ_2 be the smallest among them. The set of eigenvalues with a positive (negative) real part are numbered θ_2 up to θ_k ($\theta_{k+1}, \ldots, \theta_d$, respectively). Corresponding left eigenvectors are $x_1 = \pi, x_2, \ldots, x_d$. P_B is now given by the spectral expansion $1 - \sum_{i=1}^{d} a_i\left(\sum_{j=1}^{d} x_{ij}\right)\exp[-\theta_i B]$. Because of the boundedness of P_B, $a_{k+1} = \cdots = a_d = 0$. Furthermore, it can be shown that $a_1 = 1$. The other a_i follow from the boundary condition that says that, if $r_i > C$, the probability of an empty buffer and being in state i is zero. This leads to $\sum_{j=1}^{k} a_j x_{ji} = 0$ or $\sum_{j=2}^{k} a_j x_{ji} = -\pi_i$. We see that $P_B\exp[\theta_2 B] \to -a_2\sum_{j=1}^{d} x_{2j}$ as $B \to \infty$.

After the above brief recapitulation, we now concentrate on the asymptotic behavior of C_B. It should be realized that P_B, a_i, x_i, and θ_i ($i = 2, \ldots, d$) in the above setting depend on C. We therefore write $P_B(C), a_i(C), x_i(C)$, and $\theta_i(C)$. We first consider the following limit

$$\lim_{C\downarrow m}\left(\frac{B_C}{C} - \frac{B_C}{m}\right).$$

Multiplying both sides in $\epsilon = -\sum_{i=2}^{d} a_i \left(\sum_{j=1}^{d} x_{ij}\right) \exp[-\theta_i B]$ by $\exp[\theta_2 B]$ yields

$$B_C = \frac{1}{\theta_2(C)} \log \left(\frac{-\sum_{i=2}^{d} a_i(C) \left(\sum_{j=1}^{d} x_{ij}(C)\right) e^{-\theta_i(C) B_C} e^{\theta_2(C) B_C}}{\epsilon} \right).$$

Now we show that $\theta_2(C) \downarrow 0$ as $C \downarrow m$. It is known that $\theta_2(C)$ is the unique positive solution to $C(\theta) = C$, see Elwalid and Mitra [61]. They also showed that the effective bandwidth function $C(\cdot)$ attains m in 0, and increases. Letting $C \downarrow m$, we consequently have $\theta_2(C) \downarrow 0$. For reasons of continuity, $x_2(C) \to \pi$. Also, for $i = 3, \ldots, d$, $\theta_i(C) \to \theta_i(m)$ and $x_i(C) \to x_i(m)$. From the set of linear equations, it follows that $a_2(C) \to -1$, whereas $a_i(C) \to 0$, for $i = 3, \ldots, d$. Noting for ease $\theta(C) := \theta_2(C)$, we get

$$\lim_{C \downarrow m} \left(\frac{B_C}{C} - \frac{B_C}{m} \right) = (\log \epsilon) \lim_{C \downarrow m} \left(\frac{1}{m\theta(C)} - \frac{1}{C\theta(C)} \right) = (\log \epsilon) \lim_{C \downarrow m} \left(\frac{C - m}{mC\theta(C)} \right). \tag{6.3}$$

The right limit in (6.3) ('0 over 0') can be evaluated as follows. We recall that $C(\cdot)$ is the inverse of $\theta(\cdot)$. The effective bandwidth function $C(\cdot)$ of a Markov fluid source is an analytical, increasing function defined on $[0, \infty)$, with $C(0) = m$. Therefore, $\theta(\cdot)$ is increasing on $[m, \infty)$ and $\theta(m) = 0$. We write $\theta(\cdot)$ as a power series in a neighborhood of m: $\theta(C) = \theta(m) + (C-m)\theta'(m) + O((C-m)^2)$. We get that the limit under consideration equals $(\log \epsilon)(m^2\theta'(m))^{-1} = (\log \epsilon) m^{-2} C'(0)$. But, according to Kesidis *et al.* [105], the effective bandwidth function can alternatively be written as

$$C(\theta) = \frac{1}{\theta} \lim_{t \to \infty} \frac{1}{t} \log \mathrm{E} \exp(\theta A(t)).$$

Inserting the Taylor expansions of the log and exp provides us the first order approximation of $C(\cdot)$ in a neighborhood of 0: $C(\theta) = m + \theta v/2 + O(\theta^2)$, where v denotes the asymptotic variance of the arrival process as given in (6.2). Therefore, $C'(0) = v/2$. We conclude

$$\lim_{C \downarrow m} \left(\frac{B_C}{C} - \frac{B_C}{m} \right) = (\log \epsilon) \frac{v}{2m^2} \tag{6.4}$$

Now it is a matter of algebra to derive the behavior of C_B from the behavior of B_C:

$$\lim_{B \to \infty} (C_B - m) B = \lim_{B \to \infty} m^2 \left(\frac{B}{m} - \frac{B}{C_B} \right) = -\log \epsilon \left(\frac{v}{2} \right). \tag{6.5}$$

as desired. Analogously, $B_C \approx (-\log \epsilon) v / (2C - 2m)$ as $C \downarrow m$.

We saw that under heavy traffic $a_2(C) \to -1$ and $x_2(C) \to \pi$, implying that $\eta(C)$ in the asymptotical formula $\eta(C) \exp[-\theta(C) B]$ tends to 1, so that the Guérin *et al.* [80] results become quite accurate as $C \downarrow m$. But in that region, we found that $C(\theta)$ can be approximated by the – simpler – formula $m + \theta v/2$.

2.2 Implications of the asymptotic result

This subsection deals with some reflections on the simple asymptotic formulas derived in the previous section. In (6.1) and (6.4) it was found that

$$C_B \approx m - (\log \epsilon)\frac{v}{2B}. \tag{6.6}$$

We notice that we in fact derived an *insensitivity* result: for large buffers the required bandwidth depends on the input traffic only through the mean input rate and the asymptotic variance of the input traffic. Higher moments are irrelevant.

In case of a group of, say, M independent sources, a simple *additivity* property holds (cf. Guérin *et al.* [80]). Due to the additivity of means and variances of independent sources, the required bandwidth equals the sum of the bandwidths required by the individual sources. More concretely, if the sources have means m_i and variances v_i, C_B can be found using (6.6) with $m := \sum_{i=1}^{M} m_i$ and $v := \sum_{i=1}^{M} v_i$.

The right hand side of (6.6) has a very intuitive structure: C_B equals the mean input rate plus an additional amount to cope with the variability of the input process. This additional amount is, of course, decreasing in both ϵ and B. In most studies, the input consists of multiple (not necessarily identical) on-off sources, with on (transmitting at peak rate r) and off times that are exponentially distributed (with mean λ^{-1} and μ^{-1}, respectively). However, the Markov fluid framework allows us to choose any phase-type distribution (Coxian, hyperexponential, Erlang, deterministic, etc.). However, for most of these distributions the effective bandwidth $C(\cdot)$ of a single source cannot be given explicitly. Now notice that our dimensioning approach only requires the mean input rate m and the derivative of the bandwidth function at zero $C'(0) = v/2$. The latter can be calculated from the implicit relations between θ and $C(\theta)$, as found in Kosten [111]. For instance, in case of an Erlang(n) on-time (mean λ^{-1}) and an exponential off-time (mean μ^{-1}), we have

$$(\lambda n + \theta C(\theta))^n(\mu + \theta(C(\theta) - r)) - (\lambda n)^n \mu = 0 \quad \Rightarrow \quad v \equiv 2C'(0) = \frac{\lambda\mu r^2}{(\lambda+\mu)^3}\frac{n+1}{n},$$

after implicit differentiation. Notice that, as the number of phases of the Erlang distribution grows large, the asymptotical variance v decreases, agreeing with the fact that the arrival process loses burstiness. As $n \to \infty$, the on-times become deterministic; we get $v = (\lambda\mu r^2)/(\lambda+\mu)^3$.

Notice that in case of identical on-off sources with exponential on and off times $C(\theta)$ can be given explicitly, so there is no need for the above approximation. If, in addition, the burst lengths are much smaller than the off periods, $v/2m$ is approximately equal to the mean burst size r/λ. We get the dimensioning rule, as in Guérin *et al.* [80],

$$C_B \approx \frac{Bm}{B - (\log \epsilon)r/\lambda}.$$

2.3 Examples

EXAMPLE 2.1. We now treat a numerical illustration of the properties derived above. Suppose two types of traffic. The first (second) group of sources consists of 4 exponential on-off sources, with $\lambda^{-1} = 1$ (0.25), $\mu^{-1} = 4$ (4.75), and $r = 1$ (4). We let the service rate decrease to the mean input rate $m = 1.6$. Each group can be captured by a 5-dimensional Markov fluid source; the superposition of them yields a 25-dimensional source. We computed the full spectral expansion of the buffer level exceedance probability $P_B(C)$. In Table 1 we list the main results. Let i denote the index of the smallest positive eigenvalue larger than $\theta_2(C)$. This eigenvalue exists, because it can be shown that in this case all eigenvalues are real, and, for the values of C of interest, 23 of them are positive). Also, j denotes the argument of $\max_{j=3,\ldots,25} |a_j(C) \sum_{k=1}^{25} x_{jk}(C)|$.

Table 1: Spectral expansion

C	$\theta_2(C)$	$a_2(C) \sum_k x_{2k}(C)$	$\theta_i(C)$	$\lvert a_j(C) \sum_k x_{jk}(C) \rvert$
1.800	0.1457	-0.9014	0.7022	0.0197
1.700	0.0767	-0.9501	0.6862	0.0090
1.650	0.0394	-0.9749	0.6782	0.0043
1.620	0.0160	-0.9900	0.6734	0.0017
1.610	0.0081	-0.9950	0.6718	0.0008
1.605	0.0041	-0.9975	0.6710	0.0004
1.600	0.0000	-1.0000	0.6702	0.0000

The first observation from the table is that $\theta(\cdot)$ is more or less linear for C in the neighborhood of m. This is justified by

$$\theta(C) = \theta(m) + (C - m)\theta'(m) + O((C - m)^2) \approx \frac{2}{v}(C - m). \tag{6.7}$$

In this case, $2/v \approx 0.8110$. Furthermore, it can be checked that, using the approximation $B_C \approx (-\log \epsilon)v/(2C - 2m)$, quite accurate results are obtained (compared to exact dimensioning using the full spectral representation). Take for instance $\epsilon = 10^{-4}$ and $C = 1.7$. Our asymptotic result gives (as approximation for B_C) 120.7. The spectral expansion says that P_B equals $0.9501 \exp[-0.0767B]$ plus 23 terms, whose sum is majorized by $23 \cdot 0.0090 \exp[-0.6862B]$, leading to $B_C = 119.4$.

EXAMPLE 2.2. The fraction of time that the buffer is full in the *finite* buffer system is approximately exponential in the buffer size as well: $\zeta(C, B) \exp[-\theta(C)B]$. Notice that the amplitude ζ depends on B as well. Unfortunately, $\zeta(\cdot, B)$ does not converge to 1 as $C \downarrow m$. Consequently, the asymptotic result (6.1) of Subsection 2.1 does not apply to the fraction of time that the system is full (or to the cell loss ratio).

Consider a queue fed by the input process of the first example. Table 2 contains, for several values of B, results for the required capacity C_B to keep P_B (in the infinite-buffer model) below $\epsilon = 10^{-4}$, its approximation (6.6), and $K_B := B/m - B/C_B$, cf. (6.4). The results are obtained by fast simulation. We also list, for the finite-buffer case, the capacity yielding a loss fraction ϵ, and the corresponding K_B-value.

First we consider the infinite buffer case for which the theoretical results presented above have been proved. It is seen that the approximation results $C_B^{(\mathrm{app})}$ indeed approach the simulation results $C_B^{(\mathrm{sim})}$, for B large. It appears that the asymptotic formula, which has been derived under heavy traffic conditions, seems to be useful in situations with lower (i.e., practically relevant) loads as well. We see that $K_B^{(\mathrm{sim})}$ indeed approaches the limiting value.

Table 2: Required service rates

	infinite buffer			finite buffer	
B	$C_B^{(\mathrm{sim})}$	$C_B^{(\mathrm{app})}$	$K_B^{(\mathrm{sim})}$	$C_B^{(\mathrm{sim})}$	$K_B^{(\mathrm{sim})}$
5	5.04	3.87	2.14	4.98	2.12
10	3.31	2.74	3.22	3.12	3.04
15	2.63	2.35	3.67	2.50	3.38
20	2.32	2.17	3.87	2.20	3.41
25	2.15	2.05	4.00	2.05	3.43
30	2.05	1.97	4.11	1.95	3.37
35	1.97	1.92	4.11	1.89	3.36
40	1.925	1.88	4.22	1.84	3.26
45	1.885	1.85	4.25	1.80	3.13
50	1.855	1.83	4.30	1.75	2.68

In the finite-buffer case, $K_B^{(\mathrm{sim})}$ does not converge, but we see that it is more or less constant on a large interval (roughly speaking, $B \in [15, 40]$). This phenomenon, also appearing in the other numerical examples, leads to the following approximation idea. Suppose our goal is to dimension C for several values of B. First fix some B^*. With the quick simulation program we can dimension C_{B^*}. Once C_{B^*} is determined reasonably accurately, also for different B a value for C_B is obtained, using

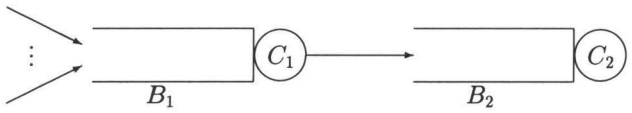
$$\frac{B}{m} - \frac{B}{C_B} = K_B \approx K_{B^*} = \frac{B^*}{m} - \frac{B^*}{C_{B^*}}.$$

Of course, this rule can be used only if B^* and the B to be inserted lie in the interval in which K_B is more or less constant.

3 Network models

In the previous section we have considered the required bandwidth for a set of traffic streams multiplexed on a single link. In this section we focus on the case that the output

Figure 1: Tandem queueing model of two consecutive network links.



stream is offered to another link. The bandwidth required for this output stream is smaller than for the original stream, due to the shaping effect of the first link. Interesting issues studied in this section are the quantification of the shaping effect and the allocation of buffer space to the links in order to minimize the total required bandwidth, subject to delay and loss constraints.

In Subsection 3.1 we examine the most simple extension of the single link, i.e., the two-link tandem model. Subsection 3.2 deals with so-called intree networks, modeling multi-stage multiplexing systems. For the sake of simplicity, in the sequel the required bandwidth of link i (associated with its finite buffer size B_i and loss fraction ϵ) is simply denoted by $\hat{C}_i$.

3.1 Two-link tandem model

In order to study the shaping effect we consider the tandem queueing system of Figure 1. The input of the system consists of M Markov modulated fluid sources. This subsection deals with two problems: (A) For given buffer sizes B_1 and B_2, determine the service rates $\hat{C}_1$ and $\hat{C}_2$, such that the fluid loss ratio equals a predefined value ϵ. (B) Find the buffer sizes B_1 and B_2, such that $\hat{C}_1 + \hat{C}_2$ is minimal, subject to the delay constraint

$$\frac{B_1}{\hat{C}_1} + \frac{B_2}{\hat{C}_2} \leq D_{\max},$$

where the left hand side of this inequality obviously is the largest possible delay.

A. Bandwidth allocation problem

For the determination of $\hat{C}_1$ given B_1, we can use results for the single link, see Section 1 and 2. To our knowledge, no explicit solutions are available for finding the required bandwidth in the second queue. In this chapter we propose a method, that uses (i) a heuristic approach, in conjunction with (ii) fast simulation, and (iii) asymptotics, that are very similar to the results derived in Section 2.

- To estimate the required capacity $\hat{C}_2$ of the second queue we will use the following *heuristic approach.* Consider the second queue in isolation and assume that the input

traffic is offered directly to this queue. Furthermore, take the buffer size equal to the actual buffer size plus the buffer size of the first queue (i.e., equal to $B_1 + B_2$). The required capacity $\hat{C}_2^{(\text{app})}$ of the resulting single queue model is used as an approximation for the required service rate of the second queue in the original tandem model. In this way the multi-link problem is reduced to a single-link problem. The main idea behind the heuristic approach is that traffic offered to the second queue can use the buffer at the first queue as well as the buffer at the second queue. A similar, though more sophisticated, heuristic approach is used in De Koster [44] to study the behavior of an intermediate queue in a model of a multi-stage production line.

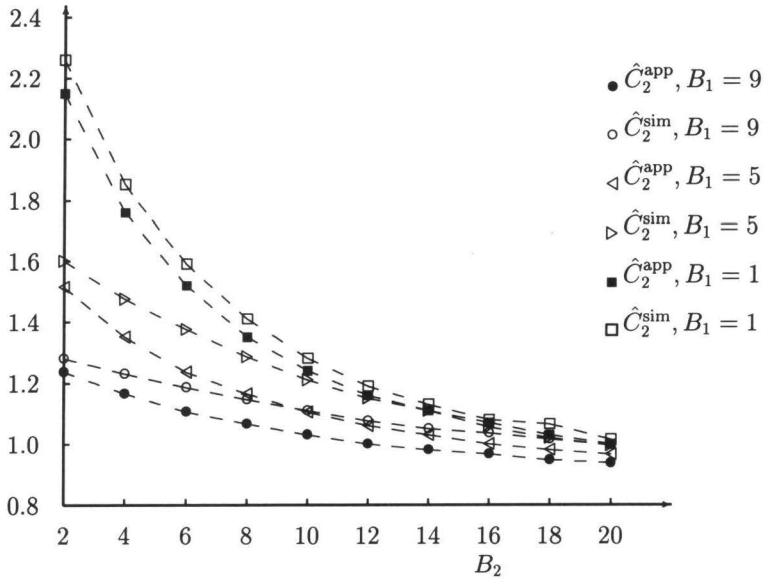
Note that the heuristic approach has some attractive properties. It yields exact results for the cases $B_1 = 0$ or $B_2 = 0$ (and therefore reasonable results for 'small' B_1 and B_2) and the approximation error decreases to zero when B_1 or B_2 becomes large. It should, however, be noted, that the approximation can be used only for the case that the required loss probabilities are equal for both queues.

• The *fast simulation program* introduced in Section 1, enables us to test the above-mentioned heuristic. Here, we consider an example with the 4 identical sources ('first group') used in Example 2.1. The mean total arrival rate m is equal to 0.8. The maximum allowed loss fraction at both queues is $\epsilon = 10^{-4}$. Figure 2 shows approximation and simulation results for the required capacity $\hat{C}_2$ of the second queue, when B_2 ranges from 2 to 20 and B_1 is constant. Results are shown for the cases that B_1 equals 1, 5 and 9. First, it is seen that the approximation certainly reflects the behavior of $\hat{C}_2$ as function of B_2. Furthermore, it appears that in all cases the approximation results $\hat{C}_2^{(\text{app})}$ are smaller than the simulation results $\hat{C}_2^{(\text{sim})}$. As expected, the relative approximation errors are maximal for moderate values of B_1 and B_2. The maximal relative error for the cases shown here is about 11% (about 25% when $\hat{C}_2 - m$ is considered).

Note that for the case $B_1 = 1$ the required bandwidth $\hat{C}_2$ of the second link decreases faster than for the other cases (i.e., $B_1 = 5$ and $B_1 = 9$). Obviously, this is due to the fact that the smoothing effect of the first link on the input traffic stream(s) is smaller when its buffer is small and, hence, the bandwidth required by the output stream is relatively large. An interesting question that arises is the following: is it more effective to buffer at the first queue or to buffer at the second queue, in order to reduce the total required bandwidth? We return to this question in the second part of this subsection.

• An interesting question is whether *asymptotics*, similar to those derived in Section 2, can be used in the case of tandem queues. Aalto [1] showed that the output process of a queue fed by a superposition of a homogeneous exponential on-off sources, is again a Markov fluid process. This suggests to use large buffer asymptotics as in Section 2 for the

Figure 2: Approximation and simulation results for the required bandwidth in the two-link tandem model.



required capacity in the second queue as well. The use of this asymptotic relation can be explained by means of the following example.

Table 3: Asymptotics for second queue

B_2	$\hat{C}_2$	K_{B_2}
4	1.475	2.29
8	1.285	3.77
12	1.150	4.57
16	1.070	5.05
20	0.995	4.90
24	0.961	5.03
28	0.935	5.05
32	0.917	5.10

Table 3 shows results for $\hat{C}_2$ and $K_{B_2} = B_2/m - B_2/\hat{C}_2$, for several values of the buffer size B_2 of the second queue. The input traffic consists of 4 exponential on-off sources, with $\lambda^{-1} = 1$, $\mu^{-1} = 4$, and $r = 1$. Furthermore, the buffer size B_1 of the first queue equals 5 yielding $\hat{C}_1 = 1.765$ (loss probability $\epsilon = 10^{-4}$). It is seen that, similar to the single link case (Example 2.2), K_{B_2} is more or less constant on a large interval of values of B_2. To determine the required bandwidth $\hat{C}_2$ for the second queue for several values of B_2 we can apply the same approach as for the single link case described in the Example 2.2.

B. Optimal buffer allocation problem

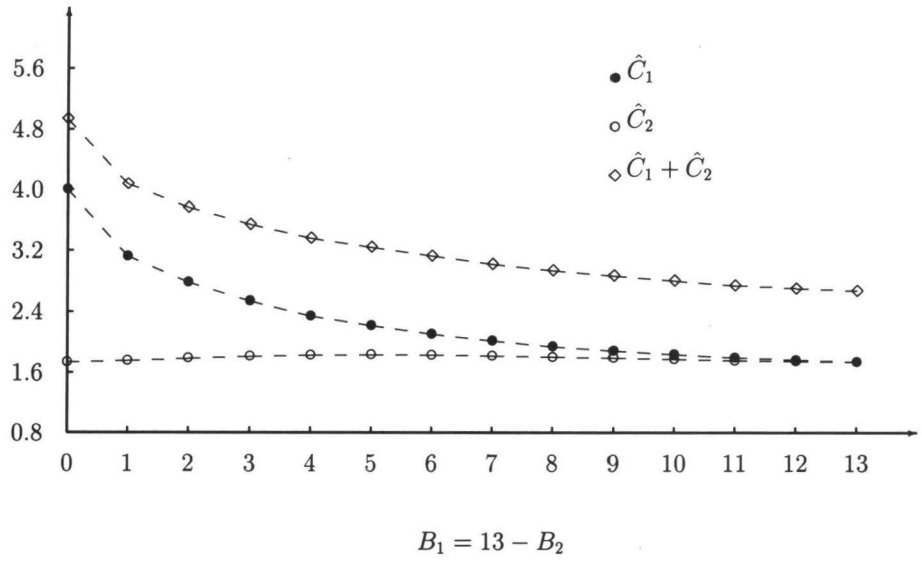
In order to study the optimal buffer allocation problem, we examined the effect of the interchange of buffers between the two links. Using fast simulation, we have determined the required bandwidths $\hat{C}_1$ and $\hat{C}_2$ in our tandem model, for several values of the buffer sizes B_1 and B_2 with $B_1 + B_2$ fixed. Figure 3 shows the results as a function of the buffer size B_1 at the first queue. In the present case B_2 has been chosen such that the total buffer size $B_1 + B_2$ equals 13. The figure also contains the curve $\hat{C}_1^{(\mathrm{sim})} + \hat{C}_2^{(\mathrm{sim})}$.

It is seen from the results that $\hat{C}_1^{(\mathrm{sim})} + \hat{C}_2^{(\mathrm{sim})}$ decreases when B_1 increases (and B_2 decreases). The $\hat{C}_1^{(\mathrm{sim})}$ curve shows a similar behavior. However, $\hat{C}_2^{(\mathrm{sim})}$ remains almost constant when B_1 increases. Note that this behavior of $\hat{C}_2^{(\mathrm{sim})}$ coincides with the idea behind the heuristic approach, which implies that the required capacity of the second queue only depends on the sum of the buffer sizes of the two queues.

From the simulation results in Figure 3 it is found that, in order to minimize $\hat{C}_1 + \hat{C}_2$, all buffering should be done at the first queue. To make this plausible, use the heuristic approach, which states that $\hat{C}_2$ remains constant when $B_1 + B_2$ is constant. Thus, as long as $B_2 > 0$, it is always beneficial to move buffer space from the second to the first queue.

This observation suggests the following dimensioning rule. Take $B_2 = 0$ and take B_1

Figure 3: Simulation results for the required bandwidth in both queues of the two-link tandem queue



such that $B_1/\hat{C}_1 = D_{\max}$, where $D_{\max}$ denotes the maximum allowed delay.

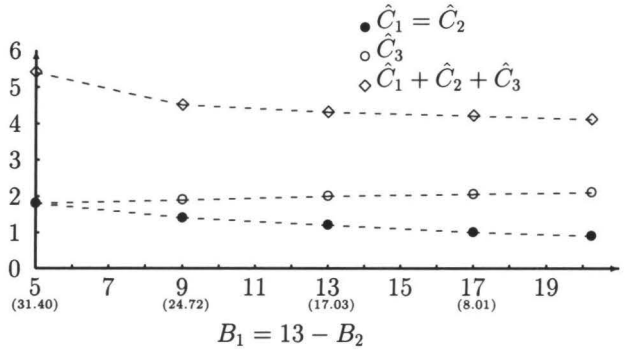
3.2 Simple intree model

The above ideas on bandwidth and buffer allocation in the two-link tandem model can be extended in a straightforward way to the case of intree models. We have done several simulations, which all showed similar results as in the case of a two-link tandem model. For example, in Figure 4 simulation results are shown for a simple intree network (depth 2) consisting of three queues. The input at both 'first-phase queues' (with buffers B_1 and B_2 and capacities C_1 and C_2) is the superposition of the four on-off streams ('first group') of Example 2.1. Their output feeds into the 'root queue' (with buffer B_3 and capacity C_3).

The first question is the determination of $\hat{C}_3$, i.e., the bandwidth dimensioning problem. This can be done by first considering a tandem network consisting of the first of both first-phase queues and the root queue. With the theory of Subsection 3.1, we can determine the capacity required in the root queue of this network, say $\hat{C}_3^{(1)}$. In the same way, we can find $\hat{C}_3^{(2)}$. Now we can approximate $\hat{C}_3$ by $\hat{C}_3^{(1)} + \hat{C}_3^{(2)}$, cf. the ideas of Guérin *et al.* [80]. Fine tuning can be done by using the importance sampling program.

Now consider the optimal buffer allocation problem. The buffer sizes are chosen such that $B_1 = B_2$ and the maximum allowed total delay $D_{\max} = B_1/\hat{C}_1 + B_3/\hat{C}_3 = 20$. The required loss rate is $\epsilon = 10^{-4}$. Indeed, it is seen from the results that, subject to these constraints, the total required capacity $\hat{C}_1 + \hat{C}_2 + \hat{C}_3$ is minimal when buffering is performed only in the first-phase queues.

Figure 4: Simulation results for the required bandwidth in the intree model. The corresponding values of B_3 are given between parentheses. $B_1 = B_2 = 20.23$ corresponds with $B_3 = 0$.



4 Overall conclusions and subjects for further research

In this chapter we have considered some issues concerning bandwidth and buffer allocation in ATM based multiservice networks. In particular we have studied the important problem of determining the required bandwidth of several (heterogeneous) traffic streams offered to a buffered link, i.e., the minimal bandwidth that is needed to achieve a predefined service level. In order to study the shaping effect, we have also investigated the required bandwidth of the output stream. An important tool in our study is a program for fast simulation based on importance sampling.

For the single link case we have derived a simple, attractive, asymptotic formula for the required bandwidth of a set of Markov fluid traffic sources. This formula shows that the bandwidth required to ensure that the probability of exceeding level B is below a predefined value ϵ, depends on the traffic characteristics only through the mean and variance, when B becomes large. Based on comparison with simulation results, it seems that the asymptotic formula is also useful for smaller values of B. Simulation results showed that the behavior of the required bandwitdh (w.r.t. a certain loss fraction) in the finite-buffer case can be described by a similar formula. However, we found that when B becomes very large it differs essentially from that in the infinite-buffer case.

In order to get insight into the way the required bandwidth of a traffic stream changes when passing several network links, we have studied a two-link tandem queueing model and an intree network model. For the tandem model we have tested a simple heuristic approach to determine the required bandwidth of the output traffic of the first queue. Simulation results showed that this approximation qualitatively reflects the behavior of the required bandwidth when the parameter values (e.g., buffer sizes, traffic characteristics) are varied, but in some cases it considerably underestimates the required bandwidth. The heuristic approach gives rise to guidelines for optimal allocation of bandwidth and buffer space to the links. In particular, extensive simulation shows that the total required bandwidth for the two-link tandem model is minimal (subject to loss and delay constraints) when buffering is done only at the first link. Empirically, it appears that the results for the two-link tandem model can be extended to intree networks. In particular, simulation results indicate that, in order to minimize the total required bandwidth, buffering should be done as much as possible (i.e., such that delay constraints are not violated) at the entrance to the network.

An interesting subject for future research is the formalization of our heuristic arguments for the solution of the optimal buffer allocation problem. One possible approach

would be to find a manageable characterization of the output stream of the first link in the tandem system, cf. Park and Perros [145]. To extend our results to more general (than intree) networks, we have to gain insight into the behavior of the individual streams flowing through the network.

Chapter 7

Call blocking in ATM networks

This chapter is concerned with the determination of blocking probabilities in loss networks. In fact our study consists of two parts. Primarily, we scale both arrival rates and link capacities, in order to derive rough asymptotical expressions. These expressions arise as the result of mathematical programming problems. Secondly, we develop a fast simulation technique to estimate the blocking probabilities. This technique is based on importance sampling, where the choice of the alternative probability model is closely related to the optimizing arguments of the above-mentioned mathematical programming problem. Some examples show that huge gain of simulation effort can be achieved.

1 Introduction

Loss networks can be used to describe various types of communication systems. For that reason, the performance evaluation of loss networks has become an important issue in applied probability.

The analysis goes back to Erlang in the beginning of this century. He considered a single-link telephone system consisting of a fixed number of circuits (or trunks). Requests for calls arrive according to a Poisson process, and use one circuit per connection. The duration of calls is stochastic. If all circuits are occupied a new request is blocked. Erlang found an elegant expression for the probability of blocking.

More recently, a number of extensions of this model have been examined. First, we can consider networks consisting of multiple links rather than one. These links consist of a fixed number of circuits, the so-called link capacity. Also, there are multiple 'customer classes'. A customer of any particular class uses a given number of circuits on any link in the network. Blocking occurs if the number of free circuits does not suffice to connect a newly arriving call.

The above formulation suggests that loss network theory applies only to conventional

circuit-switched (telephone) systems, in which a circuit is dedicated exclusively to a user during his entire call. However, loss networks can also be used to model broadband telecommunication systems in which common resources are shared by multiple connections, like in broadband ISDN. This can be explained as follows.

The transmission technology that underlies broadband ISDN is the asynchronous transfer mode (ATM). Loosely speaking, an ATM switch is fed by a number of traffic sources, while a constant number of information packets (the so-called channel capacity) can be processed per unit time. If the sources generate constant bit rate traffic streams, the model fits in the framework of loss models described above. However, in general traffic stream are of variable bit rate type: the bit rate fluctuates in time. To overcome this problem the concept of 'equivalent bandwidths' (or 'effective bandwidths') is introduced: it assigns to each traffic stream a minimum amount of the channel capacity required to achieve a prespecified grade of service. Due to the additive nature of this equivalent bandwidth concept [74], [80], [89], ATM systems can be translated into loss networks. For an extensive survey in this field, see Ritter and Tran-Gia [152].

It was shown that the steady-state probabilities of the number of customers in the network have a product form, see among others [95], [24]. Consequently, the blocking probabilities can be given explicitly. However, numerical calculation is very time-consuming since a summation over a very large number of states has to be performed. Only in special cases efficient combinatorial techniques are developed to calculate the blocking probabilities or accurate approximations. Kaufman [93] and Roberts [154] propose recursive techniques to solve the case of multiple traffic classes sharing a single link. Ross, Chung, and Tsang [177], [35] assume that the network has a specific structure (star, tree).

In order to approximate the blocking probabilities in networks with a general topology, essentially two kinds of techniques are proposed:

A. Kelly [95] applies a *scaling technique*: he multiplies the arrival rates as well as the link capacities by a factor n. He derives expressions for the blocking probabilities, asymptotically in n. As we will explain in Section 2, this method is rather unsatisfactory in many situations.

Many authors adopt this scaling to obtain asymptotic results. Gazdzicki, Lambadaris, and Mazumdar [72] give a detailed asymptotic analysis of the single link case. Using large deviations techniques Shwartz and Weiss [166, Ch. 12] aim to characterize the transient behavior of (single link) loss models. Mitra [129] determines the asymptotics of a class of tree networks. Among many other studies that apply this scaling, we also mention Zachary [190], Hunt and Kelly [90], and Whittle [187]. A summary of results is given in

Kelly's survey paper [100].

B. Also, several *simulation techniques* have been proposed to estimate blocking probabilities. Harvey and Hills [83] developed an acceptance-rejection method. However, the simulation time required is still considerable, particularly if the network grows large. Ross and Wang [158] propose an importance sampling method. Importance sampling is a technique in which the simulation is performed under an alternative measure, in order to increase the occurrence of rare events. Multiplying each observation by a likelihood ratio, unbiased estimates are obtained. However, in [158] no mathematical motivation of the choice of the alternative probability measure is given. Also, particularly in case of large networks, the simulation effort remains rather large.

As we see, there is still a need for fast and accurate techniques to capture the blocking probabilities in loss networks. The contribution of this chapter is twofold. In the first part of our study we apply Kelly's scaling and derive asymptotics of the blocking probability, also for the cases in which Kelly's analysis provides only unsatisfactory results. These (rough) estimates arise from an optimization problem. The second part deals with quick simulation techniques to find more accurate estimates. The simulation approach is based on importance sampling, where the new probability measure arises from the optimizing arguments of the above-mentioned optimization problem. From our experiments, it turns out that a huge acceleration is achieved. The simulation effort is substantially smaller than by applying the techniques proposed in [158]. We conclude that our study is in fact a link between the scaling approach and the simulation approach.

After having found accurate methods to evaluate the performance of a loss network (i.e., algorithms to find the blocking probabilities), a next step is to develop guidelines for optimal design of the system. Roughly speaking, the routing of the calls and the link capacities have to be determined, in order to meet prespecified service criteria and minimize the associated costs. In this context, the concept of trunk reservation is very important. Key references in this field are Kelly [97], [98], and Chapter 5 and 8 of Ritter and Tran-Gia [152]. These dimensioning and control issues are not in the scope of this chapter.

We organized this chapter as follows. Section 2 gives a detailed model description and focuses on the drawbacks of the existing techniques. Then, asymptotical techniques are treated in Section 3. These are used to develop the simulation method that is presented in Section 4. The chapter is concluded by a number of examples.

2 Model description and some preliminaries

This section gives a short introduction on the most important results concerning loss networks. In particular, we put emphasis on their drawbacks, which motivates the search for fast simulation methods.

2.1 Model description

The multirate Erlang loss model is described in the following way. Let $\mathcal{R}$ be the collection of customer classes, labelled by $r = 1, \ldots, R$. Customers of type $r \in \mathcal{R}$ arrive according to a Poisson process with rate $\lambda_r > 0$. All Poisson streams are mutually independent. The durations of type r calls are independent and identically distributed (positive) random variables, with finite mean μ_r^{-1}.

The set of links is $\mathcal{J}$, indexed by $j = \{1, \ldots, J\}$. On link j, C_j circuits (or trunks) are available. A customer of type r requires A_{jr} trunks on link j; $A := (A_{jr})_{j \in \mathcal{J}, r \in \mathcal{R}}$. We assume that the $A_{jr} \in \mathbb{N}_0$. (This is no real restriction. If the A_{jr} are, for instance given in three digits, multiply both the A_{jr} and C_j by 10^3.) We also assume that each type of traffic requires at least one circuit somewhere in the network. More formally: A has no null column: $\forall r \in \mathcal{R} : \exists j \in \mathcal{J} : A_{jr} > 0$.

A request for a call of type r is rejected if at that moment on any link $j \in \mathcal{J}$ there are fewer than A_{jr} circuits free. We say that the connection is blocked. Throughout this study, our attention focuses on the determination of the probability that a type r request is rejected, i.e., the type r blocking probability.

A notational comment: $(Ax)_j$ is a short notation of $\sum_{r=1}^{R} A_{jr} x_r$, i.e., the jth entry of Ax.

2.2 Equilibrium distribution – insensitivity

The blocking probability of a type r customer can be derived in the following way. The unique stationary distribution of the population of the network is given by the *product form*

$$\pi(k) \equiv \pi(k_1, \ldots, k_R) = G^{-1} \prod_{r=1}^{R} \frac{(\nu_r)^{k_r}}{k_r!}, \tag{7.1}$$

k_r denoting the number of type r customers in the system, and $\nu_r := \lambda_r / \mu_r$. The distribution is defined for all k in the 'integer polytope' S and G is a so-called normalizing constant;

$$S := \{k \mid \forall r \in \mathcal{R} : k_r \in \mathbb{N}_0, \forall j \in \mathcal{J} : (Ak)_j \leq C_j\} \text{ and } G := \sum_{k \in S} \prod_{r=1}^{R} \frac{(\nu_r)^{k_r}}{k_r!}. \tag{7.2}$$

This result is a famous *insensitivity* result: the stationary distribution depends on the call duration distributions only through their means. A key reference in this field is Burman, Lehoczky, and Lim [24]. We now concentrate on the type r blocking probability, say p_r. Let T_r define the subset of S in which a type r request cannot be accepted:

$$T_r := \{k \in S \mid \exists j \in \mathcal{J} : (Ak)_j + A_{jr} > C_j\}.$$

The overall blocking probability p is the probability that an arbitrary request cannot be accepted. Hence, from formula (7.1), the type r blocking probability and the overall blocking probability can be written as

$$p_r = G^{-1} \sum_{k \in T_r} \prod_{q=1}^{R} \frac{(\nu_q)^{k_q}}{k_q!} \quad \text{and} \quad p = \sum_{r=1}^{R} \left(\frac{\lambda_r}{\sum_{q=1}^{R} \lambda_q} \right) p_r. \tag{7.3}$$

2.3 Kelly's results and their drawbacks

In order to derive asymptotic results, Kelly [96] uses the following scaling. He parametrizes ν_r and C_j by n: $\nu_r^{(n)}$ and $C_j^{(n)}$. This parametrization is chosen such that $\nu_r^{(n)}/n$ and $C_j^{(n)}/n$ tend to a constant as $n \to \infty$. For reasons of simplicity, we will replace ν_r by $n\nu_r$ and C_j by nC_j, just as Gazdzicki *et al.* [72]. The type r blocking probability in the n-scaled process is denoted by $p_r^{(n)}$, given by (use equations (7.2) and (7.3))

$$p_r^{(n)} = \left(\sum_{k \in T_r^{(n)}} \left\{ \prod_{q=1}^{R} \frac{(n\nu_q)^{k_q}}{k_q!} \right\} \right) \left(\sum_{k \in S^{(n)}} \left\{ \prod_{q=1}^{R} \frac{(n\nu_q)^{k_q}}{k_q!} \right\} \right)^{-1}, \tag{7.4}$$

where $T_r^{(n)}$ and $S^{(n)}$ are defined in a self-evident way.

The number of customers in the system – k_r of type r – is scaled as well: $x_r := k_r/n$. As a consequence, x is contained in the simplex

$$\overline{S} := \{x \mid \forall r \in \mathcal{R} : x_r \geq 0, \forall j \in \mathcal{J} : (Ax)_j \leq C_j\}$$

Kelly considers the convex programming problem

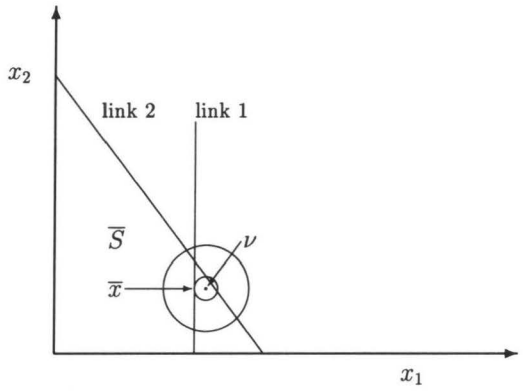
$$\inf_{x \in \overline{S}} \sum_{r=1}^{R} \left(x_r \log\left(\frac{x_r}{\nu_r}\right) - x_r \right)$$

This is a minimization of a strictly convex function on the convex set $\overline{S}$, and consequently, a unique minimum is attained, say $\overline{x}$. Kelly's main result is that

$$\lim_{n \to \infty} p_r^{(n)} = 1 - \frac{\overline{x}_r}{\nu_r}. \tag{7.5}$$

This implicitly shows that $\overline{x}_r \leq \nu_r$, for all $r \in \mathcal{R}$. He also establishes a relation with the solution of the so-called Erlang fixed point equations.

Figure 1: Heavy traffic network in which blockings of particular types are rare.



We now distinguish between two different regimes, viz. *light* and *heavy load*. In case of light load, $\nu \in \overline{S}$. One easily shows that the infimum over all non-negative x (instead of $x \in \overline{S}$) is attained in ν. Since $\nu \in \overline{S}$, this implies that, in the light traffic regime, $\overline{x} = \nu$. As a consequence, $\lim_{n\to\infty} p_r^{(n)} = 0$ for all $r \in \mathcal{R}$. However, the answer is quite unsatisfactory: we want to gain insight into the way that 0 is approached as $n \to \infty$. Gazdzicki *et al.* [72] succeed in answering that question in the single-link case. To our best knowledge, no asymptotic results are available for multiple links.

In the special case that ν is contained in the boundary of $\overline{S}$, we call the network *critically* or *moderately loaded*. Reiman [147] investigated single link critically loaded loss systems. He derived that $p_r^{(n)}\sqrt{n}$ tends to a constant.

Now let us discuss the other case: $\nu \notin \overline{S}$, i.e., a heavy traffic loss network. The minimum of the convex programming problem is attained on the boundary of $\overline{S}$. This can be seen as follows. (i) Suppose $\overline{x}$ lies in the interior of $\overline{S}$. (ii) There are convex combinations $\lambda\overline{x} + (1-\lambda)\nu \in \overline{S}$, with $\lambda \in (0,1)$. (iii) Based on the convexity of the objective function, and the fact that it is mimimal in ν, the value of the objective function in all of these convex combinations is smaller than in $\overline{x}$. Contradiction.

From Kelly's result (7.5) and the fact that $\overline{x} \neq \nu$ (apparently at least one of the entries of $\overline{x}$ is strictly smaller than the corresponding entry of ν), it follows that, for some $r \in \mathcal{R}$, $p_r^{(n)}$ has a positive limit. However, there still may exist $p_r^{(n)}$ with limit 0, as in the following example.

Suppose a loss model consisting of two types of traffic and two links, but $A_{12} = 0$. Figure 1 shows contour lines of the objective function. As said before, the objective function is minimal in ν. We get $\overline{x}_1 < \nu_1$ and $\overline{x}_2 = \nu_2$; consequently $p_1^{(n)}$ converges to a

positive constant but $p_2^{(n)}$ to 0 as $n \to \infty$. Although type 2 blocking is much rarer than type 1 blocking, the asymptotics of $p_2^{(n)}$ might still be of interest. To our knowledge, no results are available that cover this case.

As we see, both for the light and heavy traffic case there are situations in which no asymptotics of particular rare event probabilities are available. The next section addresses this open question.

3 Asymptotics of the blocking probability

In the previous section we saw that in several cases blocking probabilities in the n-scaled model tend to 0 as $n \to \infty$. This section deals with the exponential nature of this decay. We show that $(\log p_r^{(n)})/n$ tends to a constant. Apart from that, a mathematical program to calculate the corresponding limit is deduced. We scale in the same way as in Section 2. Define also the equivalent of $T_r^{(n)}$ in the scaled process:

$$\overline{T}_r := \left\{ x \in \overline{S} \mid \exists j \in \mathcal{J} : A_{jr} > 0, (Ax)_j = C_j \right\}.$$

The following theorem characterizes both the decay rate of the numerator and denominator of (7.4).

THEOREM 3.1.

$$\lim_{n\to\infty} \frac{1}{n} \log \left[\sum_{k \in S^{(n)}} \left\{ \prod_{q=1}^{R} \frac{(n\nu_q)^{k_q}}{k_q!} \right\} \right] = \lim_{n\to\infty} \frac{1}{n} \log \left[\max_{k \in S^{(n)}} \left\{ \prod_{q=1}^{R} \frac{(n\nu_q)^{k_q}}{k_q!} \right\} \right] \tag{7.6}$$

$$= - \inf_{x \in \overline{S}} \sum_{q=1}^{R} \left(x_q \log \left(\frac{x_q}{\nu_q} \right) - x_q \right), \tag{7.7}$$

and an analogous relation for $S^{(n)}$ and $\overline{S}$ replaced by $T_r^{(n)}$ and $\overline{T}_r$, respectively.

PROOF. Let us first find an upper bound to the number of elements of $S^{(n)}$. Clearly, for all $k \in S^{(n)}$, $A_{jr}k_r \leq nC_j$. Consequently,

$$k_r \leq \min_{j \in \mathcal{J}} \left(\frac{nC_j}{A_{jr}} \right) =: n\alpha_r \quad \text{for } k \in S^{(n)}.$$

Since A has no null column, α_r is finite. We get that $\#S^{(n)} \leq \prod_{r=1}^{R}(n\alpha_r + 1)$, i.e. of the order n^R. Also notice that $T_r^{(n)} \subset S^{(n)}$, so $\#T_r^{(n)}$ is smaller.

Trivial upper and lower bounds of the left hand side of (7.6) are

$$\limsup_{n\to\infty} \frac{1}{n} \log \left[\max_{k \in S^{(n)}} \left\{ \prod_{q=1}^{R} \frac{(n\nu_q)^{k_q}}{k_q!} \right\} \right] \leq \limsup_{n\to\infty} \frac{1}{n} \log \left[\sum_{k \in S^{(n)}} \left\{ \prod_{q=1}^{R} \frac{(n\nu_q)^{k_q}}{k_q!} \right\} \right] \leq$$

$$\limsup_{n\to\infty} \frac{1}{n} \log[\#S^{(n)}] + \limsup_{n\to\infty} \frac{1}{n} \log \left[\max_{k\in S^{(n)}} \left\{ \prod_{q=1}^{R} \frac{(n\nu_q)^{k_q}}{k_q!} \right\} \right].$$

Since $\#S^{(n)}$ is of the order n^R, the first term in the final expression equals 0. Therefore the inequalities can be replaced by equalities. The same procedure can be executed for the lim inf's. We can conclude that if both limits exist, the first equality of our theorem is deduced.

We will now show that these limits indeed exist and equal (7.7). This is done as follows. Approximate $k!$ by means of the Stirling formula:

$$\log(k!) = \begin{cases} 0 & \text{if } k = 0, \\ \frac{1}{2}\log(2\pi k) + k\log k - k + \theta(k) & \text{if } k \in \mathbb{N}. \end{cases}$$

Here $\theta(k)$ lies in the interval $([12k+1]^{-1}, [12k]^{-1})$. The right hand side of (7.6) therefore equals (interchanging max and log)

$$\lim_{n\to\infty} \max_{k\in S^{(n)}} \sum_{q=1}^{R} \left(-\frac{k_q}{n} \log\left(\frac{k_q}{n}\nu_q^{-1}\right) + \frac{k_q}{n} - \left(\frac{1}{2n}\log(2\pi k_q) + \frac{\theta(k_q)}{n}\right) 1\{k_q \in \mathbb{N}\} \right). \quad (7.8)$$

However, for all $k \in S^{(n)}$, we have deduced that $k_q \leq n\alpha_q$. Consequently,

$$0 \leq \sum_{q=1}^{R} \left(\frac{\log(2\pi k_q)}{2n} + \frac{\theta(k_q)}{n} \right) 1\{k_q \in \mathbb{N}\} \leq \sum_{q=1}^{R} \left(\frac{\log(2\pi n\alpha_q)}{2n} + \frac{1}{12n} \right) \to 0.$$

We get that limit (7.8) can be written as

$$\lim_{n\to\infty} \max_{k\in S^{(n)}} \sum_{q=1}^{R} \left(-\frac{k_q}{n} \log\left(\frac{k_q}{n}\nu_q^{-1}\right) + \frac{k_q}{n} \right).$$

Now notice that $\{k/n \mid k \in S^{(n)}\} \to \overline{S}$. In conjunction with the continuity of the function involved, it follows that the previous expression equals

$$-\inf_{x\in\overline{S}} \sum_{q=1}^{R} \left(x_q \log\left(\frac{x_q}{\nu_q}\right) - x_q \right),$$

The proof is finished by noticing that the same line of reasoning can be followed with $S^{(n)}$ and $\overline{S}$ replaced by $T_r^{(n)}$ and $\overline{T}_r$, respectively. ■

3.1 Interpretation of the asymptotics

The above result can be interpreted elegantly. By multiplying both numerator and denominator in (7.4) by $\exp(-n\sum_{q=1}^{R}\nu_q)$, we find the ratio of two Poisson probabilities. To be more precise, let $Z_q^{(n)}$ be the sum of n independent and identically distributed

Poisson(ν_q) random variables. Let the sequences $Z_1^{(n)}, \ldots, Z_R^{(n)}$ be mutually independent. Then probability (7.4) can be interpreted as

$$\frac{\mathcal{P}(Z^{(n)} \in T_r^{(n)})}{\mathcal{P}(Z^{(n)} \in S^{(n)})} = \mathcal{P}\left(Z^{(n)} \in T_r^{(n)} \mid Z^{(n)} \in S^{(n)}\right), \tag{7.9}$$

where the equality is because of $T_r^{(n)} \subset S^{(n)}$. In fact, this is the multivariate Poisson distribution, truncated to the polytope $S^{(n)}$. An implication of Theorem 3.1 is that the decay rate of the type r blocking probability is given by

$$\lim_{n\to\infty} \frac{1}{n} \log \left[\max_{k \in T_r^{(n)}} \left\{ \prod_{q=1}^{R} \frac{e^{-n\nu_q}(n\nu_q)^{k_q}}{k_q!} \right\} \right] - \lim_{n\to\infty} \frac{1}{n} \log \left[\max_{k \in S^{(n)}} \left\{ \prod_{q=1}^{R} \frac{e^{-n\nu_q}(n\nu_q)^{k_q}}{k_q!} \right\} \right]. \tag{7.10}$$

The maximum procedures of (7.10) in fact search the most likely state (i.e., the state with the highest value of the Poisson density) in both $T_r^{(n)}$ and $S^{(n)}$. Using Theorem 3.1 again, the value of (7.10) can be calculated by

$$\inf_{x \in \overline{S}} \sum_{q=1}^{R} \left(x_q \log\left(\frac{x_q}{\nu_q}\right) - x_q + \nu_q \right) - \inf_{x \in \overline{T}_r} \sum_{q=1}^{R} \left(x_q \log\left(\frac{x_q}{\nu_q}\right) - x_q + \nu_q \right). \tag{7.11}$$

In large deviations theory, the function $\sum_{q=1}^{R} \left(x_q \log\left(\frac{x_q}{\nu_q}\right) - x_q + \nu_q \right)$ is called the (joint) entropy function (or large deviations rate function) of independent Poisson(ν_1), ..., Poisson(ν_R) random variables. It reflects the likelihood of a sample mean x instead of ν. The function is convex and attains its minimal value in the most likely sample mean $x = \nu$. Both minimizations search the most likely states of the scaled process in the sets $\overline{S}$ and $\overline{T}_r$, respectively.

Suppose a light load. Based on the laws of large numbers, the numerator of (7.9) tends to 0, where the denominator converges to 1, $n \to \infty$. This also shows that $p_r^{(n)} \to 0$, as mentioned in Section 2. The first infimum in (7.11) is 0, attained for $x = \nu$. Consequently,

$$\lim_{n\to\infty} \frac{1}{n} \log p_r^{(n)} = - \inf_{x \in \overline{T}_r} \sum_{q=1}^{R} \left(x_q \log\left(\frac{x_q}{\nu_q}\right) - x_q + \nu_q \right).$$

Suppose on the other hand heavy traffic: $\nu \notin \overline{S}$. Then both the numerator and denominator of (7.9) go to 0. In Section 2 we argued that the minimizing argument $\overline{x}$ of

$$\inf_{x \in \overline{S}} \sum_{q=1}^{R} \left(x_q \log\left(\frac{x_q}{\nu_q}\right) - x_q \right)$$

lies on the boundary of $\overline{S}$. So, there exists at least one link j such that $(A\overline{x})_j = C_j$. We conclude that for all traffic streams r with $A_{jr} > 0$ the infimum over $\overline{T}_r$ equals the infimum over $\overline{S}$. Consequently, the decay rate of the blocking probability of these traffic streams equals zero. On the contrary, if for a traffic class r we have that $A_{jr} = 0$ (where j is again such that $(A\overline{x})_j = C_j$), then the decay rate is negative. Notice that these observations agree with Kelly's results, cf. Figure 1.

3.2 The overall blocking probability under light traffic

There is one special case in which the calculation of the decay rate is extremely simple: the case of the overall blocking probability under condition of a lightly loaded network. Recalling equation (7.3) and invoking Lemma 1.2.15 of [50], we get

$$\lim_{n\to\infty} \frac{1}{n} \log p^{(n)} = \max_{r\in\mathcal{R}} \lim_{n\to\infty} \frac{1}{n} \log p_r^{(n)}.$$

The decay rate under consideration can also be found without determining the decay rates of the individual $p_r^{(n)}$: $(\log p^{(n)})/n$ tends, by Theorem 3.1, to

$$\max_{r\in\mathcal{R}} \left[- \inf_{x\in\overline{T}_r} \sum_{q=1}^{R} \left(x_q \log\left(\frac{x_q}{\nu_q}\right) - x_q + \nu_q \right) \right] = - \inf_{x\in\overline{T}} \sum_{q=1}^{R} \left(x_q \log\left(\frac{x_q}{\nu_q}\right) - x_q + \nu_q \right),$$

where $\overline{T}$ is defined by $\left\{ x \in \overline{S} \mid \exists j \in \mathcal{J} : (Ax)_j = C_j \right\}$. This last infimum is very easy to calculate, i.e., it can be done without using complicated mathematical programming routines. This procedure consists of three steps.

STEP 1. Let P_j $(j \in \mathcal{J})$ be the jth relaxed infimum, i.e.,

$$\inf \sum_{r=1}^{R} \left(x_r \log\left(\frac{x_r}{\nu_r}\right) - x_r + \nu_r \right)$$

with $x \in [0,\infty)^R$ such that $(Ax)_j = C_j$. This problem is solved by $\overline{x}_r^{(j)} = \nu_r e^{\lambda_j A_{jr}}$ with Lagrange multiplier λ_j such that $(A\overline{x}^{(j)})_j = C_j$.

STEP 2. The set $\{\overline{x}^{(1)}, \ldots, \overline{x}^{(J)}\}$ has a non-empty intersection with $\overline{S}$. This can be seen as follows. If $\overline{x}^{(1)} \in \overline{S}$, then we are ready. Suppose now there is a $j \in \mathcal{J} \setminus \{1\}$ such that $(A\overline{x}^{(1)})_j > C_j$, as in Figure 2.

Define

$$U := \left\{ x \geq 0 \mid \sum_{r=1}^{R} \left(x_r \log\left(\frac{x_r}{\nu_r}\right) - x_r + \nu_r \right) \leq \mathrm{P}_1 \right\}.$$

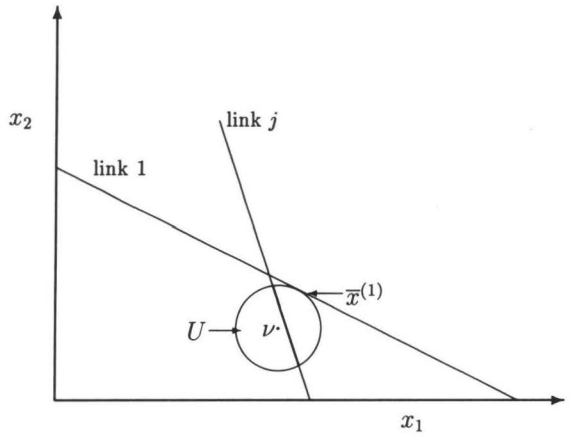
Because of the convexity of the objective function, U is convex as well. U is therefore contained in $\{x \geq 0 \mid (Ax)_1 \leq C_1\}$. Now notice that

- $\nu \in U$ and $(A\nu)_j \leq C_j$, because of the light load.
- $\overline{x}^{(1)} \in U$ (by definition) and $(A\overline{x}^{(1)})_j > C_j$ (by assumption).

It follows that U and $\{x \geq 0 \mid (Ax)_j = C_j\}$ have a non empty intersection. The argument of P_j evidently lies in this intersection. Consequently, $\overline{x}^{(j)}$ satisfies both $(A\overline{x}^{(j)})_1 \leq C_1$ and $(A\overline{x}^{(j)})_j = C_j$. Continuing this procedure we eventually find at least one $\overline{x}^{(j)} \in \overline{S}$.

STEP 3. Determine $-\inf \mathrm{P}_j$ over all $j \in \mathcal{J}$ with $\overline{x}^{(j)} \in \overline{S}$.

Figure 2: Light traffic network.



4 Fast simulation techniques

Up to this moment now we found only rough characteristics of the blocking probabilities: instead of an approximation of the probabilities themselves we only get their decay rates. As a consequence, the results of Section 3 seem to be of theoretical rather than practical interest. However, the solutions of the mathematical programming problems appear to be crucial in order to develop fast simulation techniques.

We already found that the blocking probabilities (7.4) can be interpreted as (7.9), the ratio of two probabilities (arising from the joint density of R independent Poisson random variables). We will use this interpretation throughout this section. The idea is to estimate both numerator and denominator. It is well-known how to construct confidence intervals of ratio estimators.

4.1 Importance sampling

Suppose $\mathcal{P}(Z^{(n)} \in X^{(n)})$ has to be estimated, where $\{k/n \mid k \in X^{(n)}\}$ converges to $\overline{X}$. (In our application, $X^{(n)}$ must be read as $S^{(n)}$ or $T_r^{(n)}$, and $\overline{X}$ as $\overline{S}$ or $\overline{T}_r$.) The direct way, which evidently yields an unbiased estimate, would be: (i) Sample M times Poisson$(n\nu_q)$, $q = 1, \ldots, R$, random variables. (ii) Check whether sample $m = 1, \ldots, M$ lies in $X^{(n)}$. If true, define $I_m := 1$, otherwise 0. (iii) Determine the ratio of the number of 'successes' (samples in $X^{(n)}$) and M: $\sum_{m=1}^{M} I_m/M$.

A remark, concerning the number of samples to be drawn, can be made. To obtain an estimate with a fixed confidence and relative efficiency (confidence interval half-length divided by the estimate), the number of samples required is inversely proportional to

the probability to be estimated. For instance, to get 95% confidence and 10% relative efficiency (ratio of the confidence interval half-length and the estimate), the number of samples to be drawn is about 400 over the probability of our interest. In other words, 400 successes must be generated. Suppose now that $\nu \notin \overline{X}$. Then a success is a rare event: based on the laws of large numbers $\mathcal{P}(Z^{(n)} \in X^{(n)}) \to 0$. Consequently, in order to get a fixed number of successes, possibly a huge number of samples has to be drawn.

As explained in earlier chapters, a variance reduction technique in rare event simulation is *importance sampling*. Suppose we identify the original probability model with measure $\mathcal{P}$, importance sampling is a simulation under an alternative measure $\mathcal{Q}$. The new measure $\mathcal{Q}$ must be chosen such that the rare event under consideration becomes more frequent. Sample according to $\mathcal{Q}$, weight each observation I_m by the appropriate likelihood ratio L_m, expressing the relative likelihood of the observation under the old measure with respect to the new measure. Then $\sum_{m=1}^{M} L_m I_m / M$ is an unbiased estimator as well, but with a different variance performance. Of course, we want to select that measure $\mathcal{Q}$ such that the variance of the new estimator (based on a fixed number, say K, samples) is minimal.

Our idea is to let $\mathcal{Q}$ be again a multivariate Poisson distribution, but with parameter $n\overline{x}$ instead of $n\nu$, where $\overline{x}$ is the optimizing argument in

$$\inf_{x \in \overline{X}} \sum_{q=1}^{R} \left(x_q \log\left(\frac{x_q}{\nu_q}\right) - x_q \right).$$

As we know from Section 3, $n\overline{x}$ approximates the most likely state of the multivariate Poisson distribution in the set $X^{(n)}$. In other words: the most likely state in case of being in $X^{(n)}$ under the original measure, coincides with the 'overall most likely state' under the new measure. If $\nu \in \overline{X}$, then both measures are equal. If on the contrary $\nu \notin \overline{X}$, then $\nu \neq \overline{x}$; then the rare event occurs frequently under the new measure.

This kind of changes of measure has proven to improve the variance performance considerably [85]. However, we did not succeed in proving optimality properties. Nevertheless, our choice of the alternative density seems to have a better mathematical motivation then the heuristic recipe reported in [158].

Suppose the Poisson$(n\overline{x}_q)$ random variables attain value k_q, then the likelihood ratio reads

$$\left(\prod_{q=1}^{R} \frac{e^{-n\nu_q}(n\nu_q)^{k_q}}{k_q!} \right) \left(\prod_{q=1}^{R} \frac{e^{-n\overline{x}_q}(n\overline{x}_q)^{k_q}}{k_q!} \right)^{-1} = \exp\left[-n \sum_{q=1}^{R} (\nu_q - \overline{x}_q) \right] \prod_{q=1}^{R} \left(\frac{\nu_q}{\overline{x}_q} \right)^{k_q}. \quad (7.12)$$

4.2 Computational remarks

To perform the importance sampling, the infima

$$\inf_{x\in\overline{S}} \sum_{q=1}^{R} \left(x_q \log\left(\frac{x_q}{\nu_q}\right) - x_q \right) \quad \text{and} \quad \inf_{x\in\overline{T}_r} \sum_{q=1}^{R} \left(x_q \log\left(\frac{x_q}{\nu_q}\right) - x_q \right). \tag{7.13}$$

have to be evaluated. The computation of the first infimum can be done elegantly by applying duality, see Kelly [96]. The second infimum can be handled in a similar way, as we will explain now.

First notice that the domain $\overline{T}_r$ is not convex, so we use the following decoupling:

$$\inf_{x\in\overline{T}_r} \sum_{q=1}^{R} \left(x_q \log\left(\frac{x_q}{\nu_q}\right) - x_q \right) = \inf_{j:A_{jr}>0} \left[\inf_{x\in\overline{T}'_j} \sum_{q=1}^{R} \left(x_q \log\left(\frac{x_q}{\nu_q}\right) - x_q \right) \right], \tag{7.14}$$

where $\overline{T}'_j$ is defined by $\{x \in \overline{S} \mid (Ax)_j = C_j\}$. An efficient method to calculate the inner infimum of the right hand side is the following. The objective function as well as the domain $\overline{T}'_j$ are convex, so we are in a position to exploit duality results. The minimization is equivalent to the Lagrangian form (where y is the vector of Lagrange multipliers)

$$\inf_{x\geq 0} \left[\sup_{y\in Y_j} \left[\sum_{q=1}^{R} \left(x_q \log\left(\frac{x_q}{\nu_q}\right) - x_q \right) + \sum_{i=1}^{J} y_i((Ax)_i - C_i) \right] \right].$$

Here Y_j is the orthant of $\mathbb{R}^J$ such that y_j is free and the other y_i are non-negative. Due to the duality theorem (see page 89 of Rockafellar [157]) the order of inf and sup can be reversed:

$$\sup_{y\in Y_j} \left[\inf_{x\geq 0} \left[\sum_{q=1}^{R} \left(x_q \log\left(\frac{x_q}{\nu_q}\right) - x_q \right) + \sum_{i=1}^{J} y_i((Ax)_i - C_i) \right] \right]. \tag{7.15}$$

The inner minimization can be calculated explicitly. It appears that the minimum is attained at

$$\overline{x}_q = \nu_q \exp\left(-(A^{\mathrm{T}}y)_q\right). \tag{7.16}$$

Inserting this in (7.15) we arrive, after some calculus, at

$$-\inf_{y\in Y_j} \left[\sum_{q=1}^{R} \nu_q \exp\left(-\sum_{i=1}^{J} y_i A_{iq} \right) + \sum_{i=1}^{J} y_i C_i \right].$$

This is the dual of the inner infimum in the right hand side of (7.14).

So for all links j with A_{jr} strictly positive, this dual problem has to be solved. Then the minimum over the resulting values of the objective function has to be determined. Recall that the advantages of solving the dual program (instead of the primal problem)

are twofold, as mentioned in Kelly. First, the dimensionality, i.e. the number of variables, is usually reduced considerably. In most networks $J << R$, think for instance of 'star networks' with $R = \binom{J}{2}$ and 'complete networks' with $R = (J-1)!$ Besides, the feasible region is less complicated: Y_j instead of the polytope $\overline{S}$. The minimizing y is called $\overline{y}$. We find $\overline{x}$ via (7.16).

4.3 Accelerations

In the above we described an importance sampling technique to estimate the Poisson probabilities in (7.9). However, based on a number of heuristical arguments, the simulation can be simplified. From our experiments, it appears that these simplifications still yield very accurate results.

ACCELERATION 1. First consider $\mathcal{P}(Z^{(n)} \in T_r^{(n)})$. From Theorem 3.1 it follows that this probability mass is concentrated in a neighborhood of $n\overline{x}$, where $\overline{x}$ arises from the second infimum in (7.13). Clearly, the set of links $\mathcal{J}_r$ that are most likely to block are the links j with $(A\overline{x})_j = C_j$. Based on this observation, it seems reasonable to check the occurrence of a type r blocking only on the links in $\mathcal{J}_r$, instead of all links j with $A_{jr} > 0$. More formally, we replace the set $T_r^{(n)}$ by a set $T_r^{(n)*}$ that indicates type r blocking on a 'bottleneck link':

$$T_r^{(n)*} := \{k \in \mathbb{N}_0^R \mid \forall j \in \mathcal{J}_r : (Ak)_j \leq nC_j, \quad \exists j \in \mathcal{J}_r : (Ak)_j + A_{jr} > nC_j\}.$$

Consequently, the number of checks per sample may decrease considerably, possibly saving a huge amount of simulation time. Notice that a similar simplification can be applied to estimate $\mathcal{P}\left(Z^{(n)} \in S^{(n)}\right)$.

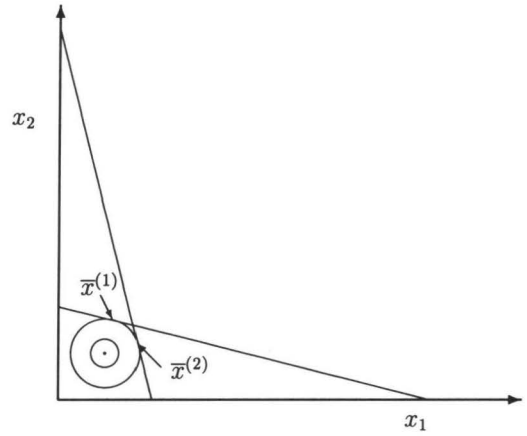
ACCELERATION 2. The optimization over $\overline{S}$ always yields a unique minimum, but the infimum over $\overline{T}_r$ does need to be determined uniquely, for instance because of symmetry, see Figure 3.

Without loss of generality we assume two minimizers, say $\overline{x}^{(1)}$ and $\overline{x}^{(2)}$. From our experiments, it turns out that $\mathcal{P}(Z^{(n)} \in T_r^{(n)})$ can be approximated very accurately by $\mathcal{P}(Z^{(n)} \in T_{r,1}^{(n)*}) + \mathcal{P}(Z^{(n)} \in T_{r,2}^{(n)*})$. Here $T_{r,i}^{(n)*}$ is defined as $T_r^{(n)*}$, but with $\overline{x}$ replaced by $\overline{x}^{(i)}$.

5 Some examples

This section deals with a number examples, that explain the importance sampling technique described in the previous section. Our estimates have 95% confidence and 10%

Figure 3: A non-unique infimum.



relative efficiency.

EXAMPLE 5.1. First we treat an example that corresponds with the heavy traffic case of Figure 1. Define a network with 3 links and 2 types of customers:

$$A := \begin{pmatrix} 2 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}, \quad C := \begin{pmatrix} 40 \\ 16 \\ 36 \end{pmatrix}, \quad \nu := \begin{pmatrix} 24 \\ 4 \end{pmatrix}.$$

First, the type 1 blocking probability is estimated. Performing the optimizations (7.13), we find that the infima over T_1 and S coincide. The optima are attained in $(16, 4)$. The 'bottleneck link' is link 2. Applying the first acceleration, we approximate T_1 by $\{k \in \mathbb{N}_0^2 \mid k_1 = 16\}$ and S by $\{k \in \mathbb{N}_0^2 \mid k_1 \leq 16\}$. Given a realization (k_1, k_2), the likelihood equals $e^{-8}(3/2)^{k_1}$, cf. equation (7.12). Performing the importance sampling simulations, we get as estimate for the numerator (denominator) $2.04 \cdot 10^{-2}$ $(5.31 \cdot 10^{-2})$, yielding 0.384 for p_1. This is more or less in agreement with Kelly's result, saying that $p_1 \approx 1/3$.

Then consider type 2 blocking. From the above and Kelly's result we have that type 2 blocking is rare, i.e., $p_2 \approx 0$. To find a numerical value, we use simulation. We already estimated the denominator, and we are therefore left with the numerator. The infimum over T_2 is attained in $(16, 8)$. Consequently, both link 1 and 2 are 'bottlenecks' and T_2 is approximated by $\{k \in \mathbb{N}_0^2 \mid 2k_1 + k_2 = 40, k_1 \leq 16\}$ The simulation leads to an estimate $1.36 \cdot 10^{-5}$ of the numerator and $2.56 \cdot 10^{-4}$ of p_2.

In both cases, the simulation time required was about 2 seconds on a 486 personal computer to obtain an estimate with our prescribed level of confidence (95% confidence,

10% relative efficiency).

EXAMPLE 5.2. The second example is taken from Ross and Wang [158]. Consider a star network with four 'leaves'. There are 12 classes of customers, and $\binom{4}{2} = 6$ routes.

Class	Route	# trunks	ν_r
1	1,2	1	9.0
2	1,3	1	9.0
3	1,4	1	9.0
4	2,3	1	9.0
5	2,4	1	9.0
6	3,4	1	9.0

Class	Route	# trunks	ν_r
7	1,2	5	1.6
8	1,3	5	1.6
9	1,4	5	1.6
10	2,3	5	1.6
11	2,4	5	1.6
12	3,4	5	1.6

Here the first column is the set of customer classes $\mathcal{R}$, and the second column indicates which links are used by the classes. The third indicates how many circuits are (per connection) needed on these links. The last is the product of the Poisson arrival rates and the mean call duration: $\nu_r = \lambda_r/\mu_r$. The link capacities are $C_1 = 90$, $C_2 = 100$, $C_3 = 110$, and $C_4 = 120$.

Now suppose that the blocking probability of a type 1 customer has to be determined. This network has a light load: the load on link i is $51/C_i < 1$. Consequently, the denominator can be estimated using the original set of parameters. Therefore, concentrate on the numerator. The theory of the previous section says that the dual objective function

$$9e^{-y_1-y_2} + \ldots + 9e^{-y_3-y_4} + 1.6e^{-5y_1-5y_2} + \ldots + 1.6e^{-5y_3-5y_4} + 90y_1 + 100y_2 + 110y_3 + 120y_4.$$

must be minimized over the Y_j with $A_{j1} > 0$, i.e.,

(i) y_1 free, $y_2, y_3, y_4 \geq 0$. The mathematical program yields 59.73, with $\overline{y}_1 = -0.1758$ and the other entries 0.

(ii) y_2 free, $y_1, y_3, y_4 \geq 0$. Now the minimal value is 57.82, attained for $\overline{y}_2 = -0.2049$ and the other entries 0.

Using (7.14), we conclude that link 1 is the link where overflow occurs most likely. The new traffic rates under the importance sampling become: $\overline{x}_1 = \overline{x}_2 = \overline{x}_3 = 9\exp(0.1758) = 10.7301$, $\overline{x}_4 = \overline{x}_5 = \overline{x}_6 = 9$, $\overline{x}_7 = \overline{x}_8 = \overline{x}_9 = 1.6\exp(5 \cdot 0.1758) = 3.8541$, $\overline{x}_{10} = \overline{x}_{11} = \overline{x}_{12} = 1.6$. Now use the first acceleration: as a consequence, T_1 is approximated by

$$\{k \in \mathbb{N}_0^6 \mid k_1 + k_2 + k_3 + 5(k_7 + k_8 + k_9) = 90\}.$$

Consequently, the Poisson variables of type 4, 5, 6, 10, 11, 12 do not even need to be sampled, and blocking on link 2 does not need to be checked. The simulation yields

blocking probability $4.42 \cdot 10^{-4}$, in about 2 seconds on a 486 personal computer. The estimate agrees with Ross and Wang [158]. They needed about 60 CPU seconds to get a 95% confidence interval with relative efficiency of only 24%.

Notice that the set of 'bottleneck links' $\mathcal{J}_1$ has only one element. Due to acceleration 1, we can approximate p_1 by the blocking probability if the network consisted of only link 1. But then we can use the results of [72], yielding the approximation $4.64 \cdot 10^{-4}$.

EXAMPLE 5.3. The third example is a large star network, with $J = 20$. There are $\binom{20}{2} = 190$ classes, all arriving according to Poisson(3.8) processes. They require one circuit on both links of their connection. The link capacities are 100. Again we are interested in the type 1 blocking probability, assuming that type 1 traffic uses link 1 and link 2.

This example is particularly suitable to show that executing the summations of (7.1) and (7.2) is not feasible. It can be checked easily that $\{0, \ldots, 5\}^{190} \subset S$, so $\#S \geq 6^{190} = 10^{148}$.

Because of the light traffic, estimating the denominator is again straightforward. To obtain the alternative measure to estimate the numerator, we have to minimize

$$\sum_{j=1}^{20} \sum_{i<j} 3.8 e^{-y_i - y_j} + \sum_{j=1}^{20} 100 y_j,$$

over Y_1 and Y_2. Evidently, due to then symmetry, these minimizations yield the same value of the objective function, with the role of y_1 and y_2 interchanged. Invoking acceleration 2, we estimate the probability of a type 1 blocking on link 1 and multiply it by 2.

It can be calculated that the minimization over Y_1 yields $\overline{y}_1 = -0.3257$ and $\overline{y}_2 = \ldots = \overline{y}_{20} = 0$. The importance sampling parameters become 5.263 for all types of traffic using link 1 and 3.8 for all other types of traffic. We approximate T_1 by

$$\left\{ k \in \mathbb{N}_0^{19} \mid \sum_{i=1}^{19} k_i = 100 \right\}.$$

This procedure yields estimate $3.45 \cdot 10^{-4}$, resulting in estimate $6.89 \cdot 10^{-4}$ for p_1. The simulation time is 0.6 second on a 486 PC. This is indeed very fast compared to 900 CPU seconds as in [158] to get a 95% confidence interval with only 22% relative efficiency.

6 Conclusions

We have studied asymptotic techniques to approximate the blocking probability in a loss network. After a scaling as in [96], the decay rate of the loss probability appears to be the solution of a convex programming problem. We established a nice relation between these

results and fast simulation techniques. In fact, the optimizing arguments of the convex programming problem provide an alternative probability measure. Importance sampling under this measure yields a huge variance reduction. Our examples show that even in networks of considerable size, the required simulation time is still relatively small.

As said before, we derive the asymptotics of the decay rate of the blocking probability $p_r^{(n)}$ of type r traffic. However, for practical purposes it is more interesting to approximate $p_r^{(n)}$ itself, for instance by finding a function $g_r(\cdot)$ such that $p_r^{(n)}/g_r(n) \to 1$ as $n \to \infty$. Another subject for further research is to develop asymptotics and fast simulation techniques in case of trunk reservation.

Chapter 8

Call blocking in cellular mobile networks

This chapter investigates cellular mobile communication networks. Its purpose is twofold. First, it is noted that the restrictive assumption of reversible routing is not required for the network population distribution to be of product form. A protocol with a specific way of handling congestion, yielding product form, is discussed. Second, the notoriously difficult task of obtaining performance measures derived from product form expressions is attacked by an efficient method based on importance sampling. This algorithm substantially speeds up the computational time required to estimate, for example, the probability that a call attempting a handover is blocked. In addition, qualitative insight is gained into the network, given blocking in a specific cell: are neighboring cells overloaded as well? The examples include networks with capacity constraints due to effective interference between cells, and a reasonably sized network containing 49 cells and 7 cell reuse groups.

1 Introduction

Mobile communications has been a rapidly growing service in the field of telecommunications. One of the problems arising in this area is that the number of frequencies available to carry mobile calls is severely limited, restricting the number of calls that can be handled. Clearly, the capacity of the mobile network can be substantially enhanced when channels can be used to carry multiple calls. This has resulted in the idea of *cellular* mobile communication networks: the area is divided into cells, that are served by 'cell transceivers'; transmission between the mobile terminal and the transceiver is at low power so that the channel can be reused in another cell. The distance between both cells must be sufficiently large, in order to guard against interference. This chapter considers performance measures for such cellular mobile communication networks.

Due to the limited number of channels available for mobile communications, the crucial performance indicators are *call blocking probabilities*. There are two types of them. On

the one hand a 'fresh call' can be blocked: there is no channel available to accommodate a new call. On the other hand, there is the so-called 'handover' blocking probability. If a mobile terminal moves from the radio coverage of one cell to the radio coverage of another cell the call is handed over from one transceiver to another. However, if there is no channel available in the new cell, the call is lost. Blocking of the latter kind is considered to be more severe than blocking of the former kind, because an *existing* call is terminated. It is these types of blocking probabilities that are the main topic of this chapter.

The literature on blocking analysis in mobile systems can be divided roughly into two major groups: models on cell-level and models on network-level. The former group considers the probabilistic behavior of a single cell. An arrival stream of fresh calls as well as an arrival stream of handover calls is modeled. The purpose is to characterize (for both streams) the probability of blocking. These models focus on mechanisms that reduce the handover blocking probability. An example is priority for handed over calls, such as trunk reservation by means of guard channels [79], [94]. Some models incorporate the possibility of rejected customers to redial, e.g. [31], [176]. The present chapter focuses on models on network level. Such models aim to find a probabilistic description of the entire cellular network population. A number of strategies has been developed, that assign channels to cells in order to decrease blocking rates. The advantage of network-level models over cell-level models is that the performance of these channel assignment policies can be evaluated. Assignment policies that are used in the literature (e.g., fixed channel allocation, dynamic channel allocation with maximum packing [66], and reuse groups [67], [64]) frequently result in a state space of the 'matrix constraint' type [100]: $S = \{k : Ak \leq C\}$, for a matrix A and a vector C. Here $k = (k_1, \ldots, k_R)$, k_r is the number of calls in cell r, $r = 1, \ldots, R$, and R is the number of cells. Two examples of assignment policies of this type can be found in Section 4.

For reasonably sized models, the size of the resulting state space usually prohibits the calculation of the probability distribution of the number of calls in the cells in the network. Several studies showed that a *product form* distribution can be deduced: the steady state probability of the number of calls in the cells is given by

$$\pi(k) = G^{-1} \prod_{r=1}^{R} \frac{{\nu_r}^{k_r}}{k_r!}, \quad k \in S = \{k : Ak \leq C\},$$

for some positive numbers ν_i; G^{-1} is the normalizing constant. The advantage of this product form expression is that only R positive numbers need to be determined to compute the state probabilities. Unfortunately, the conditions under which product form has been shown to hold for cellular mobile networks are quite restrictive. In fact, a product form has only been derived under each of the following three assumptions.

- *No capacity constraints* on the cell populations. Every call is accepted so that the

state space is $S = \mathbb{N}_0^R$. Colombo [39] analyzes a model of this type, also considering the mobiles not carrying a call.

- *No explicit modeling of mobility* of the users of the network, see e.g. Everitt *et al.* [67], [63], [64]. For every cell an arrival rate (due to handovers and newly initiated calls) and a call termination rate are defined. Since there is no explicit modeling of mobility, an arrival in a cell due to a handover does not necessarily imply a call termination in an adjacent cell. Consequently, the number of calls in the cells of the network (i.e., the k_r) are linked via a state space constraint only.

- *Reversible routing of calls among the cells* allows the introduction of state space constraints while taking into account the mobility of calls, see Pallant and Taylor [142], [143]. The assumption of reversible routing roughly implies that the average number of calls moving from one cell to another equals the average number of calls moving in the opposite direction.

The assumptions used in these models are fairly restrictive: in practice there is always a capacity restriction, mobility is too important to neglect and, particularly on smaller time scales, reversibility does not hold. A particular situation that cannot be handled by the three cases presented, are the basically unidirectional flows in traffic streams occurring during the morning and afternoon rush hours. However, in Boucherie and Mandjes [19] it is shown that product form results can also be deduced for cellular mobile communication networks that include the mobility of calls in networks with finite capacity cells in non-reversible structures. Since still some assumptions have to be satisfied, the resulting product form results might have limited use for *quantitative* analysis. However, the product form equilibrium distribution can be very useful to obtain *qualitative* insight into the network behavior.

Having deduced the product form, we have to calculate the blocking probabilities from it. Despite the explicit formulas, this calculation is substantially hampered by the resulting large summations over (part of) the state space that have to be performed. As explained in the previous chapter, only in models of small dimension, these troubles can be overcome by using efficient recursive algorithms [25], [93], [154]; in models of larger size, we can use for instance fixed point approximation techniques [100] or numerical methods using the inversion of the Laplace transform for the blocking probabilities [32]. An alternative method to evaluate the summations is the use of simulation techniques. In this chapter we will use the efficient method presented in the previous chapter, i.e., importance sampling, where the alternative distribution ('change of measure') is found by large deviations techniques. We showed that it performed considerably better than the heuristic importance sampling techniques of Ross and Wang [158]. As Monte-Carlo

simulation with an algorithm by Harvey and Hills [83] is frequently used in the mobile communication literature, cf. Everitt and Manfield [67], we will compare our method with this technique.

In addition, the large deviations results that give the change of measure, provide also insight into the state of the network, given blocking in a particular cell. Large deviations results enable us to calculate whether, upon blocking in a cell, *neighboring cells are likely to be full as well* (see Section 4). In other words: we can examine whether blocking has a local or a global character: will blocking of a particular cell imply blocking of larger parts of the network. This gives important *qualitative* insight into the behavior of cellular mobile communication networks.

This chapter is organized as follows. In Section 2, we summarize the basic product form results from [19]. In Section 3 we treat the estimation of performance measures. In Section 4 the results are illustrated through interference models [65] and reuse groups in a network of considerable size [67].

2 Product form results

In this section we review some of the product form results derived in Boucherie and Mandjes [19]. In Subsection 2.1 we describe the network in detail. Subsection 2.2 says that by introducing redial rates, the assumptions mentioned in the introduction are not necessary. Notice that in [19] two other 'realistic' protocols are described that give a product form solution as well ('soft overload' and 'push-out policy'). This section is concluded by expressing the blocking probabilities in product form terms.

2.1 Description of the cellular mobile communication network

In mobile communication networks, handovers occur typically in the border region between cells and are possible as a consequence of overlap between the areas covered by transceivers (see Figure 1). In the handover area a call can be carried by two transceivers and, due to the capacity constraints, to which the transceiver the call is assigned, has to be taken into account. In this study we aim to capture the effect of different behavior of calls: we will separate a cell in interior and handover areas but ignore the exact locations of the calls inside these areas. In the handover area between cells r and s a call is assigned to one of the two transceivers r or s. In the interior of cell r a call can be carried by transceiver r only. The explicit use of handover areas is motivated further in Remark 2.2.

Consider (a part of) a cellular mobile communication network consisting of R cells. Let m_{rr} denote the number of calls in progress in the interior of cell r, and let m_{rs} denote the number of calls in progress in the handover area between cells r and s but carried by

transceiver r. The capacity constraints on the number of calls in progress that are carried by transceiver r are on the total number of calls. Let k_r be the number of calls carried by transceiver r. Obviously,

$$k_r = m_{rr} + \sum_{s \in B_r} n_{rs},$$

where B_r contains the indices of the cells that are neighbors of cell r. In Figure 1, $B_1 = \{2, 3, 4, 5, 6, 7\}$.

For convenience, assume that new calls are generated in the interior of the cells, and that calls complete in the interior of the cells. New calls in cell r are assumed to be generated by a Poisson process with rate λ_r, $r = 1, \ldots, R$. Mobiles carrying a call remain in the interior of cell r for an $\exp(\mu_r^*)$ distributed amount of time. Then the call proceeds to the handover area with cell s with probability $p^*_{rr,rs}$. In addition, a call might also complete in cell r. This occurs after an $\exp(\mu_r)$ distributed time. Thus, the holding time of a call in the interior of cell r is exponentially distributed with mean $1/\mu_{rr} = 1/(\mu_r + \mu_r^*)$. With probability $p_{rr,0} := \mu_r/\mu_{rr}$ the call completes, and with probability $p_{rr,rs} := (\mu_r^*/\mu_{rr})p^*_{rr,rs}$ the call proceeds to the handover area with cell s. (Note that $\sum_{s \in B_r} p^*_{rr,rs} = 1$, but that $p_{rr,0} + \sum_{s \in B_r} p_{rr,rs} = 1$.) Calls remain in the handover area rs for an $\exp(\mu_{rs})$ distributed amount of time. A handover is attempted with probability $p_{rs,sr}$, and a call moves or returns to the interior of cell r with probability $p_{rs,rr}$. Acceptance of an attempted handover is determined by the handover policy.

In our analysis the so-called traffic equations play a crucial role. As can be found in Seneta [163], the following holds:

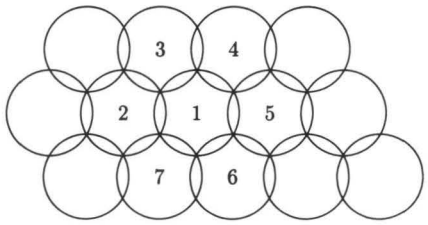
TRAFFIC EQUATIONS. *The traffic equations in unknowns* $\{\gamma_{rr}, \gamma_{rs}, s \in B_r, r = 1, \ldots, R\}$

$$\gamma_{rr} p_{rr,0} + \sum_{s \in B_r} \gamma_{rr} p_{rr,rs} = \lambda_r + \sum_{s \in B_r} \gamma_{rs} p_{rs,rr}, \quad r = 1, \ldots, R, \tag{8.1}$$

$$\gamma_{rs} p_{rs,rr} + \gamma_{rs} p_{rs,sr} = \gamma_{rr} p_{rr,rs} + \gamma_{sr} p_{sr,rs}, \quad r \in B_s, \quad r = 1, \ldots, R, \tag{8.2}$$

have a unique positive solution $\{c_{rr}, c_{rs}, s \in B_r, r = 1, \ldots, R\}$.

Figure 1: Handover areas.



2.2 The redial mechanism

Consider a mobile communication network with overlapping cells as depicted in Figure 1. The evolution of the network is described above, except for calls that cannot be admitted due to the capacity constraints of the network. New calls are rejected in cell r when $A(k+e_r) \not\leq C$: an additional call in cell r violates the capacity constraints of the network. (Note that for the state m to be admissible it must be that $Ak \leq C$.) A handover from cell s to cell r is lost when $A(k-e_s+e_r) \not\leq C$. Lost calls will attempt to re-establish their connection. This will introduce a *redial* behavior in the cells surrounding a cell r for which $A(k+e_r) \not\leq C$ (cell r has reached its capacity). This additional redial behavior is limited to the handover areas of the surrounding cells, as it is in these areas that handovers are lost. We have the following transition rates:

$$q(m,m') = \begin{cases} \lambda_r & m' = m + e_{rr},\ A(k+e_r) \leq C \\ & \text{(new call)} \\ m_{rr}\mu_{rr}p_{rr,0} & m' = m - e_{rr} \\ & \text{(end call)} \\ m_{rr}\mu_{rr}p_{rr,rs} & m' = m - e_{rr} + e_{rs} \\ & \text{(move to handover area)} \\ m_{rs}\mu_{rs}p_{rs,rr} & m' = m - e_{rs} + e_{rr} \\ & \text{(move from handover area)} \\ m_{rs}\mu_{rs}p_{rs,sr} & m' = m - e_{rs} + e_{sr},\ A(k-e_r+e_s) \leq C \\ & \text{(handover)} \\ m_{rs}\mu_{rs}p_{rs,sr} & m' = m - e_{rs},\ A(k-e_r+e_s) \not\leq C \\ & \text{(lost handover)} \\ r_{sr} & m' = m + e_{sr},\ A(k+e_r) \not\leq C,\ A(k+e_s) \leq C \\ & \text{(redial)} \end{cases}$$

The following theorem establishes a product form equilibrium distribution under a condition on the redial rates.

THEOREM 2.1. REDIAL MECHANISM. Let $\{c_{rr}, c_{rs}, s \in B_r, r = 1, \ldots, R\}$ be the unique positive solution of the traffic equations (8.1), (8.2). Assume that this solution is such that

$$c_{rs} = \frac{r_{sr}}{p_{rs,sr}}, \quad s \in B_r, \quad r = 1, \ldots, R.$$

Then under the redial mechanism the network has a unique equilibrium distribution

$$\pi(m) = G^{-1} \prod_{r=1}^{R} \left(\frac{c_{rr}}{\mu_{rr}}\right)^{m_{rr}} \frac{1}{m_{rr}!} \prod_{s \in B_r} \left(\frac{c_{rs}}{\mu_{rs}}\right)^{m_{rs}} \frac{1}{m_{rs}!}, \quad m \in S = \{m : Ak \leq C\},$$

where G^{-1} is the normalizing constant:

$$G = \sum_{m \in S} \prod_{r=1}^{R} \left(\frac{c_{rr}}{\mu_{rr}}\right)^{m_{rr}} \frac{1}{m_{rr}!} \prod_{s \in B_r} \left(\frac{c_{rs}}{\mu_{rs}}\right)^{m_{rs}} \frac{1}{m_{rs}!}.$$

PROOF. As described in [19], using the principle of partial balance [95]. ■

REMARK 2.2. HANDOVER AREAS. The behavior of calls near the boundary of a cell substantially differs from the behavior of calls in the interior: it is only in the area covered by multiple transceivers that handovers are possible. Therefore, our model explicitly takes into account these handover areas. Explicit modeling of handover areas was also suggested in Everitt [63].

Call blocking due to handovers occurs typically in the handover area between cells. However, a call that is temporarily lost as consequence of the capacity restrictions in its destination cell, say cell s, can still be in the area covered by the original transceiver, say r. This call will therefore attempt to be reconnected to transceiver r and generates a new call in handover area rs. In addition, a customer that has lost its call while entering a full cell r might continue trying to re-establish its call. While travelling through cell r this customer will not succeed, but as soon as the customer enters the area covered by a non-blocked neighboring transceiver, the call can be re-established. This will occur in the handover area between the cells. Here an additional advantage of the separate areas emerges: without the explicit modeling of the handover areas we cannot distinguish between redial-calls originating from capacity restrictions in different surrounding cells.

A further advantage of the handover areas is that the redial rate required to obtain a product form distribution is smaller than the corresponding redial rate in a model without the explicit use of handover areas, i.e., when only the total number of calls in the cells is counted. This can easily be verified in a model with homogeneous rates (equal routing probabilities from cell r to all cells $s \in B_r$).

2.3 The normalizing constant and the blocking probabilities

This subsection derives formulas for the normalizing constant and the blocking probabilities. Furthermore, we show that in a network with standard fixed channel allocation, the normalizing constant and the blocking probabilities can be computed explicitly.

THE NORMALIZING CONSTANT. The normalizing constant, G, can be expressed in k. Recall that $S = \{m : Ak \leq C\}$, and that k is defined as $k_r = m_{rr} + \sum_{s \in B_r} m_{rs}$, $r = 1, \ldots, R$. This gives $G =$

$$= \sum_{m} \prod_{r=1}^{R} \left(\frac{c_{rr}}{\mu_{rr}}\right)^{m_{rr}} \frac{1}{m_{rr}!} \prod_{s \in B_r} \left(\frac{c_{rs}}{\mu_{rs}}\right)^{m_{rs}} \frac{1}{m_{rs}!} \mathbf{1}(m \in S)$$

$$= \sum_k \left(\sum_{\{m: m_{rr} + \sum_{s \in B_r} m_{rs} = k_r,\ r=1,\ldots,R\}} \prod_{r=1}^{R} \left(\frac{c_{rr}}{\mu_{rr}}\right)^{m_{rr}} \frac{1}{m_{rr}!} \prod_{s \in B_r} \left(\frac{c_{rs}}{\mu_{rs}}\right)^{m_{rs}} \frac{1}{m_{rs}!} \right) \mathbf{1}(Ak \leq C)$$

$$= \sum_k \prod_{r=1}^{R} \frac{\nu_r^{k_r}}{k_r!} \mathbf{1}(Ak \leq C), \tag{8.3}$$

where we have defined

$$\nu_r := \frac{c_{rr}}{\mu_{rr}} + \sum_{s \in B_r} \frac{c_{rs}}{\mu_{rs}}, \quad r = 1, \ldots, R.$$

The simplification of the expression for G is a consequence of our definition of the handover areas: the calls in the handover area between cell r and cell s are separated into calls in the area rs, served by transceiver r, and calls in handover area sr, served by transceiver s.

THE FRESH CALL BLOCKING PROBABILITY. The probability that a fresh call in cell i is blocked, say B_i^I, equals the fraction of fresh calls that is blocked, i.e., the conditional probability that a fresh call arrives and is blocked given that a fresh call is generated. As fresh calls arrive in a Poisson stream, the fresh call blocking probability for a call in cell i is simply the probability of the states m such that $A(k + e_i) \not\leq C$:

$$B_i^I = \sum_{m: A(k+e_i) \not\leq C} \pi(m) = \left(\sum_{k \in T_i} \prod_{r=1}^{R} \frac{\nu_r^{k_r}}{k_r!} \right) \Big/ \left(\sum_{k \in U} \prod_{r=1}^{R} \frac{\nu_r^{k_r}}{k_r!} \right), \tag{8.4}$$

where T_i and U are defined by

$$T_i := \{k : Ak \leq C, A(k + e_i) \not\leq C\}, \quad U := \{k : Ak \leq C\}.$$

Again, observe that B_i^I is expressed in k only.

THE HANDOVER BLOCKING PROBABILITY. The probability that a handover from cell i to cell j is blocked, say B_{ij}^H, equals the fraction of handovers from cell i to cell j that is lost due to capacity constraints in cell j. This blocking probability is the conditional probability that a handover from cell i to cell j is attempted and lost given that such a handover is attempted. This is the ratio of two Palm distributions, cf. Baccelli and Brémaud [11]. The probability flux of handover attempts from cell i to cell j, say P_{ij}^H, is

$$\sum_{m \in S} \pi(m)\{q(m, m - e_{ij} + e_{ji}) + q(m, m - e_{ij})\} = G^{-1} c_{ij} p_{ij,ji} \sum_k \prod_{r=1}^{R} \frac{\nu_r^{k_r}}{k_r!} \mathbf{1}(A(k + e_i) \leq C).$$

The probability flux for lost handover attempts from cell i to cell j, say P_{ij}^L, is

$$\sum_{m \in S} \pi(m) q(m, m - e_{ij}) = G^{-1} c_{ij} p_{ij,ji} \sum_k \prod_{r=1}^{R} \frac{\nu_r^{k_r}}{k_r!} \mathbf{1}(A(k + e_i) \leq C) \mathbf{1}(A(k + e_j) \not\leq C).$$

The handover blocking probability B_{ij}^H equals the ratio of the above two expressions:

$$B_{ij}^H = P_{ij}^L / P_{ij}^H = \left(\sum_{k \in T_{ij}} \prod_{r=1}^{R} \frac{\nu_r^{k_r}}{k_r!} \right) \Big/ \left(\sum_{k \in U_i} \prod_{r=1}^{R} \frac{\nu_r^{k_r}}{k_r!} \right), \tag{8.5}$$

where T_{ij} and U_i are defined by

$$T_{ij} := \{k : Ak \leq C, A(k+e_i) \leq C, A(k+e_j) \not\leq C\}, \quad U_i := \{k : A(k+e_i) \leq C\}.$$

A similar expression was derived in Pallant and Taylor [143]. In that reference, product forms are derived via the truncation of a reversible process to the state space $S = \{k : Ak \leq C\}$. As a consequence, the transitions resulting in a blocked handover are modeled as transitions from k to k, i.e., a handover that is blocked will continue to use its current channel. In our model a blocked handover from cell r results in a transition from state k to state $k - e_r$.

EXAMPLE 2.3. FIXED CHANNEL ALLOCATION. Consider a mobile network with standard fixed channel allocation: the number of channels available in cell r equals C_r. The matrix A is the identity matrix; the state space is $S = \{m : m_r \leq C_r,\ r = 1, \ldots, R\}$. For this model the normalizing constant and the blocking probabilities can be computed explicitly from (8.3), (8.4), and (8.5):

$$\begin{aligned}
G &= \sum_{k \in S} \prod_{r=1}^{R} \frac{\nu_r^{k_r}}{k_r!} = \sum_{k \geq 0} \prod_{r=1}^{R} \frac{\nu_r^{k_r}}{k_r!} \mathbf{1}(k_r \leq C_r) = \prod_{r=1}^{R} \left(\sum_{k_r=0}^{C_r} \frac{\nu_r^{k_r}}{k_r!} \right), \\
B_i^I &= \left(\sum_{k \in S} \prod_{r=1}^{R} \frac{\nu_r^{k_r}}{k_r!} \mathbf{1}(m_i = C_i) \right) \Big/ \left(\sum_{k \in S} \prod_{r=1}^{R} \frac{\nu_r^{k_r}}{k_r!} \right) = \frac{\nu_i^{C_i}}{C_i!} \Big/ \sum_{k_i=0}^{C_i} \frac{\nu_i^{k_i}}{k_i!}, \\
B_{ij}^H &= \left(\sum_{k \in S} \prod_{r=1}^{R} \frac{\nu_r^{k_r}}{k_r!} \mathbf{1}(k_j = C_j, k_i < C_i) \right) \Big/ \left(\sum_{k \in S} \prod_{r=1}^{R} \frac{\nu_r^{k_r}}{k_r!} \mathbf{1}(k_i < C_i) \right) = \frac{\nu_j^{C_j}}{C_j!} \Big/ \sum_{k_j=0}^{C_j} \frac{\nu_j^{k_j}}{k_j!}.
\end{aligned}$$

For fixed channel allocation the handover and fresh call blocking probabilities are given by the Erlang-B formula. In general, however, the summations involved in G, B_i^I, and B_{ij}^H cannot be simplified.

3 Efficient estimation of blocking probabilities

This section is about efficient estimation from the blocking probabilities (found in Section 2) from the product form. The technique used is basically the same as the one presented in Chapter 7 and [122]. However, we will explain the choice of the importance sampling distribution somewhat more carefully. We do that by using the standard problem of estimating rare event probabilities in the context of sample means of i.i.d. random variables,

see Subsection 3.1. In the second subsection, we will translate the blocking probabilities into this framework. We also comment on the differences between this method and other simulation techniques to capture these blocking probabilities, e.g. the one proposed by Harvey and Hills [83]. Furthermore, we show that the large deviations results provide information about the network given blocking, i.e., the most likely state of the cells surrounding a blocked cell.

3.1 Estimation of large deviations probabilities

Consider a sequence of independent real-valued random variables $X_1, X_2, \ldots$, having common density function $f(\cdot)$. We assume here that this density is continuous, but the same reasoning applies to general i.i.d. random variables. Let S_n be the partial sum of the first n terms. Under the condition of a finite first moment, μ, the sample mean S_n/n converges to μ ($n \to \infty$, in probability). Cramér's theorem [23, p. 9-10] states for all $a > \mu$

$$\lim_{n\to\infty} \frac{1}{n} \log \mathcal{P}(S_n/n \geq a) = -I(a) = -\sup_{\theta}(\theta a - \log M(\theta)), \tag{8.6}$$

where $M(\theta) := \int f(x) \exp[\theta x] \mathrm{d}x$. This result can be extended to R-dimensional random variables ($R \in \mathbb{N}$), with density $f : \mathbb{R}^R \to \mathbb{R}$. The multi-dimensional rate function and moment generating function are given by

$$\begin{aligned} I(a_1, \ldots a_R) &= \sup_{\theta} \left(\sum_{i=1}^{R} \theta_i a_i - \log M(\theta) \right), \\ M(\theta) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \cdots, x_R) \exp\left[\sum_{i=1}^{R} \theta_i x_i \right] \mathrm{d}x_1 \cdots \mathrm{d}x_R. \end{aligned}$$

Cramér's theorem (8.6) states that $\mathcal{P}(S_n/n \geq a) = \eta(n) e^{-nI(a)}$ for a known number $I(a)$, and an unknown function $\eta(\cdot)$ with $\log \eta(n) = o(n)$ (where $n \to \infty$). So, in fact, the theorem yields only rough characteristics of the probability of our interest (namely its decay rate), but we have no approximation of the probability itself. Clearly, for the estimation of small probabilities it is infeasible to use crude Monte Carlo techniques, since the number of random numbers to be generated is huge. Then variance reduction techniques such as the importance sampling method can be very useful in order to develop efficient simulation techniques.

Under importance sampling, S_n is simulated by using a probability model that differs from the 'real' one. Formally, the procedure works as follows. We want to estimate $\mathcal{P}(S_n/n \geq a)$, which can be written in terms of the expected value of the indicator function $I_n = \mathbf{1}\{S_n/n \geq a\}$:

$$\mathcal{P}(S_n/n \geq a) = \int_{na}^{\infty} f_{S_n}(x) \mathrm{d}x = \mathrm{E}^{(\mathcal{P})}(I_n),$$

$f_{S_n}(\cdot)$ denoting the density of the convolution of n i.i.d. X_i's. Using crude Monte Carlo estimation, we would sample I_n a number of times (under the original measure $\mathcal{P}$). The average value is an unbiased estimator. Under importance sampling, we sample the I_n according to a probability measure $\mathcal{Q}$, i.e., S_n having density $g_{S_n}(\cdot)$. Each sample is multiplied by a likelihood, accounting for the difference between the probability measures $\mathcal{P}$ and $\mathcal{Q}$. Given that S_n attains value x under the new density $g_{S_n}(\cdot)$, when $\mathcal{P}$ is absolutely continuous w.r.t. $\mathcal{Q}$, the likelihood L can be defined as

$$L_{S_n}(x) = L := \frac{\mathrm{d}\mathcal{P}}{\mathrm{d}\mathcal{Q}}(x) = \frac{f_{S_n}(x)}{g_{S_n}(x)}.$$

It is elementary to show that indeed $\mathrm{E}^{(\mathcal{Q})}(LI_n) = \mathrm{E}^{(\mathcal{P})}(I_n)$. As a consequence, observing realizations of LI_n, which are sampled under $\mathcal{Q}$, their average is an unbiased estimator.

CHANGE OF MEASURE. The change of measure can be used to obtain variance reduction. Obviously, we want to capture the alternative measure $\mathcal{Q}$ that is endowed with the best variance properties, i.e., the $\mathcal{Q}$ that minimizes

$$\mathrm{Var}^{(\mathcal{Q})}(LI_n) = \mathrm{E}^{(\mathcal{Q})}(L^2 I_n) - (\mathrm{E}^{(\mathcal{Q})}(LI_n))^2.$$

The best choice of $\mathcal{Q}$, say $\mathcal{Q}^{\mathrm{o}}$, results in a zero variance (cf. [159, p. 122-123]). Under this optimal measure $\mathcal{Q}^{\mathrm{o}}$, the density (of S_n) is

$$g_{S_n}(x) := \begin{cases} f_{S_n}(x)/\mathcal{P}(S_n/n \geq a), & \text{if } S_n/n \geq a, \\ 0, & \text{else.} \end{cases} \tag{8.7}$$

It can be checked that $\mathrm{E}^{(\mathcal{Q}^{\mathrm{o}})}(L^2 I_n) = (\mathrm{E}^{(\mathcal{Q}^{\mathrm{o}})}(LI_n))^2 = \mathcal{P}^2(S_n/n \geq a)$, yielding variance 0. However, this new density is infeasible for two reasons. (i) The distribution of the X_i being being known does not imply that the density of S_n is explicitly known. (ii) More importantly, the probability of our interest $\mathcal{P}(S_n/n \geq a)$ is required! As the optimal $\mathcal{Q}^{\mathrm{o}}$ cannot be implemented, we relax our optimality criterion somewhat to the notion of *asymptotic optimality*. This can be introduced as follows.

Variances are non-negative. Therefore,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathrm{E}^{(\mathcal{Q})}(L^2 I_n) \geq \liminf_{n\to\infty} \frac{2}{n} \log \mathrm{E}^{(\mathcal{Q})}(LI_n) = \liminf_{n\to\infty} \frac{2}{n} \log \mathcal{P}(S_n/n \geq a) = -2I(a),$$

where the last equality is due to Cramér's theorem (8.6). This lower bound for the decay rate of the second moment of the importance sampling estimator holds for all alternative measures $\mathcal{Q}$. A measure is called asymptotically optimal (a.o.), when for this measure the above lower bound is attained.

Consider the alternative measures that keep the random walk structure, i.e., measures consisting of i.i.d. increments. We show that there is an a.o. change of measure $\mathcal{Q}^{\mathrm{ao}}$ within

the class of *exponentially twisted* densities ('θ-twisted densities') of the increments:

$$g_\theta(x) := \frac{f(x)e^{\theta x}}{M(\theta)}. \tag{8.8}$$

The mean of a single increment under the 'θ-twisted density' is

$$\int_{-\infty}^{\infty} x g_\theta(x) \mathrm{d}x = \int_{-\infty}^{\infty} \frac{x f(x) e^{\theta x}}{M(\theta)} \mathrm{d}x = \frac{M'(\theta)}{M(\theta)}.$$

For each a, under weak conditions (e.g., steepness [23], p. 13), we can find $\theta = \theta(a)$ such that this 'twisted mean' equals a, i.e., yielding a new density with mean a. It can be shown that $I(a) = \theta(a)a - \log M(\theta(a))$, and that $a > \mu$ implies $\theta(a) > 0$. If the increments are distributed according to measure $\mathcal{Q}^{\mathrm{ao}}$, under which the X_i have densities $g_{\theta(a)}(\cdot)$,

$$L(X_1, \ldots, X_n) = L := \left(\frac{f(X_1)}{g_{\theta(a)}(X_1)}\right) \cdots \left(\frac{f(X_n)}{g_{\theta(a)}(X_n)}\right) = \frac{M^n(\theta(a))}{e^{\theta(a) S_n}},$$

invoking (8.8). Consequently,

$$\mathrm{E}^{(\mathcal{Q}^{\mathrm{ao}})}(L^2 I_n) = \mathrm{E}^{(\mathcal{Q}^{\mathrm{ao}})}\left(\frac{M^{2n}(\theta(a))}{e^{2\theta(a) S_n}} I_n\right) \leq \left(\frac{M(\theta(a))}{e^{\theta(a)a}}\right)^{2n} = e^{-2nI(a)}.$$

Here we explicitly use that $\{I_n = 1\} = \{S_n \geq na\}$. This gives $\limsup \frac{1}{n} \log \mathrm{E}^{(\mathcal{Q}^{\mathrm{ao}})}(L^2 I_n) \leq -2I(a)$, yielding that the new density $g_{\theta(a)}(\cdot)$ is a.o.

The asymptotically optimal procedure explained above can be extended somewhat in the following way. Suppose that we want to estimate $\mathcal{P}(S_n/n \in W)$ for some set W not containing mean μ. Cramér's theorem states that the decay rate of this probability is $-\inf_{a \in W} I(a)$. Suppose that this infimum is attained by $a^* \in W$. Then the importance sampling has to be performed under density $g_{\theta(a^*)}(\cdot)$.

REMARK 3.1. HEURISTIC EXPLANATION. From extensive analytical/simulation studies, cf. [23, Ch. VIII], [85], it appears that this 'exponentially twisted' alternative measure performs excellent. Of course, we want to know *why* this change of measure has very good variance performance. To explain this, we return to the zero-variance importance sampling estimator mentioned above (8.7), where the event $\{S_n/n \geq a\}$ is now replaced by $\{S_n/n \in W\}$. As noted earlier, this new density cannot be used, since the probability to be estimated is required. It can be seen that $g_{S_n}(\cdot)$ is in fact the distribution of S_n, given $S_n/n \in W$. This distribution being unknown, we might use some distribution which looks very much the same.

Based on large deviations results, e.g. Sanov's theorem [166, Section 2.4], it can be argued that the increments $X_1, \ldots, X_n$ have, given $S_n/n \in W$, *asymptotically* (n large) densities $g_{\theta(a^*)}(\cdot)$. Indeed, under this density, each increment has mean a^*, so according to

the law of large numbers, S_n/n converges to $a^* \in W$. The number a^* has the interpretation of being the most likely value of S_n/n, given $S_n \in W$. We conclude that, using $g_{\theta(a^*)}(\cdot)$ for the X_i, the resulting density of S_n will be similar to (8.7). In this way, it is heuristically explained why simulation with this density works well. A more rigorous treatment of this argument can be found in [9].

3.2 Estimation of blocking probabilities

After the above explanation of efficient importance sampling, we return to the framework of product form distributions. This subsection deals with an efficient algorithm to estimate the blocking probabilities by simulation.

The crucial step is the interpretation of both blocking probabilities as ratios of two multivariate Poisson probabilities, i.e., for sets T, U they can be written as

$$\frac{\sum_{k\in T}\pi(k)}{\sum_{k\in U}\pi(k)} = \left(\sum_{k\in T}\prod_{r=1}^{R}\frac{\nu_r^{k_r}e^{-\nu_r}}{k_r!}\right)\Big/\left(\sum_{k\in U}\prod_{r=1}^{R}\frac{\nu_r^{k_r}e^{-\nu_r}}{k_r!}\right). \tag{8.9}$$

To put it more formally, let $X(\nu)$ be a R-dimensional random variable, having independent marginals, that are Poisson with mean ν_r, $r = 1, \ldots, R$. Our performance measures equal $\mathcal{P}(X(\nu) \in T)/\mathcal{P}(X(\nu) \in U)$, and they can be estimated by estimating both numerator and denominator. Methods to find confidence intervals for the ratio estimator are straightforward.

By scaling, we can transform the numerator and denominator into probabilities that can be estimated very efficiently by the importance sampling technique described in Subsection 3.1. To this end, replace the ν_i by $n\nu_i$, and replace the capacity vector C by nC. Then the sets T and U depend on n as well: $T^{(n)}$ and $U^{(n)}$. This scaling of input and capacity is used frequently in the analysis of blocking probabilities in large circuit-switched networks. Applying this scaling, asymptotic analysis ($n \to \infty$) becomes analytically tractable, see for instance Kelly [100].

The numerator of (8.9) can now be interpreted as the probability that the sum of n independent Poisson$(\nu_1, \ldots, \nu_R)$ samples, say S_n, lies in $T^{(n)}$:

$$\mathcal{P}(X(n\nu) \in T^{(n)}) = \mathcal{P}(S_n \in T^{(n)}).$$

For $T^{(n)} = T_j^{(n)}$ and $T^{(n)} = T_{i,j}^{(n)}$ we obtain the limits (for $n \to \infty$)

$$\begin{aligned}
&\{k/n : k \in T_j^{(n)}\} = \{x : x = k/n,\ Ax \le C,\ Ax + Ae_j/n \not\le C\} \\
\to\ &\overline{T}_j := \{x \ge 0 : Ax \le C,\ \exists\, k \text{ with } A_{kj} > 0 \text{ and } (Ax)_k = C_k\}, \qquad (8.10)\\
&\{k/n : k \in T_{ij}^{(n)}\} = \{x : x = k/n,,\ Ax + Ae_i/n \le C,\ Ax + Ae_j/n \not\le C\} \\
\to\ &\overline{T}_{ij} := \{x \ge 0 : Ax \le C,\ \exists\, k \text{ with } A_{kj} > A_{ki} \text{ and } (Ax)_k = C_k\}. \qquad (8.11)
\end{aligned}$$

From the theory of Subsection 3.1 we know how to efficiently estimate $\mathcal{P}(S_n/n \in \overline{T}_j)$ by using an alternative distribution. However, because of the relation (8.10) between $T_j^{(n)}$ and $\overline{T}_j$, it seems to make sense to estimate $\mathcal{P}(S_n \in T_j^{(n)})$ applying the *same alternative distribution.* The same reasoning applies to $T_{ij}^{(n)}$ and $\overline{T}_{ij}$, and also applies to the denominator $\mathcal{P}(S_n \in U^{(n)})$. Then $\overline{U} := \{x \geq 0 : Ax \leq C\}$ for both fresh call blocking and handover blocking.

The (multi-dimensional) large deviations rate function of a Poisson random variable with mean $\nu_1, \ldots, \nu_R$ can be calculated easily. First note that

$$M_r(\theta_r) = \sum_m \nu_r^m e^{-\nu_r} e^{\theta_r m}/m! = \exp(\nu_r(e^{\theta_r} - 1)).$$

Then, as the rth component of $\theta(a)$ is given by $\log(a_r/\nu_r)$,

$$I(a) = \sup_\theta \left(\sum_{r=1}^R (\theta_r a_r - \log M_r(\theta_r)) \right) = \sum_{r=1}^R \left(a_r \log\left(\frac{a_r}{\nu_r}\right) - a_r + \nu_r \right).$$

This large deviations rate function $I(\cdot)$ has to be minimized over $\overline{T}$ ($\overline{U}$, respectively) in order to find a^*, i.e., the most probable value of S_n/n in set $\overline{T}$ ($\overline{U}$). The importance sampling distribution (8.8) is a Poisson distribution with parameter a^*:

$$g_{\theta(a^*)}(k) = \prod_{r=1}^R \left(\frac{\nu_r^{k_r} e^{-\nu_r}}{k_r!} \times \frac{e^{\theta(a^*)_r k_r}}{M_r(\theta(a^*)_r)} \right) = \prod_{r=1}^R \frac{(a_r^*)^{k_r} e^{-a_r^*}}{k_r!}.$$

Typically, mobile networks have a relatively light load, i.e., $A\nu \leq C$; the network is capable of handling the normal user level. It can be easily checked that the infimum of the large deviations rate function over $\overline{U}$ is 0, attained for $a^* = \nu$. In other words: the denominator can be estimated under the original Poisson measure. In contrast, for the numerator, the infimum over $\overline{T}$ as given by (8.10) or (8.11) will in general be positive, since typically $\nu \notin \overline{T}$. Consequently, a^* (the most probable value of S_n/n in the set $\overline{T}$) will not equal ν. Moreover, given a 'type j fresh-call-blocking' (or '$i \to j$ handover blocking'), the cells corresponding to the constraints h with $(Aa^*)_h = C_h$ are the ones that are most likely to be full, recalling the interpretation of a^*. So *finding a^* in fact has two major merits:*

1. It serves as the Poisson parameter to perform the importance sampling for estimation of the numerator.

2. It gives insight into the network in case of blocking. This allows for bottleneck-analysis and enables us to identify those C_j that have to be increased to improve network performance. We will give examples on this in the next section.

REMARK 3.2. COMPARISON WITH THE HARVEY-HILLS ALGORITHM. From (8.9) we obtain that the fresh call and handover blocking probabilities can be written as

$$\mathcal{P}(X(\nu) \in T)/\mathcal{P}(X(\nu) \in U) = \mathcal{P}(X(\nu) \in T \mid X(\nu) \in U),$$

where the equality is due to $T \subset U$. A sample from $(X(\nu) \mid X(\nu) \in U)$ can be generated by an acceptance/rejection technique: samples from $X(\nu)$ are drawn until a sample lies in U. An unbiased estimator of the blocking probability is the fraction of these samples that are in T as well. Basically, this is the procedure proposed by Harvey and Hills [83] for estimating blocking probabilities in loss networks. As reported in [67] and [64], this approach is also frequently used in mobile networks.

An advantage of the Harvey-Hills technique is that one estimation is involved, whereas in our approach the numerator and denominator are estimated separately (using different probability measures). The Harvey-Hills method has one important disadvantage: it can be rather slow, both in networks with relatively light load and in networks with heavy load. Under light load ($A\nu \leq C$) a lot of samples from the Poisson(ν) distribution are in U, so it is easy to generate samples from the conditional distribution. However, since T is that boundary of U at which blocking occurs, only a very tiny fraction of the samples in U will be in T as well, due to the light load. Under heavy load ($A\nu \not\leq C$; which is rather unusual in mobile networks, but not unusual in loss networks) very few samples will be in U, so it is very time consuming to even get a sample from the conditional distribution. Our approach overcomes this drawback: due to the 'shift' of the Poisson distribution, we easily generate 'relevant samples'. We use different measures for both numerator and denominator, namely the ones that are best for their estimation.

It should be noted that the order of the blocking probabilities of interest (say 10^{-3}, usually in mobile networks) is not particularly suitable to use large deviations techniques. However, a separate (efficient) estimation of numerator and denominator, as described above, gives a very considerable speed-up. The insight into the network given overflow is an important additional merit.

More details on our simulation method can be found in Mandjes [122] and Chapter 7 of this monograph, where the estimation of blocking probabilities in circuit-switched networks is considered. There, a thorough justification of the importance sampling technique is given, including a number of additional accelerations. The importance sampling distribution can be found by solving a mathematical programming problem: the large deviations rate function must be minimized over $\overline{T}$ and $\overline{U}$. This minimization can be simplified substantially by considering its dual [122].

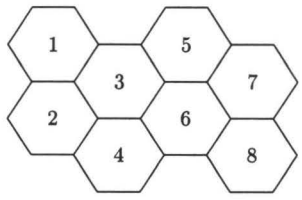
4 Numerical examples

This section deals with two numerical examples illustrating the efficiency of the simulation technique described in Section 3 for the computation of blocking probabilities. In section 2 we showed that fixed channel allocation yields trivial – Poisson – probabilities. In the examples below we turn our attention to more complicated state spaces. The first example is adopted from [65], where the amount of interference between adjacent cells is captured for certain CDMA mobile communication systems. The second example is taken from [67]. There a 49-cell layout with 7-cell reuse groups was described. For both models we will compute blocking probabilities. The underlying model is the model of Section 2. We will assume that a solution ν is obtained from the analysis as presented there. All estimated probabilities have 95% confidence and 10% relative precision (ratio of confidence interval half-length and estimate).

4.1 Interference

In the model described in Everitt and Evans [65], the authors explicitly try to capture blocking due to interference, and give numbers κ_i indicating the *effective interference* between cells that are 'neighbor in the ith degree'. The interference is a consequence of the 'overlap' between the areas covered by neighboring transceivers. On the one hand, the capacity available in the cells increases because channels from neighboring transceivers can be used to carry a call, on the other hand, the number of calls that can be carried by the transceiver in a cell decreases due to the interference with neighboring cells. The bandwidth of a neighboring transceiver cannot be used in the whole cell. Therefore, κ_2, the interference with a direct neighbor, must be less than 1. For the 8-cell layout of Figure 2 in [65], see Figure 2, Everitt and Evans derive the following constraint matrix:

Figure 2: Layout interference model.



$$A = \begin{pmatrix} \kappa_1 & \kappa_2 & \kappa_3 & \kappa_2 & \kappa_3 & \kappa_3 & 0 & 0 \\ \kappa_2 & \kappa_1 & \kappa_2 & \kappa_2 & \kappa_3 & \kappa_3 & 0 & 0 \\ \kappa_3 & \kappa_2 & \kappa_1 & \kappa_2 & \kappa_3 & \kappa_2 & \kappa_3 & \kappa_3 \\ \kappa_2 & \kappa_2 & \kappa_2 & \kappa_1 & \kappa_2 & \kappa_2 & \kappa_3 & \kappa_3 \\ \kappa_3 & \kappa_3 & \kappa_3 & \kappa_2 & \kappa_1 & \kappa_2 & \kappa_2 & \kappa_3 \\ \kappa_3 & \kappa_3 & \kappa_2 & \kappa_2 & \kappa_2 & \kappa_1 & \kappa_2 & \kappa_2 \\ 0 & 0 & \kappa_3 & \kappa_3 & \kappa_2 & \kappa_2 & \kappa_1 & \kappa_2 \\ 0 & 0 & \kappa_3 & \kappa_3 & \kappa_3 & \kappa_2 & \kappa_2 & \kappa_1 \end{pmatrix}.$$

Assume that the solution of the traffic equations give that $\nu_i = 15$ for all i (homogeneous load). Let $C_i = 65$, $i = 1, \ldots, 8$, and $\kappa_1 = 1$, $\kappa_2 = 0.5$, $\kappa_3 = 0.25$. With these parameters we obtain that $A\nu = (41.25,\ 45,\ 52.5,\ 60,\ 52.5,\ 60,\ 45,\ 41.25) < C$. Therefore, the denominator of the blocking probabilities can be estimated under the original measure, whereas the numerator must be simulated using importance sampling. As examples, we treat the blocking probability for handovers from cell 3 to cell 4 and from cell 4 to cell 3, respectively.

For the denominator, simulation under the original measure yields $\mathcal{P}(X(\nu) \in U_3) = 0.714$ and $\mathcal{P}(X(\nu) \in U_4) = 0.702$. The numerator corresponds to a rare event and must be estimated under an alternative Poisson measure. To obtain the parameter of this Poisson measure, the large deviations rate function has to be minimized over the set

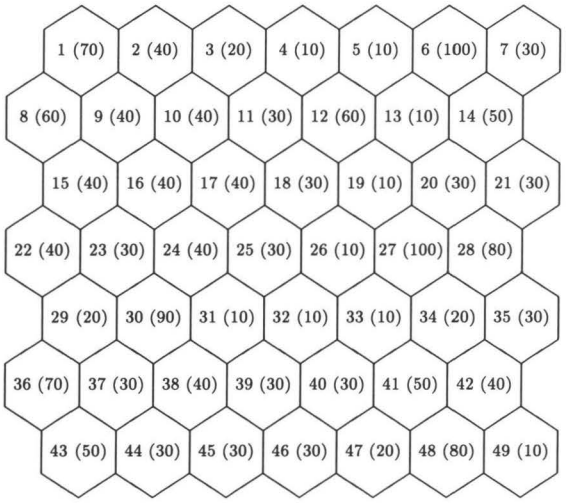
$$S(I) := \{x \geq 0 : Ax \leq C\} \cap \left(\bigcup_{r \in I} \{x \geq 0 : (Ax)_r = C_r\} \right), \tag{8.12}$$

where I is determined by (8.11): $I_i = \{k : A_{ki} > 0\}$, $I_{ij} = \{k : A_{kj} > A_{ki}\}$. For the '3→4' handover $A_{k4} > A_{k3}$ for $k = 1, 4, 5$, and thus $I_{34} := \{1, 4, 5\}$. For the '4→3' handover, $I_{43} := \{3\}$. The optima, 4.5813 and 0.3399, respectively, are attained for a^* equal to

$$\begin{array}{cccccccc} (16.038, & 16.038, & 16.038, & 17.149, & 16.038, & 16.038, & 15.511, & 15.511), \\ (14.265, & 18.124, & 27.371, & 13.566, & 13.346, & 14.841, & 14.920, & 14.920), \end{array}$$

respectively. It can be verified easily, that in the first case only $(Aa^*)_4 = 65$. Consequently, given blocking of a handover from 3 to 4, the constraint associated with cell 4 is most likely to block. Simulation yields an estimate $1.48 \cdot 10^{-2}$ of the numerator, and therefore an estimator of the blocking probability is $\hat{B}^H_{3,4} = 2.07 \cdot 10^{-2}$. With respect to blocking of a handover from cell 4 to cell 3, $(Aa^*)_3 = (Aa^*)_4 = (Aa^*)_6 = 65$. This can be interpreted as follows: given that a handover from cell 4 to cell 3 is blocked, the (most probable) cells with high load are cells 3, 4, and 6. The numerator is estimated by $1.24 \cdot 10^{-6}$, and therefore an estimator of the blocking probability is $\hat{B}^H_{4,3} = 1.77 \cdot 10^{-6}$.

Figure 3a: Reuse group model. Cells of the network. (.) represents load of the cell.



This example shows that the importance sampling technique not only allows us to estimate blocking probabilities, but also gives insight into the behavior of the network upon blocking. Depending on the value of components of the vector Aa^* blocking can be shown to be a local effect (such as for the case of $3 \rightarrow 4$ handover blocking) or a global effect (such as for $4 \rightarrow 3$ handover blocking). This will be further illustrated in the next example.

4.2 Reuse groups

Consider the 49-cell system of Figure 1 in [67]. Let the system have a total of 350 available channels. Cells are grouped in 'reuse clusters' of size 7 that share the channels, i.e., each reuse cluster of size 7 can accommodate at most 350 calls. Cells and reuse clusters are drawn in Figure 3. For example, reuse cluster 6 consists of cells 9, 10, 15, 16, 17, 23, 24. For this system the state space constraints are linear, for example $m_9 + m_{10} + m_{15} + m_{16} + m_{17} + m_{23} + m_{24} \leq 350$. Cell 7 and 49 are not present in any constraint; these cells can be deleted from our model. The resulting network has 25 constraints. Thus, the matrix A has dimension 25×47, and consists of 0's and 1's only. Notice that we did not take into account the influence of edge effects. This can be overcome by introducing a toroidal reuse structure; it results in a similar constraint matrix A, see [64].

Figure 3b: Reuse group model. Reuse groups. (.) represents load of the group.

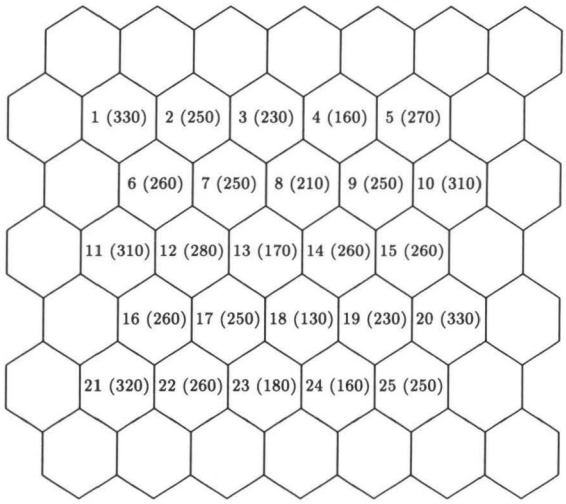


Table 1: ν parameters of Example 2.

1	70	74.242	70	62.625	26	10	10	10	10
2	40	42.242	40	35.786	27	100	100	112.251	100
3	20	20	20	20	28	80	80	89.800	80
4	10	10	10	10	29	20	20	20	20
5	10	10	10	10	30	90	90	90	90
6	100	100	100	100	31	10	10	10	10
7	30	30	30	30	32	10	10	10	10
8	60	63.636	60	53.678	33	10	10	9.863	10
9	40	42.424	40	49.478	34	20	20	19.727	20
10	40	42.424	40	49.478	35	30	30	29.590	30
11	30	30	30	30	36	70	70	70	70
12	60	60	60	60	37	30	30	30	30
13	10	10	11.381	10	38	40	40	40	40
14	50	50	56.903	50	39	30	30	30	30
15	40	42.424	40	49.478	40	30	30	30	30
16	40	42.424	40	49.478	41	50	50	49.316	50
17	40	40	40	55.305	42	40	40	39.453	40
18	30	30	30	30	43	50	50	50	50
19	10	10	11.381	10	44	30	30	30	30
20	30	30	34.142	30	45	30	30	30	30
21	30	30	34.142	30	46	30	30	30	30
22	40	40	40	40	47	20	20	20	20
23	30	30	30	41.479	48	80	80	80	80
24	40	40	40	55.305	49	10	10	10	10
25	30	30	30	30					

In contrast with Everitt *et al.* [67], [64], we assume that the network load is heterogeneous resulting in 'busy cells' (with a high ν-value, say 90 or 100) and 'quiet cells' (with a low ν-value, say 10 or 20). These parameters are listed in the first column of Table 1.

It can be checked that $A\nu < C$, so the network has a 'light load'.

The denominators of the blocking probabilities can be estimated under the original Poisson measure: they do not correspond to rare events. Straightforward simulation yields $\mathcal{P}(X(\nu) \in U) = 0.651$, $\mathcal{P}(X(\nu) \in U_i) = 0.651$, for i not in reuse cluster 1, 20, or 21, and $\mathcal{P}(X(\nu) \in U_i) = 0.636$, otherwise.

The numerators of the blocking probabilities correspond to rare event probabilities, and are estimated using an optimal change of measure. The large deviations rate function has to be minimized over $S(I)$, as defined in (8.12). For $B^H_{i,j}$ we find that $I_{ij} := \{k : A_{kj} = 1 \text{ and } A_{ki} = 0\}$, and for B^I_j we obtain that $I_j := \{k : A_{kj} = 1\}$. The minimum of the large deviations rate function can be calculated more efficiently by using the dual representation of the optimization problem [122]. Then, the number of variables is 25 instead of 47, and, apart from that, the feasible region has a more manageable form. A complete list of all blocking probabilities can now be obtained. As an illustration, we will indicate how to estimate fresh call blocking probabilities B^I_9, B^I_{14} and handover blocking probabilities $B^H_{3,10}$, $B^H_{29,23}$.

With respect to B^I_9, the set $I_9 := \{1, 2, 6\}$. Applying numerical techniques, we get minimum 0.5942, yielding the alternative Poisson parameters a^* as displayed in the second column of Table 1. We find that, given a fresh call blocking in cell 9, reuse cluster 1 is most likely to block: $(Aa^*)_1 = 350$. Of course, this is logical, since reuse cluster 1 has a higher load (330) than clusters 2 and 6 (250 and 260, respectively). In this case, blocking has an exclusively local character: given a cell 9 fresh call blocking, cluster 1 will be the only cluster to be overloaded. An estimate of the numerator is $8.56 \cdot 10^{-3}$, yielding $\hat{B}^I_9 = 1.31 \cdot 10^{-2}$.

For $B^H_{3,10}$ the minimization must be performed over $S(I)$ (as in (8.12)) with $I_{3,10} := \{1, 6, 7\}$. The parameters of the alternative measure are displayed in the second column of Table 1. We find that $\hat{B}^H_{3,10} = 1.30 \cdot 10^{-2}$. As above, given this blocking, only cluster 1 is likely to be full; blocking is local.

Both other blocking probabilities do not correspond to overload of only one reuse group. Consider first $B^H_{29,23}$, where $I_{29,23} := \{6, 12\}$. We find minimum 12.3377, attained for a^* as listed in the third column. We see that both $(Aa^*)_1$ and $(Aa^*)_6$ equal 350. Therefore, given blocking of a handover from 29 to 23, reuse group 1 as well as 6 are likely to block. The numerator is estimated by $4.31 \cdot 10^{-7}$, yielding $\hat{B}^H_{29,23} = 6.78 \cdot 10^{-7}$. Even these very small probabilities are estimated in a relative small amount of time: in the order of 10 to 30 seconds on a 586 PC.

Finally, we examined B^I_{14}. Taking $I_{14} := \{5, 10\}$, the minimum is 2.4763, for the a^* in the last column of Table 1. Again multiple reuse clusters are likely to be overloaded: cluster 10 as well as 20. The numerator is estimated by $1.72 \cdot 10^{-3}$, yielding $\hat{B}^I_{14} =$

$2.64 \cdot 10^{-3}$.

5 Conclusion

This chapter has shown that performance measures for mobile communication networks can be computed using importance sampling. To this end, product form results for the equilibrium distribution of the number of calls in the cells have been extended to also include non-reversible routing of calls, and to include explicit modeling of handover areas. A protocol with redialing in case of congestion has been studied. The resulting product form distribution is shown to be most suitable for the application of large deviations results. Although the large deviations limiting regime for the computation of blocking probabilities is not reached (the blocking probabilities are too large), estimation of these probabilities using importance sampling is considerably faster than estimation using the Harvey-Hills algorithm. In addition, large deviations results provide insight into the network in case of blocking, allowing, for example, bottleneck analysis for mobile communication networks.

Directions for further research

Performance analysis under heavy-tailed burst lengths

Consider the workload model of Chapter 2. Take for simplicity the batch size $X \equiv 1$, so we are in the classical GI/G/1 context. Then a crucial assumption for the analysis to hold is that the characteristic equation can be solved. Therefore, in order to do that, we need that the moment generating function of the work S, brought along by one customer, is finite in a neighborhood of zero. However, suppose that this is *not* the case, e.g., if the distribution of S has a regularly varying tail. This means, for all $a > 0$ and an index ζ,

$$\lim_{s\to\infty} \frac{\mathcal{P}(S > sa)}{\mathcal{P}(S > s)} = a^{\zeta}.$$

Cohen [36] and Borovkov [17] succeeded in finding asymptotics of the probability of waiting time W exceeding B, having the form $KB^{1+\zeta}$ for some constant K.

Since the calculation of K is difficult, the question arises: can an efficient simulation procedure be developed in order to estimate this rare event probabilities? Asmussen and Binswanger [7] succeed in doing that for the M/G/1 case, by actually simulating the defective renewal process that can be embedded in the workload process. Asmussen and Klüppelberg [8] investigate the GI/G/1 case. They find the most probable trajectory to overflow, but do not provide an efficient simulation scheme. Of course, it would be interesting to solve this open problem.

A next step is to consider a number of on-off sources, where the on-time distribution has a regularly varying tail. Boxma [20] gives some first results on this model. He manages to find the asymptotics of the waiting time distribution. However, his assumptions are very restrictive: he assumes that the peak rate of every individual source is larger than the service rate. Brichet *et al.* [21] found that a queue fed by a large number of sources with regularly varying burst distributions has, under some additional assumptions, the Weibull-like tail (that was mentioned in Section 1.3 of this monograph).

Of course, it is interesting to examine whether the asymptotical results found by Boxma also hold under less restrictive conditions. Apart from that, efficient simulation methods can be of interest. An other interesting subject is the relation between heavy

tailed burst lengths and long-range dependence, see e.g. Parulekar and Makowski [146] and Likhanov, Tsybakov and Georganas [116].

Markov fluid models with many sources

Recent work of Botvich and Duffield [18] shows that $I^*(B)$, as defined in Chapter 4, can be expressed as follows. Define the finite time cumulant function and its convex conjugate as

$$\lambda_t(\theta) = \frac{1}{t} \log \mathrm{E} e^{\theta(A(t)-Ct)}, \qquad \lambda_t^*(x) := \sup_\theta (\theta x - \lambda_t(\theta)),$$

where $A(t)$ denotes the amount of fluid generated by one source during $[0, t]$. Then it can be shown that $I^*(B) = \inf_{t>0} t\lambda_t^*(B/t)$. From the results of Kesidis *et al.* [105], it follows that

$$\mathrm{E} \exp[\theta(A(t) - Ct)] = \pi^{\mathrm{T}} \exp[(\Lambda + (R - CI)\theta)t]1,$$

where we adopted the notation of Chapter 3 and 4, and 1 denotes the unit vector. So, in principle, the value of $I^*(B)$ can be calculated for all values of B.

However, the optimizations involved are rather complex, and the objective function cannot be given explicitly usually. Only in the limiting regimes $B = 0$ and $B = \infty$, Botvich and Duffield succeed in giving more explicit results on $I^*(B)$. They find the same value for $I^*(0)$ as we do, and they also conclude that $I^*(B) - \theta B$ tends to a constant that can be written as $-\lim_{t\to\infty} \log \mathrm{E} \exp[\theta^*(A(t) - Ct)]$. We find $-\log(\langle \pi, x\rangle \langle \pi, y\rangle)$ for the special case of fluid sources.

Comparing Chapter 4 of this monograph and the Botvich-Duffield paper, the latter shows that the obtained results are valid for a broad class of arrival processes. We, in Chapter 4, restricted ourselves to the case of Markov fluid input. Having chosen this concrete model, we were able to find more explicit results. Apart from the approximations of $I^*(B)$ for small and large B, we gave additional characteristics as the optimum path to overflow, the most probable distribution of entering H, and gave an elegant relation to the time-reversed process.

Having two characterizations of the function $I^*(\cdot)$ (namely the above optimization program and (4.10), we might hope to be able to solve a number of open questions. Interesting subjects for future research are the following. (i) Is the relation for $I^*(B)$, as found in [18], computationally tractable? Is it essentially simpler than (4.10) or computationally equivalent? (ii) In Chapter 4, we found properties concerning the function $I^*(B)$ for small and large B. Can more properties of the function $I^*(B)$ be found? A first result in this area is by Elwalid *et al.* [59]. Under assumption of reversible sources, they succeeded in finding some (upper and lower) bounds $g_1 B + g_2$ for $I^*(B)$. (iii) Based on this asymptotic relation of the loss probability we can execute call acceptance control. As treated

in Kelly [101], some general remarks can be made on the shape of the allowed region, but can it be calculated explicitly for realistic ATM models? Does this CAC perform significantly better than the one presented in Elwalid and Mitra [59], who only consider 'large buffer effective bandwidths'? Apart from that, it can also be compared with other CAC algorithms, as Kröner *et al.* [112], Gün and Guérin [81] and Mitrou *et al.* [131].

The Available Bit Rate service

As explained in the introduction of this monograph, we can distinguish between delay-sensitive (as voice or video), and loss-sensitive traffic. In order to enhance the network efficiency, a special service for loss-sensitive traffic can be developed: the ABR service, see e.g. Bonomi and Fendick [15], Smith, Adams, and Tagg [168]. Loosely speaking, the ATM model with ABR traffic can be described as follows.

The traffic is distinguished between real-time traffic and ABR traffic. Because of the delay-sensitivity of the former group, this traffic can use all the bandwidth that is available, but does not use the buffer. The latter group uses the remaining bandwidth, and if this is not sufficient, it uses the buffer.

The rate at which the ABR sources are allowed to send traffic is determined in the following way. As soon as the buffer contents exceeds a given threshold, a message is sent to the ABR sources to decrease their offered rate; if the buffer contents drops below another threshold, it can be increased again. A complicating factor is the propagation delay τ of this 'congestion status message'.

Ritter [151] considers a situation with only ABR sources; he finds cyclic phenomena, like Bonomi, Mitra, and Seery [16]. However, a more interesting situation is, of course, a system fed by real-time as well as ABR sources. In fact, the ABR buffer can be considered as a queue with a variable service rate (depending on the current traffic rate of the real-time sources), in which the current input rate is determined by the buffer contents τ units time ago. In order to guarantee the very stringent loss requirements of ABR traffic, it is very important to get insight into the tail behavior of this queue. It is interesting to examine whether large deviations analysis and importance sampling methods can be developed.

Large deviation analysis of a cell in a cellular network

In Chapter 8, we gave a large deviations analysis of the population of a cellular mobile communications network. Applying product form results, we developed an efficient simulation algorithm to capture blocking behavior.

An other interesting point of view is to consider only *one cell* of the network. Some

research has already been done to evaluate the performance of mechanisms that prioritize handover calls with respect to fresh calls. For example, the introduction of guard channels has been discussed, and redial mechanisms or waiting room for rejected calls.

A classical reference in this area is the paper by Guérin [79], finding explicit results for blocking probabilities. Chang *et al.* [26], Yoon and Un [189], Mandjes and Tutschku [126], Tran-Gia and Mandjes [176] consider more complicated models. Basically, they provide only numerical ways to calculate relevant performance measures. However, it would of course be interesting to see how these performance criteria depend on the model parameters. A way to investigate this is by a large deviations study: after a scaling of the model, the decay rate of the blocking probabilities might be expressed in terms of these model parameters. Such a study can shed some light on the question how to effectively improve the system performance.

Bibliography

[1] AALTO, S. Output from an A-M-S type fluid queue. *Proceedings ITC 14, Antibes/Juan-les-Pins, France* (1994), 421–430.

[2] ABATE, J., CHOUDHURY, G., AND WHITT, W. Asymptotics for steady-state tail probabilities in structured Markov queueing models. *Stochastic Models 10* (1994), 99–143.

[3] ABATE, J., AND WHITT, W. The Fourier-series method for inverting transforms of probability distributions. *Queueing Systems 10* (1992), 5–88.

[4] ANANTHARAM, V. How large delays build up in a GI/G/1 queue. *Queueing Systems 5* (1988), 345–386.

[5] ANICK, D., MITRA, D., AND SONDHI, M. Stochastic theory of data-handling system with multiple sources. *The Bell System Technical Journal 61* (1982), 1871–1894.

[6] ASMUSSEN, S. Conjugate processes and the simulation of ruin probabilities. *Stochastic Processes and their Applications 20* (1985), 213–229.

[7] ASMUSSEN, S., AND BINSWANGER, K. Simulation of ruin probabilities for subexponential claims. *Preprint* (1995).

[8] ASMUSSEN, S., AND KLÜPPELBERG, C. Large deviations results in the presence of heavy tails, with applications of insurance risk. *Journal of Applied Probability* (1996).

[9] ASMUSSEN, S., AND RUBINSTEIN, R. Steady state rare event simulation in queueing models and its complexity properties. In *Advances in Queueing: theory, methods and open problems*, J. Dshalalow, Ed. CRC Press, Boca Raton, Fa., 1995, pp. 429–462.

[10] AWATER, G. *Broadband Communication - modelling, analysis and synthesis of an ATM switching element.* PhD thesis, Delft, the Netherlands, 1994.

[11] BACCELLI, F., AND BRÉMAUD, P. *Elements of queueing theory: Palm-martingale calculus and stochastic recurrences.* Springer Verlag, Berlin, 1994.

[12] BAHADUR, R., AND RAO, R. R. On deviations of the sample mean. *The Annals of Mathematical Statistics 31* (1960), 1015–1027.

[13] BAIOCCHI, A., AND BLÉFARI-MELAZZI, N. An error-controlled approximate analysis of a stochastic fluid flow model applied to an ATM multiplexer with heterogeneous on-off sources. *IEEE/ACM Transactions on Networking 1* (1993), 628–637.

[14] BLONDIA, C. A discrete-time batch Markovian arrival process as B-ISDN traffic model. *Belgian Journal of Operations Research, Statistics and Computer Science 32* (1993), 3–23.

[15] BONOMI, F., AND FENDICK, K. The rate-based flow control framework for the available bit rate ATM service. *IEEE Network* (1995), March–April, 25–39.

[16] BONOMI, F., MITRA, D., AND SEEDY, J. Adaptive algorithms for feedback-based flow control in high-speed, wide-area ATM networks. *IEEE Journal on Selected Areas in Communications 13* (1995), 1267–1283.

[17] BOROVKOV, A. *Asymptotic methods in queueing theory.* Wiley, Chichester, 1984.

[18] BOTVICH, D., AND DUFFIELD, N. Large deviations, the shape of the loss curve, and economies of scale in large multiplexers. *Queueing Systems 20* (1995), 293–320.

[19] BOUCHERIE, R., AND MANDJES, M. Computation of performance measures for product form cellular mobile communication networks. *Preprint* (1996).

[20] BOXMA, O. Fluid queues and regular variation. *COST Closing Seminar, Issy-les-Moulineaux, France* (1996).

[21] BRICHET, F., ROBERTS, J., SIMONIAN, A., AND VEITCH, D. Heavy traffic analysis of a storage model with long range dependent on/off sources. *Report FT-CNET* (1995).

[22] BRUNEEL, H., AND KIM, B. *Discrete-Time Models for Communication Systems, including ATM.* Kluwer, Dordrecht, 1993.

[23] BUCKLEW, J. *Large Deviation techniques in Decision, Simulation, and Estimation.* Wiley, New York, 1990.

[24] BURMAN, D., LEHOCZKY, J., AND LIM, Y. Insensitivity of blocking probabilities in a circuit switching network. *Journal of Applied Probability 21* (1984), 850–859.

[25] BUZEN, J. Computational algorithms for closed queueing networks with exponential services. *Communications of the ACM 16* (1973), 527–531.

[26] CHANG, C.-J., SU, T.-T., AND CHIANG, Y.-Y. Analysis of a cutoff priority cellular radio system with finite queueing and reneging/dropping. *IEEE/ACM Transactions on Networking 2* (1994), 166–175.

[27] CHANG, C.-S. Stability, queue length and delay of deterministic and stochastic queueing networks. *IEEE Transactions on Automatic Control 39* (1994), 913–931.

[28] CHANG, C.-S., HEIDELBERGER, P., JUNEJA, S., AND SHAHABUDDIN, P. Effective bandwidth and fast simulation of ATM intree networks. *Performance Evaluation 20* (1994), 45–64.

[29] CHANG, C.-S., AND THOMAS, J. Effective bandwidth in high-speed digital networks. *IEEE Journal on Selected Areas in Communications 13* (1995), 1091–1100.

[30] CHERNOFF, H. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics 23* (1952), 493–507.

[31] CHOI, B., CHOI, K., AND LEE, Y. M/G/1 retrial queueing systems with two types of calls and finite capacity. *Queueing Systems 19* (1995), 215–229.

[32] CHOUDHURY, G., LEUNG, K., AND WHITT, W. An algorithm to compute blocking probabilities in multi-rate multi-class multi-resource loss models. *Advances in Applied Probability 27* (1995), 1104–1143.

[33] CHOUDHURY, G., LUCANTONI, D., AND WHITT, W. On the effectiveness of effective bandwidths for admission control in ATM networks. *Proceedings ITC 14, Antibes/Juan-les-Pins, France* (1994).

[34] CHOUDHURY, G., LUCANTONI, D., AND WHITT, W. Squeezing the most out of ATM. *IEEE Transactions on Communications 44* (1996), 203–217.

[35] CHUNG, S., AND ROSS, K. Reduced load approximations for multirate loss networks. *IEEE Transactions on Communications 41* (1993), 1222–1231.

[36] COHEN, J. Some results on regular variation for distributions in queueing and fluctuation theory. *Journal of Applied Probability 10* (1973), 343–353.

[37] COHEN, J. *The Single Server Queue, 2nd ed.* North Holland, New York, 1982.

[38] COHEN, J., AND BOXMA, O. A survey of the evolution of queueing theory. *Statistica Neerlandica 39* (1985), 143–158.

[39] COLOMBO, G. Mobility models for mobile system design and dimensioning. *ITC Specialists Seminar, Leidschendam, the Netherlands* (1995).

[40] COOMBS, R., SAVIOTTI, P., AND WELSH, V. *Economics and Technological Change.* McMillan, London, 1987.

[41] COTTRELL, M., FORT, J.-C., AND MALGOUYRES, G. Large deviations and rare events in the study of stochastic algorithms. *IEEE Transactions on Automatic Control 28* (1983), 907–920.

[42] COURCOUBETIS, C., KESIDIS, G., RIDDER, A., WALRAND, J., AND WEBER, R. Admission control and routing in ATM networks using inferences from measured buffered occupancy. *IEEE Transactions on Communications. 43* (1995), 1778–1784.

[43] CRAMÉR, H. Sur un nouveau théorème-limite de la théorie des probabilités. In *Actualités Scientifiques et Industrielles 736. Colloque consacré à la théorie des probabilités.* Hermann, Paris, 1938, pp. 5–23.

[44] DE KOSTER, R. *Capacity oriented analysis and design of production systems.* PhD thesis, Eindhoven, the Netherlands, 1988.

[45] DE PRYCKER, M. *Asynchronous Transfer Mode, Solution for Broadband ISDN.* Prentice-Hall, New Jersey, 1995.

[46] DE VECIANA, G., COURCOUBETIS, C., AND WALRAND, J. Decoupling bandwidths for networks: a decomposition approach to resource management. *Proceedings Infocom, Toronto, Canada* (1994).

[47] DE VECIANA, G., OLIVIER, C., AND WALRAND, J. Large deviations of birth death Markov fluids. *Probability in the Engineering and Informational Sciences 7* (1993), 237–255.

[48] DE VECIANA, G., AND WALRAND, J. Effective bandwidths: Call admission, traffic policing and filtering for ATM networks. *Queueing Systems 20* (1995), 37–59.

[49] DE VRIES, R. *Switch Architectures for the Asynchronous Transfer Mode.* PhD thesis, Enschede, the Netherlands, 1992.

[50] DEMBO, A., AND ZEITOUNI, O. *Large Deviations Techniques and Applications.* Jones and Bartlett, Boston, 1993.

[51] DEUSCHEL, J., AND STROOCK, D. *Large Deviations.* Academic Press, Boston, 1989.

[52] DONSKER, M., AND VARADHAN, S. Asymptotic evaluation of certain Markov process expectations for large time, I. *Communications on Pure and Applied Mathematics 28* (1975), 1–47.

[53] DUFFIELD, N. Economies of scale in queues with sources having power-law large deviations scalings. *Journal of Applied Probability* (1996).

[54] DUFFIELD, N., LEWIS, J., O'CONNELL, N., RUSSELL, R., AND TOOMEY, F. Entropy of ATM traffic streams: a tool for estimating QoS parameters. *IEEE Journal on Selected Areas in Communications 13* (1995), 981–990.

[55] DUFFIELD, N., AND O'CONNELL, N. Large deviations and overflow probabilities for the general single server queue, with applications. *Preprint* (1993).

[56] ECKBERG, A. B-ISDN/ATM traffic and congestion control. *IEEE Network* (1992), September, 28–37.

[57] ELLIS, R. Large deviations for a general class of random vectors. *The Annals of Probability 12* (1984), 1–12.

[58] ELLIS, R. *Entropy, Large Deviations, and Statistical Mechanics.* Springer Verlag, Berlin, 1985.

[59] ELWALID, A., HEYMAN, D., LAKSHMAN, T., MITRA, D., AND WEISS, A. Fundamental bounds and approximations for ATM multiplexers with applications to videoconferencing. *IEEE Journal on Selected Areas in Communications 13* (1995), 1004–1016.

[60] ELWALID, A., AND MITRA, D. Analysis and design of rate-based congestion control of high speed networks, I: stochastic fluid models, access regulation. *Queueing Systems 9* (1991), 29–64.

[61] ELWALID, A., AND MITRA, D. Effective bandwidth of general Markovian traffic sources and admission control of high speed networks. *IEEE/ACM Transactions on Networking 1* (1993), 329–343.

[62] ERLANG, A. The theory of probabilities and telephone conversations. *Nyt Tidsskrift for Matematik 20* (1909), 33–41.

[63] EVERITT, D. Product form solutions in cellular mobile communication systems. *Fourth Australian Teletraffic Research Seminar, Paper nr 3.1.* (1989).

[64] EVERITT, D. Traffic engineering of the radio interface for cellular mobile networks. *Proceedings of the IEEE 82* (1994), 1371–1382.

[65] EVERITT, D., AND EVANS, J. Traffic variability and effective interference for CDMA cellular networks. *ITC Specialists Seminar, Leidschendam, the Netherlands* (1995).

[66] EVERITT, D., AND MACFADYEN, N. Analysis of multi-cellular mobile radiotelephone systems with loss. *British Telecom Technical Journal 1* (1983), 218–222.

[67] EVERITT, D., AND MANFIELD, D. Performance analysis of cellular mobile communication systems with dynamic channel assignment. *IEEE Journal on Selected Areas in Communications 7* (1989), 1172–1180.

[68] FELLER, W. *An Introduction to Probability Theory and Its Applications. Vol. 2, 2nd ed.* Wiley, New York, 1971.

[69] FOWLER, H., AND LELAND, W. Local Area Network traffic characteristics, with implications for broadband network congestion management. *IEEE Journal on Selected Areas in Communications 9* (1991), 1139–1149.

[70] FREIDLIN, M., AND WENTZELL, A. *Random Perturbations of Dynamical Systems.* Springer Verlag, New York, 1984.

[71] GÄRTNER, J. On large deviations from the invariant measure. *Theory of Probability and Applications 22* (1977), 24–39.

[72] GAZDZICKI, P., LAMBADARIS, I., AND MAZUMDAR, R. Blocking probabilities for large multirate Erlang loss systems. *Advances in Applied Probability 25* (1993), 997–1009.

[73] GELFAND, I., AND FOMIN, S. *Calculus of Variations.* Prentice-Hall, Englewood Cliffs, N.J., 1963.

[74] GIBBENS, R., AND HUNT, P. Effective bandwidths for the multi-type UAS channel. *Queueing Systems 9* (1991), 17–28.

[75] GLASSERMAN, P., HEIDELBERGER, P., SHAHABUDDIN, P., AND ZAJIC, T. Multilevel splitting for estimating rare event probabilities. *IBM Research Report* (1996).

[76] GLASSERMAN, P., AND KOU, S. Analysis of an importance sampling estimator for tandem queues. *ACM Transactions on Modeling and Computer Simulation 5* (1995), 22–42.

[77] GLYNN, P., AND IGLEHART, D. Importance sampling for stochastic simulations. *Management Science 35* (1989), 1367–1392.

[78] GOYAL, A., SHAHABUDDIN, P., HEIDELBERGER, P., NICOLA, V., AND GLYNN, P. A unified framework for simulating Markovian models of highly dependable systems. *IEEE Transactions on Computers 41* (1992), 36–51.

[79] GUÉRIN, R. Queueing-blocking system with two arrival streams and guard channels. *IEEE Transactions on Communications 36* (1988), 153–163.

[80] GUÉRIN, R., AHMADI, H., AND NAGHSHINEH, M. Equivalent capacity and its application to bandwidth allocation in high speed networks. *IEEE Journal of Selected Areas in Communications 9* (1991), 968–981.

[81] GÜN, L., AND GUÉRIN, R. Bandwidth management and congestion control framework of the broadband network architecture. *Computer Networks and ISDN Systems 26* (1993), 61–78.

[82] GUT, A. *Stopped Random Walks. Limit theorems and applications.* Springer Verlag, Berlin, 1987.

[83] HARVEY, C., AND HILLS, C. Determining grades of service in a network. *Proceedings ITC 9, Torremollinos, Spain* (1979).

[84] HEFFES, H., AND LUCANTONI, D. A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. *IEEE Journal on Selected Areas in Communications 4* (1986), 856–868.

[85] HEIDELBERGER, P. Fast simulation of rare events in queueing and reliability models. *ACM Transactions on Modeling and Computer Simulation 5* (1995), 43–85.

[86] HIRSCH, M., AND SMALE, S. *Differential equations, dynamical systems, and linear algebra.* Academic Press, San Diego, 1974.

[87] HÜBNER, F., AND RITTER, M. Blocking in multi-service broadband systems with CBR and VBR input traffic. *Proceedings 7th ITG/GI Conference, Aachen, Germany* (1993).

[88] HÜBNER, F., AND TRAN-GIA, P. Quasi-stationary analysis of a finite capacity asynchronous multiplexer with modulated deterministic input. *Proceedings ITC 13, Copenhagen, Denmark* (1991).

[89] HUI, J. Resource allocation for broadband networks. *IEEE Journal on Selected Areas in Communications 6* (1988), 1598–1608.

[90] HUNT, P., AND KELLY, F. On critically loaded links. *Advances in Applied Probability 21* (1989), 661–680.

[91] JACKSON, J. Networks of waiting lines. *Operations Research 5* (1957), 518–521.

[92] JAISWAL, N. *Priority Queues.* Academic Press, New York, 1968.

[93] KAUFMAN, J. Blocking in a shared resource environment. *IEEE Transactions on Communications 29* (1981), 1474–1481.

[94] KEILSON, J., AND IBE, O. Cutoff priority scheduling in mobile cellular communication systems. *IEEE Transactions on Communications 43* (19), 1038–1045.

[95] KELLY, F. *Reversibility and Stochastic Networks.* Wiley, New York, 1979.

[96] KELLY, F. Blocking probabilities in large circuit switched networks. *Advances in Applied Probability 18* (1986), 473–505.

[97] KELLY, F. Routing in circuit-switched networks: optimization, shadow prices and decentralization. *Advances in Applied Probability 20* (1988), 112–144.

[98] KELLY, F. Routing and capacity allocation in networks with trunk reservation. *The Mathematics of Operations Research 15* (1990), 771–793.

[99] KELLY, F. Effective bandwidths at multi-class queues. *Queueing Systems 9* (1991), 5–16.

[100] KELLY, F. Loss networks. *The Annals of Applied Probability 1* (1991), 319–378.

[101] KELLY, F. Notes on effective bandwidths. In *Stochastic Networks: Theory and Applications*, F. Kelly, S. Zachary, and I. Ziedins, Eds. Oxford University Press, Oxford, 1996, pp. 141–168.

[102] KEMENY, J., AND SNELL, J. *Finite Markov chains.* Van Nostrand, Princeton, N.J., 1959.

[103] KESIDIS, G., AND WALRAND, J. Quick simulation of ATM buffers with on-off multiclass Markov fluid sources. *ACM Transactions on Modeling and Computer Simulation 3* (1993), 269–276.

[104] KESIDIS, G., AND WALRAND, J. Relative entropy between Markov transition rate matrices. *IEEE Transactions on Information Theory 39* (1993), 1056–1057.

[105] KESIDIS, G., WALRAND, J., AND CHANG, C.-S. Effective bandwidths for multiclass Markov fluids and other ATM sources. *IEEE/ACM Transactions on Networking 1* (1993), 424–428.

[106] KLEINROCK, L. *Queueing Systems, Volume 1 and 2.* Wiley, New York, 1975.

[107] KNESSL, C., AND MORRISSON, J. Heavy traffic analysis of a data-handling system with many sources. *SIAM Journal of Applied Mathematics 51* (1991), 187–213.

[108] KOBAYASHI, H., AND REN, Q. A diffusion approximation analysis of an ATM statistical multiplexer with multiple types of traffic, part I: equilibrium state solutions. *Proceedings of the 1993 IEEE International Conference on Communications, Vol. 2* (1993).

[109] KOSTEN, L. Stochastic theory of a multi-entry buffer (1). *Delft Progress Report 1* (1974), 10–18.

[110] KOSTEN, L. Stochastic theory of data-handling systems with groups of multiple sources. In *Performance of Computer-Communication Systems*, H. Rudin and W. Bux, Eds. Elsevier, Amsterdam, 1984, pp. 321–331.

[111] KOSTEN, L. Liquid models for a type of information buffer problem. *Delft Progress Report 11* (1986), 71–86.

[112] KRÖNER, H., RENGER, T., AND KNOBLING, R. Performance modelling of an adaptive CAC strategy for ATM networks. *Proceedings ITC 14, Antibes/Juan-les-Pins, France* (1994).

[113] LAW, A., AND KELTON, W. *Simulation Modeling and Analysis.* McGraw-Hill, New York, 1986.

[114] LE BOUDEC, J.-Y. The Asynchronous Transfer Mode: a tutorial. *Computer Networks and ISDN Systems 24* (1992), 279–309.

[115] LELAND, W., TAQQU, M., WILLINGER, W., AND WILSON, D. On the self-similar nature of Ethernet traffic. *Proceedings Sigcomm '93 Conference, Ithaca N.Y., United States* (1993).

[116] LIKHANOV, N., TSYBAKOV, B., AND GEORGANAS, N. Analysis of an ATM buffer with self-similar ('fractal') input traffic. *Proceedings Infocom, Boston, United States* (1995).

[117] LUCANTONI, D. The BMAP/G/1 queue: a tutorial. In *Models and Techniques for Performance Evaluation of Computer and Communication Systems*, L. Donatiello and R. Nelson, Eds. Springer, New York, 1993, pp. 330–358.

[118] MANDJES, M. Fast simulation of Markov fluid models in conjunction with large deviations. *Research Memorandum 1993-58, Faculty of Economics, Vrije Universiteit Amsterdam* (1993).

[119] MANDJES, M. Large deviations and queueing applications. *Operations Research Proceedings DGOR/NSOR, Amsterdam, the Netherlands* (1993).

[120] MANDJES, M. Asymptotically optimal importance sampling for tandem queues with Markov fluid input. *Preprint* (1995).

[121] MANDJES, M. Importance sampling of buffer overflows in batch-arrivals queues. *Tinbergen Institute Discussion Paper TI95-33* (1995).

[122] MANDJES, M. Fast simulation of blocking probabilities in loss networks. *European Journal of Operational Research* (1996).

[123] MANDJES, M. Overflow asymptotics for large communication systems with general Markov fluid sources. *Probability in the Engineering and Informational Sciences 10* (1996), 501–518.

[124] MANDJES, M. Rare event analysis of batch-arrival queues. *Telecommunication Systems* (1996).

[125] MANDJES, M., AND RIDDER, A. Finding the conjugate of Markov fluid processes. *Probability in the Engineering and Informational Sciences 9* (1995), 297–315.

[126] MANDJES, M., AND TUTSCHKU, K. Efficient call handling procedures in cellular mobile networks. *Forschungsbericht, Preprintreihe nr 144. Universität Würzburg, Institut für Informatik* (1996).

[127] MANDJES, M., AND VAN DEN BERG, H. A new approach to buffer and bandwidth allocation in single- and multi-link ATM systems. *COST 242, TD (95)51* (1995).

[128] MILLER, H. A convexity property in the theory of random variables on a finite Markov chain. *The Annals of Mathematical Statistics 32* (1961), 160–182.

[129] MITRA, D. Asymptotic analysis and computational methods for a class of simple, circuit-switched networks with blocking. *Advances in Applied Probability 19* (1987), 219–239.

[130] MITRA, D. Stochastic theory of a fluid model of producers and consumers coupled by a buffer. *Advances in Applied Probability 20* (1988), 646–676.

[131] MITROU, N., KONTOVASILIS, K., KRÖNER, H., AND IVERSEN, V. Statistical multiplexing, bandwidth allocation strategies and connection admission control in ATM networks. *European Transactions on Telecommunications 5* (1994), 33–47.

[132] MOGULSKII, A. Large deviations for trajectories of multidimensional random walks. *Theory of Probability and Applications 21* (1976), 300–315.

[133] NAKAYAMA, M. A characterization of the simple biasing method for simulations of highly reliable Markovian systems. *ACM Transactions on Modeling and Computer Simulation 4* (1994), 52–88.

[134] NEUTS, M. A queue subject to exogeneous phase changes. *Advances in Applied Probability 3* (1971), 78–119.

[135] NEUTS, M. A versatile Markovian point process. *Journal of Applied Probability 12* (1979), 764–779.

[136] NEUTS, M. *Matrix-Geometric Solutions in Stochastic Models - an Algorithmic Approach.* The Johns Hopkins University Press, Baltimore, 1981.

[137] NEY, P., AND NUMMELIN, E. Markov additive processes I. Eigenvalue properties and limit theorems / Markov additive processes II. Large deviations. *The Annals of Probability 15* (1987), 561–592 and 593–609.

[138] NORROS, I. A storage model with self-similar input. *Queueing Systems 16* (1994), 387–396.

[139] NORROS, I. On the use of fractional Brownian motion in the theory of connectionless networks. *IEEE Journal on Selected Areas in Communications 13* (1995), 953–962.

[140] ONVURAL, R. *Asynchronous Transfer Mode Networks: Performance Issues.* Artech House, Boston, 1994.

[141] O'REILLY, P., AND GHANI, S. Data performance in burst switching when the voice silence periods have a hyperexponential distribution. *IEEE Transactions on Communications 35* (1987), 1109–1112.

[142] PALLANT, D., AND TAYLOR, P. Approximations of performance measures in cellular mobile networks with dynamic channel allocation. *Telecommunication Systems 3* (1994), 129–163.

[143] PALLANT, D., AND TAYLOR, P. Modeling handovers in cellular mobile networks with dynamic channel allocation. *Operations Research 43* (1995), 33–42.

[144] PAREKH, S., AND WALRAND, J. A quick simulation method for excessive backlogs in networks of queues. *IEEE Transactions on Automatic Control 34* (1989), 54–66.

[145] PARK, D., AND PERROS, H. m-MMBP characterization of the departure process of an m-MMBP/Geo/1/K queue. *Proceedings ITC 14, Antibes/Juan-les-Pins, France* (1994).

[146] PARULEKAR, M., AND MAKOWSKI, A. Tail probabilities for a multiplexer with self-similar traffic. *Proceedings Infocom, Boston, United States* (1995).

[147] REIMAN, M. A critically loaded multiclass Erlang loss system. *Queueing Systems 9* (1991), 65–82.

[148] RIDDER, A. Fast simulation of Markov fluid models. *Research Memorandum 1993-21, Faculty of Economics, Vrije Universiteit Amsterdam* (1993).

[149] RIDDER, A., AND MANDJES, M. Fast simulation of Markov modulated fluid models. *B-ISDN Teletraffic Modelling Symposium, Antwerp, Belgium* (1995).

[150] RIDDER, A., AND WALRAND, J. Some large deviations results in Markov fluid models. *Probability in the Engineering and Informational Sciences 6* (1992), 543–560.

[151] RITTER, M. Network buffer requirements of the rate-based control mechanism for ABR services. *Proceedings Infocom, San Francisco, United States* (1996).

[152] RITTER, M., AND TRAN-GIA, P. Multi-rate models for dimensioning and performance evaluation of ATM networks. *COST 242, Interim Report* (1994).

[153] ROBERT, S. A Markovian model for self-similar traffic. *Seminar on Applied Stochastic Modelling in Telecommunication and Manufacturing, Dagstuhl, Germany* (1995).

[154] ROBERTS, J. A service system with heterogeneous service requirements - applications to multi-service telecommunications systems. In *Proceedings of Performance of Datacommunications Systems and their Applications*, G. Pujolle, Ed. North Holland, Amsterdam, 1981, pp. 423–431.

[155] ROBERTS, J. *Performance evaluation and design of multi-service networks. Final report of the COST 224 Project.* Commision of the European Communities, Luxembourg, 1992.

[156] ROBERTS, J. *Methods for the performance evaluation and design of broadband multiservice networks. Final report of the COST 242 Project.* Springer Verlag, Berlin, 1996.

[157] ROCKAFELLAR, R. *The theory of subgradients and its applications to problems of optimization. Convex and nonconvex functions.* Heldermann Verlag, Berlin, 1981.

[158] Ross, K., and Wang, J. Monte Carlo summation applied to product-form loss networks. *Probability in the Engineering and Informational Sciences 6* (1992), 323–348.

[159] Rubinstein, R. *Simulation and the Monte Carlo method.* Wiley, New York, 1981.

[160] Sadowsky, J. Large deviations theory and efficient simulation of excessive backlogs in a GI/GI/m queue. *IEEE Transactions on Automatic Control 36* (1991), 1383–1394.

[161] Sanov, I. On the probability of large deviations of random variables (in Russian). *Translated in: Selected Translations in Mathematical Statistics: I* (1957).

[162] Schassberger, R. *Warteschlangen.* Springer Verlag, Berlin, 1973.

[163] Seneta, E. *Non-negative matrices and Markov chains.* Springer Verlag, Berlin, 1981.

[164] Shahabuddin, P. Importance sampling for the simulation of highly reliable Markovian systems. *Management Science 40* (1994), 333–352.

[165] Shwartz, A., and Weiss, A. Induced rare events: Analysis via large deviations and time reversal. *Advances in Applied Probability 25* (1993), 667–689.

[166] Shwartz, A., and Weiss, A. *Large Deviations for Performance Analysis, queues, communication, and computing.* Chapman and Hall, New York, 1995.

[167] Siegmund, D. Importance sampling in the Monte Carlo study of sequential tests. *The Annals of Statistics 4* (1976), 673–684.

[168] Smith, A., Adams, J., and Tagg, G. Available Bit Rate – a new service for ATM. *Computer Networks and ISDN Systems 28* (1996), 635–640.

[169] Stallings, W. *ISDN and Broadband ISDN.* McMillan, New York, 1992.

[170] Stern, T., and Elwalid, A. Analysis of separable Markov-modulated rate models for information-handling systems. *Advances in Applied Probability 23* (1991), 105–139.

[171] Takács, L. Priority queues. *Operations Research 12* (1964), 63–74.

[172] Takahashi, Y. Asymptotic exponentiality of the tail of the waiting time distribution in a Ph/Ph/c queue. *Advances in Applied Probability 13* (1981), 619–630.

[173] Tijms, H. *Stochastic Modelling and Analysis: a computational approach.* Wiley, Chichester, 1986.

[174] Tijms, H. Heuristics for finite-buffer queues. *Probability in the Engineering and Informational Sciences 6* (1992), 277–285.

[175] Tijms, H. *Stochastic Models, an algorithmic approach.* Wiley, Chichester, 1994.

[176] TRAN-GIA, P., AND MANDJES, M. Modeling of customer retrial phenomenon in cellular mobile systems. *Forschungsbericht, Preprintreihe nr 142. Universität Würzburg, Institut für Informatik* (1996).

[177] TSANG, D., AND ROSS, K. Algorithms to determine exact blocking probabilities for multirate tree models. *IEEE Transactions on Communications 38* (1990), 1266–1271.

[178] TSE, D., GALLAGER, R., AND TSITSIKLIS, J. Statistical multiplexing of multiple time-scale Markov streams. *IEEE Journal on Selected Areas in Communications 13* (1995), 1028–1038.

[179] TURNER, J. New directions in communications (or which way to the information age?). *IEEE Communications Magazine 25* (1986), 10, 8–15.

[180] VAN DIJK, N. *Queueing Networks and Product Forms, a systems approach.* Wiley, Chichester, 1993.

[181] VAN DOORN, E., JAGERS, A., AND DE WIT, J. A fluid reservoir regulated by a birth-death process. *Stochastic Models 4* (1988), 457–472.

[182] VILLEN-ALTAMIRANO, M., AND VILLEN-ALTAMIRANO, J. RESTART: A method for accelerating rare events simulations. *Proceedings ITC 13, Copenhagen* (1991).

[183] WALRAND, J. *An Introduction to Queueing Networks.* Prentice-Hall, New Jersey, 1988.

[184] WEISS, A. A new technique of analyzing large traffic systems. *Advances in Applied Probability 18* (1986), 506–532.

[185] WEISS, A. An introduction to large deviations for communication networks. *IEEE Journal on Selected Areas in Communications 13* (1995), 938–952.

[186] WHITT, W. Tail probabilities with statistical multiplexing and effective bandwidths in multi-class queues. *Telecommunication Systems 2* (1993), 71–107.

[187] WHITTLE, P. Approximation in large-scale circuit-switched networks. *Probability in the Engineering and Informational Sciences 2* (1988), 279–291.

[188] WIJNGAARD, J. The effect of interstage buffer storage on the output of two unreliable production units in series, with different production rates. *AIIE Transactions 11* (1979).

[189] YOON, C., AND UN, C. Performance of personal portable radio telephone systems with and without guard channels. *IEEE Journal on Selected Areas in Communications 11* (1993), 911–917.

[190] ZACHARY, S. On blocking in loss networks. *Advances in Applied Probability 23* (1991), 355–372.

Samenvatting

Het onderwerp van dit proefschrift is de probabilistische analyse van zeldzame gebeurtenissen in wachtrijsystemen. Afschattingen voor kansen worden ontwikkeld, en daarnaast ook efficiënte simulatiemethoden. De specifieke toepassing van de gevormde theorie is de analyse van moderne telecommunicatienetwerken.

Op dit moment staat B-ISDN (*Broadband Integrated Services Digital Network*) sterk in de belangstelling. Dit telecommunicatienetwerk wordt gekenmerkt door (a) de zeer hoge snelheid waarop data kunnen worden verstuurd ('breedband'), en (b) de mogelijkheid verschillende diensten te ondersteunen, zoals spraak, beeld en data ('geïntegreerde diensten'). De onderliggende technologie voor B-ISDN is ATM (*Asynchronous Transfer Mode*). ATM converteert het verkeer, wanneer het het netwerk binnengaat, in kleine cellen van gelijke grootte. In de 'header' van deze 'ATM-cellen' ligt alle informatie opgesloten over de route door het netwerk die door de cel gevolgd dient te worden. Wanneer het verkeer uiteindelijk op de bestemming aankomt, worden de cellen getransformeerd tot de oorspronkelijke boodschap. Het grote voordeel van ATM is flexibiliteit; zo is het relatief eenvoudig een extra dienst aan het netwerk toe te voegen.

Bij de implementatie van ATM zijn er in feite twee soorten problemen. Aan de ene kant moet het netwerk zelf ontworpen worden; er moet beslist worden waar de knooppunten gelegd worden ('topologie'), welke verwerkingssnelheden en buffergroottes gekozen dienen te worden ('dimensionering'), etc. Aan de andere kant moeten er effectieve mechanismes ontworpen worden om het netwerk tijdens het gebruik goed te laten functioneren. Bij deze laatste categorie kan men bijvoorbeeld denken aan een procedure die beslist of een nieuwe aanvraag ingewilligd kan worden, zonder de reeds aanwezige connecties al te zeer te benadelen ('toelatingsmechanisme').

De prestatie van het ATM-netwerk kan gemeten worden aan de hand van een aantal criteria. Ten eerste kunnen we, gegeven een toelatingsmechanisme, de kans bekijken dat een aanvraag voor een connectie geweigerd wordt, de zgn. blokkeringskans. Aan de andere kant zijn we, gegeven een aantal geaccepteerde connecties, geïnteresseerd in prestatiematen als vertraging van de ATM-cellen in het netwerk, en de fractie van cellen dat in het netwerk verloren gaat (als gevolg van het overlopen van buffers).

Een accurate techniek om op probabilistische wijze het netwerk te analyseren is door

middel van modellering als een netwerk van wachtrijen. Volgens een stochastisch proces komen aanvragen voor connecties aan. Als deze gehonoreerd worden, dan blijven zij een stochastische duur actief; gedurende deze periode worden volgens een bepaald kansmechanisme ATM-cellen verstuurd. Dit laatste mechanisme wordt vaak benaderd door een zgn. 'aan-uit-bron': de bron zendt ofwel niet, ofwel met een bepaalde constante snelheid.

Op het gebied van wachtrijsystemen is al veel onderzoek verricht. Met het oog op toepassingen in ATM evenwel, is men geïnteresseerd in de kans op zeer zeldzame gebeurtenissen (verliezen van 1 op de 10^9 cellen, extreme vertragingen), en hier vertoont de reeds gevormde theorie deficiënties. Weliswaar bestaan er voor sommige modellen exacte en asymptotische resultaten, maar het daadwerkelijk uitrekenen hiervan levert, voor modellen van realistische grootte, numerieke problemen op. De bijdrage van dit proefschrift is tweeledig: aan de ene kant komen we tot, relatief eenvoudig berekenbare, asymptotische relaties, aan de andere kant leveren deze resultaten impliciet de essentiële informatie voor het ontwikkelen van efficiënte simulatietechnieken.

De ontwikkelde asymptotiek valt binnen de theorie van Grote Afwijkingen (*Large Deviations*); dit is een verzamelingen van technieken/resultaten die betrekking hebben op kansen op zeldzame gebeurtenissen. Het gemeenschappelijke kenmerk is dat deze kansen op exponentiële wijze kleiner worden in een parameter van het model; in het geval van ATM-netwerken bijvoorbeeld de buffergrootte van een bepaalde rij in het netwerk of het aantal aangesloten verbindingen.

Het zij opgemerkt dat directe simulatie zeer tijdrovend is in verband met de zeer kleine kansen die geschat dienen te worden. Het idee is daarom om te simuleren onder een ander kansmodel dan het oorspronkelijke (*Importance Sampling*). Na de resultaten terugvertaald te hebben naar het eigenlijke model, krijgen we zuivere schattingen. Het is natuurlijk van belang om het 'nieuwe' kansmodel op efficiënte wijze te kiezen. Het blijkt dat de theorie van Grote Afwijkingen impliciet dit kansmodel geeft, en wel als volgt. Het is mogelijk om, *gegeven het optreden van de zeldzame gebeurtenis*, het meest waarschijnlijke pad (in de tijd) van het stochastische proces te bepalen. Het 'nieuwe' kansmodel moet nu zo gekozen worden dat dit – zeldzame – pad het 'gemiddeld gedrag' wordt. Dit kansmodel blijkt bepaalde optimaliteitseigenschappen met betrekking tot Importance Sampling te hebben: binnen een klasse van 'nieuwe kansmodellen' geeft het de grootste versnelling.

Dit procédé wordt in de hoofdstukken 2, 3 en 5 behandeld voor verschillende klassen van wachtrijsystemen. We laten zien dat de verliesfractie exponentieel daalt in de buffergrootte. Tevens vinden we dat het meest aannemelijke traject (in de tijd) van de bufferinhoud naar een volle buffer, startend in een leeg systeem, een rechte lijn is. Op deze wijze kunnen ook de parameters voor de Importance Sampling gekozen worden. In hoofdstuk 2 wordt een model onderzocht, waarin het door de bronnen aangeboden verkeer geen cor-

relatie in de tijd vertoont. Het verkeersmodel van hoofdstuk 3 doet dat wel, en is daarom relevanter in de ATM-context. Waar hoofdstuk 2 en 3 een enkele buffer van het netwerk beschouwen, wordt in hoofdstuk 5 een tandemsysteem onderzocht: twee wachtrijen in serie.

In tegenstelling tot de hoofdstukken 2, 3 en 5 (waar we asymptotiek als functie van de buffergrootte bekijken), gaat hoofdstuk 4 in op zeldzame gebeurtenissen als gevolg van een groot aantal toegelaten bronnen. Na een zekere schaling van de verwerkingssnelheid, kunnen asymptotische relaties afgeleid worden voor verliesfracties. Opvallend is dat een parallel getrokken kan worden met het zgn. tijdsomgekeerde proces. Hoofdstuk 6 gebruikt het simulatieprogramma dat ontwikkeld is in hoofdstuk 5 om te komen tot richtlijnen voor dimensionering van buffers en verwerkingssnelheden en toelatingsmechanismen voor een klasse van ATM-netwerken. Tevens wordt onderzocht hoe de stochastische aard van het verkeer verandert wanneer het verschillende wachtrijen in het netwerk passeert.

Waar in de vorige hoofdstukken meestal het aantal aangesloten connecties constant gehouden werd, bekijken we in hoofdstuk 7 het netwerk op 'connectieniveau'. Gegeven een bepaald toelatingsmechanisme kijken we naar de kans dat een nieuwe aanvraag niet geaccepteerd kan worden, de blokkeringskans. Het gebruikte toelatingsmechanisme gaat uit van het zgn. effectieve bandbreedte concept; onder dit concept kan het netwerk geïnterpreteerd worden als het klassieke verliesnetwerk. Het verliesnetwerk wordt weer geanalyseerd met asymptotische technieken en Importance Sampling-simulatie. In hoofdstuk 8 bekijken we geen ATM-netwerk maar een cellulair mobiel communicatienetwerk. Evenwel, de theorie ontwikkeld in hoofdstuk 7 kan nu ook gebruikt worden, met een paar kleine aanpassingen. We zijn primair geïnteresseerd in de kans dat, vanwege het beperkte aantal kanalen per cel, een gesprek niet kan worden toegelaten. Inzichten in de toestand van het netwerk, gegeven een blokkering in een bepaalde cel, worden verkregen.

Index

The book is no. 132 of the Tinbergen Institute Research Series. This series is established through cooperation between Thesis Publishers and the Tinbergen Institute. A list of books which already appeared in the series can be found in the back. The Tinbergen Institute is the Netherlands Research Institute and Graduate School for Economics and Business, which was founded in 1987 by the Faculties of Economics and Econometrics of the Erasmus University in Rotterdam, the University of Amsterdam and the Free University in Amsterdam. The Institute is named after the late Professor Jan Tinbergen, Dutch Nobel Prize laureate in economics in 1969. The Tinbergen Institute is located in Amsterdam and Rotterdam. If available trade editions of the books which are published in the Tinbergen Institute Research Series can be ordered through Thesis Publishers, P.O. Box 14791, 1001 LG Amsterdam, The Netherlands, phone: +3120 6255429; fax: +3120 6203395.

The following books already appeared in this series:

1. O.H. SWANK, Policy makers, voters and optimal control, estimation of the preferences behind monetary and fiscal policy in the United States.
2. J. VAN DER BORG, Tourism and urban development. The impact of tourism on urban development: towards a theory of urban tourism, and its application to the case of Venice, Italy.
3. A. JOLINK, Liberté, Egalité, Rareté. The evolutionary economics of Léon Walras.
4. R.B. BUITENDIJK, Towards an effective use of relational database management systems.
5. R.M. VERBURG, The two faces of interest. The problem of order and the origins of political economy and sociology as distinctive fields of inquiry in the Scottish Enlightenment.
6. H.P. VAN DALEN, Economic policy in a demographically divided world.
7. P.J. VERBEEK, Two case studies on manpower planning in an Airline.
8. M.W. HOFKES, Modelling and computation of general equilibrium.
9. T.C.R. VAN SOMEREN, Innovatie, emulatie en tijd. De rol van de organisatorische vernieuwingen in het economische proces.
10. M. VAN VLIET, Optimization of manufacturing system design.
11. R.M.C. VAN WAES, Architectures for Information Management. A pragmatic approach on architectural concepts and their application in dynamic environments.
12. K. NIMAKO, Economic change and political conflict in Ghana 1600-1990.
13. J.M.C. VOLLERING, Care services for the elderly in the Netherlands. The PACKAGE model.
14. S. ZHANG, Stochastic queue location problems.
15. C. GORTER, The dynamics of unemployment and vacancies on regional labour markets.
16. P. KOFMAN, Managing primary commodity trade (on the use of futures markets).
17. P.Th. VAN DE LAAR, Financieringsgedrag in de Rotterdamse maritieme sector, 1945-1960.
18. P.H.B.F. FRANSES, Model selection and seasonality in time series.
19. P.W. VAN WIJCK, Inkomensverdelingsbeleid in Nederland. Over individuele voorkeuren en distributieve effecten.
20. A.E. VAN HEERWAARDEN, Ordering of risks. Theory and actuarial applications.
21. J.C.J.M. VAN DEN BERGH, Dynamic models for sustainable development.
22. H. XIN, Statistics of bivariate extreme values.
23. C.P. VAN BEERS, Exports of developing countries. Differences between South-South and South-North trade and their implications for economic development.
24. L. BROERSMA, The relation between unemployment and interest rate. Empirical evidence and theoretical justification.
25. E. SMEITINK, Stochastic models for repairable systems.

26. M. DE LANGE, Essays on the theory of financial intermediation.
27. S.J. KOOPMAN, Diagnostic checking and intra-daily effects in time series models.
28. R.J. BOUCHERIE, Product-form in queueing networks.
29. F.A.G. WINDMEIJER, Goodness of fit in linear and qualitative-choice models.
30. M. LINDEBOOM, Empirical duration models for the labour market.
31. S.T.H. STORM, Macro-economic considerations in the choice of an agricultural policy.
32. H.E. ROMEIJN, Global optimization by random walk sampling methods.
33. R.W. VAN ZIJP, Austrian and new classical business cycle theories.
34. J.A. VIJLBRIEF, Unemployment insurance and the Dutch labour market.
35. G.E. HEBBINK, Human capital, labour demand and wages. Aspects of labour market heterogeneity.
36. J.J.M. POTTERS, Lobbying and pressure: theory and experiments.
37. P. BOSWIJK, Cointegration, identification and exogeneity. Inference in structural error correction models.
38. M. BOUMANS, A case of limited physics transfer. Jan Tinbergen's resources for re-shaping economics.
39. J.B.J.M. DE KORT, Edge-disjointness in combinatorial optimization: problems and algorithms.
40. J.F.J. DE MUNNIK, The valuation of interest rates derivative securities.
41. J.C.A. POTJES, Empirical studies in Japanese retailing.
42. J.-K. MARTIJN, Exchange-rate variability and trade: Essays on the impact of exchange-rate variability on international trade flows.
43. J.B.L.M. VERBEEK, Studies on economic growth theory. The role of imperfections and externalities.
44. R.H. VAN HET KAAR, Medezeggenschap bij fusie en ontvlechting.
45. F. KALSHOVEN, Over Marxistische economie in Nederland, 1883-1939.
46. W. SWAAN, Behaviour and institutions under economic reform. Price regulation and market behaviour in Hungary.
47. J.J. CAPEL, Exchange rates and strategic decisions of firms.
48. M.F.M. CANOY, Bertrand meets the fox and the owl. Essays in the theory of price competition.
49. H.M. KAT, The efficiency of dynamic trading strategies in imperfect markets.
50. E.A.M. BULDER, The social economics of old age: strategies to maintain income in later life in the Netherlands 1880-1940.
51. J. BARENDREGT, The Dutch Money Purge. The monetary consequences of German occupation and their redress after liberation, 1940-1952.
52. N. PIERSMA, Combinatorial optimization and empirical processes.
53. M.J.C. SIJBRANDS, Strategische en logistieke besluitvorming. Een empirisch onderzoek naar informatiegebruik en instrumentele ondersteuning.
54. H.J. VAN DER SLUIS, Heuristics for complex inventory systems. Deterministic and stochastic problems.
55. E.F.M. WUBBEN, Markets, uncertainty and decision-making. A history of the introduction of uncertainty into economics.
56. V.J. BATELAAN, Organizational culture and strategy. A study of cultural influences on the formulation of strategies, goals, and objectives in two companies.
57. R.M. DE JONG, Asymptotic theory of expanding parameter space methods and data depen-

dence in econometrics.

58. D.P.M. DE WIT, Portfolio management of common stock and real estate assets. An empirical investigation into the stochastic properties of common stock and equity real-estate.
59. A. LAGENDIJK, The internationalisation of the Spanish automobile industry and its regional impact. The emergence of a growth-periphery.
60. B.M. KLING, Life insurance, a non-life approach.
61. J.H. GROOTENDORST, De markthuur op kantorenmarkten in Nederland.
62. M. DINGENA, The creation of meaning in advertising. Interaction of figurative advertising and individual differences in processing styles.
63. R.T. LIE, Economische dynamiek en toplocaties. Locatiekarakteristieken en prijsontwikkeling van kantoren in een aantal grote Europese steden.
64. R.L.M. PEETERS, System identification based on Riemannian geometry: theory and algorithms.
65. O. MEMEDOVIC, On the theory and measurement of comparative advantage. An empirical analysis of Yugoslav trade in manufactures with the OECD countries, 1970-1986.
66. S. FISCHER, The paradox of information technology management.
67. P.A. STORK, Modelling international financial markets: an empirical study.
68. R.A. BELDERBOS, Strategic trade policy and multinational enterprises: essays on trade and investment by Japanese electronics firms.
69. I.T. VAN DEN DOEL, Dynamics in cross-section and panel data models.
70. M.W. DELL, Maximum price regulations and resulting parallel and black markets.
71. B. CHEN, Worst case performance of scheduling heuristics.
72. P.W. CHRISTIAANSE, Strategic advantage and the exploitability of information technology. An empirical study of the effects of IT on supplier-distributor relationships in the US airline industry.
73. L. LEI, User participation and the success of information system development. An integrated model of user-specialist relationships.
74. J.H. BAGGEN, Duurzame mobiliteit. Duurzame ontwikkeling en de voorzieningenstructuur van het personenvervoer in de Randstad.
75. R.A. BOSCHMA, Looking through a window of locational opportunity. A long-term spatial analysis of techno-industrial upheavals in Great-Britain and Belgium.
76. C.A.G. SNEEP, Innovation management in the Agro-food industry.
77. F.R. KLEIBERGEN, Identifiability and nonstationarity in classical and Bayesian econometrics.
78. R.F. VAN DE WIJNGAERT, Trade Unions and collective bargaining in the Netherlands.
79. M. BOOGAARD, Defusing the software crisis: Information systems flexibility through data independence.
80. E.M. VERMEULEN, Corporate risk management. A multi-factor approach.
81. R.A. ZUIDWIJK, Complementary triangular forms for pairs of matrices and operators.
82. M.H. GOEDHART, Financial planning in divisionalised firms: models and methods.
83. E. EGGINK, A symmetric approach to the labor market: an application of the sem method.
84. A.F. CORRELJÉ The Spanish oil industry: Structural change and modernization.
85. A.H.M. LELIVELD, Social security in developing countries; operation and dynamics of social security mechanisms in rural Swaziland.
86. Y.M. PRINCE, Price cost margins in Dutch manufacturing: with an emphasis on cyclical and firm size effects.

87. W.E. KUIPER, Farmers, prices and rational expectations.
88. B. ROORDA, Global total least squares. A method for the construction of open approximate models from vector time series.
89. A.I. MARTINS BOTTO DE BARROS, Discrete and fractional programming techniques for location models.
90. J.A. DOS SANTOS GROMICHO, Quasiconvex optimization and location theory.
91. R.B. KOOL, Aspects of enlarging the market effects in the Dutch health care.
92. B. LEEFTINK, The desirability of currency unification theory and some evidence.
93. A. VAN VLIET, Lower and upper bounds for on-line bin packing and scheduling heuristics.
94. C.C.J.M. HEYNEN, Models for option evaluation in alternative price-movements.
95. Y.M. VAN EVERDINGEN, Adoption and diffusion of the European currency unit. An empirical study among European companies.
96. R.G. DE VILDER, Endogenous business cycles.
97. J.J. STIBORA and A. DE VAAL, Services and services trade: A theoretical inquiry.
98. R. VAN DER BIE, "Een doorlopende groote roes". De economische ontwikkeling van Nederland, 1913-1921.
99. W.J. JANSEN, International capital mobility and asset demand. Six empirical studies.
100. N.E. STROEKER, Second-hand markets for consumer durables.
101. J.E. VAN DEN BERG, Trade union growth and decline in The Netherlands.
102. P.R.J. VAN DER LAAG, An analysis of refinement operators in inductive logic programming.
103. E. BEINAT, Multiattribute value functions for environmental management.
104. H.A. VAN KLINK, Towards the borderless mainport Rotterdam. An analysis of functional, special and administrative dynamics in port systems.
105. W.H.J. HASSINK, Worker flows and the employment adjustment of firms. An empirical investigation.
106. S. LIU, Contributions to matrix calculus and applications in econometrics.
107. C.M. VAN PRAAG, Determinants of successful entrepreneurship.
108. E.T. VERHOEF, Economic efficiency and social feasibility in the regulation of road transport externalities.
109. I. BRAMEZZA, The competitiveness of a European city and the role of urban management in improving the city's performance. The cases of the Central Veneto and Rotterdam regions.
110. G.G.A. BIESSEN, East European foreign trade and system changes.
111. A. LUCAS, Outlier robust unit root analysis.
112. C.M. FOKKEMA, Residential moving behaviour of the elderly: an explanatory analysis for the Netherlands.
113. T. EROL, Exchange rate policy in Turkey. External competitiveness and internal stability studied through a macromodel.
114. M.A. DE RUYTER VAN STEVENINCK, The impact of capital imports - Argentina 1970-1989.
115. D.R. DANNENBURG, Actuarial credibility models: evaluations and extensions.
116. C. FOLKERTSMA, On equivalence scales.
117. A. PLAAT, Research Re: search & Re-search.
118. A.W. VAN DER KROGT, Service development in less developed countries. The case of telecom services in Chile and Argentina.
119. O.W. STEENBEEK, Financial regulation in Japan. Systemic risks and the Nikkei futures market.

120. T.R.P.J. KROES, Financial intermediation and monetary transmission.
121. T.H.F. CHEUK, Exotic options.
122. K. TEPLÁ, Accounting for the quality of work-life-putting price tags on sources of job (dis)satisfaction.
123. R.H.M. EMMERINK, Information and pricing in road transport: theoretical and applied models.
124. T.J.S. OFFERMAN, Beliefs and decision rules in public good games - theory and experiments.
125. F. MERCURIO, Claim pricing and hedging under market imperfections.
126. A.W.M. ODÉ, Migrant workers in the dutch labour market today.
127. R.E. WILDEMAN, The art of grouping maintenance.
128. M.T. ROCHA DE MAGALHAES MELO, Stochastic Lot-sizing in Production Planning - Strategies for Make-to-Order and Make-to-Stock.
129. G. RUSSO, Firms' recruitment behaviour: An empirical investigation of the use of search channels, the rate of arrival of applicants, and the spatial radius of search.
130. J.N. VAN OMMEREN, Commuting and relocation of jobs and residences; a search perspective.
131. F.N. GOUWELEEUW, A general approach to computing loss probabilities in finite-buffer queues.
132. M. MANDJES, Rare event analysis of communication networks.