Capacity management for packet-switched networks with heterogeneous sources

de Jonge, L.

Award date:
2009

Capacity management for packet-switched networks with heterogeneous sources

Linda de Jonge

Master Thesis
July 29, 2009

Supervisors

Dr. Frank Roijers
Prof. dr. ir. Sem Borst
Dr. Andreas Löpker

Industrial and Applied Mathematics

Preface

This thesis is the result of an eight-month internship at the department Planning, Performance and Quality (PPQ) of TNO-ICT in Delft. It is also the conclusion of the master's program for Industrial and Applied Mathematics at the Eindhoven University of Technology, with specialization Statistics, Probability and Operations Research.

I would like to thank the people who helped me to bring this thesis to a good end. In the first place I would like to thank my daily supervisor Frank Roijers from TNO-ICT. For all his help and for the numerous discussions we had on the interpretation of the results. He gave me an introduction in telecommunication and I could always come to him with questions. He also read my thesis thoroughly and suggested many improvements.
I also want to thank my supervisors Sem Borst and Andreas Löpker from the university. They were always willing to review (parts of) my thesis and to give me tips about how to proceed with my research.
Further, I would like to thank TNO, especially the department PPQ, for giving me the opportunity to do my final project within their office and letting me attend the department meetings. In this way I was able to gain a clear insight in their daily work.

Finally I would like to thank my family and friends who were interested in the progress I made in my project.

Linda de Jonge
Nieuw-Lekkerland, July 2009

Abstract

This thesis focuses on the problem of dimensioning a link in a communication network which carries data of both consumers and a small number of corporate users. The corporate users have an access rate that is much larger than that of the consumers, but possibly they are only active a small fraction of the time. The main goal is to develop a modelling approach to determine the required link capacity in order to satisfy specific Quality of Service (QoS) levels in terms of delays (for streaming traffic) and user throughputs (for elastic traffic). In particular, the impact of the presence of corporate users and their characteristics on the required link capacity will be examined.

In the modelling approach two separate models are used for streaming and elastic traffic. For streaming traffic an exact expression for the QoS is derived and used to numerically evaluate the required capacity to satisfy the performance requirements. For elastic traffic it is not possible to derive an exact expression for the QoS and the required capacity will be determined by simulation. With the results from these models the impact of different behaviors of the corporate users and the number of corporate users is examined.

In operational capacity management, as implemented by network operators, the average workload on the link is used to determine the required capacity. If the traffic characteristics of corporate users are not known, a network operator has to provide the capacity that is required to fulfill the performance requirements in the worst-case scenario. However, the required capacity for a given value of the average workload strongly depends on the number of corporate users and their behavior, so an operator can improve the estimate for the required capacity when the behavior of a corporate user is known. We provide practical recommendations to obtain insight into the traffic characteristics at the end of this thesis.

Whether an operator should provide the required capacity for the worst-case scenario or perform the practical recommendations to get insight into the traffic characteristics depends on the trade-off between the potential profit of the knowledge of the traffic characteristics and the complexity of determining the traffic characteristics.

Contents

1 Introduction

Communication services such as electronic mail, file transfer, web browsing, online chat, IP telephony and real-time video are frequently used in today's life. These services (for both corporate users and consumers) generate traffic streams consisting of small packets that have to be transported via an underlying communication network. These traffic streams are highly variable and unpredictable. The quality of a service, as experienced by a user, degrades when packets are delayed or files are transmitted with a low throughput (average transmission rate). In order to limit packet delays and low user throughputs caused by congestion in the network, the link capacity in the communication network should be sufficiently large.

The objective of capacity management is to ensure that communication services are offered with the required quality, while minimizing the costs. The challenge of capacity management is that traffic continuously increases over time. The capacity of the network links should be increased before the service degrades due to congestion. On the contrary, when the capacity is expanded too early, this leads to unnecessary expenses.

We focus on the problem of dimensioning a link in a communication network which carries data of both consumers and a small number of corporate users. The corporate users can have an access rate that is much larger than that of the consumers. The traffic behavior of the corporate users is not specified. If a company purchases a high access rate, the traffic of this company can be the aggregate of many independent 'small' users (employees). Other types of behavior are also possible for a corporate user, e.g. the full link can be used only for a small fraction of the time when at the end of the day all data changes are transferred to update or backup a system.

In this thesis we present a modelling approach to determine the required capacity of a link in order to fulfill specific Quality of Service (QoS) levels in terms of delays and user throughputs. In particular, the impact of the presence of corporate users on the required link capacity is examined.

1.1 Outline of the thesis

This thesis starts with some background information on communication networks and modelling techniques in Section 2. In Section 3 we describe the situation considered in this thesis and translate it into a mathematical model. Two models are distinguished for the different types of services, i.e. streaming and elastic services. Performance requirements and parameters for the purpose of numerical evaluation are also stated in this section. The two models are analyzed separately in Sections 4 and 5. The analyses consist of the stationary distributions of the models, mathematical expressions for the performance requirements and the computation of quantities needed to determine the performance. The last parts of these sections contain numerical results. A comparison of the required capacities in the two models is presented in Section 6, along with some practical recommendations for determining the required capacity of a network link. Finally, the conclusions of this thesis are summarized in Section 7.

2 Background

We present a more comprehensive description of communication networks in Section 2.1 and in Section 2.2 we refer to some literature on modelling network traffic.

2.1 Communication networks

Communication services

A communication service generates traffic streams that have to be transported via an underlying communication network. These traffic streams consist of small packets (datagrams). Communication services can be divided into two categories: streaming and elastic services.

- **Streaming services** Examples of streaming services are real-time video and Voice over IP (VoIP). The latter is used to transport voice via packet-switched networks such as the internet and networks of operators. Streaming services generally have a real-time nature and continuously generate traffic. Insufficient link capacities lead to loss and delays of packets and cause degradation of the QoS. For streaming services it is important that an individual packet is delivered within certain delay restrictions. An occasional packet loss is typically allowed. Packets for streaming services are mostly transmitted using UDP (User Datagram Protocol) [15].

- **Elastic services** Examples of elastic services are web browsing and file transfer. For these services, users only notice the total transfer time of a file. As the sizes of files vary (email, web pages, backup's, mp3, movies), the average transfer rate of a file is a more adequate performance measure than the transfer time. Elastic services adapt their transmission rate to the level of congestion in the system while they are active. The protocol used to control the transmission rate is called TCP (Transmission Control Protocol) [16, 20].

Multiplexing

In a communication network the network operator does not have to provide the sum of all access rates due to so-called multiplexing gains in the core of the network. These gains arise because a user with a certain access rate does not use this total rate continuously for several reasons. The first reason is that he does not continuously use his computer during day and night. The second is that during a session, he is not always sending or receiving data. For example when he is using his computer for web browsing, data has to be transferred only when he clicks to load a new page and not when he reads a loaded page. The third reason is that not all applications use the full access rate.
For a network it suffices to provide less capacity than the sum of all access rates as active users fill the gaps of inactive users. This phenomenon is called multiplexing.

In practice multiplexing works as follows: each home has a DSL (Digital Subscriber Line) link to the DSLAM (DSL Access Multiplexer) and the DSLAM aggregates these data flows onto a single link with a high rate. After the DSLAM, the data is transferred over a link with a higher rate. For the moments that the amount of offered traffic briefly exceeds the link capacity, a buffer is used at the DSLAM to handle the excess of input data.

Quality of Service

The quality of a communication service is determined by user experience, which is a subjective measure. Moreover, the QoS (Quality of Service) requirement differs per service. For example, for email a user does not notice a few seconds delay of an individual packet, but only the total transfer time of a file. For VoIP however, a small delay already degrades the quality of the call, but an occasional packet loss is not noticed. For streaming video services, packet loss is annoying, as the loss of a single packet causes blocks in the images.

The two categories of communication services (streaming and elastic) basically have their own QoS requirements, e.g. see [17]. A typical performance measure used to describe the QoS of streaming (real-time) services formally is the fraction of packets with a large delay. For elastic services a typical performance measure is the fraction of files with a low throughput (average transfer rate).

A QoS requirement (performance requirement) states what fraction of traffic should be transferred with a sufficiently high performance. When traffic is transferred with this performance, the user does not notice QoS degradation. For streaming services the QoS requirement will state that a high fraction of packets is required to have a delay not exceeding a certain threshold. When the delay exceeds this threshold, the package is handled as if it is lost. For elastic services, a high fraction of the files is required to have a high throughput.

The communication network should have sufficient resources to offer the communication services with the required Quality of Service.

Capacity management

The objective of capacity management is to ensure that sufficient resources are available to handle the offered traffic, such that the required levels of Quality of Service (QoS) are satisfied. In terms of the performance measures mentioned above, this means determining the required capacity of a network link to fulfill the performance requirement.

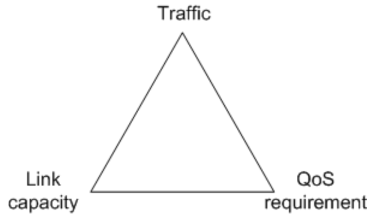


Figure 1: Connection between the three components of capacity management.

In Figure 1 the relation between the traffic, the QoS requirement and the link capacity is shown. Once two out of the three quantities are chosen, the third quantity can be determined.

The challenge of capacity management is that the amount of traffic continuously grows over time and the traffic behavior varies with the introduction of new highly-demanding services and continuously increasing access rates. The capacity of the network links should be increased before the service level degrades due to traffic congestion. On the contrary when the capacity is expanded too early this leads to unnecessary expense.

An example of a study for the dimensioning of a network link is performed by Fraleigh in his PhD thesis [7]. He studied the dimensioning for highly aggregated network links with a focus on delay-sensitive applications (streaming services). Another dimensioning approach is performed by Bonald, Olivier and Roberts [3]. They concentrate on data traffic at the flow-level and dimension the network link such that the throughput rates of TCP-connections remain above a certain threshold (for elastic traffic).

Service differentiation

Because of the diverse QoS requirements, one option is to handle packets from different services with different priorities. The packets for time-sensitive applications get a higher priority than packets from applications for which a little delay is allowed. An advantage is that the required bandwidth is smaller, because packets from services with a lower QoS can wait for a short period when there is no capacity left. A disadvantage is that the network complexity increases. Another option is to handle all packets equally within the requirements of the highest demanding service. This requires more bandwidth (overprovisioning), but the network operation is simpler, see e.g. [8].

Traffic measurements

Because of the growing use of internet services, we have to monitor the performance of the services. In practice, it is not possible to measure the direct perceived QoS. Instead we monitor the traffic load on the network links to get an indication for the performance of the services.

The traffic load can be measured at different time scales. Users already experience QoS degradation on a very small time scale, e.g. seconds for file transfers or web browsing and even less for interactive, real-time applications. But measuring the traffic on this time scale puts too much load on the network. An often used time scale for the measurements is 5 minutes. The workload that is used for determining the required capacity of the network link is the average workload in 5 minutes in the busiest hour of the day. With these 5-minute measurements we observe how the average traffic load grows in time, but actually we are interested in the tail of the traffic distribution, because data will be delayed when the amount of traffic exceeds a certain level. So we need to translate the 5-minute measurements into the traffic load on a shorter time scale. We need modelling to determine the required capacity to fulfill the performance requirements.

2.2 Modelling techniques

Modelling of network traffic

Large aggregates of internet traffic are often modelled as Gaussian traffic, which means that the amount of traffic in a period $[0, t]$ is a Gaussian process. An argument for this is provided by the Central Limit Theorem ([12], chapter 3). The assumption of Gaussian traffic is that the traffic on a link is the superposition of the traffic of many independent users. In [13] the Gaussianity of network traffic is examined.

In [2] a formula is derived for the required bandwidth as a function of the traffic load when the traffic distribution is Gaussian, i.e. under the assumption that the aggregate traffic originates from a large number of independent users. In this thesis the aggregates are not large enough to assume Gaussianity, so we cannot use the formula derived in [2].

If only a small number of users share a link, the traffic can be modelled as a superposition of ON-OFF sources. A user either is active and data is transmitted at a certain rate (ON) or the user is inactive, i.e. no data is transmitted (OFF).

Another option is to model the arrival process as a Compound Poisson process. Then requests for amounts of data arrive following a Poisson process. The difference in the interpretation of an ON-OFF arrival process and a Compound Poisson process is that with ON-OFF sources, we have a fixed pool of sources that alternate between being active or inactive. With a Compound Poisson process however, the active sources are not necessarily from a fixed pool of sources. In that case, it is possible to have a lot more sources which are all active less often than with ON-OFF sources. The arrival process cannot be modelled as a Compound Poisson process when the network consists of only a few traffic sources.
A second difference between an ON-OFF arrival process and a Compound Poisson process is that with a Compound Poisson process requests for amounts of data arrive instantaneously, while with an ON-OFF process they arrive gradually at a certain rate.

Markov fluid modelling

Standard Markov fluid queues consist of traffic sources feeding into a queue with a constant output rate. The sources are for instance of the ON-OFF type: they alternate between active and inactive periods. An overview of Markov fluid-models is presented in [9].

Streaming traffic flows into the buffer at a constant rate during a session. If the total input rate can exceed the constant output rate of the queue, every now and then the buffer of the queue fills. When the system is stable, the buffer content has a steady-state distribution. One of the first papers that contains a detailed performance analysis of the buffer content for statistically independent exponential ON-OFF sources is the paper of Anick, Mitra and Sondhi [1]. These results have been extended in many directions. In [14] for example, sources with a more general structure than exponential ON-OFF are considered.

Elastic traffic is transferred at a rate adjusted to the level of congestion in the network. Processor-sharing disciplines are used to share the capacity of the output link.

Processor-sharing disciplines

Under the processor-sharing discipline, all active flows are assumed to be processed simultaneously, each receiving a share of the server capacity. The Egalitarian PS (EPS) discipline is a basic model in which the capacity of the resource is assumed to be shared equally between all the users in the system. One of the main limitations of the EPS model is that it does not apply to heterogeneous systems, where flows may receive different service shares. A more abstract generalization of PS with a state-dependent service rate is the Generalized Processor Sharing (GPS) model as considered by Cohen [5]. In this model the total service rate is an arbitrary positive function of the total number of users in the system. As in the EPS discipline, all users receive an equal share of the total service rate. A model that allows for unequal sharing is Discriminatory Processor Sharing (DPS), where flows of different classes receive service at different rates. An analysis of the mean sojourn time conditioned on the service requirement in the M/G/1 queue with a DPS discipline is performed in [6].

3 The model

In this chapter we describe the situation we consider in this thesis in Section 3.1 and translate it into a mathematical model. We describe the properties of this model in Sections 3.2 to 3.5. In Section 3.6 we mention some reasonable values for parameters which can be used for the purpose of numerical evaluation.

3.1 Problem context

The goal of this thesis is to dimension a network link used by consumers together with one or two corporate users with an access rate significantly higher than the access rates of the consumers. The communication services of the users are either streaming or elastic services. In the first situation the packets generated by the service are transferred at a constant rate when the user is active. In the second situation the transmission rates are adapted to the level of congestion.

Traffic streams are aggregated by the network router using multiplexing (see Section 2.1). The rate (capacity) of the output link of the network router is higher than the rate of each single input link, but substantially smaller than the sum of all input rates. The required capacity of the output link depends on the traffic characteristics and the QoS requirements.

The traffic characteristics of a corporate user are not known. The traffic generated by a company can be the sum of the data traffic of the employees, but another possibility is that from time to time a large amount of data has to be transferred to the main office to update the system. In the first situation the traffic is handled as if it originates from many consumers. In the second situation we need to consider a different type of user with a larger access rate and longer interarrival and transmission times.

We want to be able to compare the required capacity in the two situations described above. To determine the required link capacity to satisfy the required Quality of Service we translate the situation described in this section into a mathematical model in the following sections.

3.2 Source behavior

We have two classes of sources representing two types of users (corporate users and consumers). In the case that the traffic of a corporate user is just the sum of the data traffic of the employees, the required capacity can be determined by only considering consumers, so then the number of corporate users is zero. In the model we assume that a corporate user transmits large files from time to time. The access rates of the corporate users are significantly higher than the access rates of the consumers. Therefore, we call the sources representing the corporate users the high-rate sources and the sources representing the consumers the low-rate sources. Define N_L as the number of low-rate sources and N_H as the number of high-rate sources. We consider a network with one or two corporate users, so $N_H \leq 2$. The model with $N_L = N$ and $N_H = 1$ is illustrated in Figure 2.

We use fluid-modelling (see Section 2.2) and model the arrival pattern as an ON-OFF process, which means that all sources alternate independently between the ON and OFF state.

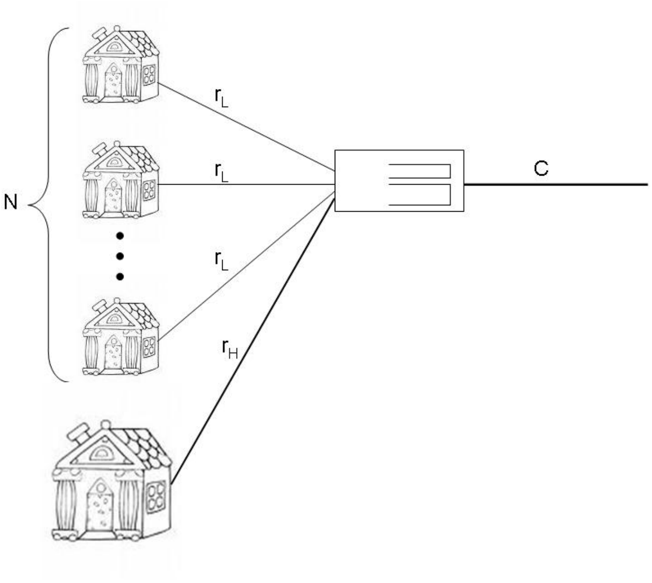


Figure 2: Model with 1 high-rate source and N low-rate sources.

We assume that the durations of the OFF periods are exponentially distributed. The rate at which a low-rate source turns on is λ_L and the rate for a high-rate source is λ_H. The durations of the ON periods are the transfer times of the files.

The file sizes are exponentially distributed, with mean file sizes f_L^{-1} and f_H^{-1} for the files originating from the low-rate and high-rate sources, respectively. The rate at which sources are allowed to transmit data can depend on the number of active sources. A low-rate source (high-rate source) can transmit data at rate $R_L(i, j)$ $(R_H(i, j))$ when i low-rate sources and j high-rate sources are active. The maximum rate at which data can be transmitted is the access rate of a source, which is denoted by r_L and r_H for the low-rate and the high-rate sources respectively.

All data is transmitted to a data-handling switch, which models the network router or switch. The rate (or capacity) of the output link of the switch is c. We vary the way the switch handles data originating from the low-rate and high-rate sources. When the total input rate into the buffer is in excess of the maximum transmission rate c, data is buffered in a single buffer or in separate buffers for each traffic class.

3.3 Models for streaming and elastic traffic

In this thesis two models are distinguished for the different types of services (streaming and elastic services). The way the data is generated and handled depends on the type of service.

3.3.1 The streaming model

Streaming services, such as VoIP and video generate data continuously. We assume that all sources transmit data at their access rates (r_L for the low-rate sources and r_H for the high-rate sources) and the data is served in order of arrival. In case of congestion data can be stored in a buffer. The switch can differentiate the treatment of the traffic of different classes, by handling traffic from a class with priority. In this case data of different classes is stored in separate buffers. If data of the high-rate source has strict priority over data of the low-rate source, the maximal output rate of the high-rate buffer is c and the output rate of the low-rate buffer is the remaining capacity.

The durations of the ON periods do not depend on the state of the system in the streaming model, so the distribution of the duration of an ON period of the sources is known. The durations of the active periods are exponentially distributed with rates $r_L f_L$ and $r_H f_H$ for the low-rate and high-rate sources respectively.

3.3.2 The elastic model

For data originating from elastic services, e.g. web browsing or file transfer, typically only the transfer time of the entire file matters. In the elastic model the output rate of the switch is shared between all sources and the sources are allowed to send data to the switch at that rate. In this way all data that arrives at the switch can immediately depart from the switch, so we do not need a buffer in the model in this situation. The data-handling discipline for the elastic model is Discriminatory Processor Sharing (DPS), which means that the capacity is shared between all active sources, but some sources receive a larger part of the capacity than other sources (see Section 2.2). The maximum rate at which data is transmitted while the source is active is r_L for the low-rate sources and r_H for the high-rate sources. Suppose that i low-rate sources and j high-rate sources are active at a certain moment. When $ir_L + jr_H \leq c$, all sources can transmit at their maximum transmission rate, but when $ir_L + jr_H > c$, the capacity has to be shared between the sources, and the transmission rates for the low-rate and high-rate sources become $R_L(i,j)$ and $R_H(i,j)$ respectively. We can vary the way the low-rate and high-rate sources are handled with the choice of $R_L(i,j)$ and $R_H(i,j)$. A few options are given:

- All sources receive a rate that is proportional to their access rate. Then the transmission rates are

$$\begin{aligned} R_L(i,j) &= \min\left(r_L, \frac{r_L}{ir_L + jr_H}c\right), & (1) \\ R_H(i,j) &= \min\left(r_H, \frac{r_H}{ir_L + jr_H}c\right). & (2) \end{aligned}$$

- Data of the high-rate sources has strict priority over data of the low-rate sources. Then the high-rate sources send their data at rate r_H or share the capacity when the number of active high-rate sources exceeds cr_H^{-1}. The low-rate sources share the remaining capacity (which can be 0). In this case

$$\begin{aligned} R_H(i,j) &= \min(r_H, \frac{1}{j}c), & (3) \\ R_L(i,j) &= \min(r_L, \frac{1}{i}(c - jR_H(i,j))) \\ &= \min(r_L, \frac{1}{i}(c - jr_H)^+). & (4) \end{aligned}$$

Where in the streaming model a buffer was used to deal with an excess of incoming data to the switch, in the elastic model an excess of incoming data is handled by decreasing the transmission rate of the sources. Hence in this case the duration of the ON period of a source depends on the number of sources in the system. As a consequence less data is transferred, because the durations of the OFF periods are unchanged. This is a realistic assumption, because for example a user who has to wait till his internet page is loaded still needs the same time to read the page.

3.4 Stability conditions

We assume that the system is stable. For the streaming model, we therefore have the following stability condition:

$$\rho := \rho_L + \rho_H < c, \tag{5}$$

where ρ_H and ρ_L represent the workload (the mean instantaneous total rate) of the high-rate and low-rate sources respectively. With ON-OFF sources this workload is given by:

$$\rho_L = \frac{N_L \lambda_L r_L}{\lambda_L + r_L f_L} \qquad \rho_H = \frac{N_H \lambda_H r_H}{\lambda_H + r_H f_H}. \tag{6}$$

The elastic model with ON-OFF values is always stable, because in that model the sources alternate slower when the server is very busy.

3.5 Performance requirements

The performance requirements that we use should guarantee that the performance of a communication service is satisfactory (see Section 2.1). However, there is no rule that prescribes which performance criterion should be used for which situation. We use the following performance criterions.

- For streaming services it is important that every individual packet is delivered on time. When a delay of d seconds does not cause a serious QoS degradation, but a higher delay does, the delay of most of the packets should not exceed d seconds. The performance requirement is then

$$\mathbb{P}(D > d) < \epsilon, \tag{7}$$

where D is the delay of a packet and ϵ is small. This performance requirement will be used for the low-rate sources and high-rate sources with d_L, ϵ_L and d_H, ϵ_H denoting the particular delays and probability thresholds respectively.
The delay of a packet depends on the buffer content at the moment the packet arrives. With a joint buffer for all traffic, the delay of a packet exceeds d when the buffer content exceeds dc. If two separate buffers are used for high-rate and low-rate traffic, the delay of a low-rate packet also depends on whether the high-rate source is active or not, because in that case the output rate of the low-rate buffer alternates.

- For elastic services it is important that the total file transmission does not take too long. Assume that the transmission time is still short enough when the average transmission rate is α times the access rate (r), but that the transmission takes too long when the average transmission rate is less than αr. Then the performance requirement is

$$\mathbb{P}(T < \alpha r) < \epsilon, \tag{8}$$

where T is the throughput (average transmission rate) of a file and ϵ is small. Let $\alpha_L, r_L, \epsilon_L$ and $\alpha_H, r_H, \epsilon_H$ denote the respective values for the low-rate and high-rate sources respectively.

3.6 Parameters for numerical evaluation

We present numerical values for the parameters of the model, which are used to obtain numerical results in Sections 4, 5 and 6. We consider the case that the corporate user is only active for a relatively short time and sends a comparably large file during his active period. The parameters are chosen as follows:

- The access rate of a high-rate source is 50 Mb/s and the access rate of a low-rate source is 5 Mb/s, so $r_H = 50$ and $r_L = 5$.
- The file that the corporate user wants to transmit is 100 MB, so $f_H = 1/800$ (note that a Byte is 8 bits). The file size for the low-rate sources is 5 MB, so $f_L = 1/40$.
- The high-rate source is active 1% of the total time. We achieve this with $\lambda_H = 1/(99{\cdot}16)$ We also consider the situation that the high-rate source is active 10% of the total time. In this case $\lambda_H = 1/144$. A low-rate source is active 2/7 of the total time, so $\lambda_L = 1/20$.
- We consider a maximum of 70 low-rate sources and two high-rate sources, so $N_L = 0, \ldots, 70$ and $N_H = 0, 1, 2$.
- At most 1% of the packets from a low-rate or high-rate source are allowed to have a delay more than 0.02 s (for streaming services). That means $d_L = d_H = 0.02$ and $\epsilon_L = \epsilon_H = 0.01$.
- At most 1% of the files for elastic services are allowed to have a throughput less than 0.8 times the access rate, so $\alpha_L = \alpha_H = 0.8$ and $\epsilon_L = \epsilon_H = 0.01$.

We now define three quantities that can be used as shorthand notations and describe some traffic characteristics.

- The first two quantities describe the fraction of time that a single source is active:

  $$\gamma_L := \frac{\lambda_L}{\lambda_L + r_L f_L}, \qquad \gamma_H := \frac{\lambda_H}{\lambda_H + r_H f_H}.$$

  Note that γ_H denotes the fraction of time a single high-rate source is active, so when there are two high-rate sources, each high-rate source is active a fraction γ_H of the time.

- The third quantity is the fraction of traffic that is from a high-rate source, which is

  $$\eta_H := \frac{\rho_H}{\rho},$$

  with ρ and ρ_H as in Equations (5) and (6).

The parameter values defined in this section are summarized in Table 1.

Low-rate Source		High-rate Source	
N_L	$0, \ldots, 70$	N_H	$0, 1, 2$
λ_L	$\frac{1}{20}$	λ_H	$\frac{1}{1584}, \frac{1}{144}$
f_L	$\frac{1}{40}$	f_H	$\frac{1}{800}$
r_L	5	r_H	50
α_L	0.8	α_H	0.8
d_L	0.02	d_H	0.02
ϵ_L	0.01	ϵ_H	0.01
γ_L	$\frac{2}{7}$	γ_H	$0.01, 0.1$

Table 1: Parameters for numerical evaluation.

4 Analysis of the streaming model

In this chapter we present an analysis of the streaming model. In the analysis of the model we assume that the buffer has infinite capacity (or both buffers have infinite capacity). This assumption provides a conservative approximation for the required capacity in order to fulfill the performance requirements. The reason is that no data is lost in the infinite-capacity buffer, so the buffer content of an infinite-capacity buffer is always greater or equal to the content of a finite-capacity buffer. The assumption of a buffer with infinite capacity provides a good approximation for the content of a buffer with finite capacity when the buffer content hardly ever exceeds the buffer capacity.

We calculate the joint distribution of the buffer content and the state of the system (the number of active high-rate and low-rate sources) in Section 4.1. In Section 4.2 we derive an expression for the performance measures we use to determine the required capacity and in Section 4.3 we derive the distribution of the delay of a low-rate packet if high-rate data has strict priority. In Section 4.4 an approximation based on time-scale decomposition is described. Finally, we show numerical results for the streaming model in Section 4.5.

4.1 Stationary distribution

We calculate the joint distribution of the total buffer content and the number of active sources. We assume that all data is handled equally such that we only have one buffer. However, the total buffer content and the state in a system with strict priority for high-rate data is identically distributed. We first consider the situation with only one high-rate source ($N_H = 1$) in Section 4.1.1 and then extend the results to the situation with multiple high-rate sources in Section 4.1.2.

4.1.1 Stationary distribution with a single high-rate source

Consider the situation $N_H = 1$, $N_L = N$. We can represent the process as a Markov process with a two-dimensional state space, representing the number of low-rate and high-rate sources that are active. Define states (i, j), where i is the number of active low-rate sources, $i = 0, \dots, N$ and j the number of active high-rate sources, $j = 0, 1$. The transition rates are shown in Figure 3.

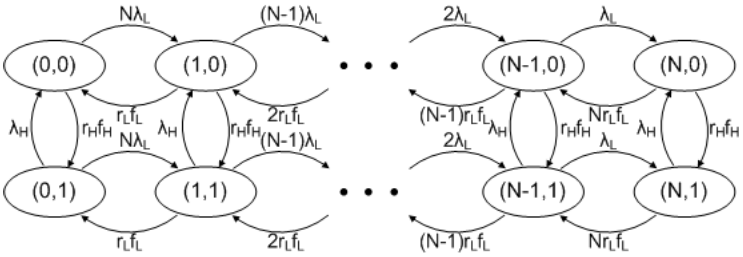


Figure 3: Transition rates.

The one-dimensional state space consists of $2N + 2$ states, where states 0 to N denote the states in which the high-rate source is inactive ($(0,0)$ to $(N,0)$) and states $N+1$ to $2N+1$ are the states in which the high-rate source is active. In this way, the states are ordered colexicographically. The generator matrix $\mathbf{Q}$ is as follows:

$$\mathbf{Q} = \begin{bmatrix} \mathbf{M} - \lambda_H \mathbf{I}_{N+1} & \lambda_H \mathbf{I}_{N+1} \\ r_H f_H \mathbf{I}_{N+1} & \mathbf{M} - r_H f_H \mathbf{I}_{N+1} \end{bmatrix}, \tag{9}$$

where

$$\mathbf{M} = \begin{bmatrix} -N\lambda_L & N\lambda_L & & & \\ r_L f_L & -((N-1)\lambda_L + r_L f_L) & (N-1)\lambda_L & & \\ & \ddots & \ddots & \ddots & \\ & & (N-1) r_L f_L & -(\lambda_L + (N-1) r_L f_L) & \lambda_L \\ & & & N r_L f_L & -N r_L f_L \end{bmatrix}$$

and $\mathbf{I}_n$ represents the identity matrix of order n.

The diagonal matrix $\mathbf{R}$ lists the net input rates into the buffer (depending on the state of the system):

$$\mathbf{R} = \operatorname{diag}\{-c, r_L - c, \ldots, N r_L - c, r_H - c, r_H + r_L - c, \ldots, r_H + N r_L - c\}. \tag{10}$$

Let N_t denote the state of the system at time t, $N_t \in \{0, \ldots, 2N+1\}$, and $V_t \in \mathbb{R}_0^+$ the buffer content at time t. The buffer content is a continuous measure, because we use fluid modelling for the data. We define the stationary joint distribution of N_t and V_t as

$$F_n(x) = \lim_{t \to \infty} \mathbb{P}(N_t = n, V_t \leq x).$$

We can calculate $F_n(x)$ as in [1]. With $(z_j, \boldsymbol{\psi}_j)$ an eigenvalue-eigenvector pair of $\mathbf{R}^{-1}\mathbf{Q}^T$, $j = 0, \ldots, 2N+1$, the solution is

$$\mathbf{F}(x) = \mathbf{F}(\infty) + \sum_{z_j < 0} a_j \boldsymbol{\psi}_j \exp(z_j x), \tag{11}$$

where $\mathbf{F}(x) = (F_0(x), \ldots, F_{2N+1}(x))^T$. The coefficients $a_j, j \in \{i | z_i < 0\}$ in Expression (11) can be obtained by using the boundary conditions $F_n(0) = 0$ for all states n for which $R_{nn} > 0$. These conditions hold because the buffer cannot be empty when the system is in a state with positive drift. Theorem 1 of [14] states that the number of negative eigenvalues is equal to the number of positive diagonal elements of R. So the number of negative eigenvalues is equal to the number of states with a positive drift. That means we have enough boundary conditions to calculate the a_j in Expression (11) uniquely.

When states exist for which the net input rate is zero, $\mathbf{R}$ has a 0 on the diagonal and $\mathbf{R}^{-1}$ does not exist. In [14] it is explained how to deal with this situation.

For $\mathbf{F}(\infty)$ in Equation (11) we know that

$$F_n(\infty) = \lim_{t \to \infty} \mathbb{P}(N_t = n, V_t \leq \infty) = \lim_{t \to \infty} \mathbb{P}(N_t = n) = \pi_n,$$

where π_n is the stationary probability distribution of the process being in state n. Observe that $\boldsymbol{\pi}$ is the normalized eigenvector that corresponds to eigenvalue 0, because the equations $\mathbf{Q}^T\boldsymbol{\pi} = \mathbf{0}$ and $\mathbf{R}^{-1}\mathbf{Q}^T\boldsymbol{\pi} = \mathbf{0} = 0 \cdot \boldsymbol{\pi}$ are equivalent. According to the definitions of stationary probabilities and eigenvectors, the former equation defines $\boldsymbol{\pi}$ to be the vector of stationary probabilities and the latter equation defines $\boldsymbol{\pi}$ to be the eigenvector that corresponds to eigenvalue 0. Moreover, the marginal numbers of active low-rate and high-rate sources are binomially distributed and independent of each other, so the stationary probability for a state n with n_L low-rate and n_H high-rate sources active is

$$\pi_n = \pi_{(n_L,n_H)} = b\left(n_L; N_L, \frac{\lambda_L}{\lambda_L + r_L f_L}\right) b\left(n_H; N_H, \frac{\lambda_H}{\lambda_H + r_H f_H}\right),$$

where

$$b(n; N, p) = \left(\begin{array}{c} N \\ n \end{array} \right) p^n (1-p)^{N-n}.$$

The authors of [1] analyzed a birth-death process and therefore the generator matrix had a tridiagonal form. In our process, we cannot only jump from n to $n-1$ and $n+1$, but also to $n+N+1$ or $n-N-1$. So now we lose the tridiagonal form of the generator matrix. The consequence is that now we do not have an explicit form for the eigenvalues and eigenvectors, so they can only be determined numerically.

The stationary marginal distribution of the buffer content follows from Expression (11) summing over all states. Define V as the buffer content at an arbitrary time instant. Then

$$\mathbb{P}(V \leq x) = \sum_{n=0}^{2N+1} F_n(x) = 1 + \sum_{z_j<0} a_j \mathbf{1}^T \boldsymbol{\psi}_j \exp(z_j x). \tag{12}$$

If data of the high-rate source has strict priority over data of the low-rate source and $r_H < c$, the high-rate buffer remains empty and the content of the low-rate buffer is equal to the total buffer content.

4.1.2 Stationary distribution with multiple high-rate sources

We can generalize the results of Section 4.1.1 to the situation with an arbitrary number of high-rate sources (N_H). The stationary distribution of the buffer content can be obtained in the same way as for one high-rate source, but now $\mathbf{Q}$ exists of N_H+1 by N_H+1 sub-matrices of size N_L+1. The (j,l)th sub-matrix of $\mathbf{Q}$ is given by

$$\mathbf{Q}[j,l] := \begin{cases} (N_H - j)\lambda_H \mathbf{I}_{N_L+1} & \text{if } l = j+1, \\ j r_H f_H \mathbf{I}_{N_L+1} & \text{if } l = j-1 \\ \mathbf{M} - \mathbf{Q}[j,j+1] - \mathbf{Q}[j,j-1] & \text{if } l = j, \\ 0 & \text{otherwise,} \end{cases} \tag{13}$$

in which

$$\mathbf{M} = \begin{bmatrix} -N_L\lambda_L & N_L\lambda_L & & & \\ r_L f_L & -((N_L-1)\lambda_L + r_L f_L) & (N_L-1)\lambda_L & & \\ & \ddots & \ddots & \ddots & \\ & & (N_L-1)r_L f_L & -(\lambda_L + (N_L-1)r_L f_L) & \lambda_L \\ & & & N_L r_L f_L & -N_L r_L f_L \end{bmatrix}.$$

The diagonal matrix $\mathbf{R}$ is

$$\mathbf{R} = \text{diag}\{\{-c, r_L - c, \ldots, N r_L - c\}, \ldots, \{N_H r_H - c, N_H r_H + r_L - c, \ldots, N_H r_H + N r_L - c\}\}.$$

The rest of the analysis in Section 4.1.1 remains the same.

4.2 Performance requirements

We derive expressions for the performance measures for streaming services as defined in Section 3.5 (Inequality (7)). Define

$$\begin{aligned} D_L &:= \text{Delay of a low-rate packet,} \\ D_H &:= \text{Delay of a high-rate packet.} \end{aligned}$$

The performance requirements for low-rate and high-rate sources are respectively:

$$\mathbb{P}(D_L > d_L) < \epsilon_L, \tag{14}$$

$$\mathbb{P}(D_H > d_H) < \epsilon_H. \tag{15}$$

Below we express these performance requirements in formula-form, for the case that a joint buffer is used in Section 4.2.1 and for the case that high-rate traffic has strict priority over low-rate traffic in Section 4.2.2.

4.2.1 Joint buffer

With a joint buffer all packets are handled equally after they arrive at the buffer. A packet that finds an amount of work B when it arrives at the buffer, has a delay of B/c seconds. Define

$$\begin{aligned} V_L &:= \text{Buffer content as observed by a low-rate packet,} \\ V_H &:= \text{Buffer content as observed by a high-rate packet.} \end{aligned}$$

The distributions of the buffer content as observed by a packet and the buffer content at an arbitrary moment are generally not equal, because the former is a packet-average measure and the latter is a time-average measure. Therefore, we have to translate the time-average measure into a packet-average measure.

The Performance Requirements (14) and (15) are now given by:

$$\mathbb{P}(D_L > d_L) = \mathbb{P}(V_L > d_L c) = \frac{\sum_{i,j}(\pi_{i,j} - F_{i,j}(d_L c))R_L(i,j)i}{\sum_{i,j}\pi_{i,j}R_L(i,j)i} < \epsilon_L, \tag{16}$$

and

$$\mathbb{P}(D_H > d_H) = \mathbb{P}(V_H > d_H c) = \frac{\sum_{i,j}(\pi_{i,j} - F_{i,j}(d_H c))R_H(i,j)j}{\sum_{i,j}\pi_{i,j}R_H(i,j)j} < \epsilon_H, \tag{17}$$

where $F_{i,j}(x) = F_{(N+1)j+i}(x)$. The numerator in Inequality (16) (Inequality (17)) is the expected number of low-rate (high-rate) packets transmitted per second while the buffer content exceeds $d_L c$ ($d_H c$). So it is the expected number of packets with a delay of more than d_L (d_H) seconds. The denominator is the total number of low-rate (high-rate) packets transmitted per seconds. $R_L(i,j)$ ($R_H(i,j)$) is the input rate of a low-rate (high-rate) source into the buffer when i low-rate sources and j high-rate sources are active. Recall that in this section we assume $R_L(i,j) = r_L$ and $R_H(i,j) = r_H$ for all i and j, thus Inequalities (16) and (17) can be simplified to

$$\sum_{i,j} F_{i,j}(d_L c)i > \gamma_L N_L(1 - \epsilon_L), \tag{18}$$

and

$$\sum_{i,j} F_{i,j}(d_H c)j > \gamma_H N_H(1 - \epsilon_H), \tag{19}$$

which are the performance requirements for the low-rate and high-rate sources in the streaming model with a joint buffer. Expressions (63) and (64) for the average packet delay in this case are derived in Appendix A.

4.2.2 Strict priority for high-rate traffic

If traffic originating from the high-rate sources has strict priority over traffic originating from the low-rate sources, we need two separate buffers. Observe that due to the assumption $c > r_H$ the high-rate buffer is always empty when $N_H = 1$. In this case the performance of the high-rate source is always excellent. In the case that $N_H = 2$ the high-rate buffer can contain data when both high-rate sources are active at the same time. The performance requirement for a high-rate source is then equal to the performance requirement of a high-rate source with a joint buffer when no low-rate sources are present.

Although a packet from a low-rate source always has to wait for at least the service of the total buffer content, which is distributed identically to the buffer content of a joint buffer, the distributions of the delay of a low-rate source are generally not equal in both situations. This is caused by the difference in the output rates of the low-rate buffer and the joint buffer without priorities. The output rate of the low-rate buffer alternates between $c - r_H$ and c when the high-rate source is active or inactive respectively. The output rate of the joint buffer without priorities is constant c.

Define

$$\begin{aligned} D_{LP} &:= \text{Delay of a low-rate packet when high-rate traffic has strict priority,} \\ D_{HP} &:= \text{Delay of a high-rate packet when high-rate traffic has strict priority,} \end{aligned}$$

and

$$G_{i,j}(d) := \lim_{t\to\infty} \mathbb{P}(D_{LP} > d, N_t = (N_L+1)j+i). \tag{20}$$

For the performance requirement for the low-rate sources in this situation we again need to translate the time-average measure $G_{i,j}(d)$ into a packet-average measure:

$$\mathbb{P}(D_{LP} > d_L) = \frac{\sum_{i,j} G_{i,j}(d_L) R_L(i,j) i}{\sum_{i,j} \pi_{i,j} R_L(i,j) i} < \epsilon_L, \tag{21}$$

The numerator in Inequality (21) is the expected number of packets per second of which the delay exceeds d_L. The denominator is the total number of low-rate packets transmitted per second. We use $R_L(i,j) = r_L$ for all i and j to simplify Inequality (21):

$$\sum_{i,j} i G_{i,j}(d_L) < \gamma_L N_L \epsilon_L. \tag{22}$$

This inequality is the performance requirement for a low-rate source in the streaming model if traffic originating from high-rate sources has strict priority over traffic originating from low-rate sources. The performance requirement for a high-rate source in this situation is:

$$\mathbb{P}(D_{HP} > d_H) = \mathbb{P}(D_H > d_H) < \epsilon_H \tag{23}$$

in a system where $N_L = 0$.

4.3 Delay of a low-rate packet with strict priority for high-rate traffic

We calculate the joint probability that the delay of a low-rate packet exceeds d_L seconds and the state of the system is (i,j) when high-rate traffic has strict priority over low-rate traffic and two separate buffers are used. The calculation in this section is for the situation with only one high-rate source ($N_H = 1$). For two high-rate sources only an expression for the left-hand side of Inequality (22) is given and the calculation can be found in Appendix B.

Consider a system with only one high-rate source. The high-rate buffer is always empty, because $r_H < c$. The output rate of the low-rate buffer alternates between $c - r_H$ and c, depending on whether the high-rate source is active or inactive. The state of the high-rate source can change during the period that a low-rate packet is in the buffer. Theoretically, the high-rate source can turn ON and OFF infinitely many times during the period that a low-rate packet is in the buffer. However, with the parameters as in Table 1, the probability that the state of the high-rate source changes two or more times is negligible. Therefore, we assume that the state of the high-rate source can change at most once while a packet is in the low-rate buffer.

The joint distribution of the buffer content and the state of the system is given by the cumulative distribution function in Expression (11). The probability that the buffer content is less than b and the state of the system is (i,j) is thus

$$F_{i,j}(b) = \pi_{i,j} + \sum_{z_k<0} a_k (\boldsymbol{\psi}_k)_{(N_L+1)j+i} \exp(z_k b),$$

and the probability density function is

$$f_{i,j}(b) = \sum_{z_k<0} a_k z_k (\boldsymbol{\psi}_k)_{(N_L+1)j+i} \exp(z_k b). \tag{24}$$

For the calculation of $G_{i,j}(d)$ (see Definition (20)), we condition on the buffer content b at the moment that a low-rate packet arrives at the buffer (when the state of the system is (i,j)). For a given buffer content b, the probability that the delay of a low-rate packet exceeds d is denoted by $G_{i,j}(d|V=b)$.

The maximal output-rate of the low-rate buffer is c (when the high-rate source is inactive). This implies that when $b > cd_L$, the delay of the packet always exceeds d_L seconds. On the contrary, when $b \le (c-r_H)d_L$, the delay of the packet never exceeds d_L seconds, even when the high-rate source is active. When $(c-r_H)d_L < b \le cd_L$, the exceedance probability depends on the time that the high-rate source changes state.

First consider $G_{i,0}(d_L|V=b)$ and look at the moment that the high-rate source becomes active, which is t seconds after the arrival of the low-rate packet at the buffer. The remaining buffer content in front of that packet is then $b-ct$, so the total delay of that packet will be $t+(b-ct)/(c-r_H)$. The delay exceeds d_L when $t < (b-(c-r_H)d_L)r_H^{-1}$, which has probability $1-\exp(-\lambda_H(b-(c-r_H)d_L)r_H^{-1})$, because the length of the OFF period of a high-rate source is exponentially (λ_H) distributed. We have

$$\begin{aligned}
&G_{i,0}(d_L)\\
&= \int_0^\infty f_{i,0}(b) G_{i,0}(d_L|V=b)\mathrm{d}b\\
&\approx \pi_{i,0} - F_{i,0}(cd_L) + \int_{(c-r_H)d_L}^{cd_L} f_{i,0}(b)\left(1-\exp\left(-\lambda_H \frac{b-(c-r_H)d_L}{r_H}\right)\right)\mathrm{d}b\\
&= \pi_{i,0} - (\pi_{i,0} + \sum_{z_k<0} a_k(\boldsymbol{\psi}_k)_i \exp(z_k c d_L))\\
&\quad + \int_{(c-r_H)d_L}^{cd_L} \sum_{z_k<0} a_k z_k (\boldsymbol{\psi}_k)_i \exp(z_k b)\left(1-\exp\left(-\lambda_H \frac{b-(c-r_H)d_L}{r_H}\right)\right)\mathrm{d}b\\
&= -\sum_{z_k<0} a_k(\boldsymbol{\psi}_k)_i \exp(z_k c d_L)\\
&\quad + \sum_{z_k<0} a_k z_k (\boldsymbol{\psi}_k)_i \int_{(c-r_H)d_L}^{cd_L}\left(\exp(z_k b) - \exp\left(z_k b - \lambda_H \frac{b-(c-r_H)d_L}{r_H}\right)\right)\mathrm{d}b\\
&= -\sum_{z_k<0} a_k(\boldsymbol{\psi}_k)_i \exp(z_k c d_L) + \sum_{z_k<0} a_k z_k(\boldsymbol{\psi}_k)_i \Big[\frac{1}{z_k}(\exp(z_k c d_L) - \exp(z_k(c-r_H)d_L))\\
&\quad - \frac{1}{z_k - \frac{\lambda_H}{r_H}}(\exp(z_k c d_L - \lambda_H d_L) - \exp(z_k(c-r_H)d_L))\Big]\\
&= \sum_{z_k<0} a_k(\boldsymbol{\psi}_k)_i \left[\frac{\lambda_H}{r_H z_k - \lambda_H}\exp(z_k(c-r_H)d_L) - \frac{r_H z_k}{r_H z_k - \lambda_H}\exp(z_k c d_L - \lambda_H d_L)\right].
\end{aligned}$$

Next consider $G_{i,1}(d_L|V = b)$ and look at the moment that the high-rate source becomes inactive, which is t seconds after the arrival of the low-rate packet at the buffer. The remaining buffer content in front of that packet is then $b-(c-r_H)t$, so the total delay of that packet will be $t + (b - (c - r_H)t)c^{-1}$. The delay exceeds d_L when $t > (cd_L - b)r_H^{-1}$, which has probability $\exp(f_H(b - cd_L))$, because the length of the ON period of a high-rate source is exponentially $(r_H f_H)$ distributed. We have

$$\begin{aligned}
&G_{i,1}(d_L)\\
&\quad= \int_0^\infty f_{i,1}(b)G_{i,1}(d_L|V = b)\\
&\quad\approx \pi_{i,1} - F_{i,1}(cd_L) + \int_{(c-r_H)d_L}^{cd_L} f_{i,1}(b)\exp\left(f_H(b - cd_L)\right)\mathrm{d}b\\
&\quad= \pi_{i,1} - (\pi_{i,1} + \sum_{z_k<0} a_k(\boldsymbol{\psi}_k)_{N_L+1+i}\exp(z_k cd_L))\\
&\qquad + \int_{(c-r_H)d_L}^{cd_L} \sum_{z_k<0} a_k z_k(\boldsymbol{\psi}_k)_{N_L+1+i}\exp(z_k b + f_H(b - cd_L))\mathrm{d}b\\
&\quad= -\sum_{z_k<0} a_k(\boldsymbol{\psi}_k)_{N_L+1+i}\exp(z_k cd_L)\\
&\qquad + \sum_{z_k<0} a_k z_k(\boldsymbol{\psi}_k)_{N_L+1+i}\frac{1}{z_k + f_H}(\exp(z_k cd_L) - \exp(z_k(c - r_H)d_L - r_H f_H d_L))\\
&\quad= -\sum_{z_k<0} a_k(\boldsymbol{\psi}_k)_{N_L+1+i}\left[\frac{f_H}{z_k + f_H}\exp(z_k cd_L) + \frac{z_k}{z_k + f_H}\exp(z_k(c - r_H)d_L - r_H f_H d_L)\right].
\end{aligned}$$

The left-hand side of Inequality (22) can now be computed:

$$\begin{aligned}
&\sum_{i,j} iG_{i,j}(d_L)\\
&\quad= \sum_{z_k<0} a_k\Bigg[\sum_i i(\boldsymbol{\psi}_k)_i\left(\frac{\lambda_H}{r_H z_k - \lambda_H}\exp(z_k(c - r_H)d_L) - \frac{r_H z_k}{r_H z_k - \lambda_H}\exp(z_k cd_L - \lambda_H d_L)\right)\\
&\qquad -\sum_i i(\boldsymbol{\psi}_k)_{N_L+1+i}\left(\frac{f_H}{z_k + f_H}\exp(z_k cd_L) + \frac{z_k}{z_k + f_H}\exp(z_k(c - r_H)d_L - r_H f_H d_L)\right)\Bigg].
\end{aligned}$$

For $N_H = 2$ the calculation of the left-hand side of Inequality (22) can be found in Appendix B and the result is:

$$\begin{aligned}
&\sum_{i,j} iG_{i,j}(d_L) \\
&\quad= \sum_{z_k<0} a_k \Bigg[\sum_i i(\boldsymbol{\psi}_k)_i \Bigg(\frac{2\lambda_H}{r_H z_k - 2\lambda_H} \exp(z_k(c-r_H)d_L) \\
&\qquad - \frac{r_H z_k}{r_H z_k - 2\lambda_H} \exp(z_k c d_L - 2\lambda_H d_L) \Bigg) - \sum_i i(\boldsymbol{\psi}_k)_{N_L+1+i} \Bigg(\frac{f_H}{z_k + f_H} \exp(z_k c d_L) \\
&\qquad + \frac{z_k}{z_k + f_H} \exp(z_k(c-r_H)d_L - r_H f_H d_L) - \exp(z_k(c-r_H)d_L) \\
&\qquad - \frac{\lambda_H}{r_H z_k - \lambda_H} \exp(z_k(c-2r_H)d_L) + \frac{r_H z_k}{r_H z_k - \lambda_H} \exp(z_k(c-r_H)d_L - \lambda_H d_L) \Bigg) \\
&\qquad - \sum_i i(\boldsymbol{\psi}_k)_{2(N_L+1)+i} \Bigg(\frac{2f_H}{z_k + 2f_H} \exp(z_k(c-r_H)d_L) \\
&\qquad + \frac{z_k}{z_k + 2f_H} \exp(z_k(c-2r_H)d_L - 2r_H f_H d_L) \Bigg) \Bigg].
\end{aligned}$$

4.4 Approximations based on time-scale decomposition

If the ON and OFF periods of the high-rate source are very long compared to the ON and OFF periods of the low-rate sources, the high-rate source alternates much slower than the low-rate sources and the buffer content in the streaming model can almost reach a stationary distribution when the high-rate source is active or inactive. To reach this situation we assume that

$$\lambda_L \gg \lambda_H, \quad r_L f_L \gg r_H f_H. \tag{25}$$

This situation can be analyzed using time-scale decomposition, because the low-rate and high-rate sources alternate at a different time-scale. In this section we assume that there is only one high-rate source. The approximation that we introduce consists of two parts, for two separate regimes.

- Regime I is for the situation $r_H + \rho_L \geq c$ (Section 4.4.2). In this case the drift for the buffer is positive when the high-rate source is active and negative when the high-rate source is inactive. We approximate the input of the low-rate sources with their mean in this regime.

- Regime II is for the situation $r_H + \rho_L < c$ (Section 4.4.3). Then the mean input rate is always less than the capacity, even when the high-rate source is active. However, the mean input rate is much higher when the high-rate source is active than when it is inactive. In this regime we use a quasi-stationary approximation, in which the stationary distribution is a combination of the stationary distributions of the system where the high-rate source is active or inactive respectively.

4.4.1 Two equivalent models

In this section we consider two equivalent manners to model the sources and the switch for the situation with $N_H = 1$. Further, assume that the rate at which the high-rate source transmits data is less than the total service rate, i.e. $r_H < c$.

- The first manner is that we consider a model in which N_L sources are of the same type and one source is of a different type. In this system, the service rate (the rate at which data can stream out of the buffer) is constant c. This manner is used in Section 4.4.2.
- The second manner models only the N_L low-rate sources, but now the service rate is time-varying. In this model, r_H is subtracted from the total service rate c when the high-rate source is active. So now we have an alternating service rate, which is c when the high-rate source is inactive and $c - r_H$ when the high-rate source is active. This manner is used in Section 4.4.3.

These two models are the same in the sense that they result in the same distribution for the total buffer content. This is because in the first model, the service rate is constant, but when the high-rate source is active, r_H Mb/s is needed to serve the data from this source. So the remaining service rate is the same as in the second model.

4.4.2 Regime I: approximation low-rate sources

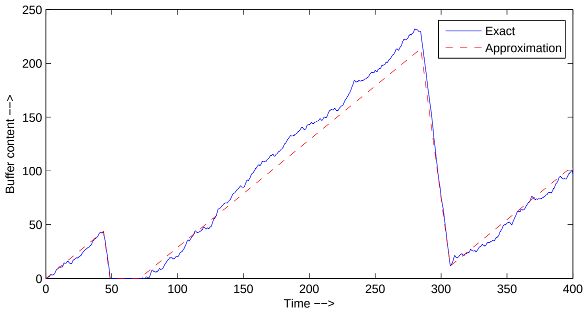


Figure 4: Sample-path of the buffer content in Regime I.

This regime covers the situation where $r_H + \rho_L > c$. When the high-rate source is active, the drift is positive, so the buffer content is expected to increase. When the high-rate source is inactive, the drift is negative, because then the mean rate of traffic that arrives is $\rho_L < c$, according to the stability condition stated in Equation (5). To fulfill Assumption (25), we fix λ_H and f_H and let λ_L and f_L tend to infinity with γ_L (as defined in Section 3.6) fixed.

The rate at which the buffer content increases or decreases is not constant during an active or inactive period of the high-rate source, because the low-rate sources are alternating (see the solid line in Figure 4). However, because we assume that the low-rate sources alternate considerably faster than the high-rate source, we approximate the rate at which data of the low-rate sources flows into the buffer with the average rate ρ_L (the dashed line in Figure 4).

This approximate model can be explicitly analyzed as in [1] with a single ON-OFF source (the high-rate source). This source has an exponentially (λ_H) distributed OFF period, an exponentially $(r_H f_H)$ distributed ON period, transmission rate of the source r_H and maximum output rate $\hat{c} := c - \rho_L$. We know that $r_H > c - \rho_L = \hat{c}$, according to the regime we examine. This means that when the high-rate source is ON, the buffer content grows. The stationary distribution of the state of the source and the buffer content in this case is denoted with $\mathbf{F}^H$.

$$\mathbf{F}^H(x) = \mathbf{F}^H(\infty) + \sum_{z_j<0} a_j \boldsymbol{\psi}_j \exp(z_j x), \tag{26}$$

where $(z_j, \boldsymbol{\psi}_j)$ is an eigenvalue-eigenvector pair of $\mathbf{R}^{-1}\mathbf{Q}^T$, $j = 0, 1$.
Let $\boldsymbol{\pi}^H$ denote the stationary distribution of the number of active high-rate sources (0 or 1):

$$\boldsymbol{\pi}^H := \mathbf{F}^H(\infty) = \begin{bmatrix} 1-\gamma_H \\ \gamma_H \end{bmatrix}.$$

Next we compute the eigenvalues of $\mathbf{R}^{-1}\mathbf{Q}^T$.

$$\mathbf{Q} = \begin{bmatrix} -\lambda_H & \lambda_H \\ r_H f_H & -r_H f_H \end{bmatrix} \text{ and } \mathbf{R} = \begin{bmatrix} -\hat{c} & 0 \\ 0 & r_H - \hat{c} \end{bmatrix}, \text{ so}$$

$$\mathbf{R}^{-1}\mathbf{Q}^T = \frac{-1}{\hat{c}(r_H - \hat{c})} \begin{bmatrix} r_H - \hat{c} & 0 \\ 0 & -\hat{c} \end{bmatrix} \begin{bmatrix} -\lambda_H & r_H f_H \\ \lambda_H & -r_H f_H \end{bmatrix} = \begin{bmatrix} \frac{\lambda_H}{\hat{c}} & -\frac{r_H f_H}{\hat{c}} \\ \frac{\lambda_H}{r_H - \hat{c}} & -\frac{r_H f_H}{r_H - \hat{c}} \end{bmatrix}.$$

The eigenvalues of $\mathbf{R}^{-1}\mathbf{Q}^T$ are the roots of the characteristic equation $\det\left(\mathbf{R}^{-1}\mathbf{Q}^T - zI\right) = 0$. So

$$\begin{aligned} 0 = \det\left(\mathbf{R}^{-1}\mathbf{Q}^T - z\mathbf{I}\right) &= z^2 - \left(\frac{\lambda_H}{\hat{c}} - \frac{r_H f_H}{r_H - \hat{c}}\right) z - \frac{\lambda_H r_H f_H}{\hat{c}(r_H - \hat{c})} + \frac{\lambda_H r_H f_H}{\hat{c}(r_H - \hat{c})} \\ &= z\left(z - \left(\frac{\lambda_H}{\hat{c}} - \frac{r_H f_H}{r_H - \hat{c}}\right)\right). \end{aligned}$$

We conclude that the eigenvalues are $z_0 = \frac{\lambda_H}{\hat{c}} - \frac{r_H f_H}{r_H - \hat{c}}$ and $z_1 = 0$. According to Expressions (5) and (6), the stability condition in this regime is $\gamma_H r_H < c - \rho_L = \hat{c}$, which implies $\frac{\lambda_H}{\hat{c}} - \frac{r_H f_H}{r_H - \hat{c}} < 0$. So z_0 is the only contributor to the sum in Expression (26). The eigenvector with eigenvalue z_0 is

$$\boldsymbol{\psi}_0 = \begin{bmatrix} \frac{r_H - \hat{c}}{\hat{c}} \\ 1 \end{bmatrix}.$$

Now Equation (26) becomes

$$\mathbf{F}^H(x) = \begin{bmatrix} 1-\gamma_H \\ \gamma_H \end{bmatrix} + a_0 \begin{bmatrix} \frac{r_H - \hat{c}}{\hat{c}} \\ 1 \end{bmatrix} \exp\left\{\left(\frac{\lambda_H}{\hat{c}} - \frac{r_H f_H}{r_H - \hat{c}}\right) x\right\}.$$

The constant a_0 can be found from the boundary condition $F_1(0) = 0$, because in state 1, the drift is upwards and the buffer cannot be empty. This gives $a_0 = -\gamma_H$, so the solution of the simplified problem is

$$\mathbf{F}^H(x) = \begin{bmatrix} 1-\gamma_H \\ \gamma_H \end{bmatrix} - \gamma_H \begin{bmatrix} \frac{r_H-\hat{c}}{\hat{c}} \\ 1 \end{bmatrix} \exp\left\{ \left(\frac{\lambda_H}{\hat{c}} - \frac{r_H f_H}{r_H - \hat{c}} \right) x \right\}.$$

Further, we get the following approximation for the buffer content distribution:

$$\mathbb{P}(V \leq x) \approx F_0^H(x) + F_1^H(x) = 1 - \frac{\gamma_H r_H}{\hat{c}} \exp\left\{ \left(\frac{\lambda_H}{\hat{c}} - \frac{r_H f_H}{r_H - \hat{c}} \right) x \right\}. \tag{27}$$

Note that with this approximation, we have

$$\mathbb{P}(V > x | V > 0) \approx \exp\left\{ \left(\frac{\lambda_H}{\hat{c}} - \frac{r_H f_H}{r_H - \hat{c}} \right) x \right\}.$$

So the distribution of the buffer content, conditioned on the buffer being non-empty, is approximated with an exponential distribution. This is explained by the exponential duration of the ON periods of the high-rate source.

In this approximation, an underlying assumption is that the buffer content does not depend on the number of active low-rate sources in the system. Therefore, to get an approximation for the joint distribution of the state of the system and the buffer content, we just have to multiply the probability that a certain number of low-rate sources are active with the right entry of $\mathbf{F}^H(x)$. Let $\hat{\mathbf{F}}(x)$ denote the approximate joint distribution of the state of the system and the buffer content. Then for state $n = (n_L, n_H)$, we have:

$$\hat{F}_n(x) = \begin{pmatrix} N_L \\ n_L \end{pmatrix} (\gamma_L)^{n_L} (1-\gamma_L)^{N_L - n_L} F_{n_H}^H(x). \tag{28}$$

We now show that the approximate distribution of the buffer content (Expression (27)) equals the limit of the exact distribution (Expression (12)) in the following way:

$$\lim_{\substack{\lambda_L \to \infty, f_L \to \infty \\ \gamma_L \text{constant}}} \mathbb{P}(V \leq x) = 1 - \frac{\gamma_H r_H}{\hat{c}} \exp\left\{ \left(\frac{\lambda_H}{\hat{c}} - \frac{r_H f_H}{r_H - \hat{c}} \right) x \right\}. \tag{29}$$

Equation (29) implies that for $\lambda_L \to \infty$ and $f_L \to \infty$ while γ_L remains constant, the contribution of the eigenvalues in the summation in Expression (12) becomes negligible, except for the dominant eigenvalue (the largest eigenvalue that is less than zero). The numerical evaluation shown in Figure 5 suggests that this dominant eigenvalue converges to $\frac{\lambda_H}{c-\rho_L} - \frac{r_H f_H}{r_H-(c-\rho_L)}$ when $\lambda_L \to \infty$ and γ_L is kept constant. For this figure, we used the parameters as listed in Table 1 with $N_L = 25$, $N_H = 1$, $\gamma_H = 0.01$ and $c = 82$.

The rate of convergence of the dominant eigenvalue depends on the parameter values. When c is small (but larger than $\rho_L + \rho_H$ for stability reasons), the numerical results indicate that the convergence is fast. When c is close to $r_H + \rho_L$, the convergence is slow.

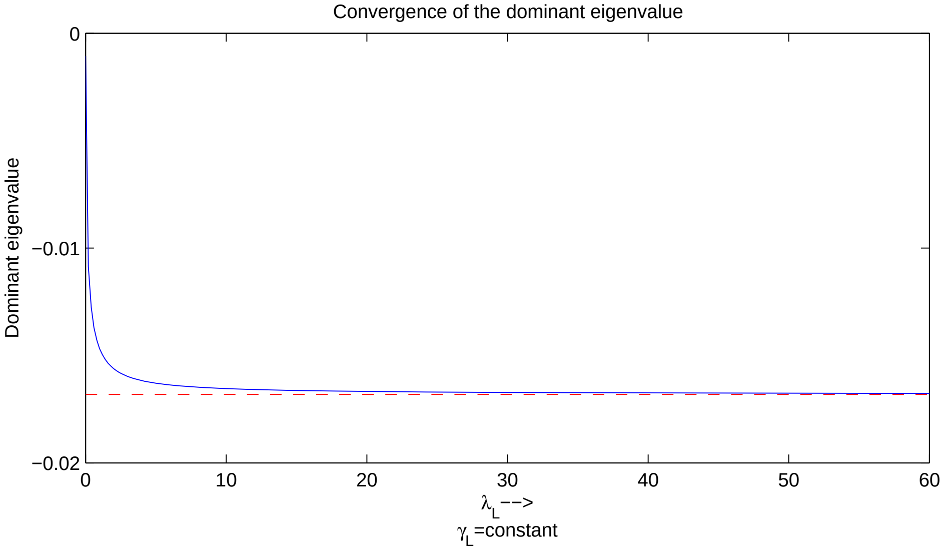


Figure 5: Convergence of the dominant eigenvalue. The solid line is the dominant eigenvalue for the exact distribution. The dashed line is $\frac{\lambda_H}{c-\rho_L} - \frac{r_H f_H}{r_H-(c-\rho_L)}$.

4.4.3 Regime II: a quasi-stationary approximation

Regime II covers the situation where $r_H + \rho_L < c$. Here, when the high-rate source is active, the rate at which data arrives at the switch is still on average less than the capacity c. This means that the buffer content does not grow very large, as can be seen in Figure 6.

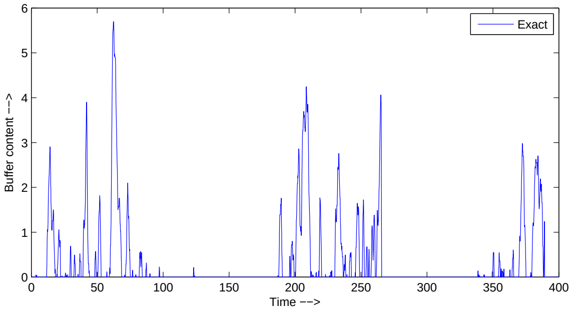


Figure 6: Sample-path of the buffer content in Regime II.

We use the equivalence between an extra high-rate source or an alternating service rate, as described in Section 4.4.1. Until now, we used the first point of view, but now we switch to the second (equivalent) manner to model the system. Hence we only consider data originating from the low-rate sources, but with an alternating service rate. The service rate alternates between $c - r_H$ when the high-rate source is ON and c when the high-rate source is OFF.

To fulfill Assumption (25) we now fix λ_L and f_L and let λ_H and f_H decrease to 0 with γ_H fixed. In the limit situation, the high-rate source cannot turn ON or OFF any more. So if the high-rate source alternates slow enough compared to the low-rate sources, we can approximate the buffer content conditioned on the high-rate source being ON or OFF with the buffer content when the high-rate source is always ON or OFF. Both cases can be explicitly analyzed separately with the theory of [1]. We describe how to compute the conditional distribution of the buffer content when the high-rate source is always inactive (service rate c).

The conditional distribution of the buffer content when the high-rate source is always OFF is:

$$\begin{aligned}\mathbb{P}(V \leq x|\text{H always OFF}) &= 1 + \sum_{j=0}^{N_L - \lfloor \frac{c}{r_L} \rfloor - 1} a_j \mathbf{1}^T \boldsymbol{\psi}_j \exp(z_j x) \\ &= 1 + \sum_{j=0}^{N_L - \lfloor \frac{c}{r_L} \rfloor - 1} a_j \Phi_j(1) \exp(z_j x),\end{aligned}$$

and the conditional joint distribution of the state of the system and the buffer content is:

$$\mathbf{F}^{OFF}(x) = \boldsymbol{\pi}^L + \sum_{j=0}^{N_L - \lfloor \frac{c}{r_L} \rfloor - 1} a_j \boldsymbol{\psi}_j \exp(z_j x)$$

where:

- $z_{N_L - \lfloor \frac{c}{r_L} \rfloor - 1} < \cdots < z_1 < z_0 < 0$ are the negative eigenvalues of $\mathbf{R}^{-1}\mathbf{Q}^T$ and $\boldsymbol{\psi}_j$ the corresponding eigenvectors, $j = 0, \ldots, N_L - \lfloor \frac{c}{r_L} \rfloor - 1$. The matrix $\mathbf{Q}$ is the generator matrix and $\mathbf{R}$ is the matrix containing the net rates at which data flows into the buffer. These matrices are as follows:

$$\mathbf{Q} = \begin{bmatrix} -N_L\lambda_L & N_L\lambda_L & & & \\ r_L f_L & -((N_L-1)\lambda_L + r_L f_L) & (N_L-1)\lambda_L & & \\ & \ddots & \ddots & \ddots & \\ & & (N_L-1)r_L f_L & -(\lambda_L + (N_L-1)r_L f_L) & \lambda_L \\ & & & N_L r_L f_L & -N_L r_L f_L \end{bmatrix}$$

and $\mathbf{R} = \text{diag}\{-c, r_L - c, 2r_L - c, \ldots, N_L r_L - c\}$.

- $a_j = -\gamma_L^{N_L} \prod_{i=0, i \neq j}^{N_L - \lfloor \frac{c}{r_L} \rfloor - 1} \frac{z_i}{z_i - z_j}, \qquad 0 \leq j \leq N_L - \lfloor \frac{c}{r_L} \rfloor - 1.$

- $\Phi_j(x)$ is the generating function of the j^{th} eigenvector, i.e.

$$\Phi_j(x) = \sum_{i=0}^{N_L} (\boldsymbol{\psi}_j)_i x^i.$$

The value of $\Phi_j(1)$ can be determined without explicitly calculating the eigenvectors [1].

- $\boldsymbol{\pi}^L$ is the stationary distribution of the number of active low-rate sources, which is binomially distributed.

The distribution of the buffer content, conditioned on the high-rate source to be always ON can be found analogously when replacing c by $c - r_H$. The same holds for $\mathbf{F}^{ON}(x)$, the joint distribution of the state of the system and the buffer content, conditioned on the high-rate source being active. Further, the stationary probability that the high-rate source is OFF is $1 - \gamma_H$ and that the high-rate source is ON is γ_H.

An approximation for the unconditional distribution of the buffer content in this case is

$$\mathbb{P}(V \leq x) \approx (1-\gamma_H)\mathbb{P}(V \leq x|\text{H always OFF}) + \gamma_H \mathbb{P}(V \leq x|\text{H always ON}). \tag{30}$$

This is the distribution of the total buffer content and also the distribution of the buffer content of the low-rate buffer in case data of the high-rate source has strict priority over the data of the low-rate source. In the latter situation, the high-rate buffer remains empty, because in this regime $r_H < r_H + \rho_L < c$.

The approximation for the joint distribution of the state of the system and the buffer content is $\hat{\mathbf{F}}(x)$, which is for state $n = (n_L, n_H)$:

$$\hat{F}_n(x) = (1-n_H)(1-\gamma_H)F^{OFF}_{n_L}(x) + n_H \gamma_H F^{ON}_{n_L}(x). \tag{31}$$

As for regime I, the exact distribution of the buffer content (Expression (12)) should converge to the approximate distribution (Expression (30)) when Assumption (25) is fulfilled. In the quasi-stationary approximation for regime II, this means

$$\begin{aligned}\lim_{\substack{\lambda_H \to 0, f_H \to 0 \\ \gamma_H \text{ constant}}} \mathbb{P}(V \leq x) &= (1-\gamma_H)\mathbb{P}(V \leq x|\text{H always OFF}) \\ &\quad + \gamma_H \mathbb{P}(V \leq x|\text{H always ON})\end{aligned} \tag{32}$$

should hold.

We now give an outline of the proof for Equation (32). We take the limit $\lambda_H \to 0$ and $f_H \to 0$ in Equation (12). In this distribution, λ_H and f_H appear in the eigenvalues z_j and eigenvectors $\boldsymbol{\psi}_j$ of the matrix $\mathbf{R}^{-1}\mathbf{Q}^T$, with $\mathbf{Q}$ and $\mathbf{R}$ as in Definitions (9) and (10). The eigenvalues of a matrix are the roots of the characteristic polynomial. The roots of a polynomial are continuous functions of its coefficients ([10], p. 539), so the eigenvalues of a matrix are continuous functions of the entries of that matrix. That means when we take the limit of a few entries going to zero, the limits of the eigenvalues are the eigenvalues of the limit of the matrix. So the eigenvalues of $\mathbf{R}^{-1}\mathbf{Q}^T$ when $\lambda_H \to 0$ and $f_H \to 0$ are the eigenvalues of $\mathbf{R}^{-1}\mathbf{Q}^T$ after taking the limit of the entries of this matrix.

In the limit, we have

$$\mathbf{Q} \to \begin{bmatrix} \mathbf{M} & \mathbf{0} \\ \mathbf{0} & \mathbf{M} \end{bmatrix} \text{ and } \mathbf{R} = \begin{bmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & r_H \mathbf{I}_{N_L+1} + \mathbf{D} \end{bmatrix},$$

with $\mathbf{M}$ as on page 22 and $\mathbf{D} = \text{diag}\{-c, r_L - c, \ldots, N_L r_L - c\}$.

The limit of $\mathbf{R}^{-1}\mathbf{Q}^T$ then consists of two submatrices with nonzero entries, i.e. $\mathbf{D}^{-1}\mathbf{M}^T$ and $(r_H \mathbf{I}_{N_L+1} + \mathbf{D})^{-1}\mathbf{M}^T$. That means the set of eigenvalues of $\mathbf{R}^{-1}\mathbf{Q}^T$ is the union of the set of eigenvalues of $\mathbf{D}^{-1}\mathbf{M}^T$ and the set of eigenvalues of $(r_H \mathbf{I}_{N_L+1} + \mathbf{D})^{-1}\mathbf{M}^T$. These two matrices are exactly the matrices we used to find the eigenvalues for the distribution of the buffer content in case that the high-rate source is always active or inactive. Hence the eigenvalues used in the exact and approximate distribution of the buffer content are equal.

Above we proved that the left and right sides of Equation (32) both are linear combinations of the same exponentials. Now it remains to prove that also the coefficients of these linear combinations coincide. The details are not contained in this thesis.

4.4.4 Usefulness of the time-scale decomposition approximations

Next we discuss whether the approximations in Sections 4.4.2 and 4.4.3 are useful for our model, with the parameters as in Table 1.

We distinguished between two regimes for the approximations based on time-scale decomposition. Regime I presented an approximation for the buffer content when $r_H + \rho_L > c$. In this regime the buffer content starts to grow when the high-rate source becomes active, caused by a positive drift in this regime. To satisfy the required QoS, the high-rate source is then only allowed to be active for a really short time. But this contradicts Assumption (25), which we need for the approximation to be accurate.

Most common in real life is the situation in regime II, where $r_H + \rho_L < c$ as described in Section 4.4.3. In this regime, the high-rate source can be active for a very long time, because there is no positive drift in this case. However, the active and inactive periods of the high-rate source do not always satisfy Assumption (25). For example, if at the end of the day a company transmits a large amount of data to the main office to update or backup the system, the inactive period is large compared to the inactive periods of the low-rate sources, but the active period is not long enough for the low-rate sources to reach stationarity.

4.5 Numerical results for the streaming model

The goal of the numerical study in this section is to examine whether the required capacity depends on the kind of traffic in the system. With a given workload of the system, we want to know whether the knowledge of the number of high-rate sources and the fraction of the time these sources are active gives us important information in order to determine the required capacity. We also present some rules-of-thumb to estimate the required capacity in several situations.

In this section we present numerical results with the parameter values as listed in Table 1 in Section 3.6 (unless mentioned otherwise). We determine both the required capacity to satisfy the performance requirement for the low-rate sources and for the high-rate sources. These capacities are denoted with:

Notation	Sources	Service differentiation	Performance requirement
c_L	low-rate	no service differentiation	(18)
c_H	high-rate	no service differentiation	(19)
c_{LP}	low-rate	strict priority for high-rate sources	(22)
c_{HP}	high-rate	strict priority for high-rate sources	(23)

To fulfill the performance requirements for all sources, the required capacity is $\max(c_L, c_H)$ (or $\max(c_{LP}, c_{HP})$). When $N_H = 0$ the capacities c_H, c_{HP} and c_{LP} are not defined, so then the required capacity is just c_L.

Required capacity as function of the workload

In Figure 7 we present the required capacity as a function of the workload in the system. We use the exact distribution of the buffer content as given in Expression (11). In the top figure a high-rate source is active 1% of the time ($\gamma_H = 0.01$, $\lambda_H = 1/1584$) and in the lower figure this is 10% of the time ($\gamma_H = 0.1$, $\lambda_H = 1/144$). The required capacity for a low-rate or high-rate source can only be determined if at least one source of that kind is present in the network. This explains the different starting values for ρ in the figure.

We observe that the required capacity for a high-rate source is higher than the required capacity for a low-rate source. This is because a high-rate source always experiences a busy system.

For a particular workload, the required capacity depends on the number of high-rate sources. This is because, although the average load is equal in all situations, the behavior is more bursty when more high-rate sources are present. When the workload is low the dependence of the required capacity on the number of high-rate sources is the largest, because then the difference in the total transmission rate when the high-rate source is active or inactive is the most extreme.

If a high-rate source is only active 1% of the time (as in the top diagram of Figure 7), only a small amount of extra capacity is required for the low-rate sources in a system with two high-rate sources, compared to a system with one high-rate source. This is because the fraction of time that both high-rate sources are active at the same time, i.e. 0.01%, is negligible. However, for the performance of a high-rate source holds that during its active period, on average 1% of the time a second high-rate source is active. So the capacity required for the high-rate source does increase when two high-rate sources are present in the network instead of one.

If a high-rate source is active 10% of the time (as in the lower diagram of Figure 7), both c_L and c_H increase significantly when two high-rate sources are present in the network instead of one.

Remarkable in Figure 7 is the difference between the shapes of the curves. In the top diagram, the curves for $\{c_L, N_H = 0\}$ and $\{c_H, N_H = 1\}$ look like square-root functions, while the other three curves have a different shape. We know that the required capacity as a function of the workload on a link with a large number of identical users has a square-root behavior [2]. Below we analyze which sources we have to take into account when determining the required

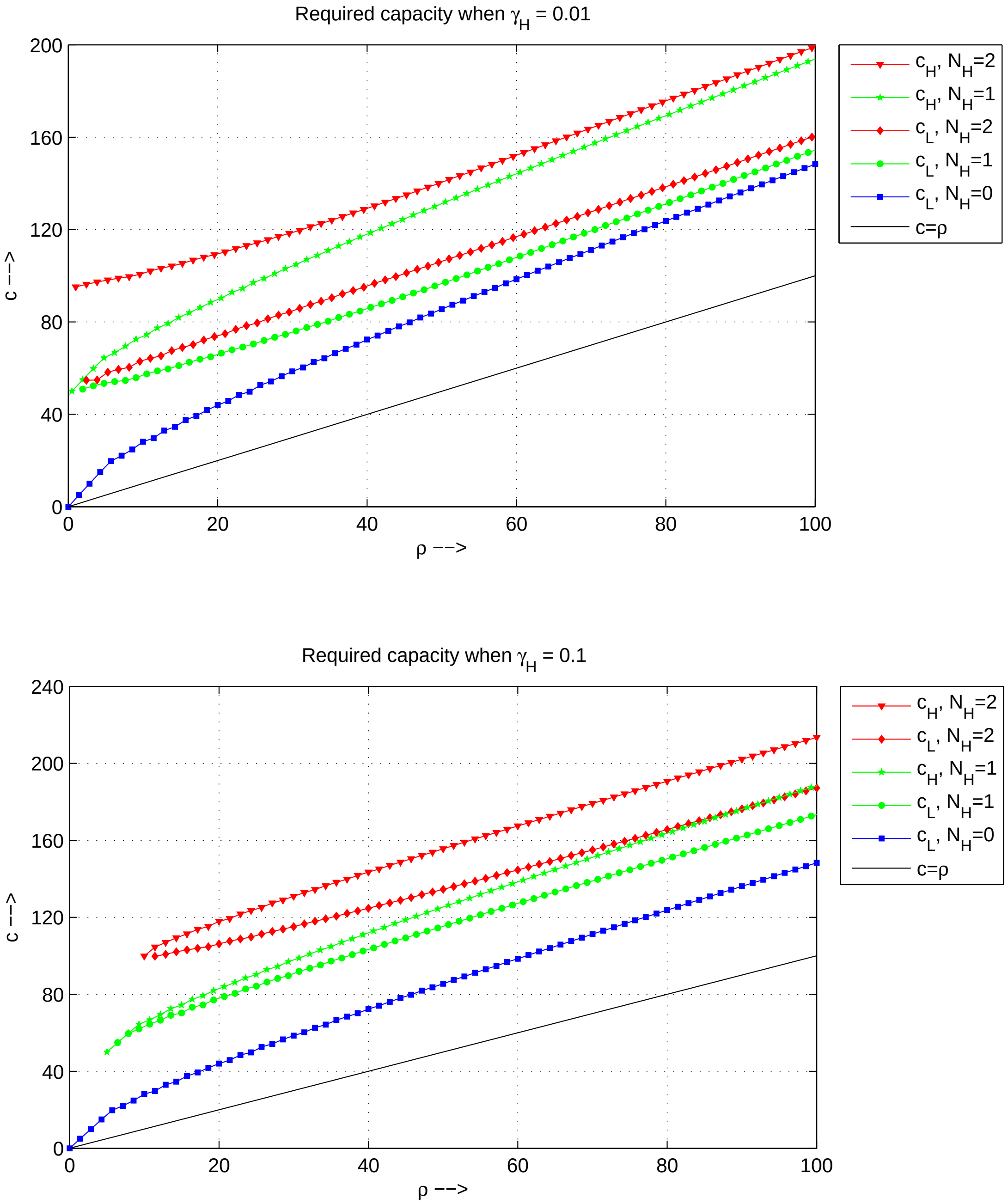


Figure 7: Required capacity depending on the workload in the system for streaming traffic.

capacity in each curve when $\gamma_H = 0.01$:

- $c_L, N_H = 0$: The required capacity for a low-rate source. No high-rate sources are present.
- $c_L, N_H = 1$: The required capacity for a low-rate source when 1% of the time a high-rate source is active.
- $c_H, N_H = 1$: The required capacity for a high-rate source. Apart from this high-rate source, only low-rate sources are present. The required capacity is approximately r_H plus the required capacity for a low-rate source with $N_H = 0$.
- $c_L, N_H = 2$: The required capacity for a low-rate source when 1.98% of the time one high-rate source is active and 0.01% of the time two high-rate sources are active. This last percentage is negligible, because it is equal to $\epsilon_L/10$.
- $c_H, N_H = 2$: The required capacity for a high-rate source. Another high-rate source is active 1% of the time. The required capacity is approximately r_H plus the required capacity for a low-rate source with $N_H = 1$.

With this information we can conclude that the capacity has a square-root behavior when we only have low-rate sources to take into account. In the lower diagram of Figure 7 the curves for $\{c_L, N_H = 1\}$ and $\{c_H, N_H = 2\}$ also look like square-root functions. In these cases, one extra high-rate source is present (apart from the source for which the capacity is determined). This high-rate source is active 10% of the time, which is 10 times ϵ_L. It follows that almost the full access rate needs to be reserved for the high-rate source, because only a fraction ϵ_L of the low-rate packets are allowed to have a large delay. The required capacity is then this reserved rate plus the required capacity for the low-rate sources. The curve for $\{c_L, N_H = 2\}$ does not look like a square-root function, because when a low-rate source is active, 18% of the time one high-rate source is active and also 1% of the time two high-rate sources are active. Now the latter percentage is not negligible.

Trade-off between γ_H and ϵ_L

While inspecting Figure 7 we got the suspicion that the required capacity to satisfy a certain QoS level strongly depends on the trade-off between the fraction of the time a high-rate source is active and the fraction of data that can be delayed. Let us consider a network with one high-rate source. In Figure 8 a sample path of the buffer content is drawn. The color of the curve shows whether the high-rate source is active or not. During the red parts of the figure, this source is active and during the blue parts it is inactive. A long delay is caused by a large buffer content.

We consider the required capacity for a low-rate source (c_L) and we distinguish between three possible situations for the relation between γ_H and ϵ_L:

- $\gamma_H \ll \epsilon_L$. In this case the high-rate source is only active for a very small fraction of the time such that data from the low-rate sources is allowed to be delayed more than d_L seconds (almost) always when the high-rate source is active. A consequence is that

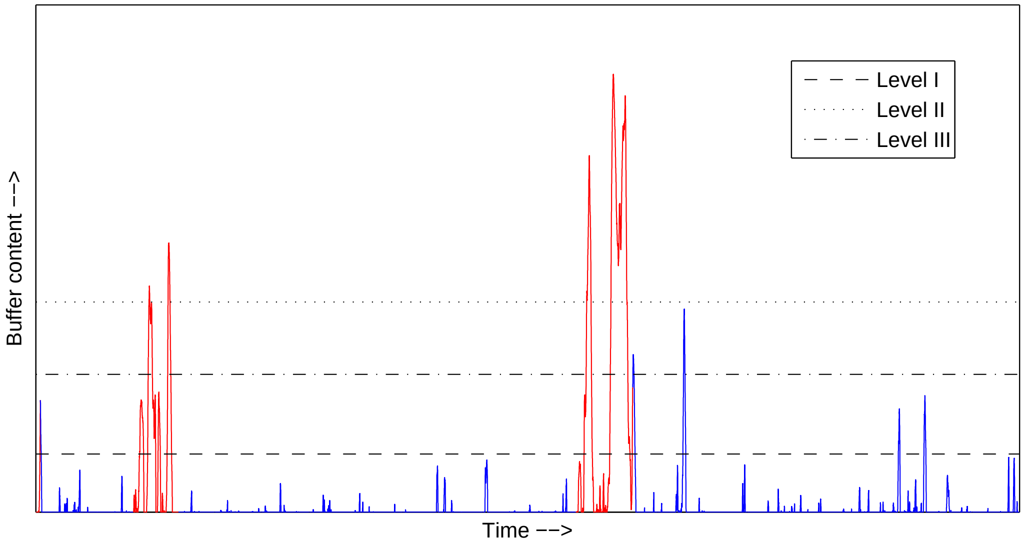


Figure 8: Sample path of the buffer content.

the buffer content is allowed to exceed the level $d_L c$ when the high-rate source is active. Then an approximation for the required capacity in this case is the required capacity for the low-rate sources only. However, ϵ_L should be decreased, because of the delayed data when the high-rate source is active. We use some formulas to explain this.

The performance requirement that we use is $\mathbb{P}(D_L > d_L) < \epsilon_L$, where D_L is the delay of a low-rate packet. We have

$$\begin{aligned}\mathbb{P}(D_L > d_L) &= \mathbb{P}(D_L > d_L|H\ ON)\gamma_H + \mathbb{P}(D_L > d_L|H\ OFF)(1-\gamma_H) \qquad (33)\\ &\leq \gamma_H + \mathbb{P}(D_L > d_L|H\ OFF)(1-\gamma_H).\end{aligned}$$

If $\gamma_H < \epsilon_L$ and the parameter settings are as stated in Table 1 then $\mathbb{P}(D_L > d_L|H\ ON)$ is close to 1, so the upper bound for $\mathbb{P}(D_L > d_L)$ is very tight. We approximate $\mathbb{P}(D_L > d_L|H\ OFF)$ with $\mathbb{P}(D_L > d_L|H$ always $OFF)$. The latter probability is less than the former, which causes that the required capacity is a little higher than the capacity obtained with the following approximate performance requirement:

$$\mathbb{P}(D_L > d_L|H \text{ always } OFF) < \frac{\epsilon_L - \gamma_H}{1-\gamma_H}.$$

An indication for the required capacity compared to the level of the buffer content such that the delay of $100(1-\epsilon_L)\%$ of the data is less than d_L seconds is given by level I in Figure 8.

- $\gamma_H > \epsilon_L$. In this case the fraction of low-rate traffic that is allowed to be delayed is less than the fraction of time that the high-rate source is active. The buffer content exceeds the level $d_L c$ mainly when the high-rate source is active, which is illustrated by level II in Figure 8. Therefore we assume that the buffer content only exceeds this level when the high-rate source is active and we approximate the required capacity with the capacity that is needed when the high-rate source is always active. However, a larger fraction of packets is allowed to be delayed in this case, because then a fraction $1-\gamma_H$ a the time the buffer content is below $d_L c$. Again, we use Formula (33) to explain this. Now we approximate $\mathbb{P}(D_L > d_L | H\ OFF)$ with 0 and get

$$\mathbb{P}(D_L > d_L) \geq \mathbb{P}(D_L > d_L | H\ ON)\gamma_H,$$

which is a tight bound with the parameter settings as stated in Table 1. We approximate $\mathbb{P}(D_L > d_L | H\ ON)$ with $\mathbb{P}(D_L > d_L | H \text{ always } ON)$. Now the latter probability is higher than the former, which causes that the required capacity is a little less than the capacity obtained with the following performance requirement:

$$\mathbb{P}(D_L > d_L | H \text{ always } ON) < \frac{\epsilon_L}{\gamma_H}.$$

- $\gamma_H \approx \epsilon_L$. In this situation the level $d_L c$ is situated somewhere between the levels in the foregoing situations as illustrated by level III in Figure 8.

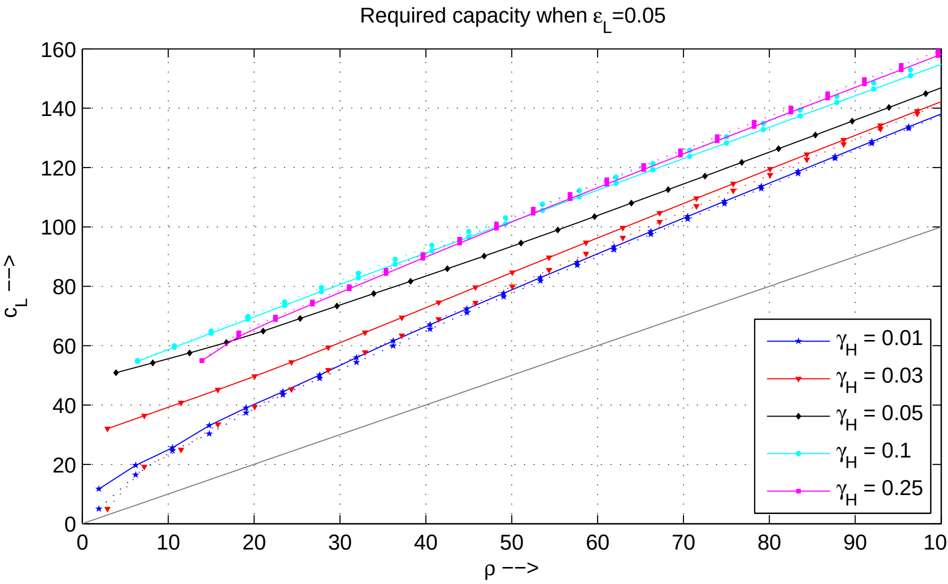


Figure 9: Required capacity depending on the fraction of time the high-rate source is active when $\epsilon_L = 0.05$.

In Figure 9 the required capacity as a function of the workload is plotted for several values of γ_H with $\epsilon_L = 0.05$. The approximations introduced above are also plotted in this figure (the dotted curves). In this figure, we can see that if $\gamma_H \approx \epsilon_L$, the required capacity as function of the workload does not look like a square-root function, but if $\gamma_H \geq 5\epsilon_L$ or $\gamma_H \leq \epsilon_L/5$, it does look like a square-root function. The explanation for this phenomenon comes from [2], because the required capacity is determined by the low-rate sources in these cases (with an adapted value for ϵ_L) and we reserve the full access-rate for the high-rate source if $\gamma_H > 5\epsilon_L$.

As we can see in Figure 9, the approximation introduced above for the situation $\gamma_H < \epsilon_L$ is a good approximation if $\gamma_H < \epsilon_L/5$, especially for higher workloads. The approximation for the situation $\gamma_H > \epsilon_L$ is immediately very good. An advantage of the latter approximation is that the capacity obtained is an upper bound for the required capacity.

Although we analyzed the trade-off between γ_H and ϵ_L with $N_H = 1$, we can also use these results for the trade-off between γ_H and ϵ_H with $N_H = 2$, because then a high-rate source observes a system with (apart from itself) only one high-rate source.

Remarkable in Figure 9 is that the curve for $\gamma_H = 0.25$ crosses the curve for $\gamma_H = 0.1$. The explanation is that the required capacity is not a monotonously increasing or decreasing function of the fraction of traffic originating from a high-rate source (η_H), as can be seen in Figure 10 and is explained in the next part of this section.

Required capacity as function of the fraction of traffic from high-rate sources

The top diagram of Figure 10 illustrates the required capacity for a varying number of high-rate sources and the fraction of traffic that originates from the high-rate sources. The average workload is kept at a constant value of 60 Mb/s. Hence when we have a system with the high-rate sources sending a larger amount of data to the buffer, there are less low-rate sources. In this figure, the red lines (corresponding to three high-rate sources) are drawn only for illustration. The curves corresponding to one high-rate source end at $\eta_H = 5/6$, because one high-rate source (with access rate 50) is not enough to reach $\rho = 60$.

First consider the required capacity for a low-rate source. If there are no high-rate sources at all, the required capacity is a little less than 100, as can be seen in the top diagram of Figure 10 (zero traffic from high-rate sources). If only a small fraction of the total traffic originates from the high-rate source(s), the required capacity increases rapidly. Then it reaches a maximum value and after that the capacity decreases. With two or three high-rate sources, we see some inflection points. This is because if the high-rate sources are only active a very little fraction of the time, the probability that two or more of them are active at the same time is negligible. But if the high-rate sources are active more frequently, this probability grows and becomes significant. With $N_H = 3$, the same holds for 3 high-rate sources active at the same time. The location of the inflection points can be explained by considering the probability that multiple high-rate sources are active as is displayed in the second diagram of Figure 10. The inflection points in the top diagram are located approximately at the point where this probability is ϵ_L.

The fact that the required capacity is not a monotone increasing or decreasing function of the fraction of traffic originating from the high-rate source can be explained by analyzing the variance of the instantaneous input rate from the sources into the buffer. If the high-rate source is only active for a small fraction of the time, The workload originates from a higher number

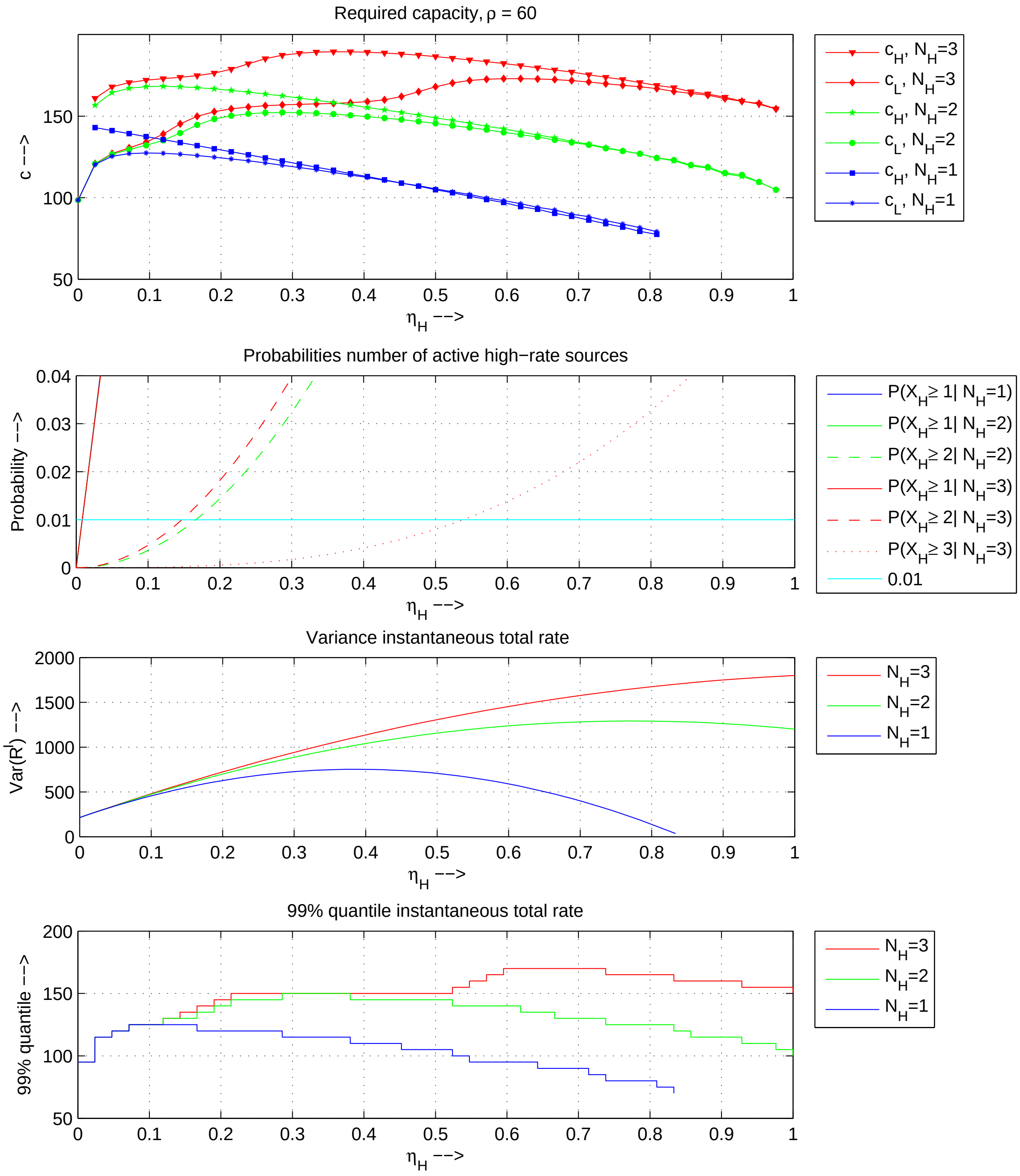


Figure 10: Required capacity depending on the fraction of traffic that is from high-rate sources.

of low-rate sources. With many independent sources with low access rates, the instantaneous input rate is less variable. If the high-rate sources are responsible for a larger amount of data, there is more variability in the system. However, when a large part of the data is from the high-rate source(s), there are less low-rate sources and the peak rates are lower. In Appendix C.1 the variance of the instantaneous rate is calculated. The variance function is plotted in the third diagram of Figure 10.

Although the variance of the instantaneous rate explains the fact that the required capacity is not a monotonous function of the fraction of traffic originating from a high-rate source, the maximum value for the required capacity is reached for a fraction of traffic from the high-rate source that is lower than the fraction that gives the maximum variance.

The delay restrictions ($d_L = 0.02$ and $\epsilon_L = 0.01$) are very stringent. When the total input rate exceeds the output rate of the buffer, the level $0.02c$ is reached almost immediately. When we would use $d_L = 0.0001$ instead of $d_L = 0.02$, the shapes of the curves as in the top diagram of Figure 10 do not change. The curves are only shifted up a little. In the bottommost diagram of Figure 10, the 99% quantile of the instantaneous input rate is displayed. This quantile is the solution of $0.01 = \mathbb{P}(R^I > x_{0.99})$. The calculation of this quantile can be found in Appendix C.2. The quantile is a step function of η_H, because the number of low-rate sources in the system is discrete and so the instantaneous rate also is a step function. We observe that the required capacity for a low-rate source is close to the 99% quantile.

Next consider the required capacity for a high-rate source. When apart from the high-rate source only low-rate sources are present, the required capacity depends on the number of low-rate sources in the system, which decreases when η_H increases. In a system with two or three high-rate sources, a high-rate source observes a system with the low-rate sources and (apart from itself) one or two high-rate sources respectively. This explains that the shape of a curve for c_H is very much the same as the shape of a curve for c_L with one high-rate source less.

Numerical results time-scale approximation

Next we examine the performance of the approximations based on time-scale decomposition (see Section 4.4). In the top diagram of Figure 11 the required capacity for both a low-rate and a high-rate source is drawn. The solid curves show the required capacities computed with the exact formula for the buffer content (given in Expression (11)). The dotted curves show the capacities as a result of the approximation based on time-scale decomposition. For this approximation we use Expressions (28) and (31) for the joint distribution of the buffer content and the state of the system when $c < r_H + \rho_L$ (regime I) and $c > r_H + \rho_L$ (regime II) respectively. The parameters used for this figure are stated in Table 1.

In the two other diagrams of Figure 11 the fraction of delayed packets is plotted as a function of the capacity of the switch. We observe a singular point in the approximated curves. This singular point is situated at $c = r_H + \rho_L$ and separates the approximation in regime I from the approximation in regime II. The existence of this singular point can be easily seen in Expression (27). When $r_H + \rho_L = c$, we have $r_H - \hat{c} = r_H - (c - \rho_L) = 0$. So in the

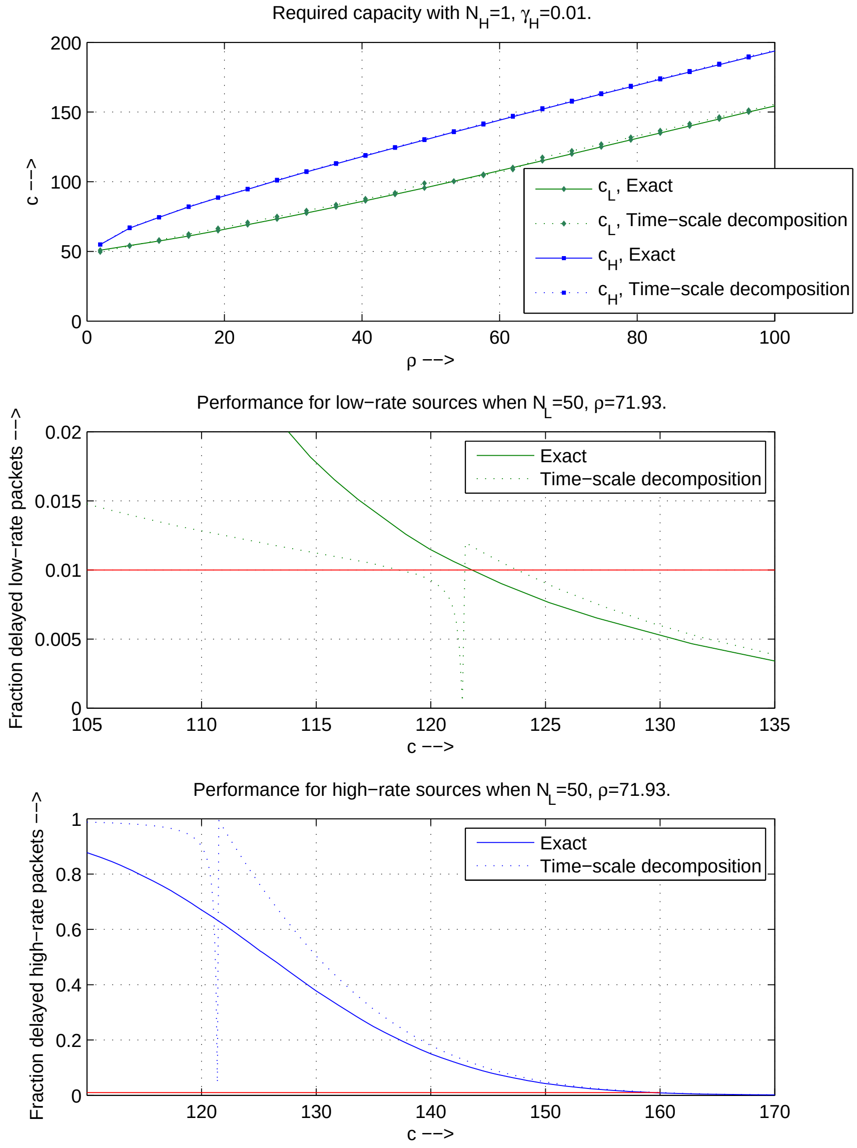


Figure 11: Approximation based on time-scale decomposition when $\gamma_H = 0.01$.

approximation this point is not defined, which causes a singular point.
The intuitive explanation for the existence of the singular point is that in regime I we approximate the input rate of the low-rate sources with their average input rate, so when $c = r_H + \rho_L$ the buffer is always empty in this approximation.

In the two lower diagrams of Figure 11, we can see how the capacity is determined with the exact and approximate distributions for the buffer content. The required capacity is the intersection of the fraction of delayed packets with the horizontal line at level $\epsilon_L = \epsilon_H = 0.01$. With the exact distribution this intersection point is unique. However, with the approximated buffer content there are multiple points at which the fraction of delayed packets can cross the line 0.01. To obtain a good approximation for the required capacity we start looking for an intersection point in regime II.

As we can see in the middle diagram of Figure 11, the capacity obtained with the approximate buffer distribution and the exact capacity differ most when $c \approx r_H + \rho_L$. The required capacity for the low-rate sources is less than $r_H + \rho_L$ on the left side of $\rho = 70\,\text{Mb/s}$ and more than $r_H + \rho_L$ on the right side. Although in this case $c \approx r_H + \rho_L$, the capacity that follows from the approximated buffer content already is quite good. If $c \gg r_H + \rho_L$ as is the case in the lowermost diagram of Figure 11, the approximation is very good.

Strict priority for high-rate traffic

We examine the required capacity in case high-rate traffic has strict priority over low-rate traffic and compare this capacity to the required capacity with a joint buffer. Figure 12 is equal to Figure 7 with extra curves which show the required capacities c_{LP} and c_{HP} if traffic originating from the high-rate sources has strict priority over traffic originating from the low-rate sources. We observe that

$$c_L \leq c_{LP} \leq c_H. \tag{34}$$

The leftmost inequality of Statement (34) can be explained by considering the amount of data that is served before the low-rate packet can be served. With a joint buffer, the delay of a low-rate packet is the transfer time of the amount of data that is present in the buffer upon arrival. With two separate buffers and strict priority for a high-rate source, a low-rate packet still has to wait until the total buffer content at arrival is transmitted and also for the high-rate packets that arrive during the waiting period of the low-rate packet. Therefore, the delay of a low-rate packet in this case is greater or equal to the delay in a system with a joint buffer, so $c_{LP} \geq c_L$.

The explanation for the rightmost inequality of Statement (34) is that a high-rate source always experiences a busy system. If traffic originating from high-rate sources gets priority and the buffers are empty most of the time, only a fraction γ_H of the time (for one high-rate source) low-rate packets are extra delayed and the delay of the rest of the low-rate packets remains small. This is also the reason that the difference between c_L and c_{LP} is negligible, together with the fact that the delay of most of the low-rate packets already exceeds d_L with a joint buffer when the high-rate source is active. So when the delay of a packet is high with a joint buffer, the delay would be very high when using priorities. However, the fraction of traffic for which the delay exceeds d_L does not increase significantly.

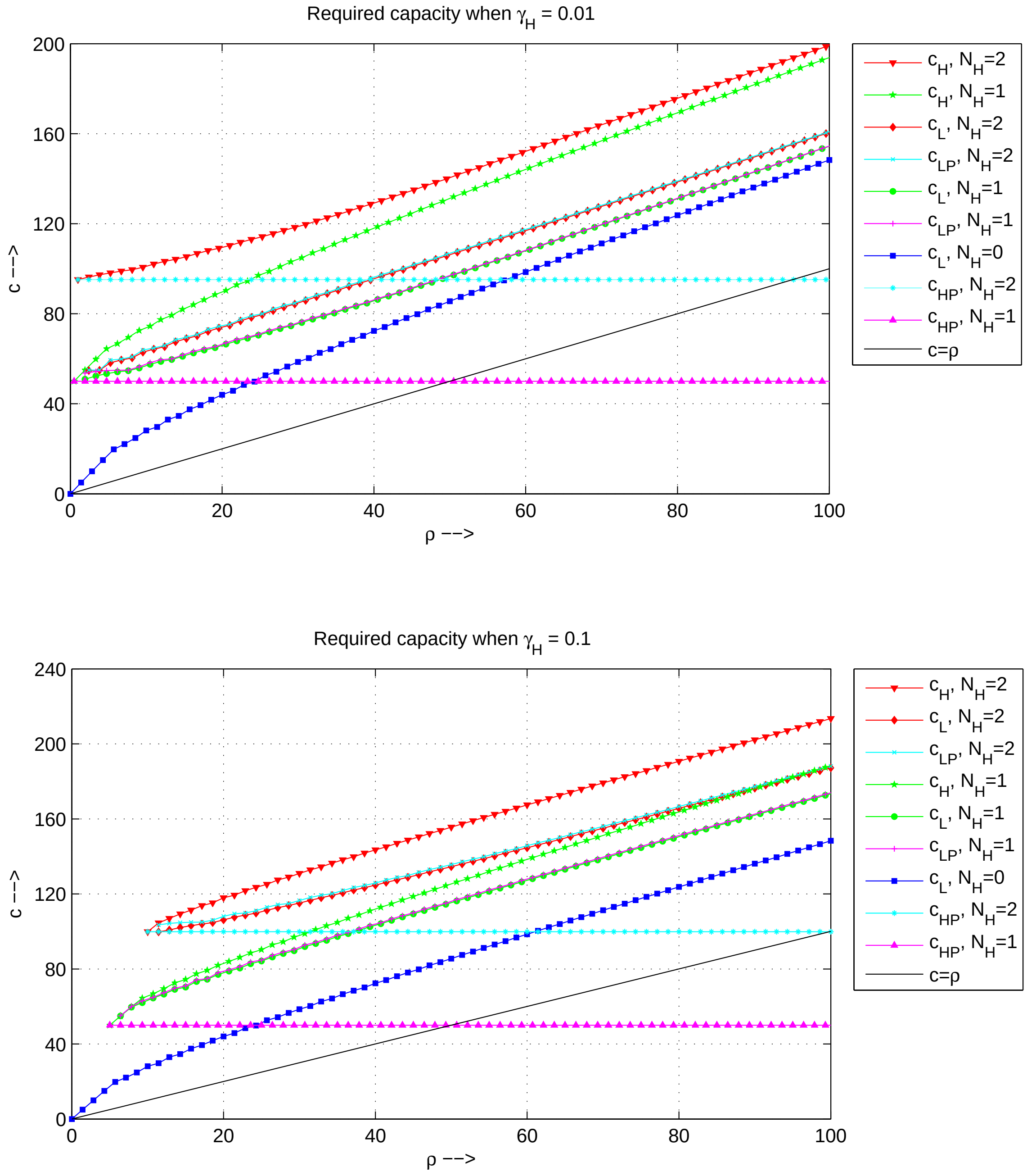


Figure 12: Required capacity with strict priority for high-rate sources.

Comparison of the required capacities

We compare the required capacities in the different situations with a workload $\rho = 100\,\text{Mb/s}$. Define the relative capacity with respect to the workload as the extra capacity that is required to take care of the bursty traffic behavior:

$$\Delta c := 100\frac{c-\rho}{\rho}\%. \tag{35}$$

We examine the relative capacity with $\rho = 100\,\text{Mb/s}$ in Table 2.

		$N_H = 0$		$N_H = 1$		$N_H = 2$	
		c	Δc	c	Δc	c	Δc
	c_L	148	48%				
$\gamma_H = 0.01$	$\max(c_L, c_H)$			194	94%	199	99%
	$\max(c_{LP}, c_{HP})$			155	55%	161	61%
$\gamma_H = 0.1$	$\max(c_L, c_H)$			188	88%	214	114%
	$\max(c_{LP}, c_{HP})$			174	74%	188	88%

Table 2: Required capacity in the streaming model (absolute values c and relative values Δc with $\rho = 100\,\text{Mb/s}$).

In Table 2 the required capacities are collected for the different values for N_H and γ_H we considered in the numerical study. When $N_H \geq 1$ the required capacity is the maximum of the required capacity for the low-rate and the high-rate sources, because the performance requirements of both the low-rate and the high-rate sources should be fulfilled. Without service differentiation $\max(c_L, c_H) = c_H$, because the high-rate sources always experience a busy network. With strict priority for the high-rate source $\max(c_{LP}, c_{HP}) = c_{LP}$ when the workload is high, because in this case low-rate data has to wait for high-rate data in case of congestion. If $N_H = 0$ only low-rate sources are present, so in that case c_H, c_{HP} and c_{LP} are not defined and the required capacity is c_L.

In Section 3.1 we mentioned two scenarios for the traffic characteristics of a corporate user. The first scenario was that traffic generated by a company is the sum of the data traffic of the employees. In this case, traffic is handled as if it originates from many consumers ($N_H = 0$) and the required relative capacity is 48% when $\rho = 100\,\text{Mb/s}$. The other scenario for the traffic characteristics of a corporate user was that from time to time a large amount of data has to be transferred. In that case the required capacity is much higher. If a high-rate source is active only 1% of the time, the required relative capacity is 94% or 99% when one or two corporate users are present in the network, respectively. So when $\gamma_H = 0.01$, the relative capacity is approximately doubled (from 48% to 94% or 99%). The absolute difference between the required capacity in a network with a few corporate users and a network with only consumers is approximately 50, which is the access rate of a corporate user. If a high-rate source is active 10% of the time, the required relative capacity in a network with two corporate users (114%) is significantly higher than in a network with one corporate user (88%).

We also considered the required capacity of the network link in case traffic from corporate users is handled with strict priority over traffic of consumers. If a corporate user only transmits data 1% of the time, this leads to a significant decrease of the required capacity. Then the presence of one or two corporate users leads to an increase of the required relative capacity from 48% to 55% or 61% respectively, instead of the doubling without service differentiation. If a corporate user transmits data 10% of the time, service differentiation also decreases the required capacity, but the gain is less than with $\gamma_H = 0.01$.

5 Analysis of the elastic model

In this chapter we analyze the elastic model, as described in Section 3.3.2. We derive the stationary distribution of the number of active sources in Section 5.1. In Section 5.2 we introduce the performance requirements for the elastic model and we define quantities to express these performance requirements. In the subsequent sections we derive expressions for these quantities. Finally, we present numerical results in Section 5.6.

5.1 Stationary distribution

For the two-dimensional process, where in state (i, j) i low-rate sources and j high-rate sources are active, the transition rates are:

From state	To state	Transition rate
(i, j)	$(i+1, j)$	$(N_L - i)\lambda_L$
(i, j)	$(i, j+1)$	$(N_H - j)\lambda_H$
(i, j)	$(i-1, j)$	$iR_L(i, j)f_L$
(i, j)	$(i, j-1)$	$jR_H(i, j)f_H$

Throughout this section, we assume that $R_L(i, j) > 0$ and $R_H(i, j) > 0$ for all $0 \le i \le N_L$, $0 \le j \le N_H$.

When we translate the process into a one-dimensional process as we did in Section 4.1.1, we obtain a process with generator matrix $\mathbf{Q}$, which consists of $N_H + 1$ by $N_H + 1$ sub-matrices of size $N_L + 1$ by $N_L + 1$. The (j, l)th sub-matrix of $\mathbf{Q}$ is given by

$$\mathbf{Q}[j, l] := \begin{cases} \mathbf{M}_j & \text{if } l = j, \\ (N_H - j)\lambda_H \mathbf{I}_{N_L+1} & \text{if } l = j+1, \\ \mathbf{R}_j & \text{if } l = j-1 \\ 0 & \text{otherwise.} \end{cases} \tag{36}$$

The (i, k)th element $(i \neq k)$ of $\mathbf{M}_j$ is given by

$$\mathbf{M}_j(i, k) := \begin{cases} (N_L - i)\lambda_L & \text{if } k = i+1, \\ iR_L(i, j)f_L & \text{if } k = i-1, \\ 0 & \text{otherwise.} \end{cases}$$

The matrices $\mathbf{R}_j$, $j = 1, \ldots, N_H$, are diagonal matrices, with

$$\mathbf{R}_j(i, i) = jR_H(i, j)f_H.$$

The diagonal elements of $\mathbf{Q}$ are such that the row sums are zero.

The stationary distribution $(\boldsymbol{\pi})$ of this process is given by $\mathbf{Q}^T\boldsymbol{\pi} = \mathbf{0}$ together with $|\boldsymbol{\pi}| = 1$. So $\boldsymbol{\pi}$ is the normalized eigenvector of $\mathbf{Q}^T$ with eigenvalue 0. We retranslate this one-dimensional stationary distribution into the two-dimensional distribution as follows:

$$\boldsymbol{\pi}_{i,j} = \boldsymbol{\pi}_{(N_L+1)j+i}.$$

A closed-form expression for the stationary distribution cannot be computed in general. However, if $r_H = r_L$ the stationary distribution has a product-form and a closed-form expression for the distribution can be found. The calculation is given in Appendix D.

With the stationary distribution, we can easily compute the fraction of time that the instantaneous transmission rate of an arbitrary high-rate or low-rate source is below a certain value. However, a user does not necessarily notice service degradation when the transmission rate is very low for a short time, because he only experiences the average transmission rate during the transmission of a file. In the next section a performance requirement for the elastic model is stated.

Later on in this thesis we will need the distribution of the process at an arrival moment of a low-rate or high-rate file. In a closed product-form network, the distribution of the process at an arrival moment is equal to the stationary distribution of the same closed network with one source less (e.g. see Proposition 8.3 of [19]). The network considered in this thesis is not product-form, but a reasonable approximation of the arrival distribution would be the stationary distribution in a system with one low-rate or high-rate source less, which we denote with $\hat{\boldsymbol{\pi}}^L$ and $\hat{\boldsymbol{\pi}}^H$ respectively. These distributions can be found earlier in this section, using $N_L - 1$ low-rate sources and N_H high-rate sources for $\hat{\boldsymbol{\pi}}^L$ and N_L low-rate sources and $N_H - 1$ high-rate sources for $\hat{\boldsymbol{\pi}}^H$.

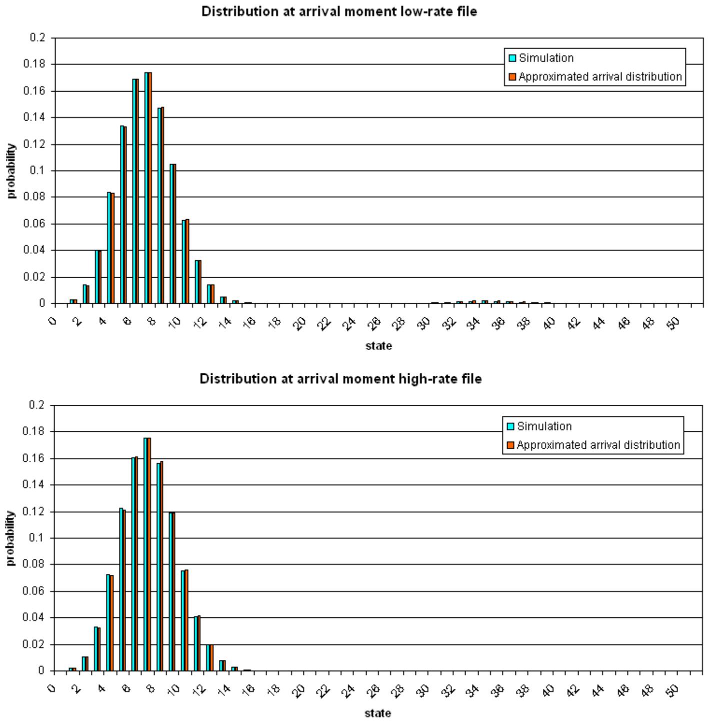


Figure 13: Distribution at an arrival moment of a low-rate file (top) and a high-rate file (bottom).

In Figure 13 the approximated arrival distributions $\hat{\boldsymbol{\pi}}^L$ and $\hat{\boldsymbol{\pi}}^H$ are plotted against the simulated distribution of the state of the system at an arrival moment of a low-rate or a high-rate file. The parameter values used for this figure can be found in Table 1 ($N_L = 25$, $N_H = 1$, $\gamma_H = 0.01$). The required capacity to fulfill the performance requirement for the low-rate (high-rate) sources for these parameter settings is $c = 66$ ($c = 88$). Figure 13 shows that at least for these parameter values, $\hat{\boldsymbol{\pi}}^L$ and $\hat{\boldsymbol{\pi}}^H$ are very good approximations for the arrival distribution of a low-rate file (high-rate file).

5.2 Definitions and performance requirements

We introduce the performance requirements for the elastic model and define quantities for the throughput and sojourn time (transfer time) in Section 5.2.1. We also give a roadmap to calculate the required capacity in Section 5.2.2.

5.2.1 Performance requirements based on tail probabilities

The performance requirement in the elastic model is based on the throughput during the transmission of a file, because the delay of a file is caused by a degradation of the rate at which the file can be transmitted. Define T_L (T_H) as the throughput during the transmission of a file originating from a low-rate (high-rate) source. This is the average rate at which the source can transmit data during the file transmission. Now we can formulate the performance requirements as follows:

$$\mathbb{P}(T_L < \alpha_L r_L) < \epsilon_L, \tag{37}$$

$$\mathbb{P}(T_H < \alpha_H r_H) < \epsilon_H, \tag{38}$$

where $0 < \alpha_L, \alpha_H \leq 1$ and ϵ_L and ϵ_H are very small positive numbers. The probabilities in Performance Criterions (37) and (38) should be seen as the fractions of files that are transmitted with an insufficient throughput.

Now we define and relate some quantities for the throughput and sojourn time. Let $\mathbb{S}_L$ ($\mathbb{S}_H$) be the state space containing the possible states of the process without the low-rate source (high-rate source) for which we calculate the throughput. That means

$$\begin{aligned} \mathbb{S}_L &= \{(i,j) | 0 \leq i < N_L, 0 \leq j \leq N_H\} \\ \mathbb{S}_H &= \{(i,j) | 0 \leq i \leq N_L, 0 \leq j < N_H\}. \end{aligned}$$

Assume that we have a low-rate file of size x and define

$$T^L_{i,j}(x) := \begin{array}{l} \text{Throughput of an amount of data } x \\ \text{when the process is in state } (i,j) \in \mathbb{S}_L \text{ at arrival,} \end{array}$$

and

$$S^L_{i,j}(x) := \begin{array}{l} \text{Sojourn time (transfer time) of an amount of data } x \\ \text{when the process is in state } (i,j) \in \mathbb{S}_L \text{ at arrival.} \end{array}$$

The throughput during the transmission of an amount of data x is x divided by the total transmission time of that file, so

$$T_{i,j}^L(x) = \frac{x}{S_{i,j}^L(x)}.$$

An expression for the probability of a low throughput (unconditional on the state of the system at arrival and the file size) is then

$$\mathbb{P}(T_L < \alpha_L r_L) = \sum_{i=0}^{N_L-1} \sum_{j=0}^{N_H} \hat{\pi}_{i,j}^L \int_0^\infty f_L e^{-f_L x} \mathbb{P}\left(\frac{x}{S_{i,j}^L(x)} < \alpha_L r_L \right) \mathrm{d}x.$$

To calculate the probability $\mathbb{P}(T_L < \alpha_L r_L)$ in Performance Requirement (37), we need to know

$$\mathbb{P}\left(\frac{x}{S_{i,j}^L(x)} < \alpha_L r_L \right) = \mathbb{P}\left(S_{i,j}^L(x) > \frac{x}{\alpha_L r_L} \right),$$

so we need the distribution of $S_{i,j}^L(x)$. In Section 5.3 we will see that it is very hard to determine this distribution in a way that it can be used numerically. Therefore we proceed with the calculation of the mean throughput instead.

5.2.2 Performance requirements based on means

The mean throughput conditional on the state of the system at arrival and the file size is

$$\mathbb{E}\left[T_{i,j}^L(x)\right] = \mathbb{E}\left[\frac{x}{S_{i,j}^L(x)}\right] \geq \frac{x}{\mathbb{E}\left[S_{i,j}^L(x)\right]}.$$

An expression for the exact mean conditional throughput is given in Section 5.3, but this expression is not numerically tractable either. Therefore, we use the approximate expression for the mean throughput conditional on the state of the system at arrival and the file size in the following alternative definition for the average throughput.

$$\overline{T}_L := \sum_{i=0}^{N_L-1} \sum_{j=0}^{N_H} \hat{\pi}_{i,j}^L \int_0^\infty f_L e^{-f_L x} \frac{x}{\overline{S}_{i,j}^L(x)} \mathrm{d}x, \tag{39}$$

where

$$\overline{S}_{i,j}^L(x) = \mathbb{E}\left[S_{i,j}^L(x)\right]$$

is the mean conditional sojourn time. The approximation for the conditional throughput is a conservative approximation, because the exact throughput is larger than the approximation of the throughput.

To determine the required capacity with Formula (39) for the average throughput, we need to translate Performance Requirement (37) into a target value for the average throughput. So we need to know the relation between $\mathbb{P}(T_L < \alpha_L r_L)$ and $\overline{T}_L$. There is no simple formula to relate these two quantities, because the shape of the throughput distribution depends on the parameters of the model. Instead, we will go through the following steps.

1. Approximate the throughput of a file with the instantaneous transmission rate of a file just after arrival. The assumption in this approximation is that the transmission rate does not change during the transmission of the file. With this approximation we have
$$\mathbb{P}(T_L < \alpha_L r_L) \approx \sum_{i=0}^{N_L-1} \sum_{j=0}^{N_H} \hat{\pi}_{i,j} \mathbb{1}_{\{R_L(i+1,j)<\alpha_L r_L\}} \tag{40}$$
and
$$\overline{T}_L \approx \sum_{i=0}^{N_L-1} \sum_{j=0}^{N_H} \hat{\pi}_{i,j} R_L(i+1,j). \tag{41}$$
2. Now compute the minimum capacity that is required to satisfy $\mathbb{P}(T_L < \alpha_L r_L) < \epsilon_L$ with the approximation for the throughput as in Expression (40). Denote this capacity with c^a.
3. Compute $\overline{T}_L$ with this capacity c^a and the approximation as in Expression (41). Denote the average throughput obtained in this way with $\overline{T}_L^a$.
4. Now $\overline{T}_L^a$ can be used as a target value for the average throughput of a low-rate file. The capacity obtained following this procedure is denoted with $\bar{c}$.

With this procedure we determine the required capacity with the formula for the average throughput ($\bar{c}$) in step 4, but we also determine an intermediate capacity c^a in step 2 that is used to find the relation between α_L and the average throughput.

All expressions introduced for low-rate sources are defined for high-rate sources analogously.

5.3 The conditional sojourn time

In this section we define matrices for the computation of the conditional sojourn time and we compute the Laplace transform of the conditional sojourn time distribution.

Consider a low-rate file. When this file finds the process in state (i, j) at arrival, there are $i+1$ low-rate sources and j high-rate sources active directly after arrival of the new file. Then the first transition of the process is either a departure of the source corresponding to the new file (with rate $R_L(i+1, j)f_L$) or a transition of the rest of the system. The transition rates of the system without the new file are $(N_L - i - 1)\lambda_L$, $(N_H - j)\lambda_H$, $iR_L(i+1, j)f_L$ and $jR_H(i+1, j)f_H$ to the states with one low-rate or high-rate source more and one low-rate or high-rate source less, respectively.

In Figure 14 the process is drawn. The time until a transition of the process without the new low-rate source is exponentially distributed with rate
$$A(i,j) := (N_L - i - 1)\lambda_L + (N_H - j)\lambda_H + iR_L(i+1,j)f_L + jR_H(i+1,j)f_H. \tag{42}$$
Let $\mathbf{Q}_L^*$ be the generator matrix of the (one-dimensional version of the) process with state space $\mathbb{S}_L$ and an extra permanent low-rate source. In $\mathbf{Q}_L^*$ the transition rates from state

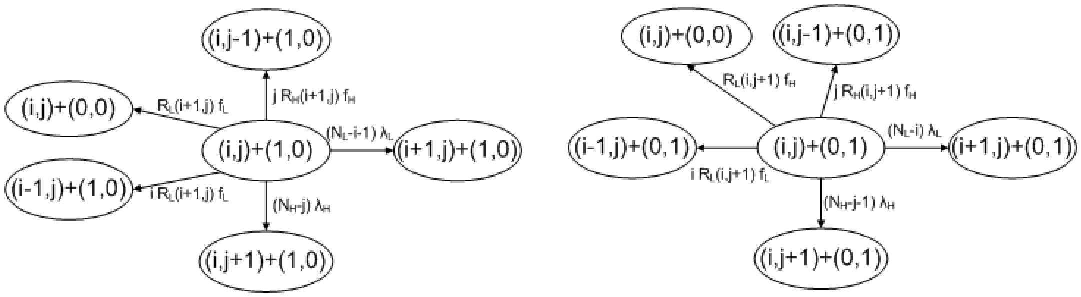


Figure 14: Transition rates in a system with a special low-rate source (left) or high-rate source (right).

$(i, j) \in \mathbb{S}_L$ are the transition rates from (i, j) as displayed in $\mathbf{Q}$ (Expression (13)), with $N_L - i$ replaced by $N_L - i - 1$ and the transmission rates $R_L(i, j)$ and $R_H(i, j)$ replaced by $R_L(i+1, j)$ and $R_H(i+1, j)$ respectively. The diagonal elements of $\mathbf{Q}_L^*$ are such that the row sums are zero.

Let $\mathbf{R}_L$ be the diagonal matrix in which the rates at which the extra low-rate source can transmit data are stored for the states in $\mathbb{S}_L$. When i low-rate sources, j high-rate sources and the permanent low-rate source are active, this rate is $R_L(i + 1, j)$.

The vector $\widetilde{\mathbf{S}}^L(x, \omega)$, containing the Laplace transforms of the distributions of $S_{i,j}^L(x)$ for all $(i, j) \in \mathbb{S}_L$, is derived in Appendix E and is given by

$$\widetilde{\mathbf{S}}^L(x, \omega) = \exp\left(\mathbf{R}_L^{-1}\left(\mathbf{Q}_L^* - \omega \mathbf{I}\right) x\right) \mathbf{1}.$$

Note that $\exp(\mathbf{A}x)$ with $\mathbf{A}$ a matrix is the matrix exponential. Unfortunately, we cannot find a closed-form expression for the distribution of the conditional sojourn time having this Laplace transform.

With the Laplace transform, we can calculate the mean conditional throughput:

$$\mathbb{E}\left[T_{i,j}^L(x)\right] = x \int_0^\infty \widetilde{S}_{i,j}^L(x, \omega) \mathrm{d}\omega. \tag{43}$$

The calculation can be found in Appendix E.

The Laplace transform of the conditional sojourn time of a high-rate file can be found analogously by using $\mathbf{Q}_H^*$ and $\mathbf{R}_H$ instead of $\mathbf{Q}_L^*$ and $\mathbf{R}_L$. The matrix $\mathbf{Q}_H^*$ is the generator matrix of the (one-dimensional version of the) process with state space $\mathbb{S}_H$ with an extra permanent high-rate source. In $\mathbf{Q}_H^*$ the transition rates from state $(i, j) \in \mathbb{S}_H$ are the transition rates from (i, j) as displayed in $\mathbf{Q}$, with $N_H - j$ replaced by $N_L - j - 1$ and the transmission rates $R_L(i, j)$ and $R_H(i, j)$ replaced by $R_L(i, j + 1)$ and $R_H(i, j + 1)$ respectively. The diagonal elements of $\mathbf{Q}_H^*$ are such that the row sums are zero.
The matrix $\mathbf{R}_H$ is a diagonal matrix in which the rates at which the extra high-rate source can transmit data are stored for the states in $\mathbb{S}_H$. When i low-rate sources, j high-rate sources

and the permanent high-rate source are active, this rate is $R_H(i, j+1)$.

The solution for the mean conditional throughput as given in Expression (43) is not numerically tractable, so we proceed with the calculation of the average throughput as defined in Expression (39).

5.4 Computation of the mean conditional sojourn time

In this section we derive an expression for the mean sojourn time (total transfer time), conditional on the file size and the state of the system at arrival. We do this along the lines of [11], page 116. Define

$\overline{S}^L_{i,j}(x) =$ The mean transmission time of a file of size x from a low-rate source when the process is in state (i, j) at arrival.

Consider an amount of low-rate data x and a time interval of length $\Delta > 0$, with Δ sufficiently small such that the transfer of the file for which we determine the sojourn time cannot finish within this time, i.e. $\Delta < xR_L(i+1, j)^{-1}$. We condition on all possible events occurring during this interval. These events are the transitions as displayed in Figure 14, except for the departure of the special low-rate source.

$$\begin{aligned}
\overline{S}^L_{i,j}(x) &= \mathbb{E}[S^L_{i,j}(x)] \\
&= \Delta + iR_L(i+1,j)f_L\Delta\overline{S}^L_{i-1,j}(x-\mathcal{O}(\Delta)) + (N_L-i-1)\lambda_L\Delta\overline{S}^L_{i+1,j}(x-\mathcal{O}(\Delta)) \\
&\quad + jR_H(i+1,j)f_H\Delta\overline{S}^L_{i,j-1}(x-\mathcal{O}(\Delta)) + (N_H-j)\lambda_H\Delta\overline{S}^L_{i,j+1}(x-\mathcal{O}(\Delta)) \\
&\quad + (1-A(i,j)\Delta)\overline{S}^L_{i,j}(x-R_L(i+1,j)\Delta) + o(\Delta).
\end{aligned}$$

Rearranging terms gives

$$\begin{aligned}
&\frac{\overline{S}^L_{i,j}(x) - \overline{S}^L_{i,j}(x-R_L(i+1,j)\Delta)}{R_L(i+1,j)\Delta} \\
&= \frac{1}{R_L(i+1,j)} + \frac{iR_L(i+1,j)f_L}{R_L(i+1,j)}\overline{S}^L_{i-1,j}(x-\mathcal{O}(\Delta)) + \frac{(N_L-i-1)\lambda_L}{R_L(i+1,j)}\overline{S}^L_{i+1,j}(x-\mathcal{O}(\Delta)) \\
&\quad + \frac{jR_H(i+1,j)f_H}{R_L(i+1,j)}\overline{S}^L_{i,j-1}(x-\mathcal{O}(\Delta)) + \frac{(N_H-j)\lambda_H}{R_L(i+1,j)}\overline{S}^L_{i,j+1}(x-\mathcal{O}(\Delta)) \\
&\quad - \frac{A(i,j)}{R_L(i+1,j)}\overline{S}^L_{i,j}(x-\mathcal{O}(\Delta)) + \frac{1}{R_L(i+1,j)}\frac{o(\Delta)}{\Delta}.
\end{aligned}$$

Now let $\Delta \to 0$. Then for all $(i, j) \in \mathbb{S}_L$

$$\begin{aligned}
\frac{\mathrm{d}\overline{S}^L_{i,j}(x)}{\mathrm{d}x} &= \frac{1}{R_L(i+1,j)} + \frac{iR_L(i+1,j)f_L}{R_L(i+1,j)}\overline{S}^L_{i-1,j}(x) + \frac{(N_L-i-1)\lambda_L}{R_L(i+1,j)}\overline{S}^L_{i+1,j}(x) \\
&\quad + \frac{jR_H(i+1,j)f_H}{R_L(i+1,j)}\overline{S}^L_{i,j-1}(x) + \frac{(N_H-j)\lambda_H}{R_L(i+1,j)}\overline{S}^L_{i,j+1}(x) - \frac{A(i,j)}{R_L(i+1,j)}\overline{S}^L_{i,j}(x).
\end{aligned}$$

In matrix notation this is

$$\frac{\mathrm{d}}{\mathrm{d}x}\overline{\mathbf{S}}^L(x) = \mathbf{R}_L^{-1}\left(\mathbf{1} + \mathbf{Q}^*\overline{\mathbf{S}}^L(x)\right), \tag{44}$$

where $\overline{\mathbf{S}}^L(x)$ is the vector with entries $\overline{S}^L_{i,j}(x)$, $(i,j) \in \mathbb{S}_L$, ordered colexicographically.

The time needed to transfer no data is 0, so

$$\overline{\mathbf{S}}^L(0) = \mathbf{0}. \tag{45}$$

Now we make a substitution and write Equations (44) and (45) in terms of

$$\mathbf{W}(x) := \mathbf{1} + \mathbf{Q}_L^*\overline{\mathbf{S}}^L(x). \tag{46}$$

The differential equation then becomes

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbf{W}(x) = \mathbf{Q}_L^*\frac{\mathrm{d}}{\mathrm{d}x}\overline{\mathbf{S}}^L(x) = \mathbf{Q}_L^*\mathbf{R}_L^{-1}\mathbf{W}(x), \tag{47}$$

with

$$\mathbf{W}(0) = \mathbf{1}. \tag{48}$$

To get a numerical solution for Equations (47) and (48) we use the approach of Anick, Mitra and Sondhi [1], which gives

$$\mathbf{W}(x) = a_0\boldsymbol{\phi}_0 + \sum_{k=1}^{(N_H+1)N_L-1} a_k e^{z_k x}\boldsymbol{\phi}_k, \tag{49}$$

where z_k are the eigenvalues of $\mathbf{Q}_L^*\mathbf{R}_L^{-1}$ and $\boldsymbol{\phi}_k$ the corresponding eigenvectors.

From Theorem 1 of [14], we know that the number of negative eigenvalues is $(N_H+1)N_L - 1$ and the multiplicity of the eigenvalue 0 is 1. So the eigenvalues of $\mathbf{Q}^*\mathbf{R}_L^{-1}$ are $z_{(N_H+1)N_L-1} < z_{(N_H+1)N_L-2} < \cdots < z_1 < z_0 = 0$. Note that the size of the matrices $\mathbf{Q}^*$ and $\mathbf{R}_L$ is $(N_H+1)N_L \times (N_H+1)N_L$, because we consider a process with one low-rate source less.

Further, from Equation (48), we have

$$\sum_{k=0}^{(N_H+1)N_L-1} a_k\boldsymbol{\phi}_k = \mathbf{1}. \tag{50}$$

Equation (50) gives enough boundary conditions to obtain a unique solution for the coefficients $a_k, k = 0, \ldots, (N_H+1)N_L - 1$. However, we compute these coefficients in a different manner. We use the eigenvalue decomposition of the matrix $\mathbf{Q}^*\mathbf{R}_L^{-1}$. Then

$$\mathbf{Q}^*\mathbf{R}_L^{-1} = \mathbf{V}\mathbf{D}\mathbf{V}^{-1}, \tag{51}$$

in which $\mathbf{D}$ is a diagonal matrix with the eigenvalues of $\mathbf{Q}^*\mathbf{R}_L^{-1}$ and $\mathbf{V}$ is the matrix with the eigenvectors of $\mathbf{Q}^*\mathbf{R}_L^{-1}$ as columns. So

$$\mathbf{D} = \mathrm{diag}\{0, z_1, \ldots, z_{(N_H+1)N_L-1}\}$$

and

$$\mathbf{V} = [\boldsymbol{\phi}_0, \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{(N_H+1)N_L-1}].$$

From Equations (47), (48) and (51) we obtain

$$\mathbf{W}(x) = \exp(\mathbf{Q}^* \mathbf{R}_L^{-1} x)\mathbf{1} = \mathbf{V} \exp(\mathbf{D}x)\mathbf{V}^{-1}\mathbf{1}. \tag{52}$$

Now we compare Equations (49) and (52) and find for a (which is the vector with coefficients a_k, $k \in \{0, \ldots, (N_H+1)N_L - 1\}$):

$$a = \mathbf{V}^{-1}\mathbf{1}.$$

We proceed with determining an expression for $\overline{\mathbf{S}}^L(x)$. Combining Equations (44), (46) and (49) yields

$$\frac{\mathrm{d}}{\mathrm{d}x}\overline{\mathbf{S}}^L(x) = \mathbf{R}_L^{-1}\mathbf{W}(x) = a_0 \mathbf{R}_L^{-1}\boldsymbol{\phi}_0 + \sum_{k=1}^{(N_H+1)N_L-1} a_k e^{z_k x}\mathbf{R}_L^{-1}\boldsymbol{\phi}_k. \tag{53}$$

Note that $\mathbf{R}_L^{-1}\boldsymbol{\phi}_0 = R_L(N_L, N_H)^{-1}\mathbf{1}$, since $\boldsymbol{\phi}_0$ is the eigenvector of $\mathbf{Q}_L^*\mathbf{R}_L^{-1}$ corresponding to the eigenvalue 0. So $\mathbf{Q}_L^*\mathbf{R}_L^{-1}\boldsymbol{\phi}_\mathbf{0} = \mathbf{0}$, which means that $\mathbf{R}_L^{-1}\boldsymbol{\phi}_\mathbf{0}$ is the eigenvector of $\mathbf{Q}_L^*$ corresponding to the eigenvalue 0. Further, the row sums of $\mathbf{Q}_L^*$ are 0, so an eigenvector with eigenvalue 0 should be a multiple of the unit vector. We assumed that the last entry of $\boldsymbol{\phi}_0$ is 1, so the last entry of $\mathbf{R}_L^{-1}\boldsymbol{\phi}_\mathbf{0}$ is $R_L(N_L, N_H)^{-1}$.

Now we take integrals of both sides of Equation (53) and get

$$\overline{\mathbf{S}}^L(x) = \frac{a_0}{R_L(N_L, N_H)} x\mathbf{1} + \sum_{k=1}^{(N_H+1)N_L-1} \frac{a_k}{z_k} e^{z_k x}\mathbf{R}_L{}^{-1}\boldsymbol{\phi}_k + \kappa,$$

where κ is a constant factor which should be chosen such that $\overline{\mathbf{S}}^L(0) = 0$.

The solution for the mean conditional sojourn time of a low-rate file (depending on the state of the process at arrival) is then

$$\overline{\mathbf{S}}^L(x) = \frac{a_0}{R_L(N_L, N_H)} x\mathbf{1} + \sum_{k=1}^{(N_H+1)N_L-1} \frac{a_k}{z_k}\left(e^{z_k x} - 1\right)\mathbf{R}_L{}^{-1}\boldsymbol{\phi}_k, \tag{54}$$

with $z_k, k \in \{1, \ldots, (N_H+1)N_L - 1\}$ the nonzero eigenvalues of $\mathbf{Q}_L^*\mathbf{R}_L^{-1}$, $\boldsymbol{\phi}_k$ the corresponding eigenvectors (with last element 1) and $[a_0, \ldots, a_{(N_H+1)N_L-1}]^T = [\boldsymbol{\phi}_0, \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{(N_H+1)N_L-1}]^{-1}\mathbf{1}$.

The solution for the mean conditional sojourn time of a high-rate source (depending on the state of the process at arrival) can be found analogously and is given by

$$\overline{\mathbf{S}}^H(x) = \frac{b_0}{R_H(N_L, N_H)} x\mathbf{1} + \sum_{k=1}^{N_H(N_L+1)-1} \frac{b_k}{v_k}\left(e^{v_k x} - 1\right)\mathbf{R}_H{}^{-1}\boldsymbol{\psi}_k, \tag{55}$$

with $v_k, k \in \{1, \ldots, N_H(N_L+1)-1\}$ the nonzero eigenvalues of $\mathbf{Q}_H^* \mathbf{R}_H^{-1}$, $\boldsymbol{\psi}_k$ the corresponding eigenvectors (with last element 1) and $[b_0, \ldots, b_{N_H(N_L+1)-1}]^T = [\boldsymbol{\psi}_0, \boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_{N_H(N_L+1)-1}]^{-1}\mathbf{1}$.

We observe that the solution for the mean conditional sojourn time contains a linear term. Further, the eigenvalues in the powers of the exponentials are negative, so when x becomes very large only the linear term remains:

$$\lim_{x\to\infty} \frac{\overline{\mathbf{S}}_{i,j}^L(x)}{x} = \frac{a_0}{R_L(N_L, N_H)}.$$

This limit is the inverse of the average transmission rate during the transfer of a file of infinite size from a low-rate source, so it is the average transmission rate of a low-rate source that is permanent active. From this point of view we have

$$\lim_{x\to\infty} \frac{\overline{\mathbf{S}}_{i,j}^L(x)}{x} = \left(\sum_{i=0}^{N_L-1} \sum_{j=0}^{N_H} \pi_{i,j}^{L*} R_L(i+1, j) \right)^{-1}, \tag{56}$$

where $\pi_{i,j}^{L*}$ is the stationary distribution of the state of the system with an extra permanent low-rate source, so $\pi_{i,j}^{L*}$ is the normalized eigenvector of ${\mathbf{Q}_L^*}^T$ with eigenvalue 0. Now we have found

$$\frac{a_0}{R_L(N_L, N_H)} = \left(\sum_{i=0}^{N_L-1} \sum_{j=0}^{N_H} \pi_{i,j}^{L*} R_L(i+1, j) \right)^{-1}, \tag{57}$$

and for the high-rate sources it follows analogously

$$\frac{b_0}{R_H(N_L, N_H)} = \left(\sum_{i=0}^{N_L} \sum_{j=0}^{N_H-1} \pi_{i,j}^{H*} R_H(i, j+1) \right)^{-1}. \tag{58}$$

Expression (54) has the same form as the expression in Lemma 2 of [6], in which Poisson arrivals are used and the maximum transmission rate of a source is infinite.

The method used in this section is not the only method to obtain a numerical solution for the mean conditional sojourn time. Another method is to derive an iterative relation for the Laplace transforms and retranslate that relation into a numerical solution for the conditional mean sojourn time. This method can be found in Appendix F. It is also shown in that appendix that the two methods lead to the same solution.

5.5 Computation of the mean unconditional sojourn time

In this section we derive an expression for the mean transfer time of a file (the mean sojourn time), dependent on the state of the system at arrival. We can find the mean unconditional transfer time of a file from the file size distribution and the conditional transfer time as given in Expression (54), but a simpler expression can be found when deriving the unconditional sojourn time directly. Define

$\overline{S}^L_{i,j} =$ The mean transmission time of a file from a low-rate source when the process is in state (i,j) at arrival.

Since all file sizes are assumed to be exponentially distributed, we can use the memoryless property. The transitions of the process are drawn in Figure 14. The time until a transition of the process without the source we are looking at is exponentially distributed with rate $A(i,j)$ (see Expression (42)). If nothing changes in the number of active sources, the transmission of the file will be finished after an exponential $(R_L(i+1,j)f_L)$ period. Therefore, the mean time until the total process makes a transition is $(A(i,j)+R_L(i+1,j)f_L)^{-1}$. The probability that this transition is to the state with one low-rate source less or more or one high-rate source less or more is $\frac{iR_L(i+1,j)f_L}{A(i,j)+R_L(i+1,j)f_L}$, $\frac{(N_L-i-1)\lambda_L}{A(i,j)+R_L(i+1,j)f_L}$, $\frac{jR_H(i+1,j)f_H}{A(i,j)+R_L(i+1,j)f_L}$ or $\frac{(N_H-j)\lambda_H}{A(i,j)+R_L(i+1,j)f_L}$ respectively. Otherwise, the transmission of the file we are looking at is finished. This can be summarized in the following equation:

$$\begin{aligned}
\overline{S}^L_{i,j} &= \frac{1}{A(i,j)+R_L(i+1,j)f_L} + \frac{iR_L(i+1,j)f_L}{A(i,j)+R_L(i+1,j)f_L}\overline{S}^L_{i-1,j} \\
&\quad + \frac{(N_L-i-1)\lambda_L}{A(i,j)+R_L(i+1,j)f_L}\overline{S}^L_{i+1,j} + \frac{jR_H(i+1,j)f_H}{A(i,j)+R_L(i+1,j)f_L}\overline{S}^L_{i,j-1} \\
&\quad + \frac{(N_H-j)\lambda_H}{A(i,j)+R_L(i+1,j)f_L}\overline{S}^L_{i,j+1}.
\end{aligned}$$

Rearranging terms gives

$$\begin{aligned}
-iR_L(i+1,j)f_L\overline{S}^L_{i-1,j} - (N_L-i-1)\lambda_L(N_L-i-1)\lambda_L\overline{S}^L_{i+1,j} & \\
+(A(i,j)+R_L(i+1,j)f_L)\overline{S}^L_{i,j} - jR_H(i+1,j)f_H\overline{S}^L_{i,j-1} - (N_H-j)\lambda_H\overline{S}^L_{i,j+1} &= 1,
\end{aligned}$$

for all $(i,j) \in \mathbb{S}_L$. In matrix form, this is

$$\left(-\mathbf{Q}^*_L + f_L\mathbf{R}_L\right)\overline{\mathbf{S}}^L = \mathbf{1},$$

with solution

$$\overline{\mathbf{S}}^L = \left(-\mathbf{Q}^*_L + f_L\mathbf{R}_L\right)^{-1}\mathbf{1}. \tag{59}$$

The mean conditional sojourn time for a high-rate source can be found analogously:

$$\overline{\mathbf{S}}^H = \left(-\mathbf{Q}^*_H + f_L\mathbf{R}_H\right)^{-1}\mathbf{1}. \tag{60}$$

5.6 Numerical results for the elastic model

In this section we present numerical results with the parameter values as listed in Table 1 of Section 3.6. We determine the required capacity in order to fulfill the performance requirements for the low-rate sources and the high-rate sources in two situations. In the first situation the transmission rates are proportional to the access rates and given in Equations (1) and (2). In the second situation, traffic originating from a high-rate source has strict priority over traffic originating from a low-rate source and the transmission rates are given in Equations

(3) and (4). We use the notations c_L, c_H, c_{LP} and c_{HP} as defined in Section 4.5, with the performance requirements for low-rate and high-rate sources replaced by Inequalities (37) and (38) respectively. The required capacity in order to fulfill the performance requirements for both the low-rate and the high-rate sources is $\max(c_L, c_H)$ or $\max(c_{LP}, c_{HP})$.

Required capacity as function of the workload

In Figure 15 we present the required capacity as a function of the realized workload in the system. Both the required capacity and the realized workload are determined by simulation. If sources can always transmit data at their access rates, the workload of the system is $N_L\gamma_L r_L + N_H\gamma_H r_H$. However, in case of congestion the transmission rates decrease and the durations of the active periods of the sources increase, with the total data volume during an active period remaining equal. The durations of the inactive periods of the sources do not change. Then the average amount of data that is transmitted per second decreases (although sources are active for a longer time period). The realized workload is thus always less or equal to $N_L\gamma_L r_L + N_H\gamma_H r_H$. This phenomenon is caused by the ON-OFF modelling of the sources, together with the adaptation of the transmission rates.

In the top diagram of Figure 15 a high-rate source is active 1% of the time ($\gamma_H = 0.01$, $\lambda_H = 1/1584$) and in the lower diagram this is 10% of the time ($\gamma_H = 0.1$, $\lambda_H = 1/144$). The explanation for the different starting values of the curves in Figure 15 is already given in Section 4.5 in the explanation of Figure 7. The shapes of the curves and the order of the curves for c_L and c_H in the figure are also identical to the shapes and order of the curves in Figure 7 for the streaming model. Therefore, we refer to Section 4.5 for the explanation.

Strict priority for high-rate traffic

Now we examine the required capacity in case high-rate traffic has strict priority over low-rate traffic (c_{LP} and c_{HP} in Figure 15) and compare this capacity to the required capacity when the transmission rates are proportional to the access rates (c_L and c_H in Figure 15).

We observe that $c_{LP} < c_L$ with $N_H = 1$ and $\gamma_H = 0.01$ when the workload is not too high. This is against our intuition, because the remaining capacity for low-rate sources is less when high-rate traffic has strict priority. This phenomenon is caused by the ON-OFF modelling of the sources and the fact that the priority for high-rate source is strict, as explained below.
If $c < r_H$ the transmission rate of a low-rate source is zero when the high-rate source is active, so low-rate sources cannot turn OFF when the high-rate source is active. This implies that at most one file from each low-rate source observes that the high-rate source is active and all other files from a low-rate source are transmitted during an inactive period of the high-rate source. The fraction of low-rate files that is transmitted during an active period of a high-rate source is less than $\epsilon_L = 0.01$ when $\gamma_H = 0.01$, so the required capacity remains constant when the throughput of low-rate files is sufficiently high when the high-rate source is inactive (N_L is small). When more low-rate sources are present, the performance of a low-rate source can also be degraded when the high-rate source is inactive, so then c_{LP} increases when the workload increases.
If $c > r_H$ and $c - r_H$ is small, the low-rate files also receive positive transmission rates and

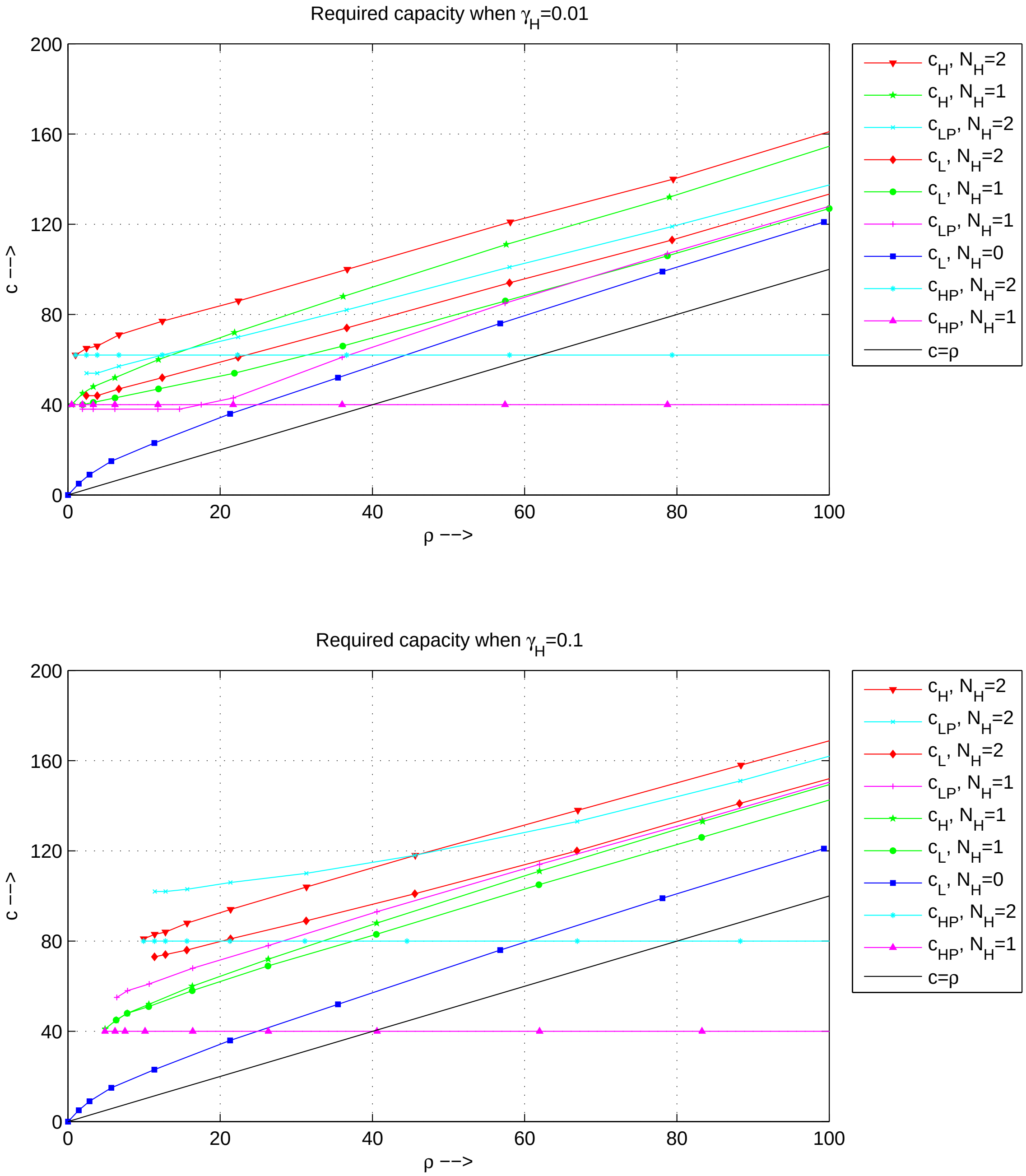


Figure 15: Required capacity depending on the realized workload in the system for elastic traffic.

can finish a file transmission when the high-rate source is active. Then more low-rate files experience a high level of congestion. The required capacity c_{LP} is still less than c_L, because without priority for the high-rate source the throughput of a low-rate file already is less than $\alpha_L r_L$ when the high-rate source is active. Furthermore, when the high-rate source has priority, the transmission time of a high-rate file is shorter and the low-rate sources alternate slower, so less low-rate files experience an active period of the high-rate source.
With a high workload we observe $c_{LP} > c_L$ for $N_H = 1$. Apparently the low-rate sources are not really slowed down any more when the workload is high and the fraction of low-rate files with a throughput less than $\alpha_L r_L$ is higher when high-rate sources have strict priority than when the transmission rates are proportional to the access rates.

For $\gamma_H = 0.01$ with $N_H = 2$ we have $c_L < c_{LP} < c_H$, as we would expect. In this case $c_{LP} > r_H$ for all values of ρ, so the low-rate files can always be transmitted (at an adjusted rate) and they experience a high level of congestion when the high-rate sources are active.

If $\gamma_H = 0.1$ we observe that $c_{LP} > c_H$ with $N_H = 1$ for all values of ρ. This is again against our intuition, because we would expect that the fraction of low-rate files with a low throughput (if high-rate sources have strict priority) is less, because many low-rate files are transmitted during the inactive period of the high-rate source. The explanation is that when proportional transmission rates are used, the throughputs of all sources decrease in case of congestion, but with strict priority for high-rate traffic the throughput for this source is always very high. The performance for the high-rate source is then higher than necessary and more capacity is required to satisfy the performance requirements of the low-rate sources.
Another explanation for $c_{LP} > c_H$ in this case is that the high-rate source transmits a significantly larger amount of data when this source gets strict priority. This is because the high-rate source is active for a significant fraction of the time and with strict priority, the source alternates faster.

For $\gamma_H = 0.1$ with $N_H = 2$ we observe $c_{LP} > c_H$ when the workload is low and $c_{LP} < c_H$ otherwise. The first relation can be explained with the same argument as given for $N_H = 1$. The fact that the capacity c_H increases faster than c_{LP} can be explained by examining the link capacity that is used for high-rate data. If high-rate traffic has strict priority, it remains constant. However, if traffic rates are proportional to the access rates, the capacity used for high-rate data decreases when ρ increases (N_L increases). So then the high-rate sources are active a longer fraction of the time and more low-rate files experience a busy system.

Required capacity determined with the average throughput

We present numerical results for the required capacity as computed via the procedure described at the end of Section 5.2.2. As we mentioned this procedure leads to two values for the capacity, namely the required capacity $\overline{c}$ as determined with Expression (39) and the intermediate capacity c^a (in step 2 of the procedure), which is used to translate α into a target value for the average throughput. The capacity c^a is the required capacity in case the throughput of a file is equal to the transmission rate of that file just after arrival. In Figure 16 the required capacities as determined by simulation are plotted against the two capacities determined in the procedure when the transmission rates are proportional to the access rates (see Expressions (1) and (2)).

The curves for c^a and $\overline{c}$ have a staircase behavior, because the instantaneous transmission rate at arrival only depends on the state of the system at arrival and there are only a finite number of states. Therefore we cannot always find a capacity c^a such that the fraction of files, for which the instantaneous throughput just after arrival is insufficient high, exactly equals ϵ.

In Figure 16, we observe that c^a exceeds the actual required capacity c (as determined by simulation) in most cases. The only exception is when $N_H = 2$ with $\gamma_H = 0.01$. The difference between c and c^a is caused by two opposing effects:

- $c^a > c$: The capacity c^a is determined by the instantaneous transmission rate just after the arrival of a file. However, a requirement based on an instantaneous transmission rate is more stringent than a requirement based on the average transmission rate during the transmission of a file. A user does not necessarily notice service degradation when the transmission rate is very low for a short time, because he only experiences the total transmission time of the file. Therefore, the required capacity c can be a little less than c^a.

- $c^a < c$: After the arrival of a file the number of active sources is increased, which implies that the transmission rates of all sources are decreased. As a result of this all files stay in the system longer. Therefore, the number of new file arrivals will exceed the number of file departures and the link becomes busier than at an arrival instant. This causes that the throughput of a file is less than the instantaneous arrival rate. By this effect, the required capacity c can exceed c^a.

The relative magnitude of the above-mentioned effects determines the order of c and c^a. From Figure 16 we conclude that the second effect only dominates the first when a high-rate file arrives and another high-rate file is present with $\gamma_H = 0.01$.

Another observation in Figure 16 is that always $\overline{c} > c^a$. The capacity $\overline{c}$ is the required capacity such that the average throughput is equal to the average instantaneous transmission rate at arrival. This is also caused by the second effect mentioned above.

In the curve for $\overline{c}$ with $N_H = 2$, $\gamma_H = 0.01$, we observe that the slope of parts of the curve is negative. This is when c^a is constant for a few subsequent values of N_L. In this case the fraction of files with a degraded performance is less when N_L is less (because c^a could not be chosen such that this fraction was closer to ϵ). The average throughput $\overline{T}^a$ is then based on a more stringent performance requirement, which causes that the capacity $\overline{c}$ is higher.

From Figure 16 we conclude that the capacity c^a is a very good approximation of the required capacity c, except for c_H when $N_H = 2$ and $\gamma_H = 0.01$, because in that case the capacity is underestimated. The capacity $\overline{c}$, as determined with the procedure at the end of Section 5.2, leads to an approximation of the required capacity c that is in general worse than the approximation c^a. Furthermore, the computation of the capacity c^a is much simpler than the computation of $\overline{c}$. The computation time for $\overline{c}$ is very large, because numerical integration is needed. To avoid this, we can also use the average unconditional throughput or the average

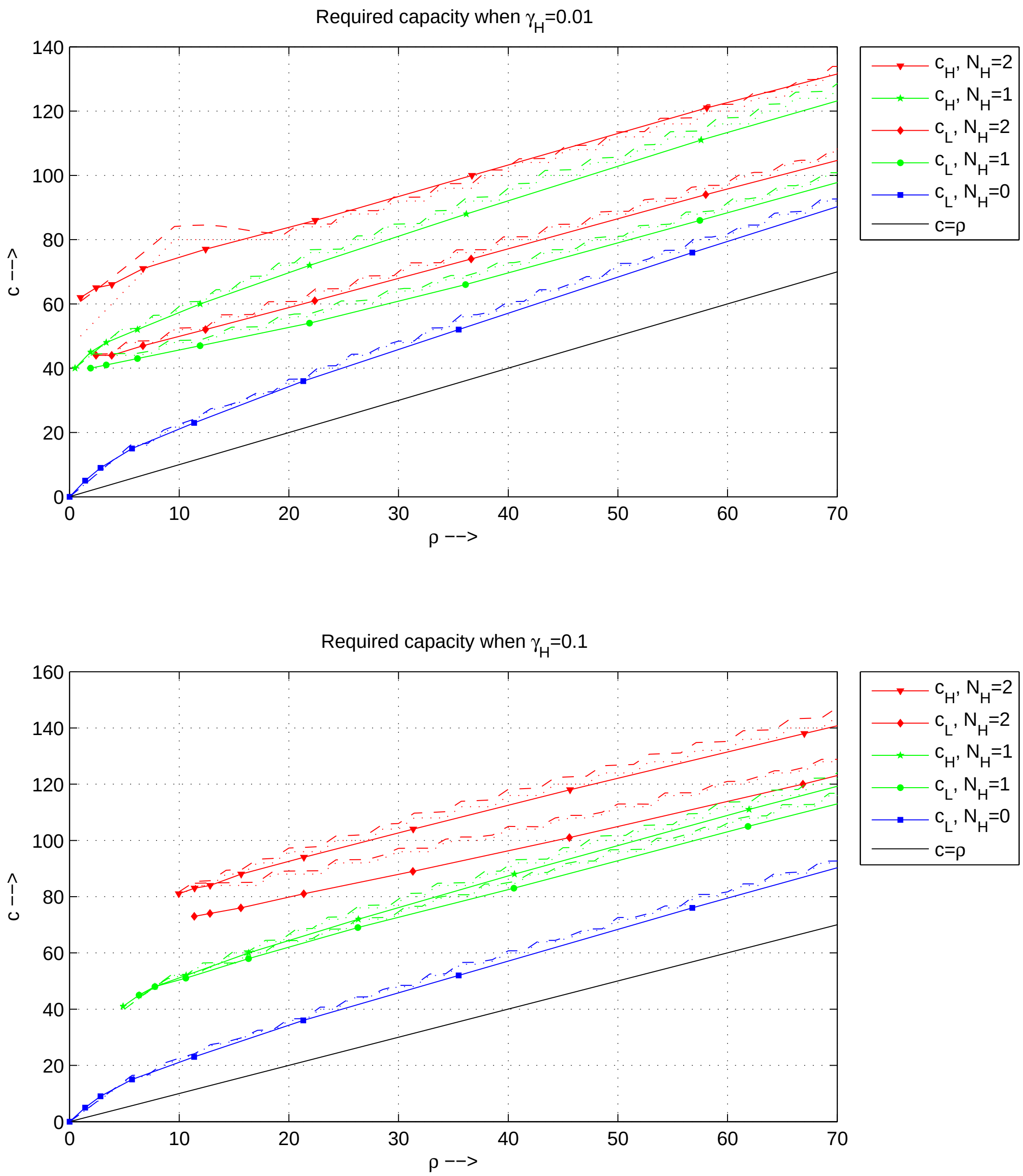


Figure 16: Required capacity determined with simulation (c, solid), instantaneous transmission rate at arrival (c^a, dotted) and average throughput ($\bar{c}$, dashed).

long-term throughput. These can be computed with the formulas for the average unconditional sojourn time (Expression (59) or (60)) or the average transmission rate of a source that is permanently active (the inverse of Expression (57) or (58)). The resulting capacities are slightly higher than the ones obtained with the average conditional throughput, so these approximations are also worse than c^a.

Comparison of the required capacities

We compare the required capacities in the elastic model for the different situations with a workload $\rho = 100\,\mathrm{Mb/s}$. In Table 3 the required relative capacities (as defined in Definition (35)) are collected for the different values for N_H and γ_H we considered in the numerical study.

		$N_H = 0$		$N_H = 1$		$N_H = 2$	
		c	Δc	c	Δc	c	Δc
	c_L	121	21%				
$\gamma_H = 0.01$	$\max(c_L, c_H)$			154	54%	161	61%
	$\max(c_{LP}, c_{HP})$			128	28%	137	37%
$\gamma_H = 0.1$	$\max(c_L, c_H)$			149	49%	168	68%
	$\max(c_{LP}, c_{HP})$			150	50%	161	61%

Table 3: Required capacity in the elastic model (absolute values c and relative values Δc with $\rho = 100\,\mathrm{Mb/s}$)

In Section 3.1 we mentioned two scenarios for the traffic characteristics of a corporate user. The first scenario was that traffic generated by a company is the sum of the data traffic of the employees. In that case, traffic is handled as if it originates from many consumers ($N_H = 0$) and the required relative capacity is 21% when $\rho = 100\,\mathrm{Mb/s}$. The other scenario for the traffic characteristics of a corporate user was that from time to time a large amount of data has to be transferred. In that case the required capacity is much higher. If a high-rate source is active only 1% of the time, the required relative capacity is 54% or 61% when one or two corporate users are present in the network, respectively. So if $\gamma_H = 0.01$, the relative capacity is more than twice as much in this case. The absolute difference between the required capacity in a network with two corporate users and a network with only consumers is 40, which is the minimal transmission rate of a corporate user that does not cause service degradation. If a high-rate source is active 10% of the time, the required relative capacity in a network with one corporate user is 49%, which is less than in a network with one corporate user which is active 1% of the time. The required relative capacity in a network with two corporate users that are both active 10% of the time is more than when they are only active 1% of the time.

We also considered the required capacity of the network link in case traffic from corporate users is handled with strict priority over traffic of consumers. If a corporate user only transmits data 1% of the time, this leads to a significant decrease of the required capacity. If a corporate user transmits data 10% of the time, service differentiation does not significantly decrease the required capacity. With only one corporate user, the required capacity in case of service differentiation is even more than the required capacity without service differentiation.

6 Comparison of the streaming and the elastic model

In this chapter we compare the numerical results for the streaming model and the elastic model and we explain the differences.

6.1 Required capacity as function of the workload

The results of the required capacities in the streaming and the elastic model are presented in Figures 12 and 15. The shapes and the order of the curves for c_L and c_H in the diagrams are almost identical in the streaming and the elastic model, because these depend on the trade-off between γ_H and ϵ (as explained in Section 4.5). The required capacities c_{LP} have a different behavior in the two models. The main reason is that the durations of the active periods are influenced by the level of congestion in the system in the elastic model (as explained in Section 5.6), while the sources can always transmit data at their access rate in the streaming model. In the streaming model the amount of data that is transmitted is constant, while less packets are generated in the elastic model in case of congestion.

The required capacities in the streaming model are significantly higher than the required capacities in the elastic model. To determine the required capacity of a network link, the only quantity that can be measured on the link is the average workload. Given the average workload, we want to know the required capacity in order to fulfill the performance requirements for both the consumers and the corporate users. If $N_H = 0$ only low-rate sources are present, so the required capacity is c_L. If $N_H > 0$ the required capacity is the maximum of the required capacity for the low-rate and the high-rate sources, because the performance requirements of both the low-rate and the high-rate sources should be fulfilled. Without service differentiation $\max(c_L, c_H) = c_H$, because the high-rate sources always experience a busy network. With strict priority for high-rate sources $\max(c_{LP}, c_{HP}) = c_{LP}$ when the workload is high, because low-rate data has to wait for high-rate data in case of congestion.

			$N_H = 0$		$N_H = 1$		$N_H = 2$	
			c	Δc	c	Δc	c	Δc
Streaming		c_L	148	48%				
	$\gamma_H = 0.01$	$\max(c_L, c_H)$			194	94%	199	99%
		$\max(c_{LP}, c_{HP})$			155	55%	161	61%
	$\gamma_H = 0.1$	$\max(c_L, c_H)$			188	88%	214	114%
		$\max(c_{LP}, c_{HP})$			174	74%	188	88%
Elastic		c_L	121	21%				
	$\gamma_H = 0.01$	$\max(c_L, c_H)$			154	54%	161	61%
		$\max(c_{LP}, c_{HP})$			128	28%	137	37%
	$\gamma_H = 0.1$	$\max(c_L, c_H)$			149	49%	168	68%
		$\max(c_{LP}, c_{HP})$			150	50%	161	61%

Table 4: Required capacities (absolute values c and relative values Δc with $\rho = 100\,\text{Mb/s}$)

In Table 4 the relative capacities (see Definition (35)) are collected for the streaming and the elastic model with a workload $\rho = 100\,\text{Mb/s}$. This table contains the data of Table 2 (for streaming services) and Table 3 (for elastic services). In the sections corresponding to these tables the required capacities for different behaviors of a corporate user are compared for streaming and elastic services separately. Now we only need to compare the required capacities between the two models.

We observe that the relative capacities for the streaming model are significantly higher than for the elastic model in all situations. The explanation is that the performance requirement for the streaming model (with $d_L = d_H = 0.02$) is more stringent than the performance requirement in the elastic model (with $\alpha_L = \alpha_H = 0.8$). A comparison of the performance requirements for the streaming and the elastic model is presented in Section 6.2.

With stringent performance requirements (very small d and ϵ, large α) the behaviors of the streaming and the elastic model are very much the same. In the streaming model the buffer is empty most of the time and in the elastic model the transmission rates are close to the access rates in this situation.

6.2 Comparison of the performance requirements

The goal of this section is to demonstrate that the performance requirements in the elastic model are more stringent than the performance requirements in the streaming model.

The performance requirement for low-rate sources in the elastic model is

$$\mathbb{P}(T_L < \alpha_L r_L) < \epsilon_L.$$

The performance requirements in the streaming model are formulated in terms of packet delays. A one-to-one relation between the throughput of a file and the delay of a packet does not exist. However, we can relate the throughput of a file to the delay of the last packet of a file to make an estimate.

Consider a low-rate file with size f_L^{-1} and throughput $\alpha_L r_L$. The sojourn time of this file is $(f_L \alpha_L r_L)^{-1}$. Without any congestion the sojourn time of the file would have been $(f_L r_L)^{-1}$. The delay d_L of the last packet of this file is thus

$$d_L = \frac{1}{f_L \alpha_L r_L} - \frac{1}{f_L r_L} = \frac{1 - \alpha_L}{f_L \alpha_L r_L}, \tag{61}$$

and for a high-rate source it follows analogously:

$$d_H = \frac{1 - \alpha_H}{f_H \alpha_H r_H}. \tag{62}$$

When the same amount of data would have been transmitted in the streaming model, a delay of d_L (d_H) as in Equation (61) (Equation (62)) for the last packet would thus coincide with a throughput $\alpha_L r_L$ ($\alpha_H r_H$) in the elastic model. Note that the delay of the last packet of a file is in general larger than the delay of an arbitrary packet.

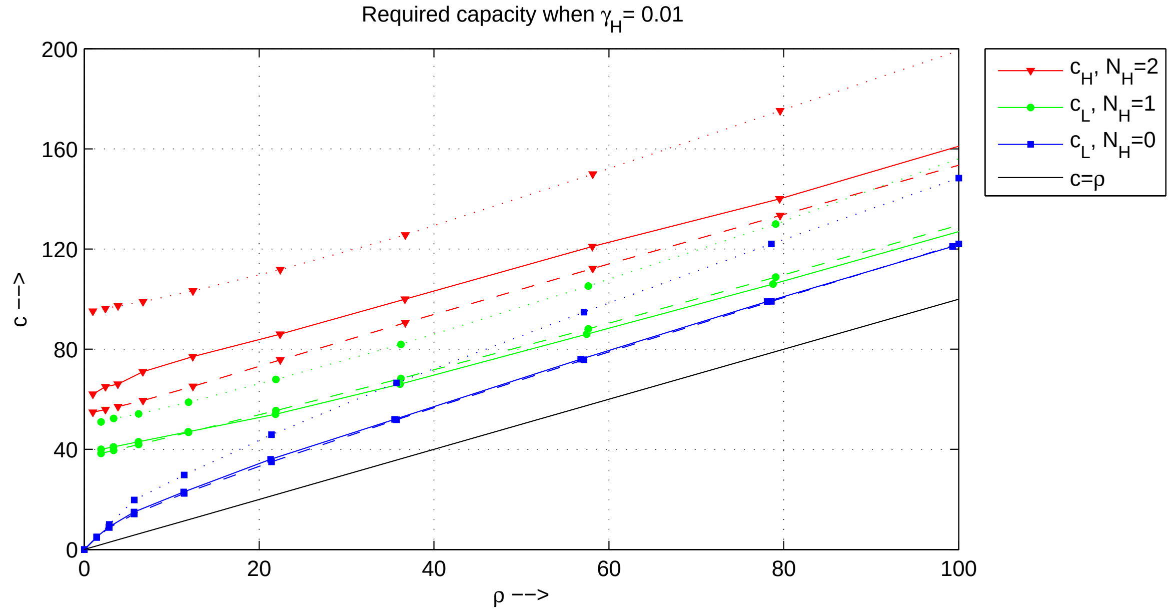
In Table 5 we show the results of the comparison above for the parameters as stated in Table 1 and in Figure 17 the required capacity in the streaming model is plotted with $d_L = 2$ and

Elastic	Streaming
$\alpha_L = 0.8$	$d_L = 2$
$\alpha_H = 0.8$	$d_H = 4$
$\alpha_L = 400/401 \approx 0.9975$	$d_L = 0.02$
$\alpha_H = 800/801 \approx 0.9988$	$d_H = 0.02$

Table 5: Comparison of α and d.

$d_H = 4$. Only the curves for c_L with $N_H = 0, 1$ and for c_H with $N_H = 2$ are displayed. The solid curves are the required capacities in the elastic model with $\alpha_L = \alpha_H = 0.8$ (obtained by simulation). The dashed curves are the required capacities in the streaming model with $d_L = 2$ and $d_H = 4$. The dotted curves are the required capacities in the streaming model with $d_L = d_H = 0.02$.
We observe that with $d_L = 2$ and $d_H = 4$ the required capacities in the streaming model are indeed close to the required capacity in the elastic model with $\alpha_L = \alpha_H = 0.8$.

With Table 5 and Figure 17 we conclude that the performance requirements for streaming traffic (with $d_L = d_H = 0.02$) are considerably more stringent than the performance requirement for elastic traffic (with $\alpha_L = \alpha_H = 0.8$).



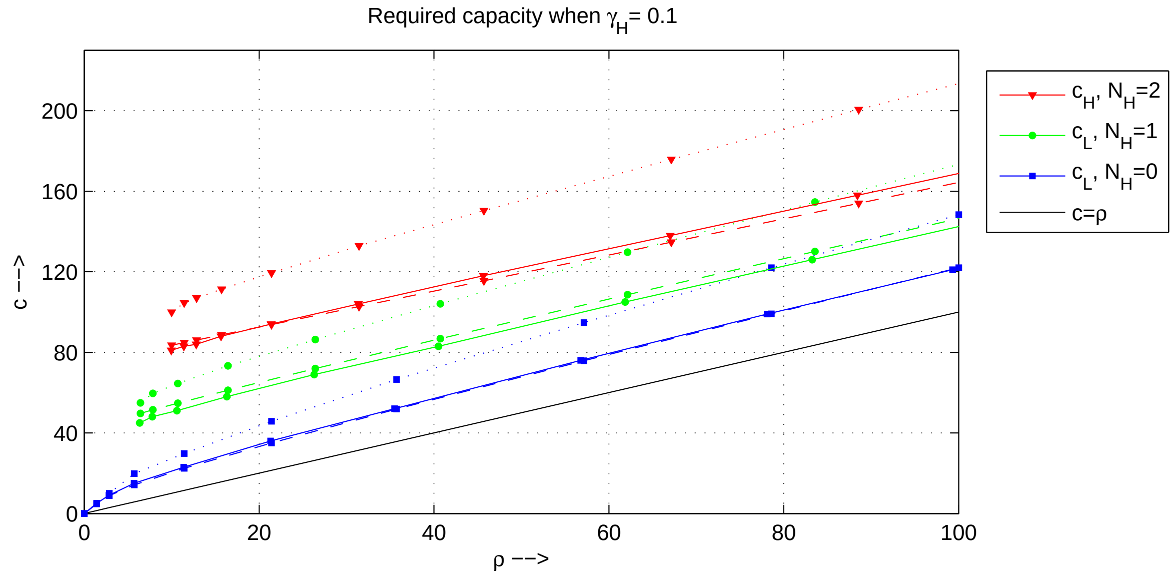


Figure 17: Required capacity when $d_L = 2$, $d_H = 4$. Top: $\gamma_H = 0.01$, bottom: $\gamma_H = 0.1$.

6.3 Practical recommendations for determining the required capacity of a network link

In operational capacity management, as implemented by network operators, the 5-minute workloads in the busiest hour of the day (see Section 2.1) are used to determine the required capacity. However, in Sections 4.5, 5.6 and 6.1 we have seen that the required capacity for a particular workload also depends on other characteristics.

If the traffic characteristics are not known, a network operator has to provide the capacity that is required to fulfill the performance requirements in the worst-case scenario, such that the performance requirements are satisfied in all cases. Then the provided capacity can be significantly higher than necessary in a particular situation (see Table 4), so it is useful to get more insight into the traffic characteristics. Important facts to know about the data traffic are:

1. The category of the communication service by which data packets are generated (streaming or elastic). In Section 6.1 we saw that the required capacity to fulfill the performance requirements for streaming traffic is significantly higher than the required capacity for elastic traffic.
2. The number of corporate users. These are the users with a high access rate.
3. The behavior of corporate users. Is traffic generated by a company the sum of the data traffic of the independently behaving employees or does the corporate user behave differently, e.g. he transfers a large amount of data from time to time? In Table 4 we observed that the required capacity strongly depends on the behavior of a high-rate source. If a corporate user transfers a large amount of data from time to time the required relative capacity with respect to the workload can be twice as much as if traffic from a corporate user is the sum of the data traffic of the employees.
4. The fraction of time that a corporate user transmits data in case he transfers a large amount of data from time to time. The required capacity strongly depends on the maximum number of sources that is active for a fraction of time exceeding ϵ, where ϵ is the fraction of traffic with an insufficient high performance.

In practice, some of the characteristics mentioned above can be obtained by an operator. The first characteristic is known by an operator in many cases, because service providers use virtual network connections per service (web, VoIP, TV) to avoid that the performance of a particular service is degraded by an excess of traffic from other services. In this way the category of traffic on a (virtual) network link can be obtained.

Also the second characteristic can be obtained. For a given link, the access rates are stored in a network administration system. From this administration system the number of users with a large access rate can be obtained.

The third and fourth traffic characteristics (about the behavior of corporate users) cannot easily be obtained by an operator. However, in Table 4 we observe that an operator can strongly improve the estimate for the required capacity when the behavior of a corporate user is known.

We continue with some practical recommendations that can be used to get insight into the traffic characteristics (also for the third and fourth characteristics):

- For the first traffic characteristic we mentioned that different (virtual) network connections are used. To obtain whether the traffic on a link has a streaming or elastic behavior, we can monitor the protocols over the link. If e.g. UDP is used, we know that the traffic has a streaming behavior and if e.g. TCP is used, the traffic is elastic (see Section 2.1).

- The number of users with a large access rate can be obtained from the network administration system. When all users have a small access rate, we know that the network traffic is only generated by consumers and we can obtain the required capacity with $N_H = 0$.

- When a few users have a large access rate, we consider two scenarios for the behavior of these corporate users as explained in Section 3.1. To obtain the behavior of the corporate users we can monitor the data traffic of these users. When we observe that the full access rate is used during a small fraction of the time and the amount of traffic transmitted during the remaining part of the time is negligible, we know that the corporate user should be modelled as a high-rate source. Otherwise, the data traffic of a corporate user is not different from the data traffic of many independently behaving consumers and we can use $N_H = 0$ to obtain the required capacity.

- Another option to get insight into the traffic characteristics is to use incidental 1-second measurements to obtain the peak rates. The level of these peak rates can then be compared with the peak rates in the model for the different situations to determine which model matches with the traffic characteristics on the link. With this model the required capacity for other values of the workload can also be determined.

Many measurements are required to perform the above-mentioned recommendations. However, the estimate for the required capacity can be significantly improved with the knowledge of these traffic characteristics. An operator can choose between performing these measurements to make a better estimate of the required capacity or providing the required capacity for the worst-case scenario. What the best option is depends on the trade-off between the potential profit of the knowledge of the traffic characteristics (which can be determined with the model) and the complexity of determining the traffic characteristics with the recommendations mentioned above.

7 Conclusions

This thesis presents a modelling approach for the dimensioning of an (IP) network link which carries data of both consumers and a small number of corporate users. The users make use of several applications, each having their own traffic characteristics. Two separate models are used for streaming and elastic traffic. We summarize the most important mathematical results for these models. Next we provide an overview of the most important results obtained in the numerical studies for both models and in the comparison between the two models. Finally we present practical recommendations to determine the required capacity on a network link.

For streaming traffic we derived exact expressions for the Quality of Service of the users for a given capacity. This is done both for the situation that all data traffic is handled equally and for the situation that data originating from corporate users is handled with strict priority. The required capacities to satisfy the performance requirements can be evaluated numerically with these expressions. Besides the exact expressions for the QoS, an approximation based on time-scale decomposition is used to reduce the computation time of the required capacity. It is shown that the approximation performs very well.
The numerical results for this model illustrate that for streaming traffic it is advantageous to handle data originating from corporate users with strict priority, because in that case the required capacity for low-rate sources hardly increases and the required capacity for high-rate sources significantly decreases.

For elastic traffic it is not possible to derive an exact expression for the QoS, because only the Laplace transform of the throughput distribution can be calculated and not the distribution itself. Therefore the performance requirement in the elastic model (which is based on the probability that the throughput of a file is insufficiently high) is translated into a target value for the average throughput. An approximation for the average throughput of a file is used to evaluate the required capacity. This approximation is time-consuming and performs worse than a simple approximation based on the instantaneous transmission rate of a file. The last approximation is quite good within the range of the realistic parameters used in this thesis.
Handling data originating from corporate users with strict priority is less advantegeous for elastic traffic than for streaming traffic, because the performance requirements for elastic services are less stringent. This causes that the performance for corporate users is unnecessarily high at the expense of the performance for the consumers (on which the required capacity is based in this case).

In the numerical studies in this thesis the required capacities to fulfill the performance requirements are compared in several situations. In particular, the impact of the presence of one or two corporate users on the required link capacity is examined. An observation for both models is that a corporate user that transmits a large amount of data from time to time, experiences a higher level of congestion in the network than consumers. Further the required capacity to satisfy the performance requirements of all users strongly depends on the traffic characteristics. We mention the most important traffic characteristics. The first one is the category of the communication service by which data packets are generated (streaming or elastic). The performance requirements for streaming traffic are more stringent than those for elastic traffic. The second characteristic is the number of corporate users in the network.

The third is the behavior of corporate users. When a corporate user transfers a large amount of data from time to time the required capacity to fulfill the performance requirements is higher than when traffic generated by a company is the sum of the data traffic originating from independent employees. The final important traffic characteristic is the fraction of time that a corporate user transmits data in case he transfers a large amount of data from time to time. The required capacity to satisfy the performance requirements strongly depends on the trade-off between the fraction of time that a corporate user transmits data and the fraction of traffic that is allowed to be transmitted with an insufficiently high performance.

In operational capacity management, as implemented by network operators, the 5-minute workloads in the busiest hour of the day are used to determine the required capacity. If the traffic characteristics are not known, a network operator has to provide the capacity that is required to fulfill the performance requirements in the worst-case scenario. We presented some practical recommendations to determine the traffic characteristics in the network, such that the required capacity can be better estimated. The first recommendation is to monitor the protocols over the link to determine whether the traffic on a link has a streaming or elastic behavior as elastic traffic requires less capacity. The second recommendation is to determine the number of users with a large access rate from the network administration system as this number has a strong impact on the required capacity. If a few users have a large access rate, the third recommendation is to monitor the data traffic of these users to determine their traffic characteristics.. Another option to get insight into the traffic characteristics is to use incidental 1-second measurements to obtain the peak rates and match these to a corresponding model such that the results of the model can be used to forecast the capacity for other links and workloads.

Whether an operator should provide the required capacity for the worst-case scenario or perform the recommendations mentioned above to get insight into the traffic characteristics depends on the trade-off between the potential profit of the knowledge of the traffic characteristics and the complexity of determining the traffic characteristics.

A Average packet delay

We calculate the average packet delay if a joint buffer is used. From Expression (16) in Section 4.2.1 it follows that the exceedance probability of the delay of a low-rate packet is given by

$$\mathbb{P}(D_L > x) = \frac{\sum_{i,j}(\pi_{i,j} - F_{i,j}(cx))i}{\gamma_L N_L},$$

where D_L is the delay of a low-rate packet when using a joint buffer. Then the probability density function of D_L is

$$\begin{aligned} f_{D_L}(x) &= -\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{P}(D_L > x)\frac{1}{\gamma_L N_L}\sum_{i,j} icf_{i,j}(cx) \\ &= \frac{c}{\gamma_L N_L}\sum_{i,j} i \sum_{z_k<0} a_k z_k (\boldsymbol{\psi}_k)_{(N_L+1)j+i}\exp(z_k cx) \\ &= \frac{c}{\gamma_L N_L}\sum_{z_k<0} a_k z_k \Big(\sum_{i,j} i(\boldsymbol{\psi}_k)_{(N_L+1)j+i}\Big)\exp(z_k cx), \end{aligned}$$

where $F_{i,j}(b)$ is given in Expression (24). Now we can compute the average packet delay of a low-rate packet.

$$\begin{aligned} \mathbb{E}[D_L] &= \int_0^\infty x f_{D_L}(x)\mathrm{d}x \\ &= \frac{c}{\gamma_L N_L}\sum_{z_k<0} a_k z_k \Big(\sum_{i,j} i(\boldsymbol{\psi}_k)_{(N_L+1)j+i}\Big)\int_0^\infty x\exp(z_k cx)\mathrm{d}x \\ &= \frac{1}{\gamma_L N_L c}\sum_{z_k<0}\frac{a_k}{z_k}\Big(\sum_{i,j} i(\boldsymbol{\psi}_k)_{(N_L+1)j+i}\Big). \end{aligned} \tag{63}$$

The average delay of a high-rate packet is

$$\mathbb{E}[D_H] = \frac{1}{\gamma_H N_H c}\sum_{z_k<0}\frac{a_k}{z_k}\Big(\sum_{i,j} j(\boldsymbol{\psi}_k)_{(N_L+1)j+i}\Big). \tag{64}$$

B Delay of a low-rate packet with two high-rate sources with strict priority

We calculate the joint probability that the delay of a low-rate packet exceeds d_L seconds and the state of the system is (i, j) when $N_H = 2$ and the high-rate sources have strict priority. We use the same method as for the situation with $N_H = 1$ in Section 4.3. We assume that the state of the high-rate source can change at most once while a packet is in the low-rate buffer, because the probability that the state of the high-rate source changes two or more times is negligible with the parameters as in Table 1.

- $j = 0$. The delay of a packet exceeds d_L when the buffer content exceeds cd_L and possibly when the buffer content is between $(c - r_H)d_L$ and cd_L and a high-rate source becomes active too soon. In the latter case the output rate of the low-rate buffer is c until the moment a high-rate source becomes active (after t seconds) and $c - r_H$ for the rest of the time. The total delay of a packet is then $t + (b - ct)/(c - r_H)$, which exceeds d_L with probability $1 - \exp(-2\lambda_H(b - (c - r_H)d_L)r_H^{-1})$.
- $j = 1$. The delay of a packet exceeds d_L when the buffer content exceeds cd_L and possibly when the buffer content is between $(c - r_H)d_L$ and cd_L and the high-rate source departs too late or the buffer content is between $(c - 2r_H)d_L$ and $(c - r_H)d_L$ and the second high-rate source arrives too soon. With a source departure, the output rate of the low-rate buffer is $c - r_H$ until the moment the high-rate source departs (after t seconds) and $c - r_H$ for the rest of the time. The total delay of a packet is then $t + (b - (c - r_H)t)c^{-1}$, which exceeds d_L with probability $\exp(f_H(b - cd_L))$. With an arrival of the second high-rate source, the output rate of the low-rate buffer is $c - r_H$ until the moment the high-rate source arrives (after t seconds) and $c - 2r_H$ for the rest of the time. The total delay of a packet is then $t + (b - (c - r_H)t)/(c - 2r_H)$, which exceeds d_L with probability $1 - \exp(-\lambda_H(b - (c - 2r_H)d_L)r_H^{-1})$.
- $j = 2$. The delay of a packet exceeds d_L when the buffer content exceeds $(c - r_H)d_L$ and possibly when the buffer content is between $(c - 2r_H)d_L$ and $(c - r_H)d_L$ and a high-rate source departs too late. In the latter case the output rate of the low-rate buffer is $c - 2r_H$ until the moment a high-rate source departs (after t seconds) and $c - r_H$ for the rest of the time. The total delay of a packet is then $t + (b - (c - 2r_H)t)/(c - r_H)$, which exceeds d_L with probability $\exp(2f_H(b - (c - r_H)d_L))$.

The expression for $G_{i,0}(d_L)$ when $N_H = 2$ is equal to the expression for $N_H = 1$ with λ_H replaced by $2\lambda_H$.

$$\begin{aligned} &G_{i,0}(d_L) \\ &\quad \approx \sum_{z_k<0} a_k(\boldsymbol{\psi}_k)_i \left[\frac{2\lambda_H}{r_H z_k - 2\lambda_H} \exp(z_k(c - r_H)d_L) - \frac{r_H z_k}{r_H z_k - 2\lambda_H} \exp(z_k c d_L - 2\lambda_H d_L) \right]. \end{aligned}$$

For $G_{i,1}(d_L)$ we have

$$\begin{aligned}
&G_{i,1}(d_L)\\
&\quad\approx\quad \pi_{i,1}-F_{i,1}(cd_L)+\int_{(c-r_H)d_L}^{cd_L} f_{i,1}(b)\exp\left(f_H(b-cd_L)\right)\mathrm{d}b\\
&\qquad\quad +\int_{(c-2r_H)d_L}^{(c-r_H)d_L} f_{i,1}(b)\left(1-\exp\left(-\lambda_H\frac{b-(c-2r_H)d_L}{r_H}\right)\right)\mathrm{d}b\\
&\quad=\quad -\sum_{z_k<0} a_k(\boldsymbol{\psi}_k)_{N_L+1+i}\exp(z_kcd_L)\\
&\qquad\quad +\sum_{z_k<0} a_k(\boldsymbol{\psi}_k)_{N_L+1+i}\frac{z_k}{z_k+f_H}(\exp(z_kcd_L)-\exp(z_k(c-r_H)d_L-r_Hf_Hd_L))\\
&\qquad\quad +\sum_{z_k<0} a_k(\boldsymbol{\psi}_k)_{N_L+1+i}\Bigg[\exp(z_k(c-r_H)d_L)-\exp(z_k(c-2r_H)d_L)\\
&\qquad\qquad\qquad -\frac{r_Hz_k}{r_Hz_k-\lambda_H}(\exp(z_k(c-r_H)d_L-\lambda_Hd_L)-\exp(z_k(c-2r_H)d_L))\Bigg]\\
&\quad=\quad -\sum_{z_k<0} a_k(\boldsymbol{\psi}_k)_{N_L+1+i}\Bigg[\frac{f_H}{z_k+f_H}\exp(z_kcd_L)+\frac{z_k}{z_k+f_H}\exp(z_k(c-r_H)d_L-r_Hf_Hd_L)\\
&\qquad\qquad\qquad -\exp(z_k(c-r_H)d_L)-\frac{\lambda_H}{r_Hz_k-\lambda_H}\exp(z_k(c-2r_H)d_L)\\
&\qquad\qquad\qquad +\frac{r_Hz_k}{r_Hz_k-\lambda_H}\exp(z_k(c-r_H)d_L-\lambda_Hd_L)\Bigg].
\end{aligned}$$

The expression for $G_{i,2}(d_L)$ when $N_H=2$ is equal to the expression for $G_{i,1}(d_L)$ when $N_H=1$ with f_H replaced by $2f_H$, c by $c-r_H$ and $(\boldsymbol{\psi}_k)_{N_L+1+i}$ by $(\boldsymbol{\psi}_k)_{2(N_L+1)+i}$.

$$\begin{aligned}
G_{i,2}(d_L)\quad\approx\quad &-\sum_{z_k<0} a_k(\boldsymbol{\psi}_k)_{2(N_L+1)+i}\Bigg[\frac{2f_H}{z_k+2f_H}\exp(z_k(c-r_H)d_L)\\
&+\frac{z_k}{z_k+2f_H}\exp(z_k(c-2r_H)d_L-2r_Hf_Hd_L)\Bigg].
\end{aligned}$$

C Variance and quantile of the instantaneous total input rate in the streaming model

In this section we calculate the variance and the 99% quantile of the instantaneous total input rate of the sources to explain the required capacity as plotted in the top part of Figure 10 in Section 4.5. Let R^I be the instantaneous total input rate of all active sources together at an arbitrary moment. Then $R^I = r_L X_L + r_H X_H$, where X_L and X_H are the number of active sources at a given moment. With ON-OFF sources, we have $X_L \sim BIN\left(N_L, \gamma_L\right)$ and $X_H \sim BIN\left(N_H, \gamma_H\right)$.

In Figure 10, the workload ρ is kept at a constant value of 60 Mb/s. In this figure the required capacity is plotted against the fraction of the traffic that originates from the high-rate source (η_H). The workload is defined as the mean instantaneous total input rate, so we have

$$\rho = \mathbb{E}[R^I] = r_L \gamma_L N_L + \eta_H \rho. \tag{65}$$

The workload is kept at a constant value by decreasing the number of active low-rate sources when η_H increases. We can find this number from Equation (65):

$$N_L = \frac{\rho(1-\eta_H)}{\gamma_L r_L}.$$

Further, we can use the relation

$$\gamma_H = \frac{\rho}{r_H N_H} \eta_H \tag{66}$$

to express the variance and quantile as a function of η_H.

C.1 Variance

The number of active high-rate sources is independent of the number of active low-rate sources, so we can add the variances.

$$\mathrm{Var}(R^I) = r_L^2 \gamma_L (1-\gamma_L) N_L + r_H^2 \gamma_H (1-\gamma_H) N_H, \quad 0 \le \gamma_H \le 1. \tag{67}$$

Now we substitute Equations (66) and (66) in Expression (67) and obtain:

$$\begin{aligned}
\mathrm{Var}(R^I) &= r_L(1-\gamma_L)\rho(1-\eta_H) + r_H^2 \frac{\rho}{r_H N_H}\eta_H\left(1 - \frac{\rho}{r_H N_H}\eta_H\right) N_H \\
&= r_L(1-\gamma_L)\rho + \rho\eta_H \left(r_H - r_L(1-\gamma_L) - \frac{\rho}{N_H}\eta_H \right) \\
&= \frac{1500}{7} + 60\eta_H \left(\frac{325}{7} - \frac{60}{N_H}\eta_H \right), \quad 0 \le \eta_H \le \frac{r_H N_H}{\rho} = \frac{5}{6} N_H.
\end{aligned}$$

In the last step, we substituted the parameters as used for Figure 10.

The variance as a function of the fraction of the traffic originating from a high-rate source is a parabolic function. The maximum variance is achieved when the fraction of traffic from the high-rate source is $N_H\left(r_H - r_L(1-\gamma_L)\right)/(2\rho)$. In our situation, this is when $\eta_H = 65/168 N_H$.

Note that this maximum is always reached with the same value for the fraction that a single high-rate source is active. The fraction that gives the maximum variance is $\gamma_H = 13/28$ in our situation or $\gamma_H = 1/2 - (1-\gamma_L)r_L/(2r_H)$ in general.

C.2 Quantile

We denote the 99% quantile with $x_{0.99}$. This quantile is the solution of the following equation.

$$\begin{aligned}
0.01 &= \mathbb{P}(R^I > x_{0.99}) \\
&= \mathbb{P}(r_L X + r_H Y > x_{0.99}) \\
&= \sum_{\substack{r_L n_L + r_H n_H \\ > x_{0.99}}} \begin{pmatrix} N_L \\ n_L \end{pmatrix} \gamma_L^{n_L} (1-\gamma_L)^{N_L - n_L} \begin{pmatrix} N_H \\ n_H \end{pmatrix} \gamma_H^{n_H} (1-\gamma_H)^{N_H - n_H} \\
&= \sum_{\substack{r_L n_L + r_H n_H \\ > x_{0.99}}} \begin{pmatrix} \frac{\rho(1-\eta_H)}{\gamma_L r_L} \\ n_L \end{pmatrix} \gamma_L^{n_L} (1-\gamma_L)^{\frac{\rho(1-\eta_H)}{\gamma_L r_L} - n_L} \\
&\quad \cdot \begin{pmatrix} N_H \\ n_H \end{pmatrix} \left(\frac{\rho\eta_H}{r_H N_H}\right)^{n_H} \left(1 - \frac{\rho\eta_H}{r_H N_H}\right)^{N_H - n_H} .
\end{aligned}$$

In the last step, we substituted Equations (66) and (66). Next we substitute the parameters as used for Figure 10. Then the 99% quantile is the solution of the following equation.

$$\begin{aligned}
0.01 &= \sum_{\substack{5n_L + 50n_H \\ > x_{0.99}}} \begin{pmatrix} 42(1-\eta_H) \\ n_L \end{pmatrix} \left(\frac{2}{7}\right)^{n_L} \left(\frac{5}{7}\right)^{42(1-\eta_H) - n_L} \\
&\quad \cdot \begin{pmatrix} N_H \\ n_H \end{pmatrix} \left(\frac{6\eta_H}{5N_H}\right)^{n_H} \left(1 - \frac{6\eta_H}{5N_H}\right)^{N_H - n_H} .
\end{aligned}$$

D Product-form stationary distribution

We compute the stationary distribution of the elastic model of which the transition rates are given in Section 5.1, with the extra assumption that $r_L = r_H = r$. This is not a realistic assumption for the model, but with this assumption we can model the network as a closed network of queues with a product-form distribution. The modelling of the network is shown in Figure 18.

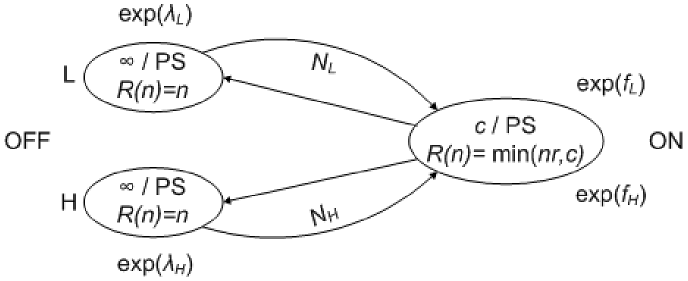


Figure 18: Closed network for the elastic model with $r_L = r_H = r$.

The queues on the left side model the inactive periods of the sources. These queues are infinite-server queues. The inactive periods of the sources have exponential (λ_L) and exponential (λ_H) durations, and the server rate is $R(n) = n$ when n sources of the low-rate (high-rate) source are inactive. The queue on the right side models the active periods of the sources. This is a processor-sharing server with maximum server rate c. When n sources are active (low-rate and high-rate sources together), the server rate is $R(n) = \min(nr, c)$. The durations of the active periods are exponential (f_L) and exponential (f_H) for the low-rate sources and high-rate sources respectively.

The stationary distribution of this network has a product-form. Define

$$\pi(n_L^{OFF}, n_H^{OFF}, n_L^{ON}, n_H^{ON})$$

as the stationary probability that n_L^{OFF} low-rate and n_H^{OFF} high-rate sources are inactive and n_L^{ON} low-rate and n_H^{ON} high-rate sources are active. In total we have N_L (N_H) low-rate (high-rate) sources, so $n_L^{OFF} + n_L^{ON} = N_L$ and $n_H^{OFF} + n_H^{ON} = N_H$. From these relations it follows that the state of the process is determined by only n_L^{ON} and n_H^{ON}. The stationary distribution has a product-form, so

$$\pi(n_L^{ON}, n_H^{ON}) := \pi(n_L^{OFF}, n_H^{OFF}, n_L^{ON}, n_H^{ON}) = G\pi_L(n_L^{OFF})\pi_H(n_H^{OFF})\pi_{ON}(n_L^{ON}, n_H^{ON}), \qquad (68)$$

where $\boldsymbol{\pi}_L$, $\boldsymbol{\pi}_H$ and $\boldsymbol{\pi}_{ON}$ are the stationary distributions of the infinite-server or processor-sharing queues and G is the normalization constant. We have

$$\begin{aligned}
\pi_L(n_L^{OFF}) &= G_L \frac{(\rho_L^{OFF})^{n_L^{OFF}}}{n_L^{OFF}!}, \\
\pi_H(n_H^{OFF}) &= G_H \frac{(\rho_H^{OFF})^{n_H^{OFF}}}{n_H^{OFF}!}, \\
\pi_L(n_L^{ON}, n_H^{ON}) &= G_{ON} \binom{n_L^{ON} + n_H^{ON}}{n_H^{ON}} \frac{(\rho_L^{ON})^{n_L^{ON}} (\rho_H^{ON})^{n_H^{ON}}}{R(1)R(2)\cdots R(n_L^{ON} + n_H^{ON})},
\end{aligned}$$

where G_L, G_H and G_{ON} are constants and ρ_L^{OFF}, ρ_H^{OFF}, ρ_L^{ON} and ρ_H^{ON} are the server utilizations. These are

$$\rho_L^{OFF} = \frac{v_L}{\lambda_L},\ \ \rho_H^{OFF} = \frac{v_H}{\lambda_H},\ \ \rho_L^{ON} = \frac{v_L}{f_L},\ \ \rho_H^{ON} = \frac{v_H}{f_H},$$

where v_L (v_H) is the rate at which low-rate (high-rate) sources turn ON or OFF.

Substituting $\boldsymbol{\pi}_L$, $\boldsymbol{\pi}_H$ and $\boldsymbol{\pi}_{ON}$ in Equation (68) and using $n_L^{OFF}+n_L^{ON} = N_L$ and $n_H^{OFF}+n_H^{ON} = N_H$ gives

$$\begin{aligned}
\pi(n_L^{ON}, n_H^{ON}) &= G\frac{(\rho_L^{OFF})^{N_L-n_L^{ON}}}{(N_L-n_L^{ON})!}\frac{(\rho_H^{OFF})^{N_H-n_H^{ON}}}{(N_H-n_H^{ON})!}\left(\begin{array}{c} n_L^{ON}+n_H^{ON} \\ n_H^{ON} \end{array}\right) \\
&\quad \cdot \frac{(\rho_L^{ON})^{n_L^{ON}}(\rho_H^{ON})^{n_H^{ON}}}{R(1)R(2)\cdots R(n_L^{ON}+n_H^{ON})} \\
&= G\left(\frac{v_L}{\lambda_L}\right)^{N_L}\left(\frac{v_H}{\lambda_H}\right)^{N_H}\frac{1}{N_L!\,N_H!}\left(\frac{\lambda_L}{f_L}\right)^{n_L^{ON}}\left(\frac{\lambda_H}{f_H}\right)^{n_H^{ON}} \\
&\quad \cdot \left(\begin{array}{c} N_L \\ n_L^{ON} \end{array}\right)\left(\begin{array}{c} N_H \\ n_H^{ON} \end{array}\right)\frac{(n_L^{ON}+n_H^{ON})!}{R(1)R(2)\cdots R(n_L^{ON}+n_H^{ON})} \\
&= G'\left(\begin{array}{c} N_L \\ n_L^{ON} \end{array}\right)\left(\frac{\lambda_L}{f_L}\right)^{n_L^{ON}}\left(\frac{\lambda_H}{f_H}\right)^{n_H^{ON}}\left(\begin{array}{c} N_H \\ n_H^{ON} \end{array}\right)\frac{(n_L^{ON}+n_H^{ON})!}{R(1)R(2)\cdots R(n_L^{ON}+n_H^{ON})}.
\end{aligned}$$

The normalization constant G' follows from the fact that $\boldsymbol{\pi}$ is a probability distribution:

$$G' = \left(\sum_{n_L=0}^{N_L}\sum_{n_H=0}^{N_H}\left(\begin{array}{c} N_L \\ n_L^{ON} \end{array}\right)\left(\frac{\lambda_L}{f_L}\right)^{n_L^{ON}}\left(\frac{\lambda_H}{f_H}\right)^{n_H^{ON}}\left(\begin{array}{c} N_H \\ n_H^{ON} \end{array}\right)\frac{(n_L^{ON}+n_H^{ON})!}{R(1)R(2)\cdots R(n_L^{ON}+n_H^{ON})}\right)^{-1}.$$

E Closed-form Laplace transform of the conditional sojourn time distribution

We derive a closed-form expression for the Laplace transform of $S^L_{i,j}(x)$ along the lines of [11], pages 105-108. The Laplace transform of the conditional sojourn time of a high-rate source can be computed analogously.

Consider an amount of low-rate data x and a time interval of length $\Delta > 0$, with Δ sufficiently small such that the transfer of the file for which we determine the transfer time cannot finish within this time, i.e. $\Delta < xR_L(i,j)^{-1}$. We condition on all possible events occurring during this interval. These are the events illustrated in Figure 14, except for the departure of the special source.

$$
\begin{aligned}
&\widetilde{S}^L_{i,j}(x,\omega)\\
&\quad= \mathbb{E}[e^{-\omega S^L_{i,j}(x)}]\\
&\quad= iR_L(i+1,j)f_L\Delta e^{-\omega\Delta}\widetilde{S}^L_{i-1,j}(x-R_L(i+1,j)(\Delta-\mathcal{O}(\Delta))-R_L(i,j)\mathcal{O}(\Delta),\omega)\\
&\qquad+(N_L-i-1)\lambda_L\Delta e^{-\omega\Delta}\widetilde{S}^L_{i+1,j}(x-R_L(i+1,j)(\Delta-\mathcal{O}(\Delta))-R_L(i+2,j)\mathcal{O}(\Delta),\omega)\\
&\qquad+jR_H(i+1,j)f_H\Delta e^{-\omega\Delta}\widetilde{S}^L_{i,j-1}(x-R_L(i+1,j)(\Delta-\mathcal{O}(\Delta))-R_L(i+1,j-1)\mathcal{O}(\Delta),\omega)\\
&\qquad+(N_H-j)\lambda_H\Delta e^{-\omega\Delta}\widetilde{S}^L_{i,j+1}(x-R_L(i+1,j)(\Delta-\mathcal{O}(\Delta))-R_L(i+1,j+1)\mathcal{O}(\Delta),\omega)\\
&\qquad+(1-A(i,j)\Delta)e^{-\omega\Delta}\widetilde{S}^L_{i,j}(x-R_L(i+1,j)\Delta,\omega)+o(\Delta).
\end{aligned}
$$

Rearranging terms and summarizing the decrease of the remaining amount of data with $\mathcal{O}(\Delta)$ gives

$$
\begin{aligned}
&\frac{\widetilde{S}^L_{i,j}(x,\omega)-\widetilde{S}^L_{i,j}(x-R_L(i+1,j)\Delta,\omega)}{R_L(i+1,j)\Delta}\\
&\quad=\frac{iR_L(i+1,j)f_L}{R_L(i+1,j)}e^{-\omega\Delta}\widetilde{S}^L_{i-1,j}(x-\mathcal{O}(\Delta),\omega)+\frac{(N_L-i-1)\lambda_L}{R_L(i+1,j)}e^{-\omega\Delta}\widetilde{S}^L_{i+1,j}(x-\mathcal{O}(\Delta),\omega)\\
&\qquad+\frac{jR_H(i+1,j)f_H}{R_L(i+1,j)}e^{-\omega\Delta}\widetilde{S}^L_{i,j-1}(x-\mathcal{O}(\Delta),\omega)+\frac{(N_H-j)\lambda_H}{R_L(i+1,j)}e^{-\omega\Delta}\widetilde{S}^L_{i,j+1}(x-\mathcal{O}(\Delta),\omega)\\
&\qquad-\frac{A(i,j)}{R_L(i+1,j)}e^{-\omega\Delta}\widetilde{S}^L_{i,j}(x-\mathcal{O}(\Delta),\omega)+\frac{e^{-\omega\Delta}-1}{R_L(i+1,j)\Delta}e^{-\omega\Delta}\widetilde{S}^L_{i,j}(x-\mathcal{O}(\Delta),\omega)+\frac{o(\Delta)}{\Delta}.
\end{aligned}
$$

Now let $\Delta \to 0$, then

$$
\begin{aligned}
\frac{\partial\widetilde{S}^L_{i,j}(x,\omega)}{\partial x} &= \frac{iR_L(i+1,j)f_L}{R_L(i+1,j)}\widetilde{S}^L_{i-1,j}(x,\omega)+\frac{(N_L-i-1)\lambda_L}{R_L(i+1,j)}\widetilde{S}^L_{i+1,j}(x,\omega)\\
&\quad+\frac{jR_H(i+1,j)f_H}{R_L(i+1,j)}\widetilde{S}^L_{i,j-1}(x,\omega)+\frac{(N_H-j)\lambda_H}{R_L(i+1,j)}\widetilde{S}^L_{i,j+1}(x,\omega)\\
&\quad-\frac{\omega+A(i,j)}{R_L(i+1,j)}\widetilde{S}^L_{i,j}(x,\omega).
\end{aligned}
$$

In matrix notation this is

$$
\frac{\partial}{\partial x}\widetilde{S}^L(x,\omega)=\mathbf{R}_L^{-1}\left(\mathbf{Q}^*_L-\omega\mathbf{I}\right)\widetilde{S}^L(x,\omega), \tag{69}
$$

where $\tilde{\mathbf{S}}^L(x,\omega)$ is the vector with entries $\widetilde{S}^L_{i,j}(x,\omega)$ and $\mathbf{Q}^*_L$ and $\mathbf{R}_L$ as defined in Section 5.1.

The time needed to transfer no data is 0, so $S^L_{i,j}(0,\omega)=0$ for all $(i,j)\in\mathbb{S}_L$, which gives the initial condition

$$\widetilde{S}^L(0,\omega)=\mathbf{1}. \tag{70}$$

The unique solution to Equations (69) and (70) is

$$\widetilde{S}^L(x,\omega)=\exp\left(\mathbf{R}_L^{-1}\left(\mathbf{Q}^*_L-\omega\mathbf{I}\right)x\right)\mathbf{1}. \tag{71}$$

With the Laplace transform, we can calculate the mean (conditional and unconditional) throughput. The mean conditional throughput is the mean throughput of a file with a known file size and conditioned on the state of the process at arrival. This quantity can be calculated in the following way ([11], page 191):

$$\begin{aligned}
\mathbb{E}\left[T^L_{i,j}(x)\right] &= \mathbb{E}[\frac{x}{S^L_{i,j}(x)}] \\
&= \int_0^\infty \frac{x}{t}\mathrm{d}\Phi_{i,j,x}(t) \\
&= x\int_0^\infty \left(\int_0^\infty \exp(-\omega t)\mathrm{d}\omega\right)\mathrm{d}\Phi_{i,j,x}(t) \\
&= x\int_0^\infty \left(\int_0^\infty \exp(-\omega t)\mathrm{d}\Phi_{i,j,x}(t)\right)\mathrm{d}\omega \\
&= x\int_0^\infty \widetilde{S}^L_{i,j}(x,\omega)\mathrm{d}\omega,
\end{aligned}$$

where $\Phi_{i,j,x}(t)$ is the cumulative distribution function of $S^L_{i,j}(x)$ given file size x and state (i,j) at arrival.

Deconditioning on x and (i,j) gives the unconditional mean throughput:

$$\mathbb{E}[T^L]=\sum_{j=0}^{N_H}\sum_{i=0}^{N_L-1}\hat{\pi}^L_{i,j}\int_0^\infty f_L e^{-f_L x}\left(x\int_0^\infty \widetilde{S}^L_{i,j}(x,\omega)\mathrm{d}\omega\right)\mathrm{d}x, \tag{72}$$

where $\hat{\boldsymbol{\pi}}$ is the distribution of the state of the system at an arrival moment of a low-rate file.

F An alternative method to compute the mean conditional sojourn time

We describe an alternative method to compute the mean conditional sojourn time, as given in Expression (54). This method is along the lines of [18] (page 137), which in turn relies on [4].

We derive a recursive relation for the Laplace transform of the mean transfer time $\overline{S}_{i,j}^L(x)$. To do this, we look at the first transition of the process and condition on the time t of this first transition. We know that the remaining amount of data that has to be transmitted by the source we are looking at is x, so when no sources change from active to inactive or the other way round, the transmission will be completed after $xR_L(i+1,j)^{-1}$ seconds. Otherwise, the transition of the rest of the process is the first event and we can express the average sojourn time in terms of the average sojourn time of the remaining data.

A recursive relation for the mean conditional transfer time of a file is:

$$\begin{aligned}
\overline{S}_{i,j}^L(x) &= \int_{\frac{x}{R_L(i+1,j)}}^{\infty} A(i,j)e^{-A(i,j)t}\frac{x}{R_L(i+1,j)}\mathrm{d}t \\
&\quad + \int_0^{\frac{x}{R_L(i+1,j)}} A(i,j)e^{-A(i,j)t}\Big\{ t + \frac{iR_L(i+1,j)f_L}{A(i,j)}\overline{S}_{i-1,j}^L(x-R_L(i+1,j)t) \\
&\qquad\qquad + \frac{(N_L-i-1)\lambda_L}{A(i,j)}\overline{S}_{i+1,j}^L(x-R_L(i+1,j)t) \\
&\qquad\qquad + \frac{jR_H(i+1,j)f_H}{A(i,j)}\overline{S}_{i,j-1}^L(x-R_L(i+1,j)t) \\
&\qquad\qquad + \frac{(N_H-j)\lambda_H}{A(i,j)}\overline{S}_{i,j+1}^L(x-R_L(i+1,j)t) \qquad \Big\}\mathrm{d}t \\
&= \frac{x}{R_L(i+1,j)}e^{-A(i,j)\frac{x}{R_L(i+1,j)}} + \int_0^x \frac{A(i,j)}{R_L(i+1,j)}e^{-A(i,j)\frac{x-t}{R_L(i+1,j)}} \\
&\quad \cdot\Big\{ \frac{x-t}{R_L(i+1,j)} + \frac{iR_L(i+1,j)f_L}{A(i,j)}\overline{S}_{i-1,j}^L(t) + \frac{(N_L-i-1)\lambda_L}{A(i,j)}\overline{S}_{i+1,j}^L(t) \\
&\quad + \frac{jR_H(i+1,j)f_H}{A(i,j)}\overline{S}_{i,j-1}^L(t) + \frac{(N_H-j)\lambda_H}{A(i,j)}\overline{S}_{i,j+1}^L(t)\Big\}\mathrm{d}t.
\end{aligned}$$

By $\psi_{i,j}(\omega)$ we denote the Laplace transform of $\overline{S}_{i,j}^L(x)$, i.e.

$$\psi_{i,j}(\omega) = \int_0^\infty e^{-\omega\overline{S}_{i,j}^L(x)}\mathrm{d}x.$$

Now we take Laplace transforms of the recursive relation for the mean conditional transfer time and get for all $(i,j) \in \mathbb{S}_L$ (after some algebra):

$$\begin{aligned}
\psi_{i,j}(\omega) &= \frac{1}{\omega}\frac{1}{R_L(i+1,j)\omega + A(i,j)} + \frac{1}{R_L(i+1,j)\omega + A(i,j)}\Big(iR_L(i+1,j)f_L\psi_{i-1,j}(\omega) \\
&\quad +(N_L-i-1)\lambda_L\psi_{i+1,j}(\omega) + jR_H(i+1,j)f_H\psi_{i,j-1}(\omega) + (N_H-j)\lambda_H\psi_{i,j+1}(\omega)\Big).
\end{aligned}$$

This leads to the following system of linear equations:

$$\begin{aligned}\frac{1}{\omega} &= -iR_L(i+1,j)f_L\psi_{i-1,j}(\omega)(N_L-i-1)\lambda_L\psi_{i+1,j}(\omega)+(R_L(i_1,j)\omega+A(i,j))\psi_{i,j}(\omega)\\ &\quad -jR_H(i+1,j)f_H\psi_{i,j-1}(\omega)-(N_H-j)\lambda_H\psi_{i,j+1}(\omega),\end{aligned}$$

for all $(i,j)\in\mathbb{S}_L$. In matrix notation, this is

$$\mathbf{1}=\omega\mathbf{B}(\omega)\boldsymbol{\psi}(\omega),$$

where $\mathbf{B}(\omega):=-\mathbf{Q}^*+\omega\mathbf{R}_L$ and the states are ordered colexicographically to make the state space one-dimensional.

The Laplace transform can be solved from the linear system by applying Cramer's rule to $\omega\boldsymbol{\psi}(\omega)=(\mathbf{B}(\omega))^{-1}\mathbf{1}$. Then we get

$$\omega\psi_n(\omega)=\frac{\det(\mathbf{B}_{-n}(\omega))}{\det(\mathbf{B}(\omega))}, \tag{73}$$

where $\mathbf{B}_{-n}(\omega)$ is defined as $\mathbf{B}(\omega)$ with the column corresponding to state n replaced by $\mathbf{1}$. The roots of $\det(\mathbf{B}(\omega))=0$ are precisely the eigenvalues of $\mathbf{R}_L^{-1}\mathbf{Q}^*$, because $\det(\mathbf{B}(\omega))=0$ coincides with $\det(\mathbf{R}_L^{-1}\mathbf{Q}^*-\omega\mathbf{I})=0$. From Theorem 1 of [14], we know that the number of negative eigenvalues is equal to the number of positive diagonal elements of $\mathbf{R}$ minus 1. Further, the multiplicity of the eigenvalue 0 is 1. So the eigenvalues of $\mathbf{R}_L^{-1}\mathbf{Q}^*$ are $z_{(N_H+1)N_L-1}<z_{(N_H+1)N_L-2}<\cdots<z_1<z_0=0$. Note that the size of the matrices $\mathbf{Q}^*$ and $\mathbf{R}_L$ is $(N_H+1)N_L\times(N_H+1)N_L$, because we consider a process with one low-rate source less.

With the roots of $\det(\mathbf{B}(\omega))=0$, we can write down a partial-fraction representation of Expression (73). We obtain

$$\omega\psi_n(\omega)=\frac{U_{n,0}}{\omega}+\sum_{k=1}^{(N_H+1)N_L-1}\frac{U_{n,k}}{\omega-z_k}, \tag{74}$$

where the constants $U_{n,k}$ are such that Expression (74) corresponds to Expression (73). Now the solution for the Laplace transform is

$$\begin{aligned}\psi_n(\omega) &= \frac{U_{n,0}}{\omega^2}+\sum_{k=1}^{(N_H+1)N_L-1}\frac{U_{n,k}}{\omega(\omega-z_k)}\\ &= \frac{U_{n,0}}{\omega^2}+\sum_{k=1}^{(N_H+1)N_L-1}\frac{U_{n,k}}{z_k}\left(\frac{1}{\omega-z_k}-\frac{1}{\omega}\right).\end{aligned}$$

Inverting this Laplace transform yields the result for $\overline{S}_{i,j}^L(x)$, namely

$$\overline{S}_{i,j}^L(x)=U_{N_Lj+i,0}\,x+\sum_{k=1}^{(N_H+1)N_L-1}\frac{U_{N_Lj+i,k}}{z_k}(e^{z_kx}-1).$$

We can check that this method leads to the same solution as the one given in Section 5.4 by taking the Laplace transform of both sides of differential Equation (44). Then we get

$$\omega\boldsymbol{\psi}(\omega) = \mathbf{R}_L^{-1}(\frac{1}{\omega} + \mathbf{Q}^*\boldsymbol{\psi}(\omega)),$$

which can be rewritten as

$$\mathbf{1} = \omega(-\mathbf{Q}^* + \omega\mathbf{R}_L)\boldsymbol{\psi}(\omega)$$

and is equal to Equation (69).

References

[1] D. Anick, D. Mitra, M.M. Sondhi (1982), Stochastic theory of a data-handling system with multiple sources, Bell System Technical Journal, Vol. 61, p. 1871–1894.

[2] H. van den Berg, M. Mandjes, R. van de Meent, A. Pras, F. Roijers, and P. Venemans (2006), QoS-aware bandwidth provisioning for IP network links, Computer Networks, Vol. 50, p. 631–647.

[3] T. Bonald, P. Olivier and J. Roberts (2003), Dimensioning high speed IP access networks, in Proceedings of ITC 18, p. 241–251.

[4] S. Borst, O. Boxma, N. Hegde (2005), Sojourn times in finite-capacity processor-sharing queues, in Proceedings of the 1st EURO-NGI Conference, p. 53–60.

[5] J.W. Cohen (1979). The multiple phase service network with generalized processor sharing, Acta Informatica, Vol. 12, p. 245-284.

[6] G. Fayolle, I. Mitrani, R. Iasnogorodski (1980), Sharing a processor among many job classes, Journal of the ACM, Vol. 27(3), p. 519–532.

[7] C. Fraleigh (2002), Provisioning Internet Backbone Networks to Support Latency Sensitive Applications, PhD Thesis, Stanford University, USA.

[8] C. Fraleigh, F. Tobagi, C. Diot (2003), Provisioning IP backbone networks to support latency sensitive traffic, in Proceedings of IEEE Infocom, p. 375–385.

[9] M. Gribaudo, M. Telek (2007), Fluid models in performance analysis, in Formal Methods for Performance Evaluation, p. 271-317.

[10] R.A. Horn, C.R. Johnson (1985), Matrix analysis, Cambridge University Press.

[11] R. Litjens (2003), Capacity allocation in wireless communication networks - models and analyses, PhD Thesis, University of Twente.

[12] M. Mandjes (2007), Large Deviations for Gaussian Queues: Modelling Communication Networks, Wiley.

[13] R. van de Meent, M. Mandjes, A. Pras (2006), Gaussian traffic everywhere?, in Proceedings of the IEEE International Conference on Communications.

[14] D. Mitra (1988), Stochastic theory of a fluid model of producers and consumers coupled by a buffer, Adv. in Appl. Probab. 20, p. 646–676.

[15] J. Postel (1980), User Datagram Protocol, RFC 768.

[16] M. del Rey (1981), Transmission Control Protocol, RFC 793.

[17] J.W. Roberts (2001), Traffic theory and the internet, IEEE Communications Magazine, Vol. 39(1), p. 94–99.

[18] F. Roijers (2009), Fluid models for QoS provisioning in communication networks, PhD Thesis, University of Amsterdam.

[19] S.M. Ross (2003), Introduction to Probability Models, Eighth Edition, Academic Press.

[20] W. Stevens (1994), TCP/IP Illustrated, Volume 1: The protocols, Addison-Wesley.