Stochastic modelling and control of communication networks
Zuraniewski, P.W.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Chapter 1

Introduction

1 Mathematical models for communication networks

During the last two decades we have witnessed an unprecedented growth of the information technologies (IT) sector [1]. New services emerged, resulting in the evolution of communication networks from platforms which carried mostly voice towards an infrastructure capable of carrying data, voice and video. Customers’ demands grow continuously: companies want to send their mission-critical data in a secure and reliable way, individual users want to be (almost) always on-line. The technology did not evolve at the same pace. Technologies, that are now the foundation of the Internet, were not designed to handle all the current requirements. Therefore, a challenging task for the research society and IT industry is to develop solutions for new problems faced by network users, providers and administrators.

This motivates the increasing interest in the use of mathematical modelling to develop sound mechanisms for network traffic control and management. Currently, such routines are often based on an empirical approach, incorporate ad-hoc methods or ‘rules of thumb’, and their general validity may be questioned. The development of reliable procedures requires a systematic and rigorous approach, which is enabled by adequate use of the mathematical apparatus. Good mathematical models may help for example in reducing costs of the infrastructure investments (think of dimensioning issues), or allow to offer ‘better’ services (i.e., with higher reliability level, less latency, etc.) which may be a crucial factor in the highly competitive market of new technologies.

Recently, we observe that Ethernet, which was originally designed as a Local Area Network (LAN) protocol, has attracted more attention of the service providers and equipment vendors as a technology which can also be used in Wide Area Networks.
(WANs) [2]. There are several advantages of having the same technology at both the LAN and WAN level, such as lower purchase and operation costs and avoidance of ‘translation’ between protocols or additional encapsulation. Besides purely technology-dependent challenges, one may also identify many challenges which have to be addressed to successfully run the network, which are not concerned with a particular technology but could be considered generic. These problems emerge at different timescales, cf. the rough description highlighted in e.g. [3, 4].

- fine or subsecond timescale – traffic management which may be associated with mechanisms like policing or queue management. Policing is a mechanism which ensures that traffic stream complies with a certain contract. If, for example, the source transfer rate exceeds the agreed value, a policer detects this and drops (or marks) traffic that is sent above the limit. Lack of such procedures may result in one client using all the available resources, leaving no chance for the others to transmit anything.

Queue management mechanisms assume non-equal treatment of all incoming traffic. It may prevent a threat of delay sensitive data packets to be treated the same way as delay-tolerant ones. The queue manager may recognize packets which need special service and for example put them in front of the queue.

- medium (orders of seconds to minutes) timescale – traffic control. One example here is admission control, i.e., the decision whether a particular request should be admitted or rejected. Rejection may be necessary if we foresee that an increase of the load imposed on the system will have an unacceptable negative impact on the currently transmitted traffic, or if a requested quality level for the new connection cannot be guaranteed. Developing appropriate assessment procedures is an ambitious goal due to the large number of aspects which have to be taken into account.

Another example is the anomaly detection problem, i.e., a notification to the system administrator that there are certain problems developing in the network. They may be due to an equipment failure or intentional malicious activity (an ‘attack’), but in any case it is extremely important to know early about it. Finding a balance between false positives and misses is a classical but notoriously hard problem here.

- large and very large timescale (weeks to months) – dimensioning and capacity planning would be the typical tasks performed on coarser timescales. They are usually strategic decisions, often with substantial financial consequences as they may require purchasing new equipment, renting more optical fibers or hiring additional
staff. Before any action is taken, a careful analysis is conducted to assess to what extent an expansion of the system is needed.

Additionally, resource allocation could be placed somewhere in between the medium and large timescale. Setup and teardown of high speed and secure connections dedicated for a certain task (like massive backup) may serve as a good example here. The availability of good statistical information is an important factor if we wish to predict resource consumption: if the traffic is highly variable (‘bursty’) much more capacity may be required compared to a situation when the traffic is ‘smooth’.

In order to deal with these topics, several strategies can be employed. The first one could be using hands-on experience and some rules of thumb along with, for example, default configuration parameters for the network equipment. While it may be attractive due to the low overhead and possibly works well for some cases (small networks, limited set of services), this strategy tends to work poorly if the scale of the system grows. One may argue that an other approach may be more beneficial in which the use of measurements and mathematical models should play a crucial role in many traffic management mechanisms. Once we define a technical problem and reformulate it as a mathematical question, we are able to analyze it in a rigorous way using powerful existing mathematical techniques (and/or developing new techniques). Furthermore, once we propose the model, real data traces allow us to fit its parameters and verify the goodness of fit to ensure that the model approximates the real life situation sufficiently well.

2 The use of measurement data for management and control

A main question addressed in this thesis is: once we have measurement data, what is the right way of using it for traffic management and control? The methods we propose have a solid mathematical support, and rely on the use of (on-line) measurements.

We will now outline some situations, to be further elaborated in this thesis, in which the use of such a data analysis framework may be beneficial. They are:

- link load estimation;
- anomaly detection;
- resource allocation.
We now elaborate on these in greater detail.

A correct estimate of the link usage is one of the most important factors in capacity planning or expansion. Contrary to common beliefs, it may not be enough to perform just coarse measurements (for instance over windows of, say, 15 minutes), to estimate the traffic load from these, and then add ‘some’ safety margin to estimate capacity required [5]. Variability of the traffic (‘burstiness’) is surely a non-negligible factor and its estimation is a highly non-trivial task. Furthermore, depending on the timescale, it may be necessary to also take deterministic components into consideration, such as daily or weekly patterns. Finally, it may be that from a practical point of view the deviation from the ‘typical’ system behavior [6] is highly relevant as well. One therefore needs to determine, for example, the probability that the load imposed on the system significantly differs from the expected value. Such situations may severely degrade network performance and thus should not be disregarded despite its possibly relatively low chance of happening.

Another question faced by the communication systems administrators concerns the detection of the sudden problems in the network, preferably not by the feedback from dissatisfied customers calling the help-desk. Ideally, the notification would be based on measurements, performed in a (quasi) on-line fashion with low false-alarm and high detection ratios.

The importance of resource allocation should not be underestimated either. One of the issues it sheds light on is the question whether a given set of customers can be connected, so as to offer them a given minimum quality level. This question is critical to network operators who currently have a (logically) separated infrastructure for their premium-grade service but want to benefit from so called statistical multiplexing gains. Imagine that several customers (Fig. 1.1) share the same link, but each of them has a guaranteed portion of bandwidth which means that even if one of them is idle, none of the others can use its spare capacity. Such an approach may therefore lead to a waste of the resources as the total link utilization may be low. An alternative could be to multiplex all the traffic streams in hope that, at times that some customers send their data at the high rate, the others are more quiet, which will ultimately lead to an ‘average’ state of the whole mix, i.e., a reduced variability, and consequently some bandwidth gain (Fig. 1.2).

There are, however, certain conditions to be met for this effect to kick in. First of all, to have such a gain, we have to allow some data loss. For the overall system performance, usually it is not a problem if the packet loss fraction is of the order of, say, $10^{-6}$. As it happens relatively infrequently, depending on the type of the transmission, either the missing data will be efficiently retransmitted (e-mail, WWW) or its loss is acceptable.
(voice, video). A second condition that needs to be met to obtain a statistical multiplexing gain relates to the type of the traffic itself. Not all the data streams can be efficiently mixed together: (nearly) constant-bit-rate flows (think of some voice or video streams) hardly yield any gain. In general, gains are more significant when the variability is large.

Different types of traffic control require different measurements procedures and techniques. For example, if information on nanoseconds timescale is needed, dedicated and highly specific equipment must be used (DAG cards with GPS synchronized clocks may serve as an example). The broad and important question how traffic should be measured is, however, not a subject of this thesis. Similarly, we do not discuss here in detail how to implement the proposed algorithms in an operational network; see however, Chapter 6 for some remarks on a possible migration scenario.

In the rest of this chapter we will define in more detail the research questions addressed in this thesis and outline our contributions. To prepare this, we first provide a brief overview of the current market situation.

3 Market, demand, services

The term ‘converged network’ became one of the keywords since the mid 1990s when it was realized that running data and voice on a single network could lead to drastic cost reduction. Today, many service providers offer their customers so-called quadruple play service [7], combining high-speed Internet connectivity, telephone, television (both live and on-demand) and wireless (mobile) access. All the services can be used at the same
time without any degradation of performance. We are not going to discuss in detail the business aspects related to quadruple play (such as billing schemes or customer care costs) but rather focus on technology-related questions which were the inspiration for the research performed in this thesis.

Broadband Internet access, the first component of quadruple play, requires primarily a link of high capacity so as to accommodate large file transfers from e.g. so called peer-to-peer networks. Packet loss, unless really excessive, is not regarded as an issue here due to the retransmission capabilities of transport protocols. Similarly, neither delay nor jitter (delay variation) has significant impact. Some content-rich, interactive web pages might suffer because of these negative effects but it may be to a large extent mitigated by buffering mechanism at the user’s machine.

With telephony, however, the situation is very different. The capacity required to transfer digitized voice is relatively low, but there are two concerns. The first concern is a delay which would be regarded as negligible for file transfer, but which may, in case of voice, have severe impact on the user’s experience. The second problem is jitter: if the variability of packet delays is high, then many packets are discarded and the user experiences serious degradation of the call quality.

On-line video transfer is even more demanding: high quality video streaming requires large capacity and, due to the coding technique, it may be prone to packet loss. This is because not all the video frames carry equally important information: some are fully specified pictures while some may only be responsible for updating information carried by the other ones. Losing a frame of the first type has of course a much higher negative impact than just losing the update.

4 Methodology – an overview

As we briefly stated earlier, this thesis advocates the use of mathematical modelling in communication networking. A proper reformulation of the ‘engineering question’ into mathematical terms allows for precise analysis by using available or newly developed theory. As the underlying process that drives the analyzed system is inherently random (traffic flows, user behavior, file size distribution etc.), we can use a stochastic modelling approach in order to address the problems mentioned in the previous sections. We will now give a brief overview of the stochastic models and methods used in the following chapters.

- Queueing theory is one of the most frequently used branches of mathematics when dealing with communication networks problems, and its origin dates back
to the early 1900s with the pioneering works of A.K. Erlang [8]. A basic ‘building block’ in this approach is a system which serves arriving customers. It is assumed that their interarrival times are governed by a certain random mechanism and similarly, the time required to serve each of the customers stems from a certain distribution. Finally, a number of servers needs to be specified. The customers who are not currently being served form a queue. A commonly used notation is due to Kendall [9] which, in one of its versions, takes a three factor form A/B/C. Here A describes an arrival process, B is a service time distribution, and C is a number of servers. For example, M/M/1 means that there is only one server and both interarrival and service times are exponentially distributed, as traditionally M stands for ‘Markovian’. The model denoted as M/G/∞ describes a situation in which there is no limit on the number of servers, the interarrival times are exponentially distributed but the service time can be governed by any distribution (G stands for ‘general’ here). In Chapters 2, 3 and 4 we will deal with some questions regarding M/M/∞ and M/G/∞ models.

- Large deviations theory. While building and using statistical models, one is often interested in ‘average’ (typical) behavior of the analyzed system. However, while evaluating the performance of communication networks this may not be enough. It may be in fact highly relevant to analyze events which happen relatively infrequently but have a potentially high impact. An example could be a buffer overflow, leading to packet drops. This may then (depending on the communication protocol used) lead to retransmission of the missing data by the sources, thus further contributing to congestion. Large deviations theory is a suitable tool for analyzing the occurrence of such rare events, being essentially due to a cumulation of several ‘sub-events’ at the same time (e.g., arrival of a large number of high-volume data streams, forming a ‘burst’). We extensively use results from large deviations theory in Chapters 2, 3, 4 and 6.

- Long range dependent models. Many probabilistic theorems and statistical data analysis methods rely on the assumption that the observations under consideration are independent. Real data analysis, however, often reveals dependence in the sample. While sometimes this may be a sort of artifact due to, for example, a deterministic trend present in the data, (a strong) stochastic dependence (e.g., correlation) is also often observed in the sample. Ignoring dependence may lead to serious mistakes, like over-optimistic forecasts or too narrow confidence intervals. If the dependence is weak then the errors may be acceptable but one may also face a situation in which dependence is not only strong but occurs even between observations which are very distant in time. This is named long range dependence, and is often quantified by the so-called Hurst parameter. Such effects
(also known as ‘stochastic fractality’) were previously observed in many situations, including teletraffic and finance, with the levels of the river Nile analyzed by H.E. Hurst being the seminal example [10]. Certain aspects of long range dependent modelling are subject of our interest in Chapter 5.

- Wavelets. Fourier analysis is a well known and established tool used in many domains, including statistical data analysis. One of its drawbacks, however, is the lack of ability to analyze both spatial and temporal aspects of the sample at the same time. Wavelets and Fourier analysis share the major idea of decomposition by looking for the correlation between the analyzed sample and a given basis function. Unlike, however, sinusoidal functions, wavelets are localized in frequency and time. Furthermore, apart from the possibility to perform spatio-temporal analysis, wavelets are very well capable of representing functions with sharp spikes or edges. Additionally, wavelets have turned out to be a very useful tool in analyzing long range dependent samples, and this fact is exploited in Chapter 5.

5 Overview of research topics

In this thesis we address several research questions which arise when running modern communication networks. As indicated above, our objective is to resolve these issues relying on the interplay between data measurements and mathematical theory. In more detail, the problems we consider in this monograph are the following.

- In Chapters 2 and 3 we devise changepoint detection procedures that are capable of detecting load changes in a link used by Voice-over-IP users. The methodology used borrows elements from both queueing theory (as the number of lines occupied is modelled by $M/M/\infty$ and $M/G/\infty$ queueing models), but also from statistics (more specifically, sequential analysis which helps in developing a formal load change detection test). Our procedures are backed by simulations of the underlying models, as well as experiments with real data traces. Chapter 2 is based on [11, 12] while Chapter 3 uses material from [12, 13].

- While the procedures presented in Chapters 2 and 3 can be used when each single user consumes roughly the same amount of bandwidth, things complicate when this assumption is lifted. An important situation relates to detecting changepoints in the traffic generated by large aggregates of data users. In the early 1990s it was discovered that those aggregates often exhibit fractal properties, summarized by the ‘long range dependence parameter’ $H$. Chapter 5, based on [14], presents...
a changepoint detection procedure of the Hurst parameter in such a model; it intensively relies on the machinery of wavelets.

• As indicated earlier, having a good estimate for the system load is a crucial pre-requisite when sizing a network link. It is also important, though, to get insight into the likelihood of the momentary load attaining excessively high values. Large deviations techniques can then be used to assess these tail probabilities. In Chapter 4 we consider this issue for one of the most fundamental models in networking: the M/G/∞ model. The resulting approximations are tested by simulation. This chapter has appeared as [6].

• To offer multiple performance levels, network elements are often equipped with several queues, that are served in a hierarchical manner. For example, traffic which is prone to delay should be treated with service priority over delay-tolerant traffic. In order to get a full understanding of the related resource allocation issues, we perform in Chapter 6 an analysis of a multiple-priority queueing system. The methodological framework predominantly relied on large deviations theory. This chapter is based on [15].

In the remainder of this section we provide a more detailed description of these research questions. First a motivation is given, which directly stems from our networking context, and then an outline of the mathematical theory used to provide a solution, along with our specific contribution.

5.1 Changepoint detection for selected queueing models

Background and motivation

Consider the issue of dimensioning a link in a communication network. The crucial question we face relates to providing enough resources to assure a certain minimal performance level, but at the same time not too much capacity should be wasted. If certain assumptions are fulfilled (more specifically, each user consumes roughly the same amount of bandwidth), then one of the possible approaches is to model the link occupancy by using the celebrated Erlang loss model. This yields the distribution of the number of users present in the system in steady state; for a given load value this distribution enables us to size our link.

This explains why it is crucial to have a good estimate of the load. A first question is: is the load constant in time? Due to the dynamic nature of the networks (day/night patterns, etc.) this is probably not the case. On the other hand, however, since we
assumed a stochastic model, not every change of the system behavior is due to a change in the underlying model. The key question is therefore: can we distinguish between variations that are essentially due to stochastic fluctuations within a fixed model, and variations that indicate a change of the model?

**Mathematical description**

Let us state the aforementioned problem in detail for a specific scenario. We assume that users arrive according to a Poisson process with intensity $\lambda$, while the service times are independent and identically distributed, with mean $1/\mu$. The load $\varrho$ is then defined as the unit-less number $\varrho := \lambda/\mu$. Moreover, we assume each user requires the same amount of resources, which we normalize to 1. If there are $C$ units of resources available in total, then the probability of blocking in this model is known to be $[8]$

$$p(C \mid \varrho) := \frac{\varrho^C}{C!^\varrho} \left/ \sum_{c=0}^{C} \frac{\varrho^c}{c!} \right.$$  

In the lingo of queueing theory this model is referred to as the M/G/C/C queue (where the fourth parameter in Kendall’s notation corresponds to the maximum number of customers in the system, including those being served; if omitted, no limit is assumed). To simplify the analysis one may use a similar model in which arrival and service processes are defined as above, but the number of servers and hence the queue size is not limited (M/G/∞). Then an approximation of $p(C \mid \varrho)$ is the probability that the number of customers exceeds $C$ in this corresponding infinite-server model.

One may now calculate the minimal value of $C$ for which the blocking probability is still below some predefined number $\varepsilon$; this procedure thus facilitates link sizing, given the service criterion under consideration. To use such a procedure, it is evidently crucial to know whether the current load estimate is still valid. As mentioned above, this calls for statistically sound techniques that detect changes in the load parameter $\varrho$.

Using a statistical framework, we can re-express this in terms of a hypothesis testing problem. Our objective is then to test whether all samples correspond to load $\varrho$ (which we associate with hypothesis $H_0$), or whether there has been a changepoint within the data set, such that before the changepoint the data was in line with load $\varrho$, and after the changepoint with $\bar{\varrho}$ (which is hypothesis $H_1$).

To be able to perform such a test, an appropriate test statistic is needed and its distributional properties (under $H_0$) must be known. Specifically, to enable such tests, we have to know what the probability is, under $H_0$, that our test statistic is larger than
a certain threshold value. For the case the dynamics of the number of lines occupied follow an M/M/∞ model, we have succeeded in conducting an explicit analysis in a large-deviations regime, leading to approximations of the test statistic of interest.

Such analysis is not possible when the service times have a general distribution (and even in the exponential case, computing the test statistic is numerically rather involved). In view of this, we propose an algorithm for changepoint detection in the M/G/∞, provided that the number of customers in the system is recorded at equidistant points in time, under the additional assumption that these times are sufficiently separated to safely assume that these samples are independent. A detailed account of this procedure is provided in Chapter 2.

Chapter 3 is devoted to the experimental validations of the changepoint detection procedures of Chapter 2. Notably, the performance of the proposed algorithm was checked not only using simulations, but also extensive experimentation with real data has been carried out. Since we assumed that each user requires roughly the same amount of resources (bandwidth in our case), we have focused on tests that use the number of concurrent calls in a Voice-over-Internet-Protocol (VoIP) context. The experiments indicate the limitations of the detection procedure, but show that under quite general circumstances it performs well.

As one may expect, an immediate application of our methods to raw data is not always possible, as some of the assumptions are violated. Essentially by its very construction, it is evident that the algorithm requires to have gone through an (almost) stationary period first, in order to detect any change. Therefore the timescale at which the traffic is analyzed can be considered up to, say, one or multiple hours. Considering longer timescales, there are evidently the ‘regular’ intra-day patterns that should not play a role in the changepoint detection: one should not issue an alarm when a ‘normal’ (that is, predictable) load change is taking place. An example could be a sharp load increase after 8 AM during a working day. One could think, however, of filtering techniques that remove the regular diurnal pattern, in order to return to the stationarity setting. It is noted, though, that the device of such filtering techniques is non-trivial. We comment on these as well.

5.2 Changepoint detection under long range dependence

Background and motivation

Above we motivated our interest in changepoint detection procedures, i.e., algorithms capable of identifying points in time where we have statistical evidence that the current
model no longer applies. While Chapters 2 and 3 (from the engineering point of view) dealt with the number of calls in a VoIP system, in Chapter 5 we focus on the situation where the traffic stream under consideration results from a large aggregate of data users (think of Local Area Network traffic).

One of the most influential papers in the networking community [16] drew attention to classes of models that capture the persistent correlation exhibited in large traffic aggregates, often referred to as long range dependence.

Roughly speaking, Chapter 5 deals with models where a current state heavily depends on a large set of previous states. If the system load was high in the past it may be very likely that it will also remain significantly higher in the future. This setup is in stark contrast to the traditional assumption of samples being independent (or weakly dependent, for instance in a Markovian way); neglecting the strong dependence may lead to over-optimistic traffic management and control procedures. Therefore, not only the correct estimates of the parameters quantifying long range dependence are of great importance; equal emphasis has to be put on techniques to assess whether these estimates remain valid over time.

**Mathematical description**

A commonly used definition (but there are others!) is that a stationary stochastic process \( \{X_n\}, n \in \mathbb{N} \), is called long range dependent if its autocorrelation function decays at a rate slower than a negative exponential. In the frequency domain long range dependence appears as a \( 1/f \)-like spectrum around the origin: denoting the spectral density of \( \{X_n\} \) by \( S(f) \), we have

\[
S(f) \sim \frac{c_f}{|f|^\alpha}
\]

for frequency \( |f| \to 0 \). The parameters that describe the long range dependence are \( \alpha \) and \( c_f \). The scaling parameter \( \alpha \) is related to the ‘intensity’ of the long range dependence (a qualitative measure), and is usually expressed in terms of the Hurst parameter \( H = (1 + \alpha)/2 \). On the other hand, \( c_f \) can be interpreted as a more quantitative measure of the long range dependence.

We are interested in detecting a certain type of nonstationarity in the observed data \( \{x_n\}, n \in \{1, \ldots, N\} \), viz. a statistically significant change in the value of \( H \). A direct application of some known changepoint detection algorithms (such as those based on information theory and entropy, or those based on cumulative sums) is unfortunately not possible, due to the fact that these assume the samples to be independent. The
remedy for the Gaussian case could be to transform the original time series such that in a new representation the correlation vanishes almost completely; then the aforementioned known changepoint detection methods can be applied. It can be shown that particular wavelet transforms have this decorrelating property.

Wavelet analysis requires to calculate so-called detail coefficients, which can be regarded as a measure of correlation between an analyzed function (or discrete sample) and a given basis function — they may be compared to Fourier coefficients in standard spectral analysis. Knowledge of wavelets detail coefficients allows the estimation of the Hurst parameter $H$, and in addition a change in the value of $H$ is reflected by a change in the coefficients. As the coefficients (depending on the transform) tend to be ‘almost uncorrelated’, it is possible then to use known changepoint detection algorithms.

While such a method looks appealing, still several issues need to be addressed. An example is the choice of the wavelet transform: there are several of these, each with its own specific properties. Some have excellent abilities to precisely localize the events of interest, but this may be at the expense of producing highly correlated sequences of coefficients, which may eventually degrade the performance of the detection method significantly. We have evaluated three transforms, paying attention to the trade-off between various aspects, such as the detection ratios and the precision of the changepoint localization.

5.3 Load estimation

Background and motivation

As we pointed out in the previous section, the M/G/$\infty$ system is a cornerstone model in performance evaluation. Some of its properties have been known for a long time, such as the distribution of the number of customers in steady state. Other important measures have been studied considerably less, though. Specifically, one may be interested in analyzing the behavior of the total load imposed on the M/G/$\infty$ system over a certain time interval of given length. While its mean value is known (and equals the product of the arrival rate and mean job size), less is known about the deviations from this mean. These deviations, even if they occur very infrequently, may have severe impact on the performance, as perceived by the customers. This motivates our interest in analyzing the corresponding tail probabilities.
Mathematical description

Consider an $M/G/\infty$ queueing system, where the arrival process is Poissonian with intensity $\lambda$ and their duration is according to some generic distribution with finite mean of $1/\mu$. An infinite buffer is assumed, which means that no job is rejected due to insufficient system capacity. Let $N(t)$ denote the number of jobs in the system at time $t$ and define $A(t) := \int_0^t N(s)ds$, i.e., the load imposed on the system in time interval $[0,t]$.

A classical result for this system describes the distribution of the number of jobs in equilibrium: this is Poisson distributed, where $\rho := \lambda/\mu$ is its expected value. Moreover, we have $\mathbb{E}A(t) = \rho t$ (if the system was in stationarity at time 0). In Chapter 4 we are interested in deviations from the mean — more precisely: we are looking for an explicit function $f(\cdot)$ such that

$$f(t) \sim p_\rho(t) := \mathbb{P}(A(t) > \rho t(1+\varepsilon)),$$

for $t$ large and some $\varepsilon > 0$. Put differently, we are interested in the exact asymptotics associated with the tail distribution of $A(t)/t$. A major difficulty we face is that $A(t)$ does not have i.i.d. increments so it is not possible to directly use the well-known Bahadur-Rao theorem. However, relying on a specific decomposition of $A(t)$, we demonstrate that it can be written as the sum of i.i.d. random variables, but these increments depend on $t$. We then appropriately modify Bahadur-Rao to cover this case, which enables us to establish the desired result. Our proof relies on a change-of-measure argument, in conjunction with Berry-Esseen-type estimates. Two asymptotic regimes are considered which we refer to as large timescale ($t \to \infty$) and large load ($\rho \to \infty$). A numerical study is conducted to validate the theoretical results, and to compare various approximations.

5.4 Resource allocation

Background and motivation

An important problem in network planning relates to the allocation of resources, subject to given performance requirements. One strategy is to reserve (for a certain period) dedicated resources for any individual customer, and to not allow anyone else to use them even if there are resources left unused. Obviously, we may end up in a situation in which the total utilization of the system is low. An alternative approach would be to be less strict about the reservation policy and to allow more customers, anticipating that not all of them will maximally utilize the system at the same time. In the latter policy
reservations are essentially done for aggregates, hoping that they still yield acceptable performance for the individual customers.

Related to the latter approach (which can be regarded as some sort of statistical multiplexing), immediately several questions arise. A first question is: how many users can be allowed? And: do we always benefit and if yes, is the gain really worth it? Can we still offer multiple performance levels despite the fact that the traffic streams from all users may be mixed together? — this is highly desirable, as different applications have different performance requirements (think of delay-sensitive users versus delay-tolerant users). Perhaps a better strategy then is to keep the flows with different performance levels completely separated and use different multiplexing rules for them in order to serve their requirements accordingly? And last but not least: does the technology currently employed support the delivery of multiple performance levels?

Queueing theory offers a set of valuable tools which help address the aforementioned questions. We can estimate a maximum number of users that may be allowed under a given quality-of-service constraint. What is more, by using priority queues we can model a situation in which traffic streams of different types can be treated in different way — think of regular Internet traffic (mail, WWW,...) versus business-critical data (like stock exchange quotes).

**Mathematical description**

In Chapter 6 we consider a queueing system in which the service capacity $C$ is shared by two input streams: a high priority (hp) stream and a low priority (lp) stream. The hp input is expected to be served essentially immediately, and therefore it uses a queue with no buffer (or a very small buffer, just to absorb packet-scale fluctuations). Denoting by $C_{hp}$ the random variable describing the capacity consumed by the hp input, then $C - C_{hp}$ can be interpreted as the service rate which is available to the lp flows. If the lp input rate temporarily exceeds this available service rate, then the excess may be stored in a buffer of size $B$.

An important observation is that the interaction between the queues goes in just one direction: the lp queue is affected by the hp queue, but not the other way around. The situation of a single queue with constant service rate is well studied, so the hp queue is straightforward to analyze. Therefore, the major complication of our model relates to the lp queue: the random capacity being seen by the lp queue has to be carefully taken into account. To assess the performance of the described system, one would typically be interested in techniques to evaluate the loss (overflow) probability as a function of model parameters. A classical question would be in the spirit of admission control: how
many lp streams can be allowed given that we do not want the loss probability to be higher than some predefined number \( \varepsilon \) (given that a certain set of hp streams is active).

In Chapter 6 we identify several possible approaches to attack the problem described above. To give a flavor for these, one of them will be outlined here. Assume that the rate generated by the hp input is modelled by a Gaussian random variable with parameters \( \mu \in (0, C) \) and \( \sigma^2 \), which is then truncated as the input can be neither negative nor higher than the system capacity:

\[
C_{hp}^N := \max\{0, \min\{C, \mathcal{N}(\mu, \sigma^2)\}\}.
\]

Moreover, we abstract from the buffer \( B \), i.e., loss occurs as soon as the lp input rate exceeds \( C - C_{hp} \) (which may evidently be regarded as a conservative approach). Then suppose that the lp source transmits traffic at rate \( \beta \) during a random time \( T_{on} \), and is silent during a random time \( T_{off} \); let the sequence of on- and off-times be i.i.d. random variables, and let both sequences be mutually independent. We define the probability of an lp source transmitting traffic by \( \pi := T_{on}/(T_{on} + T_{off}) \). Supposing there are \( \ell \) lp jobs, then the aggregate rate required is \( \beta Y \), where \( Y \) has a binomial distribution with parameters \( \ell \) and \( \pi \). It is concluded that \( \ell \) jobs can be admitted if

\[
q^N := \frac{1}{\mathbb{B}^N(C)} \int_0^C \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathbb{P}(\beta Y > C - x) \, dx \leq \varepsilon,
\]

where

\[
\mathbb{B}^N(C) := \int_0^C \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \, dx.
\]

To avoid direct evaluation of this integral one may rescale the model by \( n \), i.e., replace \( C \mapsto nC, \ell \mapsto n\ell, \mu \mapsto n\mu \) and \( \sigma^2 \mapsto n\sigma^2 \). Under this scaling it is possible to use large deviation theory to derive accurate asymptotic (\( n \) large) approximations of the loss probability. Specifically, the Bahadur-Rao theorem allows us to obtain the following result for the described model:

\[
q^N(n) \sim \frac{K}{\sqrt{n}} e^{-n\Psi(C)},
\]

where the constant \( K \) is explicitly given.

This is just the ‘base model’; several more sophisticated modifications of the model are treated in Chapter 6 as well. For example, the hp traffic stream may be modelled such that \( C_{hp} \) is a (highly) correlated time series. More flexibility may also be included in the description of the lp streams: for example we may have \( J \) types (classes) of lp streams, each characterized by its own parameters. Finally, we have also considered
models that allow the lp queue to have a buffer, and that also incorporate some kind of
lp input traffic regulating mechanism (which is often referred to as dual leaky bucket).