Chapter 1

Introduction

Performance Analysis, Queuing Theory, Large Deviations. Performance analysis of communication networks is the branch of applied probability that deals with the evaluation of the level of efficiency the network achieves, and the level of (dis)satisfaction enjoyed by its users. Clearly, there is a broad variety of measures that characterize these two aspects. Focusing on the efficiency of the use of network resources, one could think of the throughput, i.e., the rate at which the network effectively works–in the case of a single network element, this could be the rate (in terms of, say, bits per second) at which traffic leaves. Another option is to use a relative measure, such as the utilization, commonly defined as the ratio of the throughput and the available service speed of the network element. Also the (dis)utility experienced by users can be expressed by a broad variety of measures. Realizing that at any network element traffic can be stored in a buffer when the input rate temporarily exceeds the available service rate, it seems justified to study performance indicators that describe the delay incurred when passing the network node. Buffers have a finite size, so there is the possibility of losing traffic, and as a result the fraction of traffic lost becomes a relevant metric.

Performance analysis is a probabilistic discipline, as the main underlying assumption is that user behavior is inherently *random*, and therefore described by a statistical model. This statistical model defines the probabilistic properties of the arrival process (or, input process) of traffic at the network. Traffic could arrive in a smooth way, but highly irregular patterns also occur; in the latter case, communication engineers call the arrival process *bursty*.

Justified by the above description of network elements as storage systems, we could model a communication network as a network of *queues*; at any node traffic arrives, is stored if it cannot be handled immediately, and is served. Performance analysis often relies heavily on results from the theory that describes the performance of these queues, i.e., *queuing theory*. A key element of performance analysis

Large deviations for Gaussian queues M. Mandjes

is the characterization of the impact of 'user parameters' on the performance offered by the network (how is the delay affected by the arrival rate? what is the impact of increased variability of the input traffic? etc.). On the other hand, one often studies the sensitivity of the performance in the system parameters (what is the impact of the buffer size on the loss probability? how does the service speed affect the mean delay? etc.)

A substantial part of the defined performance metrics relates to *rare events*. Often network engineers have the target to design the system such that the loss probability is below, say, 10^{-6}. Another common objective is that the probability that the delay is larger than some predefined excessive value is of the same order. This explains why we heavily rely on a subdomain of probability theory that exclusively focuses on the analysis of rare events: *large deviations theory*. This theory has a long history, but has been applied intensively for performance analysis purposes only during the last, say, two decades.

Traffic management, dimensioning. Once one is capable of evaluating the performance of a static situation (i.e., calculating performance metrics for a given arrival process and given network characteristics), the next step is often to choose the set of design parameters such that a certain condition is met, or such that some objective function is optimized. For instance, a requirement imposed upon the network element could be that just (on average) a fraction ϵ of the incoming traffic is lost. Evidently, when increasing the buffer size B, the loss probability decreases, and therefore it is legitimate to ask for which minimal B the loss probability is at most ϵ. Of course, there is often a cost incurred when increasing B. As a result one could imagine that one should maximize an objective function that consists of a 'utility part', minus a 'cost part', where both parts increase in B. Selecting an appropriate value for B is usually called *buffer dimensioning*; similarly the choice of a suitable service speed is referred to as *link rate dimensioning* (or, shortly, link dimensioning).

On the other hand, knowledge of the static situation enables the computation of conditions on the arrival process (both in terms of average input traffic rate and the variability of the arrival process) under which the network can offer some required performance level:

- In this way, one could develop mechanisms that decide what the maximum number of users is such that the mean delay stays within some predefined bound; such a mechanism is usually called *admission control*. To implement an admission control, one needs to be able to characterize the so-called admissible region, which is, in a situation of two classes of users, the combination of all numbers of users of both classes (n_1, n_2) for which for both classes the performance requirement is met.

- Also, insight into the static situation may tell us how to 'smooth' traffic (i.e., decrease the variability of the arrival process), such that the traffic stream becomes more 'benign', and the loss probability in some target queue can

meet some set requirement (a technique known as *traffic shaping*). Traffic shaping is usually done by inserting an additional queue between the traffic sources and the target queue that is emptied at a service rate C' that is lower than the peak rate of the original stream (but higher than the service rate C of the target queue). Then the traffic stream arriving at the second queue is smoother than the original traffic stream, and therefore easier to handle, but this is at the expense of introducing additional delay.

This traffic shaping example explains the interest in *tandem queues*, i.e., systems of queues in series (in which the output of the first queue feeds into the second queue). In such a situation one would, for instance, like to dimension the shaping rate: given a buffer B and service rate C in the target queue, how should one choose the shaping rate C' to ensure that the loss probability in the target queue is below ϵ (where it assumed that the shaper queue has a relatively big buffer).

In the literature, the set of control measures that affect the network's efficiency or the user's (dis-)satisfaction is often called *traffic management*. Clearly, dimensioning is a traffic management action that relates to a relatively long timescale: one can choose a new value for the buffer size or the link rate only at a very infrequent rate; the process of updating the resource capacities is known as the *planning cycle*. Mechanisms like admission control serve to control fluctuations of the offered traffic at a relatively short timescale: admission control is done on the timescale that new users arrive (and hence the decision to accept or reject a new user has to be done essentially in real time).

Performance differentiation. We have described above the situation in which we wished to guarantee some performance requirement that is uniform across users; for instance, all users should be offered the same maximum loss probability. In practice, however, all applications have their own specific performance requirements. Think of a voice user, who tolerates a substantial amount of loss (up to the order of a few percents, if certain codecs are used) but whose delay is critical, versus a data user, who has very stringent requirements with respect to loss, but is less demanding with respect to delay. Of course one could treat all traffic in the same fashion, e.g., by using first-in-first-out (FIFO) queues; clearly, to meet the performance requirements of all users, the requirement of the most stringent users should be satisfied. Such an approach will, however, inevitably lead to a waste of resources, and therefore one has developed queuing disciplines that actively discriminate. An example of such a scheme is the (two-class) *priority queue*, in which one class has strict priority over another class. The high-priority class does not 'see' the low-priority class, so its performance can be evaluated as in the FIFO case. The low-priority class, however, sees a fluctuating service capacity, and therefore its performance is considerably harder to analyze.

Strict priority has the intrinsic drawback of 'starvation', i.e., the low-priority class can be excluded from service for relatively long periods of time (namely, the periods in which the high-priority class uses all the bandwidth). To avoid this

starvation effect, one could guarantee the low-priority class at least some minimal service rate. This thought led to the idea of *generalized processor sharing* (GPS). In GPS, both classes have their own queue. Class i can always use a fraction ϕ_i of the total service rate C (where $\phi_1 + \phi_2 = 1$). If one of the classes does not use its full capacity, then the remaining capacity is allocated to the other class (thus making the service discipline work conserving). Note that the priority queue is a special case of GPS (choose $\phi_i = 1$ to give class i strict priority over the other class). One of the crucial engineering questions here is, for two user classes with given traffic arrival processes and performance targets, how should the weights be set?

Scope of this book. In view of the above, one could say that traffic management is all about the interrelationship between

(N) the network traffic offered (not only in terms of the average imposed load, but also in terms of its fluctuations, summarized in a certain arrival process);

(R) the amount of network resources available (link capacity, buffers, etc.);

(P) the performance level achieved.

With this interrelationship in mind, we conclude that there are three indispensable prerequisites for appropriate traffic management.

In the first place, we should have accurate traffic models at our disposal (i.e., N). Part A of this book is devoted to a class of models that has proven to be suitable in the context of communication networks: Gaussian traffic processes. An interesting feature of this class is that it is highly versatile, as it covers a broad class of correlation structures. We introduce this class and provide a number of generic properties. Then we explain why Gaussian models are likely to be an adequate statistical descriptor, and how this can be empirically verified. We also present a number of standard Gaussian models that are used throughout the book.

Secondly, we show in part B how to assess the performance of the network, for a given Gaussian traffic model, and for given amounts of available resources (i.e., (N,R) $\mapsto$ P). In other words, we analyze Gaussian queues, i.e., queues with Gaussian input. It turns out that only for a very limited subclass of inputs exact analysis is possible, and this explains why we resort to asymptotics. We present and explain several asymptotic results. Emphasis is on the so-called many-sources framework, which is an asymptotic regime in which the number of users grows large (where the traffic streams generated by these users have more or less similar statistical properties), and where the resources are scaled accordingly. Single queues are relatively easy to deal with in this framework, but we also focus on problems that are significantly harder, such as the analysis of a tandem queue, and a queue operating under GPS.

The final subject of the book is how these Gaussian queues can be used for traffic management purposes. Essentially, these problems all amount to questions of the type (N,P) $\mapsto$ R: given a traffic model and some performance target, how

much resources are needed? Specific attention will be paid to link dimensioning in the single queue, the weight setting problem in generalized processor sharing, and bandwidth trading.

Bibliographical notes

This book focuses on large deviations for Gaussian queues, with applications to communication networking. There is a vast body of related literature, which we will cite at several occasions. Here, we briefly list a number of textbooks that can be used as background.

The literature on performance analysis is vast, and the key journals include *IEEE/ACM Transactions on Networking*, *Computer Networks*, and *Performance Evaluation*. A textbook that gives an excellent survey on performance evaluation techniques is by Roberts *et al*. [253], albeit with a focus on somewhat out-of-date technologies. We also recommend the book by Kurose and Ross [167], and the classical book by Bertsekas and Gallager [32].

There are several strong textbooks on queuing theory – without attempting to provide an exhaustive list, here we mention the books by Baccelli and Brémaud [17], Cohen [52], Prabhu [246], and Robert [250]. The beautiful survey by Asmussen [13] deserves some special attention, as it gives an excellent account of the state of the art on many topics in queuing theory. The leading journal in queuing is *Queueing Systems*, but there are many nice articles scattered over several other journals (including *Advances in Applied Probability*, *Journal of Applied Probability* and *Stochastic Models*).

During the last two decades a number of books on large deviations appeared with a focus on applications in performance and networking. In this context we mention the book by Bucklew [42] as a nice introduction to large deviations and the underlying intuition. The book by Shwartz and Weiss [267] is technically considerably more demanding, but the reader's efforts pay off when working through a beautiful series of appealing examples. Interestingly, Chang [46] connects deterministic network calculus methods with large deviations techniques. The book that is perhaps most related to the present book is Ganesh, O'Connell, and Wischik [109]. Also there the emphasis is on the application of large-deviations techniques in a queuing setting, albeit without focusing on Gaussian inputs, and without applying it (explicitly) in a communication networks context.

Apart from these books, there are a number of books on large deviations, but without a focus on queuing. Ellis [91] approaches large deviations from the angle of statistical mechanics, whereas in Dupuis and Ellis [87] control-theoretic elements appear. Perhaps the most complete, rigorous introductory book is by Dembo and Zeitouni [72]. Other useful textbooks include Deuschel and Stroock [75] and den Hollander [132]. Articles on large deviations appear in a broad variety of journals; besides the Applied Probability journals mentioned above, this also includes *Stochastic Processes and their Applications* and *Annals of Applied Probability*.