Chapter 1

Introduction

1.1. Motivation

Network performance analysis, and the underlying queueing theory, was born at the beginning of the 20th Century when two Scandinavian engineers, Erlang[1] and Engset[2], independently found very close formulas for calculating the reject probability of a telephone call. Their results have since proved instrumental in dimensioning telephone networks, to find the optimal capacity given some expected demand and target call reject rates.

Nowadays, the engineering of communication networks and computer systems, which consists of both dimensioning and designing resource-sharing algorithms and traffic control schemes, relies on mathematical tools derived from the queueing theory. The objective of this book is to describe some of these tools and to show how they are used in solving the practical engineering and performance issues.

---

1 Agner Krarup Erlang, Danish engineer and mathematician (1878–1929).

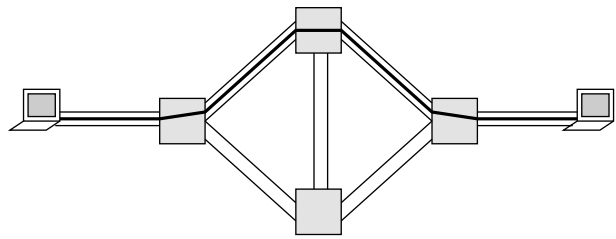
2 Tore Olaus Engset, Norwegian engineer and mathematician (1865–1943).

1.2. Networks

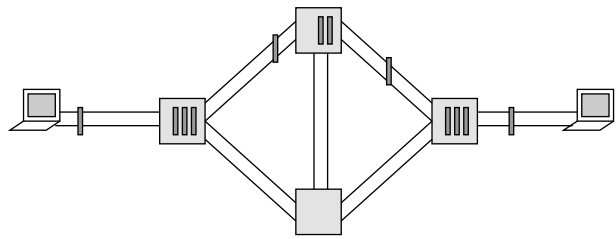
Roughly, there are two techniques for sharing the resources of communication networks:

– the "circuit" technique, which consists of reserving the resources prior to any communication and transferring information once the reservation is completed, along the established circuit;

– the "packet" technique, by which communications occur without any prior reservation, information being transferred in the form of independent packets subject to congestion (delay, loss) on their path to the destination.



(a) Circuit mode



(b) Packet mode

Figure 1.1. *Communication techniques*

In short, this is the main difference between the (public switched) telephone network and the IP network: the principle of bandwidth reservation versus that of bandwidth sharing, the questions of accessibility (call reject rate) against those of speed (bit rate) and integrity (packet delay, packet loss rate).

The boundary between circuit mode and packet mode is not so distinct in practice. The multi-protocol label switching (MPLS) technology uses virtual circuits in IP, for instance: 3G radio access networks use both circuit and packet modes; an Internet service provider can block some video streams in case of congestion, each stream then constituting a virtual circuit in the IP network. There are many such examples. However, this broad classification between the circuit mode and the packet mode is very useful. It corresponds to two types of traffic models we shall study:

– in the circuit mode, the Erlang model and its extensions, described in Chapter 8;

– in the packet mode, the IP traffic models, described in Chapter 9 for real-time traffic (voice, video) and Chapter 10 for elastic traffic (file transfers).

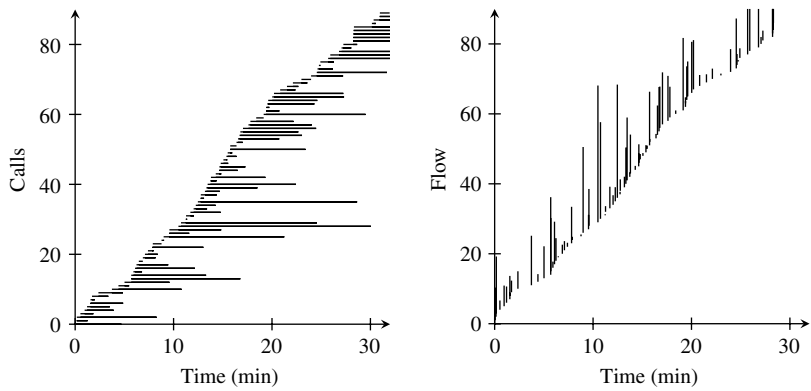
1.3. Traffic

Network performance is mainly driven by the random traffic fluctuations caused by the user behavior. To find his formula in 1917, Erlang assumed, for instance, that calls arrived according to a Poisson process[3] and had exponential durations[3]. Figure 1.2(a) shows such a sequence of calls, whose durations are represented by the lengths of the horizontal bars. These assumptions allowed Erlang to apply the novel theory of Markov[3] and derive the call reject probability with respect to the number of available circuits and the traffic intensity.

Moreover, Erlang noticed that his formula was “insensitive” to the distribution of call durations beyond the mean.

3 We will come back to these notions in detail in Chapters 2–5.

This property, which was formally proved 40 years later[4], shows the simplicity and robustness of the Erlang formula, which depends on traffic intensity only and not on fine traffic statistics like the distribution of call durations. This also explains why the formula is still used today, though today's telephone traffic has nothing to do with that of Erlang's epoch.



(a) Telephone traffic (b) Data traffic

Figure 1.2. *The random nature of traffic*

Similarly, performance of the IP networks depends on the random nature of traffic. Figure 1.2(b) shows for instance data flows arriving according to a Poisson process, with exponential sizes (volumes in bytes) represented by the lengths of the vertical bars. We shall see that, under some assumptions of the way data flows share bandwidth, most performance metrics are also insensitive to the distribution of flow sizes beyond the mean. They depend on traffic statistics

4 B.A. Sevastyanov, *An Ergodic Theorem for Markov Processes and its Application to Telephone Systems with Refusals*, 1957.

through the traffic intensity only. These results may be viewed as the natural extensions of the Erlang formula of IP networks, with the same desirable characteristics of simplicity and robustness.

1.4. Queues

Queues are omnipresent in packet-switched networks. They are at the heart of any computer, switch, router, and access point. This is the place where sharing policies are implemented through packet scheduling and active queue management. More generally, a set of data flows sharing the same link may be viewed as a virtual queue due to the link capacity constraint, the service required by each flow corresponding to the transfer of some data volume.

By extension, the models of circuit-switched networks where calls are either admitted or rejected may be viewed as specific queues, in which customers do not wait but may be lost. Formally, we should refer to either "loss" or "waiting" queues; the simpler term of *queues* is commonly used.

1.5. Structure of the book

The book is structured as follows:

Chapter 1: Introduction;

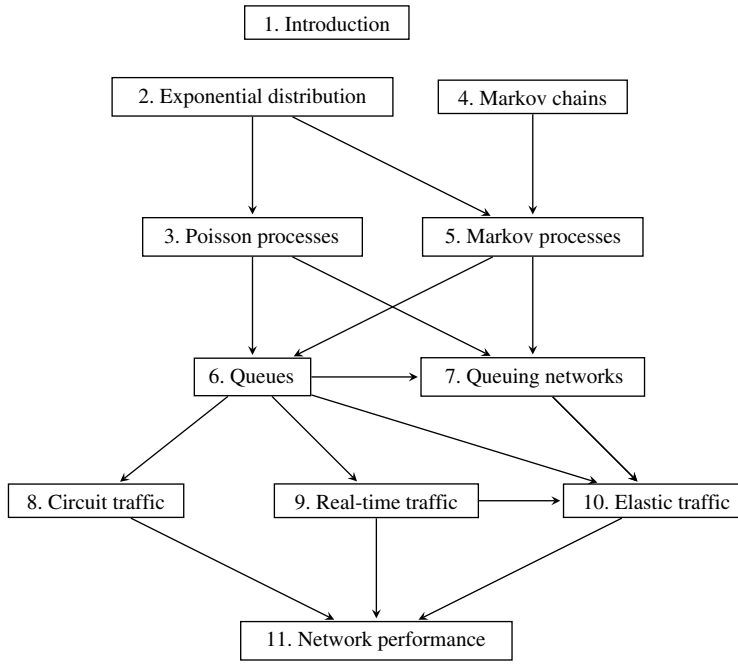
Chapters 2–5: Poisson processes and Markov theory;

Chapters 6 and 7: Elements of queueing theory;

Chapters 8–10: Traffic models;

Chapter 11: Application to networks.

The relationship between the chapters is as follows:



Each chapter (except for this one) contains a series of exercises with solutions. Throughout the book, we use the acronyms a.s. for "almost surely" and i.i.d. for "independent and identically distributed". We denote by $1(\cdot)$ the indicator function, by $\mathrm{P}(\cdot)$ the probability, and by $\mathrm{E}(\cdot)$ the expectation.

1.6. Bibliography

For further information, the interested reader is referred to the following books:

BRÉMAUD P., *Markov Chains, Gibbs Fields, Monte Carlo Simulation, and Queues*, Springer-Verlag, 1999.

KELLY F., *Reversibility and Stochastic Networks*, Wiley, 1979.

KLEINROCK L., *Queueing Systems: Volume I – Theory*, Wiley Interscience, 1975.

ROSS K.W., *Multiservice Loss Networks for Broadband Telecommunications Networks*, Springer-Verlag, 1995.

SERFOZO R., *Introduction to Stochastic Networks*, Springer-Verlag, 1999.