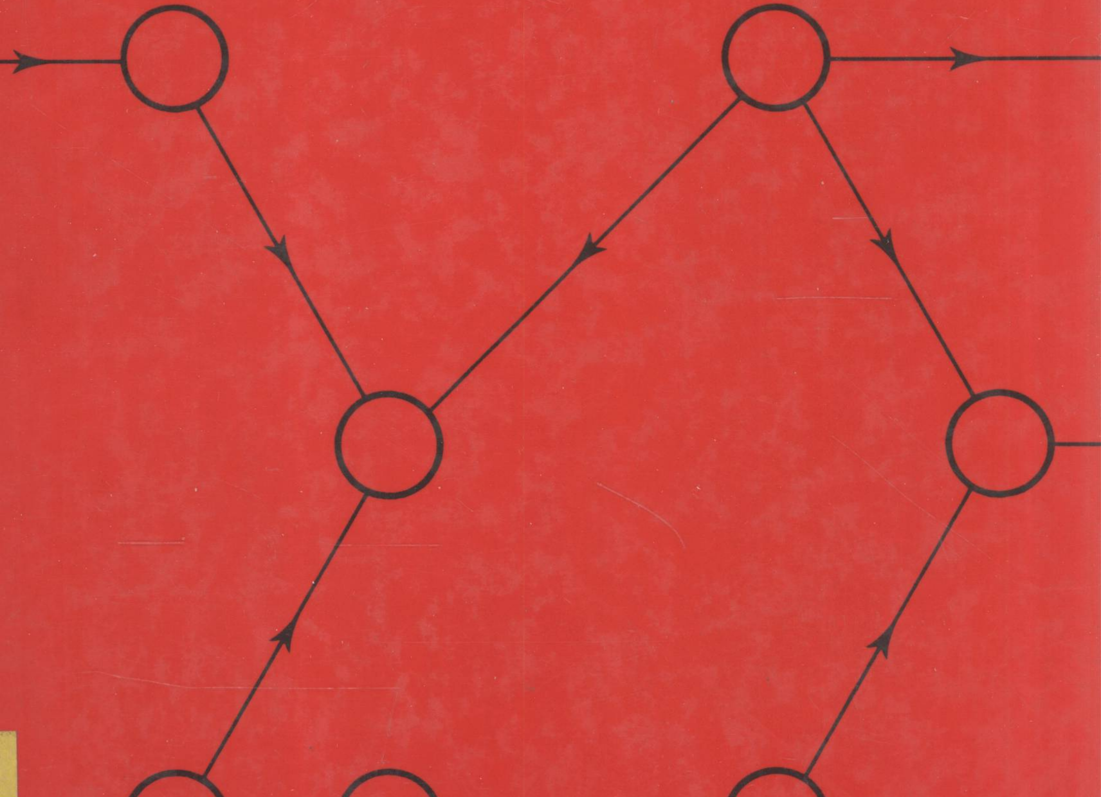
E. GELENBE • G. PUJOLLE

INTRODUCTION TO QUEUEING NETWORKS



Introduction to Queueing Networks

E. Gelenbe

Université de Paris Sud, Paris, France

and

G. Pujolle

Université Pierre et Marie Curie, Paris, France

Translated by

J. C. C. Nelson

University of Leeds, Leeds, UK

JOHN WILEY & SONS

Chichester · New York · Brisbane · Toronto · Singapore

This work is the translation of the French book
by Gelenbe and Pujolle:
Introduction aux réseaux de files d'attente
published by Editions Eyrolles, Paris

Library of Congress Cataloging-in-Publication Data:

Gelenbe, E., 1945–
Introduction to queueing networks.

Translation of: Introduction aux réseaux de files
d'attente.
Includes bibliographies and index.
1. Queueing theory. I. Pujolle, G., 1949–
II. Title.
T57.9.G4513 1987 519.8′2 86–13197

ISBN 0 471 90464 3

British Library Cataloguing in Publication Data:

Gelenbe, E.
Introduction to queueing networks.
1. Computer networks 2. Queueing
theory
I. Title II. Pujolle, G. III. Introduction
aux résaux de files d'attente. *English*
004.6 TK5105.5

ISBN 0 471 90464 3

Phototypesetting by Thomson Press (India) Limited, New Delhi
Printed and bound in Great Britain

Introduction to Queueing Networks

INTRODUCTION

Queues and queueing systems have been the subject of considerable research since the appearance of the first telephone systems. In addition to models originating in biology and genetics (branching processes), they have been the principal examples of realistic discrete state space random processes.

In the years immediately following the Second World War, the problems of operational research, that is inventory and production control, aroused a new interest in this subject area. It was rapidly discovered that models of the reliability of complex systems could well be formulated in terms of queues (arrivals of breakdowns and repair sevices). Moreover, these two aspects have given rise to an abundant literature on the optimization problems for particular queueing models.

The modelling of computer systems and data transmission systems opened the way, in the sixties, to studies of queues characterized by complex service disciplines and have created the need to analyse interconnected systems. Progress in this area has been rapid and industrial applications have been widely accepted since the seventies. At present in the computer industry, queuing network models have resulted in software packages for the automatic solution of problems arising in the design of new computers and in the evaluation and improvement of existing systems.

The methods of queueing networks have always been a basic component of the study of communication systems. The widespread introduction of computers into these systems has introduced the use, in a systematic manner, of new results on queueing networks in studies of the performance of large communication networks.

There has also been a renewal of interest by mathematicians in a subject which has developed outside the traditional mainstream of probability theory. The combination of fundamental considerations and practical problems is the best guarantee of the vitality of the subject.

This book makes no pretension to be exhaustive in an area which includes a large number of significant results and contributions. In the choice of subjects which are presented it has been necessary to make a selection which has caused us to exclude interesting theoretical aspects (such as a detailed study of a queue with one server, which is well treated elsewhere [6]).

The presentation of this book is biased towards the study of queueing networks since they lead, in our opinion, to the most interesting possibilities for application to data processing and communication systems. It is also necessary to mention the limited number of works on the subject: in the list of thirty basic books which we cite, there are only four or five which treat the subject and in a manner which is often superficial. With regard to the application examples presented in this work, we have very often chosen them in the area of data transmission networks. There, also, we have taken into consideration the fact that books which treat the performance evaluation of computer systems are more numerous.

Concerning the presentation of this book, it should be mentioned that Chapter 1 has been written with the object of giving the reader an outline of the methods used in the following chapters. We have developed, in a tutorial manner, the complementarity between the *deterministic* and *probabilistic* approaches; a large place has been reserved for the use, within the simple context of a queue with a single server, of methods arising in the theory of regenerative processes developed in [5]. A complete reading of this chapter, with the possible exception of Sections 1.6 and 1.7, is recommended even to readers interested primarily in applications.

Chapter 2 is devoted to the simplest queueing networks. A deterministic approach allows the introduction of the subject in an elementary but rigorous manner, and a substantial part is devoted to numerical algorithms which permit practical application. The probabilistic aspect is then introduced and the formal equivalence (that is to say in the sense of the equations obtained) with the deterministic case is shown.

Chapter 3 plays a less important role; here different particular cases (such as systems having a limited storage capacity) are treated to illustrate the theory. Nevertheless, we introduce the use of diffusion processes to the analysis of queueing systems.

Chapter 4 is concerned with virtually the most general queueing networks which it is known can be treated in an exact manner with solutions in 'product form'. This form of solution is very important since it permits decomposition of the joint probabilities of the states of the model into products of marginal probabilities.

This leads naturally to the examination of several different methods of solution which are presented in Chapter 5. It concerns, in particular, diffusion processes, the methods of decomposition and isolation, mean value analysis and also numerous examples of application examples.

The first five chapters are devoted almost entirely to a study of the *state* of queueing network as described by the position of customers at the service stations. A complementary approach, adopted in Chapter 6, concerns an examination of *flow* of customers through the network. This very recent and less well known approach is of a more theoretical nature and paves the way for new research directions.

Each chapter is accompanied by a bibliography containing books and articles of reference. In an appendix at the end of the book, numerous formulae suitable for practical use are proposed.

In the present chapter, we give two biographical lists, the first containing basic books, while the second deals with more specialized books or sources of additional information.

For the first list, references [1, 2, 8, 14, 17, 23, 26] are introductory works to the subject of which [1] is the easiest. The simple queue is treated in depth [6] and equally in [10, 13, 19, 21, 29, 30]. Relations with problems of operational research and reliability are treated in [15, 25, 27, 28] and relations with general methods of random processes in [3, 4, 7, 18] (see also [5]). Methods adapted to the study of computer systems (and in particular the method of decomposition) are presented in [9, 11, 12, 20, 22] and form a large part of the applications concerning data transmission networks. Results concerning systems with priorities appear in [16, 22].

Problems of evaluating the properties of computer systems by queueing networks are described in [34, 35, 39]. Models related to telephone systems appear in [32, 37, 40]. Aspects relative to the theory of point processes appear in [33] and other, more technical, aspects are treated for instance in [31, 36, 38, 41].

BIBLIOGRAPHY

1. Allen, A. O. (1978). *Probability, Statistics and Queueing Theory with Computer Science Applications,* Washington D. C.: Academic Press.
2. Beckman, P. (1968). *Introduction to Elementary Queueing Theory*, Boulder, Colo.: Golem Press.
3. Benes, V. E. (1963). *General Stochastic Processes in the Theory of Queues*, Mass: Addison, Wesley.
4. Borovkov, A. A. (1976). *Stochastic Processes in Queueing Theory*, New York: Springer-Verlag.
5. Cinlar, E. (1975). *Introduction to Stochastic Processes*, Englewood Cliffs, NJ: Prentice-Hall.
6. Cohen, J. W. (1969). *The Single Queue*, New York: American Elsevier.
7. Cohen, J. W. (1976). *On Regenerative Processes in Queueing Theory*, New York: Springer-Verlag.
8. Cooper, R. B. (1972). *Introduction to Queueing Theory*, New York: Macmillan.
9. Courtois, P. J. (1977). *Decomposability Queueing and Computer System Applications*, New York: Academic Press.
10. Cox, D. R., and Smith, W. L. (1961). *Queues*, London: Methuen.
11. Gelenbe, E., Labetoulle, J., Marie, R., Metivier, M., Pujolle, G., and Stewart, W. (1980). *Réseaux de files d'attente*, Paris: Éditions Hommes et Techniques.
12. Gelenbe E., and Mitrani I. (1980). *Analysis and Synthesis of Computer Systems*, London: Academic Press.
13. Genedenko, B. V., and Kovalenko, I. N. (1968). *Introduction to Queueing Theory, Israel*, Program for Scientific Translations, Jerusalem, Israel.
14. Gross D., and Harris C. M. (1974). *Fundamentals of Queueing Theory*, New York: Wiley.
15. Hillier, F. S., and Lieberman, G. J. (1967). *Introduction to Operations Research*, San Francisco: Holden-Day.
16. Jaiswal, N. (1968). *Priority Queues*, New York: Academic Press.
17. Kaufmann, A. (1972). *Méthodes et modèles de la recherche opérationnelle tome 1*, Paris: Dunod.
18. Kelly, F. P. (1979). *Reversibility and Stochastic Networks*, John Wiley.
19. Khintchine, A. Y. (1960). *Mathematical Methods in the Theory of Queueing*, London: Griffin.
20. Kleinrock, L. (1964). *Communication Nets*, McGraw-Hill, New York.
21. Kleinrock, L. (1975). *Queueing Systems, – Volume I: Theory*, John Wiley.
22. Kleinrock, L. (1976). *Queueing Systems, – Volume II: Computer Applications*, John Wiley.
23. Kobayashi, H. (1978). *Modeling and Analysis*, Addison Wesley.
24. Le Gall, P. (1962). *Les systèmes avec ou sans attente et les processus stochastiques, Tome I*, – Paris, France: Dunod.

25. Morse, P. M. (1958). *Queues, Inventories and Maintenance*, John Wiley.
26. Newell, G. F. (1972). *Applications of Queueing Theory*, London: Chapman and Hall.
27. Prabhu, N. U. (1965). *Queues and Inventories*, New York: Wiley.
28. Riordan, J. (1962). *Stochastic Service Systems*, New York: Wiley.
29. Saaty, T. L. (1961). *Elements of Queueing Theory with Applications*, New York: McGraw-Hill.
30. Takacs, L. (1962). *Introduction to the Theory of Queues*, Oxford University Press.
31. Bagchi, T. P., and Templeton, J. G. C. (1972). *Numerical Methods in Markov Chains and Bulk Queues*, Berlin: Springer-Verlag.
32. Benes, V. E. (1965). *Teletraffic Queueing Problem*, Mass.: Addison, Wesley.
33. Bremaud, P. (1981). *Dynamical Point Processes and Ito Systems in Communications and Queueing*, Berlin: Springer-Verlag.
34. Descloux, A. (1962). *Delay Tables for Finite- and Infinite-Source Systems*, New York: McGraw-Hill.
35. Ferrari, D. (1978). *Computer Systems Performance Evaluation*, Prentice-Hall, Engelwood Cliffs, N.J.
36. Ghosal, A. (1970). *Some Aspects of Queueing and Storage Systems*, Berlin: Springer-Verlag.
37. Haight, F. (1963). *Mathematical Theory of Traffic Flow*, New York: Academic Press.
38. Kaufmann, A. (1972). *Méthodes et modèles de la recherche opérationnelle – Tome I*, Paris: Dunod.
39. Svobodova, L. (1976). *Computer Performance Measurement and Evaluation Methods: Analysis and Applications*, New York: Elservier.
40. Syski, R. (1960). *Introduction to Congestion Theory in Telephone System*, London: Olivier and Boyd.
41. Teghem, J., Loris-Teghem J., and Lambotte, J. P. (1969). *Modèles d'attente M/M/1 et G1/M1 à arrivées et services en groups.* Berlin: Springer-Verlag.

Contents

CHAPTER 1

Queues with a Single Server

1.1 - INTRODUCTION

Most problems involving modelling of computer systems or data transmission networks deal with systems having multiple resources (central processing units, channels, memories, communication circuits, etc.) to be taken into account. This complex structure leads to the study of queueing networks, rather than simple queues with a single server.

Nevertheless, the study of a single queue is interesting for several reasons. On the one hand, understanding queueing phenomena is easier in the context of the simplest model. On the other hand, the simple queue is a useful framework for the development of the mathematical tools used in queues. Finally, 'seen from the outside' the entire computer or transmission system can be regarded as a unique server with a queue having a complex service discipline.

It is this last point of view which will be adopted in Section 1.2 where we shall examine, in a *deterministic* manner, a queue with a single server without specifying the service discipline (or the order in which service is allocated to customers). This deterministic approach will be used again in Section 1.4 where the links between deterministic and probabilistic results will be established from the properties of regenerative random processes. In Section 1.2 a general formula is obtained which computes the proportion of time spent by the queue in a given state, with very weak assumptions. Section 1.3 is devoted to the elementary 'M/M/1' queue—exponential distributions, Poisson processes, the Chapman–Kolmogorov equation and its stationary solution; here results similar to those of the preceding section are found again.

Section 1.4 starts with a proof of Little's formula for the deterministic case, as well as the case of a regenerative process, which establishes a simple and very useful link between the mean rate of arrival, the mean response time and the mean number in the queue. Before introducing Kendall's now classic notation for describing the properties of a simple queue, simple but important characteristics such as the distribution of the length of the queue at arrival and departure times and the probability that the queue is empty are obtained. These results depend very little on the precise properties of the arrival and service processes.

Sections 1.5 and 1.6 show the classic results concerning queues with Poisson arrivals and a general service distribution or, conversely, a general arrival distributions but exponential services. The computation methods used make use of the properties of Markov renewal processes. Finally, in Section 1.7, certain elementary results are established for a queue with a single server where the arrival and service distributions are general (a 'GI/GI/1' system).

1.2 - DETERMINISTIC APPROACH TO A QUEUE WITH A SINGLE SERVER

In this section, we are interested in a very simple system and a deterministic analysis of its behaviour. An analogy will also be established between certain properties of its deterministic behaviour and those of its random behaviour.

A series of 'customers' carrying successive numbers $1, 2, 3, \ldots$ etc., arrive at instants $a_1 < a_2 < a_3 < \cdots$ at the queue of Figure 1.1. The server deals with them in some order; the customers line up in the queue and it is always the one at the head who is being served.

Let N(t) be the number of customers waiting in the queue plus the one who is being served at time t and consider a time interval $[a, b]$ such that N(a) = N(b) = 0 (see Figure 1.2).

For this time interval let T(n) be the time spent by the queue in the state N(t) = n;

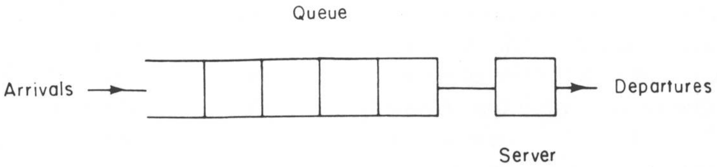


Figure 1.1 Simple queue

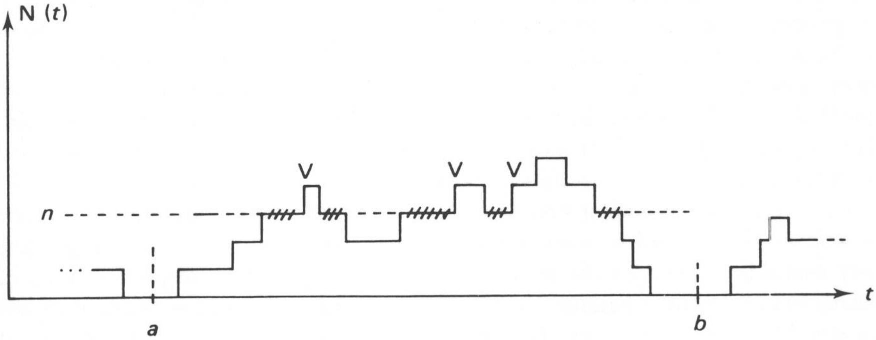


Figure 1.2 Behaviour of the length of a queue

that is the hatched regions in Figure 1.2. $\alpha(n)$ is the number of arrivals of customers when the length of the queue is n (indicated by V) and $\beta(n)$ is the number of departures when the state is n, still in the same interval $[a, b]$. Since $\mathrm{N}(a) = \mathrm{N}(b) = 0$, it is necessary that:

$$\alpha(n) = \beta(n+1).$$

Let $\mathrm{T} = b - a$ be the duration of this interval. Hence:

$$\frac{\alpha(n)}{\mathrm{T}} = \frac{\beta(n+1)}{\mathrm{T}}, \quad n = 0, 1, 2, \ldots.$$

Let $p(n)$ be the proportion of time spent in state n, in the interval in question:

$$p(n) = \frac{\mathrm{T}(n)}{\mathrm{T}}.$$

Hence:

$$p(n)\frac{\alpha(n)}{\mathrm{T}(n)} = p(n+1)\frac{\beta(n+1)}{(n+1)}.$$

To simplify these formulae, let $\lambda(n)$ be the number of arrivals in unit time when the queue is of length n in the interval $[a, b]$:

$$\lambda(n) = \alpha(n)/\mathrm{T}(n)$$

and $\mu(n)$ the number of departures in unit time:

$$\mu(n) = \beta(n)/\mathrm{T}(n).$$

Hence we have the recurrence:

$$p(n+1) = [\lambda(n)/\mu(n+1)]p(n), \quad n = 0, 1, \ldots$$

of which the solution is:

$$p(n) = p(0)\prod_{i=1}^{n}\frac{\lambda(i-1)}{\mu(i)}, \quad n = 1, 2, \ldots \tag{1.2.1}$$

Therefore the proportions of time spent in each state must satisfy (1.2.1) and this formula will be applicable to all time intervals which start and finish with an 'empty' state of the queue. $p(0)$ can easily be determined since we must obtain:

$$\sum_{n=0}^{\infty}\frac{\mathrm{T}(n)}{\mathrm{T}} = \sum_{n=0}^{\infty}p(n) = 1$$

which gives:

$$p(0) = \left[1 + \sum_{n=1}^{\infty}\prod_{i=1}^{n}(\lambda(i-1)/\mu(i))\right]^{-1}. \tag{1.2.2}$$

It must be emphasized that the quantities $\lambda(n)$ and $\mu(n)$ are *measurable* experimentally; but formula (1.2.1) does not *predict* the queue's behaviour outside the precise interval which starts at time a and finishes at time b.

Nevertheless, property (1.2.1) occurs in certain probabilistic models which we shall examine subsequently and which are used to *predict* behaviours in known statistical conditions.

1.3 - THE EXPONENTIAL DISTRIBUTION AND QUEUES WITH A SINGLE SERVER

When it is desirable to *predict* the performance of a system outside an observation period during which all the data concerned are accessible, it becomes necessary to make certain assumptions concerning its behaviour. The probabilistic assumptions which we shall make in this section and in those which follow lead to predictions concerning the systems which we shall analyse. The link between these probabilistic assumptions and measurements on actual systems is established with the aid of statistics.

The results of Section 1.2 concern a particular interval $[a, b]$ and a deterministic behaviour. The results which we obtain in Section 1.3 concern an infinite period of time and a (probable) set of realizations.

In this section, we assume that the durations $\mathrm{I}_1 = a_1 - 0$, $\mathrm{I}_2 = a_2 - a_1, \ldots,$ are random variables, that is they consist of quantities whose exact values are not known but for which a probability distribution can be defined. The variables $\mathrm{S}_1, \mathrm{S}_2, \mathrm{S}_3, \ldots$ represent successive service times and are also assumed to be random.

Also, we assume that the time intervals between successive arrivals, or *interarrivals*, $\mathrm{I}_1, \mathrm{I}_2, \mathrm{I}_3, \ldots$ are all distributed according to the exponential distribution:

$$\mathrm{P}\{\mathrm{I}_j < x\} = 1 - e^{-\lambda x}, \quad \text{for all } j \geqslant 1,$$

the same assumption is made for the service times:

$$\mathrm{P}\{\mathrm{S}_i < x\} = 1 - e^{-\mu x}, \quad i \geqslant 1,$$

where λ and μ (the parameters of the distributions) are real positive and finite. On the other hand we suppose that for $i \neq j$:

$$\mathrm{P}\{\mathrm{I}_i < x \quad \text{and} \quad \mathrm{I}_j < y\} = \mathrm{P}\{\mathrm{I}_i < x\}\mathrm{P}\{\mathrm{I}_j < y\}$$

that is to say the interarrivals are *independent*. We also assume that the service times are independent of each other and that they are independent of the interarrivals.

1.3.1 'Memoryless' property of the exponential distribution

The exponential distribution has one particularly interesting property which is one of the factors explaining its popularity. Suppose that we are dealing with a time bomb which explodes automatically after a time X distributed according to an exponential distribution.

$$\mathrm{P}\{\mathrm{X} < x\} = 1 - e^{-\lambda x}, \quad \infty > \lambda > 0.$$

We trigger the mechanism at time $t = 0$ to cause an explosion at time $t = X$. At an intermediate time $t = y$, before the explosion occurs, we would like to know the time remaining before the explosion!

This simply means that we wish to know the distribution of $X - y$ knowing that $X > y$ since the explosion has not occurred at time $t = y$. We calculate:

$$\begin{aligned} P\{X - y < x | X > y\} &= P\{y < X < y + x\}/P\{X > y\}, \\ &= \frac{1 - e^{-\lambda(y+x)} - (1 - e^{-\lambda y})}{e^{-\lambda y}}, \\ &= 1 - e^{-\lambda x} = P\{X < x\} \qquad (1.3.1) \end{aligned}$$

and we discover that the fact that the explosion has not occurred up to time t allows us to establish simply that $X - y$ has the same distribution as X. This is called the *Markovian* or '*memoryless*' property of the exponential distribution.

In fact it can be proved that if a continuous positive random variable has the property (1.3.1), then its distribution is exponential.

1.3.2 Analysis of a queue with exponential interarrivals and services

The queue being studied has been defined as a system in which the times between successive arrivals and the service times are mutually independent random variables distributed according to the exponential distribution. We approach its analysis in a manner analogous to that adopted in Section 1.2 and we let $N(t)$ be the instantaneous number of customers in the queue, including the one who is being served.

Consider the existence of distinct times a and b at which the queue is empty ($N(a) = N(b) = 0$), but chosen such that $N(t) > 0$ for $t = a^+$ and $t = b^+$: that is, at times a^+ and b^+, an arrival occurs. We require also that the queue becomes empty once between a and b. This leads to the arrangement presented in Figure 1.3.

Let $\pi_{i,j}$ be the probability of passing from state i to state j after an arrival at or departure from the queue.

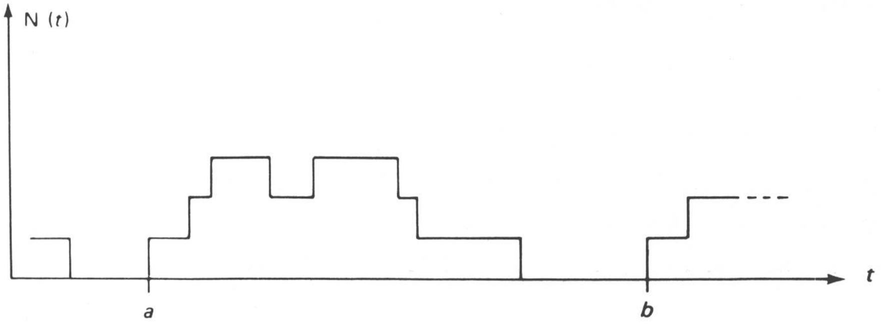


Figure 1.3 Example of the evolution of the number of customers in the queue