NORTH-HOLLAND SERIES IN

APPLIED MATHEMATICS AND MECHANICS

THE SINGLE SERVER QUEUE

BY

J. W. COHEN

THE SINGLE SERVER QUEUE

BY

J. W. COHEN

Professor of Operational Analysis
Mathematical Institute
University of Utrecht, Utrecht

Revised edition

1982

NORTH-HOLLAND PUBLISHING COMPANY — AMSTERDAM · NEW YORK · OXFORD

First printing 1969
Revised edition 1982

PUBLISHERS:
NORTH-HOLLAND PUBLISHING COMPANY
AMSTERDAM · NEW YORK · OXFORD

SOLE DISTRIBUTORS FOR THE U.S.A. AND CANADA:
ELSEVIER NORTH-HOLLAND, INC.
52 VANDERBILT AVENUE, NEW YORK, N.Y. 10017

Library of Congress Cataloging in Publication Data

Cohen, Jacob Willem.
The single server queue.

(North-Holland series in applied mathematics and mechanics ; v. 8)
Bibliography: p.
Including indexes.
1. Queuing theory. I. Title. II. Series.
T57.9.C63 1981 519.8'2 81-14032
ISBN 0-444-85452-5 AACR2

PRINTED IN THE NETHERLANDS

To Annette B.

EDITORIAL NOTE

The enormous increase in the amount of research information, published in an ever-growing number of scientific journals, has stimulated the demand for authoritative monographs on well-defined topics. Such monographs have become virtually indispensable to young research workers and students working in a particular field, who are either bewildered by the vast quantity of literature in existence, or are in danger of duplicating research that has already been published, but is not readily accessible. Specialists also may benefit from the availability of competent surveys by fellow experts in their own field.

The North-Holland Series in Applied Mathematics and Mechanics is intended to help meet this demand. The editors believe that a continuing close relationship between applied mathematics and mechanics, having proved so fruitful in the past, will continue to benefit both subjects in the future. The series will include original monographs as well as translations of outstanding works which would otherwise have remained inaccessible to many readers.

PREFACE

Queueing theory is one of the most important branches of modern probability theory applied in technology and management. As far as the future development of technology and management may be extrapolated from the past and present state of affairs the need for a deeper insight into queueing theory will increase rapidly. It is hardly necessary to point out the many actual queueing situations encountered in every-day life. Production lines, the theory of scheduling and transportation (both surface and air traffiic), the design of automatic equipment such as telephone and telegraph exchanges, and particularly the rapidly growing field of information handling and data processing are but a few fields in which queueing situations are encountered. To characterize: it arises in every situation, in which a facility for common use is provided, where waiting and queueing may and usually do arise. The organisation and the performance of the facility on the one hand and the behaviour of the users on the other determine the queueing system.

Although the broad field of applications amply justifies an intensive study of queueing models, it turned out that these models are also of great use outside the field of queueing theory, e.g. in inventory and maintenance theory. Moreover, the analytical problems encountered in the study of queueing models are often very interesting from a mathematical point of view and, consequently, obtained much attention from probability theorists.

As with so many branches of applied mathematics the task of queueing theory consists of the classification of the more fundamental models, of the design of analytical methods for the study of these models and of the mathematical analysis of those quantities which describe the essential features and properties of the model. On the basis of this knowledge, eventually supplemented by experimental studies usually performed by simulation techniques, the designer of systems with queueing situations should obtain the information and feeling needed to predict the behaviour of the actual queueing situation. The step from the model and its mathematical description to the actual situation is usually the most difficult one, the more so if the actual situation is too complicated to allow a fruitful theoretical or experimental investigation of its model. A sound knowledge of the fundamental models and their properties is often the best guide in making this step.

The present book concentrates on the most basic model of queueing theory, i.e. the single server model. Its aim is two-fold. Firstly, a description of those mathematical techniques which have been proved to be the most fruitful for the investigation of queueing models, and secondly, an extensive analysis of the single server queue and its most important variants. Even within this limited range restrictions had to be made, but the author's goal will be reached if the reader acquires an understanding of the models, purposes and methods of the theory of queues.

The book is divided into three parts. The first part deals with those topics of the theory of stochastic processes which have been successfully used in queueing theory. Part II is devoted to the simplest single server model. A number of analytical techniques for this model is discussed. The most powerful one is that based on Pollaczek's approach, which is closely related to the Wiener-Hopf technique, combined with renewal theory. This technique is often very intricate and in simple cases it is possible to use a much more elegant argument, which is often shorter. However, it is the author's conviction that the primary duty of applied mathematics is the development of sharp and powerful tools leading to useful results. For this reason in the derivations elegance is often sacrificed to utility. In part III several of the more important variants of the single server model are discussed. It concludes with a chapter on asymptotic relations and limit theorems.

The book is intended for applied mathematicians concerned with system design or interested in applied probability theory. A sound knowledge of Volume I of Feller's "An Introduction to Probability Theory and its Applications" is needed, while some knowledge of advanced probability theory, e.g. subjects discussed in Feller's Volume II, is desirable; the Laplace-Stieltjes transform and the theory of functions are tools extensively used in the text. A short review of literature is added to each chapter. Completeness of these reviews is not claimed.

Relations which are not separated by text have only one reference number placed at the end of the first one; a reference to relation (5.24) in a section of part I refers to relation 5.24 of that part, whereas references to relations outside a part are prefixed by a roman numeral indicating the part, so (I.5.24) means relation (5.24) of chapter 5 of part I. The name of an author followed by a date refers to the list of references. Much trouble has been taken to give every symbol a unique meaning. Occasionally, deviations from this ideal could not be avoided, but sufficient provisions have been made to avoid confusion. The symbol '$\overset{\text{def}}{=}$' stands for the defining equality sign. All symbols indicating stochastic variables are printed in bold type.

Thanks are due to a number of colleagues, collaborators and students for reading chapters and suggesting improvements, in particular to Mr. S. J. de Lange, Mr. P. B. M. Roes and Mr. J. H. A. de Smit. I express my gratitude to Mrs. N. Zuidervaart and Miss M. Berenschot for their efficient typing of the manuscript and to Mr. B. Broere for his help with drawing the figures. Special thanks are also due to my friend Richard Syski. The stimulating discussions and correspondence I had with him about the writing of the book were a real contribution to its completion.

The Hague, 1968 J. W. COHEN

PREFACE TO THE SECOND EDITION

Several new and important developments have taken place during the ten years after the appearance of the first edition. To provide the interested reader with some information concerning these developments a survey chapter has been incorporated; it may hopefully serve as a guide. A new section on some topics of the theory of regenerative processes has been added; the relevant theorems have appeared to be extremely useful for the general analysis of stationary processes.

A number of changes stem from remarks made by helpful referees, colleagues and students. The author is grateful to them, and also to Miss M. Visser for her careful typing of the revised parts. He is especially indebted to Dr. O. J. Boxma for his help and constructive criticism in preparing this second edition.

Doubts about the value of Queueing Theory and its applicability have been expressed occasionally; a closer look at the relevant engineering literature quickly leads to the conclusion that such doubts are not based on facts.

The Hague, September 1979 J. W. Cohen

CONTENTS

PART I. STOCHASTIC PROCESSES

PART II. THE SINGLE SERVER QUEUE

Part III. Some Variants of the Single Server Queue

Part IV. Some Recent Developments

THE SINGLE SERVER QUEUE

BY

J. W. COHEN

Professor of Operational Analysis
Mathematical Institute
University of Utrecht, Utrecht

Revised edition

1982

NORTH-HOLLAND PUBLISHING COMPANY — AMSTERDAM · NEW YORK · OXFORD

PART I

STOCHASTIC PROCESSES

I.1. INTRODUCTION

The theory of stochastic processes is a rather recently developed branch of probability theory. One of the first processes investigated was that of the Brownian motion; this process has been extensively studied and from its investigation fundamental contributions to probability theory have originated. The first important results concerning this process date back to the beginning of this century. At the same time telephone engineers were confronted with a type of stochastic process, today called a birth and death process, which turned out to be of fundamental importance for designing telephone exchanges. The theory of stochastic processes originating from needs in physics and technology is at present a rather well developed theory; a large number of basic processes have been classified and the most important properties of these processes are known (cf. DOOB [1953], LOÈVE [1960], BLANC-LAPIERRE et FORTET [1953], PARZEN [1962]).

The concept of "stochastic variable" or "random variable" is fundamental in modern probability theory and its applications; we assume that the reader is acquainted with it. Let t denote a parameter assuming values in a set T, and let x_t represent a stochastic variable for every $t \in T$. We thus obtain a family $\{x_t, t \in T\}$ of stochastic variables. Such a family will be called a stochastic process if the parameter t stands for time, and from now on t will be interpreted as such. The elements of T are hence time points, and T will be a linear set, denumerable or non-denumerable. For queueing theory the most important cases are that T is the interval $[0, \infty)$ or that $T = \{t: t = t_n, n = 0, 1, 2, \ldots\}$ with $t_0 = 0$, and particularly with $t_n = ne$, where e is the time unit. If $T = \{t: t \in [0, \infty)\}$ then the stochastic process $\{x_t, t \in T\}$ is said to be a process with continuous time parameter, whereas in the second case the process is said to have a discrete time parameter.

For arbitrary $t \in T$ the set of all possible realisations of the stochastic variable $\boldsymbol{x}_t$ is the sample space or state space of $\boldsymbol{x}_t$. From now on it will be assumed that all $\boldsymbol{x}_t, t \in T$ have the same state space. The state space may be a denumerable or non-denumerable set. For instance $\boldsymbol{x}_t$ may represent the number of customers waiting at a service station at time t; or $\boldsymbol{x}_t$ represents, at time t, the total time a server has been busy since $t=0$. However, it may also happen that the state space of the process is a vector space. For instance $\boldsymbol{x}_t$ is the vector variable $(\boldsymbol{\xi}_t, \boldsymbol{\tau}_t)$, where at time t the variable $\boldsymbol{\xi}_t$ stands for the number of customers waiting, and $\boldsymbol{\tau}_t$ for the length of time between t and the arrival of the next customer.

What information should be known to describe a stochastic process $\{\boldsymbol{x}_t, t \in T\}$ completely? We shall not study this question here but refer the reader to the existing literature on this subject (cf. e.g. DOOB [1953]). For our purpose it is (in general) sufficient to know the n-dimensional joint distribution of the variables $\boldsymbol{x}_{t_1}, \boldsymbol{x}_{t_2}, \ldots, \boldsymbol{x}_{t_n}$,

$$F_{t_1, \ldots, t_n}(x_1, \ldots, x_n),$$

for every finite positive integer n, for every point set $(t_1, t_2, \ldots, t_n)$ belonging to T and for all $x_1, x_2, \ldots, x_n$ belonging to the state space of the process. When investigating a stochastic process our aim will be to find these n-dimensional joint distributions. It should be noticed that the functions $F_{\ldots}(., \ldots, .)$ just mentioned cannot be completely arbitrary multidimensional distribution functions. As has been shown by Kolmogorov, these distribution functions have to satisfy two consistency conditions, viz. (*i*) if $m_1, \ldots, m_n$ is a permutation of $1, 2, \ldots, n$ then

$$F_{t_1, \ldots, t_n}(x_1, \ldots, x_n) = F_{t_{m_1}, \ldots, t_{m_n}}(x_{m_1}, \ldots, x_{m_n}),$$

and (*ii*) for $m = 1, 2, \ldots, n-1$,

$$F_{t_1, \ldots, t_m}(x_1, \ldots, x_m) = \lim_{x_{m+1} \to \infty} \ldots \lim_{x_n \to \infty} F_{t_1, \ldots, t_n}(x_1, \ldots, x_n).$$

That there are questions concerning a stochastic process which cannot be answered from the mere knowledge of all finite dimensional distributions may be illustrated by the following example. For the stochastic process $\{\boldsymbol{x}_t, \ t \in [0, \infty)\}$ with state space the real line the event

$$\max_{0<t<\tau} \boldsymbol{x}_t < \alpha,$$

where α and τ are given numbers, is an event involving more than a finite, even more than a denumerably infinite, number of stochastic variables. The