MATHEMATICS AND ITS APPLICATIONS
series editor: G.M.Bell, Professor of Mathematics,
Chelsea College, University of London

Lecture Notes on

QUEUEING SYSTEMS

BRIAN CONOLLY
Professor of Mathematics [Operational Research]
Chelsea College, University of London

ELLIS HORWOOD · PUBLISHER · CHICHESTER

Lecture Notes on

QUEUEING SYSTEMS

BRIAN CONOLLY

Professor of Mathematics (Operational Research)
Chelsea College, University of London

ELLIS HORWOOD LIMITED
Publisher · Chichester

Halsted · a division of
JOHN WILEY & SONS
New York · London · Sydney · Toronto

The publisher's colophon is reproduced from
James Gillison's drawing of
The ancient Market Cross, Chichester

First published in 1975 by

ELLIS HORWOOD LIMITED
Coll House, Westergate, Chichester, Sussex, England

DISTRIBUTORS
Australia, New Zealand, South-east Asia:
JOHN WILEY & SONS AUSTRALASIA PTY LIMITED
110 Alexander Street, Crow's Nest, N.S.W., Australia

Europe, Africa:
JOHN WILEY & SONS LIMITED
Baffins Lane, Chichester, Sussex, England

Canada:
JOHN WILEY & SONS CANADA LIMITED
22 Worcester Road, Rexdale, Ontario, Canada

N. & S. America and the rest of the World:
Halsted Press a division of
JOHN WILEY & SONS INC.
605 Third Avenue, New York, N.Y. 10016, U.S.A.

Library of Congress Catalog No. 75 7,788
ISBN (Ellis Horwood) 85312 0358
ISBN (Halstead Press) 0470 168579

Printed in Great Britain

MATHEMATICS & ITS APPLICATIONS

Series Editor: Professor G.M. Bell,
Chelsea College, University of London

Mathematics and its applications are now awe-inspiring in their scope, variety and depth. Not only is there rapid growth in pure mathematics and its applications to the traditional fields of the physical sciences, engineering and statistics, but new fields of application are emerging in biology, ecology and social organisation. The user of mathematics must assimilate subtle new techniques and also learn to handle the great power of the computer efficiently and economically.

The need of clear, concise and authoritative texts is thus greater than ever and our series will endeavour to supply this need. It aims to be comprehensive and yet flexible. Works surveying recent research will introduce new areas and up-to-date mathematical methods. Undergraduate texts on established topics will stimulate student interest by including applications relevant at the present day. The series will also include selected volumes of lecture notes which will enable certain important topics to be presented earlier than would otherwise be possible.

In all these ways it is hoped to render a valuable service to those who learn, develop and use mathematics.

The Foundation Programme includes:

MODERN APPLIED MATHEMATICS
Dr. D.N. Burghes, University of Newcastle-upon-Tyne
Dr. A.M. Downs, University of Sheffield

VECTOR & TENSOR METHODS
Frank Chorlton, University of Aston, Birmingham

Lecture notes on
QUEUEING SYSTEMS
Professor Brian Conolly, Chelsea College, University of London

AUTHOR'S PREFACE

These lecture notes on congestion theory stem from a thirty hour course given by the author at the Virginia Polytechnic Institute and State University during the Summer of 1969. The course was part of an advanced level seminar supported by the U.S. National Science Foundation and addressed to postgraduate students of mathematics and statistics with no prior knowledge of the subject. It was an objective of the course to expose the student to some research.

Although mostly theoretical in content the frank objective of the course is a practical one - the analysis and evaluation of methods available for the reduction of congestion arising when demands for service overwhelm the capacity to satisfy it. For this reason the original course was advertised under the title Congestion Theory, rather than Queueing Theory which is possibly more familiar.

In this version the basic stochastic processes of the theory are introduced and analysed in the context of the single server system M/M/1 which, from the practical point of view, may be regarded as operationally the least satisfactory. The results thus provide a standard against which the improvements offered by such devices as multiplication of servers, partial determinism, adaptive service or demand, can be measured.

It is an objective to provide limited, yet adequate, mathematical tools for the analysis of most systems and their ramifications. The choice presented (almost entirely differential and integro-difference equations, use of the Laplace transformation when advantageous, elementary notions from complex analysis) may be criticized but it is simple and versatile.

With the wealth of material available it could have been a problem to decide what to omit. The choice finally made can also be criticized, but the guiding principle expounded in the second paragraph tends to resolve many of the difficulties. Here attention is drawn, without apology, to the allocation of more space than is usual to systems with infinite service capacity. They are felt to be important and neglected, and the results have relevance to the problems addressed. The final chapter devoted to adaptive systems reflects the research interest of the writer at the time of giving the lectures, an interest not yet exhausted.

These are truly lecture notes, intended both as *aide mémoire* and *vade mecum*, generally speaking telegraphic (not, it is hoped, a euphemism for "obscure"). It has been found practical to make selections from the notes for courses at Chelsea College, University of London, suitable for undergraduates and graduates (with adequate mathematical background) studying the theory as part of operational research techniques courses, as part of theoretical courses in stochastic processes, or simply for students of the subject in its own right. The separate Appendix concerning systems with fixed deterministic arrival, or service pattern has been included specifically for those students without time to progress beyond Chapter 4. It is repeated that a minimum of thirty hours is needed to do justice to the course as a whole.

COMMONLY USED NOTATION AND ABBREVIATIONS

G/G/N:	Kendall's [7] notation for N server queueing facility with general independent arrivals and service times.
M:	Denotes a negative exponential distribution.
E_k:	Denotes the Erlangian k distribution with probability density function, say $\lambda e^{-\lambda t}(\lambda t)^{k-1}/(k-1)!$.
D:	Denotes deterministic arrivals (appointment system) or fixed length service.
Example:	D/M/1 denotes single server system with deterministic arrivals and negative exponentially distributed service times.
λ:	Mean arrival rate, and when arrivals are M the interarrival times have probability density functions $\lambda e^{-\lambda t}$.
μ:	Mean service rate, and when service times have M distribution the probability density function is $\mu e^{-\mu t}$.
ρ:	Ratio of mean service time to mean interarrival interval, a measure of traffic intensity. Non-dimensional but often quoted in Erlangs. For M/M/1 $\rho = \lambda/\mu$.
Ƶ:	In M/M/1 analysis $Ƶ = z+\lambda+\mu$ where z is in general complex with $R\ell z \geq 0$. In M/M/2 analysis $Ƶ = z+\lambda+2\mu$.
R:	$(Ƶ^2 - 4\lambda\mu)^{\frac{1}{2}}$. For M/M/2 $R = (Ƶ-8\lambda\mu)^{\frac{1}{2}}$.
α,β:	The roots with larger and smaller modulus, respectively, of the quadratic $\mu x^2 - Ƶx + \lambda = 0$, of constant recurrence in M/M/1 analysis. In analysis of M/M/2 the same notation is used for roots of $2\mu x^2 - Ƶx + \lambda = 0$ with $Ƶ = z+\lambda+2\mu$.

These letters may sometimes have different meanings according to the context. For example :

$\alpha(z)$: is also the Laplace transform of general interarrival interval probability density function $a(t)$ in G/M/1 analysis, and

$\beta(z)$: is likewise used frequently to denote the Laplace transform of service time probability density function $b(t)$ in M/G/1 analysis.

The following are standard:

iff: if and only if.

pdf: probability density function.

df: distribution function.

iid: independently and identically distributed.

LT: Laplace transform.

GF: generating function.

$f(t)\underset{L}{\rightarrow}\phi(z)$: $f(t)$ is mapped into $\phi(z)$ by Laplace transformation, and

$\phi(z)\underset{L}{\leftarrow}f(t)$ emphasises the inverse operation.

THE BOOK

QUEUEING SYSTEMS is a concise analysis and evaluation of conventional and adaptive queueing and infinite capacity service systems. It provides an advanced account of underlying theory without excessively heavy mathematics, and has strong practical motivation. It emphasizes the development of simple yet adequate mathematical tools and the calculation of effective numerical measures for assessing queueing systems. Hence the book will serve as a companion to a self-contained course of study to advanced academic level, and for research; or as a practitioner's reference manual.

An unusual feature is the evaluation of traditional and newer methods now available to combat "congestive" situations where capacity has become overwhelmed by excessive demands of service. In this context the "untraditional" methods are the adaptive systems, where service responds to uneven patterns of demand, or where demand monitors service and adjusts to variant patterns.

QUEUEING SYSTEMS offers a presentation of the current research available in these and important areas of stochastic processes. It includes material on adaptive systems and multi-server systems not previously available in book form. Much of it was evolved during the author's professional career as a scientist with the British Admiralty, and subsequently, as a Group Leader in NATO anti-submarine warfare research programme. More recently he taught the material at Virginia Polytechnic Institute and State University and at Chelsea College, London University.

THE AUDIENCE

Advanced undergraduates and postgraduates, practitioners, researchers in statistics, stochastic processes, operational research, computer science, applied mathematics, industrial engineering. Industries involved include computer system development, transport systems, medical systems, the processing of fuels, steel, chemicals etc.

THE AUTHOR

BRIAN CONOLLY graduated with a first class honours degree in mathematics from Reading University in 1944 and, in 1946, gained his MA from King's College, London University for work on hypergeometric functions.

He was a scientist with the British Admiralty 1944-1959, applying mathematics to the solution of problems in engineering, physics and military operational research. He joined NATO Senior Scientific staff for work on anti-submarine warfare problems (1959-1973), becoming Group Leader in 1964.

In 1967, as full Professor of Statistics at Virginia Polytechnic and State University during sabbatical leave, he taught the subject of this book in a course he was asked to develop. In 1973 he took up his present appointment (made in 1971) to the Chair of Mathematics [Operational Research] at Chelsea College in the University of London.

The author is a member of the Royal Statistical Society, London Mathematical Society, Cambridge Philosophical Society, American Mathematical Society, Mathematical Association of America, and the Society for Industrial and Applied Mathematics. He is actively interested in any field of mathematical application which offers prospects for meaningful modelling and analysis.

TABLE OF CONTENTS

TABLE OF CONTENTS

Folio No.

TABLE OF CONTENTS

Folio No.

TABLE OF CONTENTS

Folio No

Chapter 1

INTRODUCTION & DEFINITIONS

1.1 Aims

For practical purposes congestion theory is identical with queueing theory in that it is concerned with mathematical models of "queueing" situations where service is demanded by a customer from the appropriate service point, and the customer must wait in some kind of queue, or waiting line, if service is not immediately available. A typical example is the telephone exchange which services callers requesting connection with some distant point. Supermarkets, restaurants, car parks, many aspects of hospital and airport operations, are self-evident demand/supply situations. Reservoirs, inventory, traffic at intersections (more examples) may seem less obvious, but can be seen to belong to the same family by appropriate interpretation of the notions of demand and service. In all instances common basic mechanisms are at play.

A feature of all situations is the increasing delay suffered by customers as the mean demand rate approaches the mean capability of the service to satisfy it. This is the congestive regime, and because the "slant" of the course is the evaluation of methods for combating congestion, its subject matter has been baptized "Congestion Theory".

The theory is essentially stochastic; that is to say it considers a stream of demands occurring in a chance-dependent manner, serviced by a mechanism such that the duration of each service is also chance-dependent. The theory provides a description of the consequent chance fluctuations in queue length (in our version queue size is unlimited), a customer's waiting time, busy period, output intervals, and other features of interest to both users and operators of a service-providing facility.

The theory is of interest to mathematicians in that it is a complete and on the whole successful and realistic application of mathematics to a familiar non-physical situation with many interpretations, predominantly of social concern. The treatment given is a unified one and has been chosen and presented in such a way as to make the minimum of mathematical demands. Non-mathematicians may contest this statement but, even if not familiar, the special tools needed are limited: some familiarity with (sometimes tortuous) probability argument, formulation and solution of differential and integral-difference equations, the calculus of the Laplace transformation, some elementary complex and real variable theory. Algebraic details have usually been omitted and it should be the task of the student to follow them through, mostly a tedious, but straight-forward task.

The elements described statistically by the theory are stochastic processes and for this reason it is of appeal to pure probabilists. And there are many who feel, with some justification, that the theory belongs to operational research, that corpus of scientific investigation devoted to the understanding of complex operations, often by mathematical modelling, in order to optimize and to aid decision-making. The practitioners of operational research do indeed have much on their side in feeling that queueing is their prerogative. Was it not after all A.K. Erlang, that percipient Danish engineer, who formulated the first models in pursuance of his studies of the Copenhagen telephone system in the early years of the century?

In summary the course is designed to give, by means comprehensible to all likely to be interested, a working knowledge of the theory which theorists and practicians may wish to pursue to the more profound and specialized depths appropriate to their lines of enquiry.

The specialist literature is vast. The reader is referred to [1] - [5] in the reference list at the end of the text for introductory and collateral reading. [4] and [5] provide general background on the theory of probability and of stochastic processes.

1.2 Terms of Reference and Definitions

We deal with mathematical models for situations where service is demanded.

Keywords are:

Customer or client: The unit, animate or inanimate, demanding some form of service.

Server(s): The facility that provides service.

Queue or Waiting Line: Customers who cannot be served immediately on arrival and are prepared to wait are considered to form a queue.

(Am. "waiting line"; Fr. "file d'attente"; It. "fila di attesa" or "coda"; Ge. "Warteschlange").

The component customers need not be served in order of arrival, though *first come, first served* is the most usual form of *queue discipline*. Other disciplines are *last come, first served*, *random service*, and there are possible various systems of *priority*. *Customer behaviour* is another factor which may include cheating, passing from one queue to another in the case of a multiple queue situation, collusion to retain earlier service than that ordained by prevailing discipline, and so forth. (See [2] for further details).

Size of waiting room can limit maximum queue size. In principle all sizes of waiting room can be considered varying from zero (no waiting permitted and hence clients *lost*, when service is occupied). Throughout this treatment no limitation is imposed on queue size (infinite waiting room). This slight loss of generality is compensated by clarity of exposition and ease of understanding. Demands for service are thought of as occurring in the form of a stream of arrivals each requiring attention. Intervals between arrivals are called *inter-arrival intervals*. These will normally be treated as random variables, independently and identically distributed (iid) with properly behaved distribution function (df). Similarly *duration of service*, or *service time*, will normally be a random variable. A major objective of this course is to compare situations in which service is independent of demand with situations in which attention is given to demand.

Traffic intensity is measured by the ratio of mean service time to mean inter-arrival interval. This parameter, usually denoted by ρ, has an important rôle in the theory. Instinct tells us to expect that service can "on average" contain demand without congestion when $\rho < 1$, the reverse when $\rho > 1$. Instinct proves to be unreliable in cases where $\rho = 1$. ρ, though non-dimensional, is measured by international agreement in *erlangs* to commemorate A.K. Erlang, the father of queuing theory. An account of his work may be found in [6].

A notation introduced by Kendall [7] sums up the salient features of a queuing situation. It is GI/G/N. GI refer to arrivals, the letter signifying that the inter-arrival intervals have a general distribution, each interval being independent of every other. Nowadays the I is often dropped. G refers to service and the letter again means general. N is the number of servers. No indication is given of discipline, customer behaviour, etc. $G/G/\infty$ may appear a fiction but it does model systems with so many service facilities that they can be regarded as unlimited with no need for a client ever to wait. This system provides an approximate model for very large car parks, ships at sea, customers in a large supermarket. It also gives an upper limit of performance of a system with prescribed parameters. Common forms of GI or G are denoted by M, D and E_k. M means *negative exponential* and the letter refers to the Markov lack of memory property. D means *deterministic*, the interval in question having then fixed duration (e.g. appointment system). E_k refers to the distribution with probability density function (pdf)