Computer Science
and Applied Mathematics

ANALYSIS AND SYNTHESIS OF COMPUTER SYSTEMS

Gelenbe/Mitrani

Analysis and Synthesis of Computer Systems

E. GELENBE

Laboratoire de Recherche en Informatique
Université de Paris-Sud, Orsay, France

and

I. MITRANI

Computing Laboratory
University of Newcastle upon Tyne,
England

1980 ACADEMIC PRESS

A Subsidiary of Harcourt Brace Jovanovich, Publishers

London New York Toronto Sydney San Francisco

ACADEMIC PRESS INC. (LONDON) LTD.
24/28 Oval Road
London NW1

United States Edition published by
ACADEMIC PRESS INC.
111 Fifth Avenue
New York, New York 10003

British Library Cataloguing in Publication Data
Gelenbe, Erol
Analysis and synthesis of computer systems.
1. Electronic data processing – Evaluation
2. Stochastic processes
I. Title II. Mitrani, Israel
001.6′4 QA76.9.E94 80-49992

ISBN 0-12-279350-1

Analysis and Synthesis of Computer Systems

This is a volume in

COMPUTER SCI

A Series of Monog

Editor: WERNER

A complete list of titles in this series appears at the end of this volume.

Preface

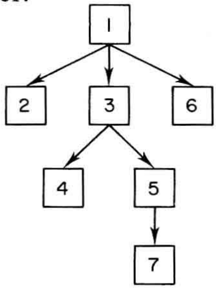
The subject of computer performance evaluation has grown steadily in the last decade or so. As the area of computer applications expanded out of the scientific domain and into those of business, commerce and administration, it became more and more important to know in advance what a given system can do and how it should be operated. After all, an inefficient computer system can cause no small amount of chaos—from pay-slips not arriving on time, to traffic jams in a busy town centre. There are also cost considerations: large computer systems represent considerable investments, both for the manufacturers who design and build them, and for the users who buy them. It is in the interest of everyone concerned to be able to answer questions about the performance of such systems—preferably before they are built or bought. This is our interest, too.

The approach we adopt is that of probabilistic modelling; we deal with the construction and analysis of models, with their use in evaluating the performance of computer systems and with their application to system design and control. Why use probabilistic modelling in relation to objects which are man-made and deterministic by nature? The answer is, of course, that while the machines are deterministic, the demands placed upon them are not. The arrival patterns and the resource requirements of the tasks comprising the demand are, in general, random. Each task execution involves such a complex sequence of operations (machine instructions) that it is, for all practical purposes, unpredictable. We model these random phenomena by introducing probabilistic assumptions.

The problems that we consider fall into two broad categories: problems of analysis and problems of synthesis. If all system parameters—number and speed of processors, size of memory, characteristics of demand, system software, etc.—are given, and the question is "how does this system perform?", then we have a problem of analysis. If, on the other hand, some components of the system are regarded as variable (typically these are concerned with job scheduling and resource allocation algorithms), and one is asked to find a configuration that achieves a certain pre-specified performance objective, then we say that the problem is one of synthesis. Often, although not always, analysis and synthesis go hand in hand. Suppose, for example, that we are offered a family of job scheduling strategies and are asked to choose the best one according to some criterion (a synthesis problem). In some cases, the answer can be obtained directly from general principles; in others, it is necessary first to analyse the strategies in the family and to evaluate their performance. We shall encounter both situations.

The reader is introduced to the methodology and the necessary tools of probabilistic modelling in Chapter 1. Some basic results from the theory of Markov processes and queueing theory are derived there. In Chapter 2, a class of queueing systems known as "systems with a server of walking type" are analysed and the results are applied to the performance evaluation of several peripheral storage devices. Chapter 3 is devoted to the analysis of queueing network models. Cases with known closed-form solutions are examined, involving fixed and variable job populations, and allowing multiple job classes. Chapters 4 and 5 are concerned with the approximate analysis of models for which exact solutions are not available. Chapter 4 describes diffusion approximation techniques; these are applied to both single-node and network models. Chapter 5 deals with decomposition and iterative methods aimed at reducing the complexity of models. Chapters 6 and 7 address synthesis problems in single and multiple resource systems: conservation laws, achievability of response time objectives, minimisation of certain cost functions, control of the degree of multiprogramming in virtual memory systems with one or several job classes and control of performance by memory allocation.

The box diagram illustrates the dependencies among chapters. An arrow from one box to another indicates that the corresponding chapters should be read in that order.



Chapters 2 and 4 have a somewhat higher mathematical content than the others.

The book is aimed at students, researchers and practitioners in the field of performance evaluation. However, it is not intended as a "book of recipes"; we hope, rather, that sufficient knowledge and understanding are imparted to the reader so that he can write his own recipes. Knowledge of basic computer system concepts is assumed, as well as some mathematical background, including undergraduate calculus and probability theory.

January 1980

E.G.
I.M.

Acknowledgements

We would like to express our gratitude to all those who have helped us during the writing and preparation for publication of this book. Special thanks are due to Ewan S. Page who read several chapters of the manuscript and made many valuable suggestions, comments and corrections. Constructive criticism was offered also by Peter J. Denning, Guy Fayolle, Peter J. B. King and Guy Pujolle.

The manuscript was competently typed on both sides of the English Channel by Claudine Décalf, Marie-Thérèse Bouvier, Diane Pelley and Elaine Herron. Hazel Mitrani helped out in the unrewarding task of preparing the index. Academic Press, in the person of Diana Beaven, guided us through the production stages with patience and understanding.

Last but not least, we wish to thank our two Universities, of Paris-Sud and Newcastle upon Tyne, and the Institut National de Recherche en Informatique et en Automatique. Without the stimulating environments in those institutions, and their financial support, this work would have been impossible.

E.G.
I.M.

Computer Science and Applied Mathematics

A SERIES OF MONOGRAPHS AND TEXTBOOKS

Editor
Werner Rheinboldt
University of Pittsburgh

HANS P. KÜNZI, H. G. TZSCHACH, and C. A. ZEHNDER. Numerical Methods of Mathematical Optimization: With ALGOL and FORTRAN Programs, Corrected and Augmented Edition

AZRIEL ROSENFELD. Picture Processing by Computer

JAMES ORTEGA AND WERNER RHEINBOLDT. Iterative Solution of Nonlinear Equations in Several Variables

AZARIA PAZ. Introduction to Probabilistic Automata

DAVID YOUNG. Iterative Solution of Large Linear Systems

ANN YASUHARA. Recursive Function Theory and Logic

JAMES M. ORTEGA. Numerical Analysis: A Second Course

G. W. STEWART. Introduction to Matrix Computations

CHIN-LIANG CHANG AND RICHARD CHAR-TUNG LEE. Symbolic Logic and Mechanical Theorem Proving

C. C. GOTLIEB AND A. BORODIN. Social Issues in Computing

ERWIN ENGELER. Introduction to the Theory of Computation

F. W. J. OLVER. Asymptotics and Special Functions

DIONYSIOS C. TSICHRITZIS AND PHILIP A. BERNSTEIN. Operating Systems

ROBERT R. KORFHAGE. Discrete Computational Structures

PHILIP J. DAVIS AND PHILIP RABINOWITZ. Methods of Numerical Integration

A. T. BERZTISS. Data Structures: Theory and Practice, Second Edition

N. CHRISTOPHIDES. Graph Theory: An Algorithmic Approach

ALBERT NIJENHUIS AND HERBERT S. WILF. Combinatorial Algorithms

AZRIEL ROSENFELD AND AVINASH C. KAK. Digital Picture Processing

SAKTI P. GHOSH. Data Base Organization for Data Management

DIONYSIOS C. TSICHRITZIS AND FREDERICK H. LOCHOVSKY. Data Base Management Systems

JAMES L. PETERSON. Computer Organization and Assembly Language Programming

WILLIAM F. AMES. Numerical Methods for Partial Differential Equations, Second Edition

ARNOLD O. ALLEN. Probability, Statistics, and Queueing Theory: With Computer Science Applications

ELLIOTT I. ORGANICK, ALEXANDRA I. FORSYTHE, AND ROBERT P. PLUMMER. Programming Language Structures

ALBERT NIJENHUIS AND HERBERT S. WILF. Combinatorial Algorithms. Second edition.

JAMES S. VANDERGRAFT. Introduction to Numerical Computations

AZRIEL ROSENFELD. Picture Languages, Formal Models for Picture Recognition

ISAAC FRIED. Numerical Solution of Differential Equations

ABRAHAM BERMAN AND ROBERT J. PLEMMONS. Nonnegative Matrices in the Mathematical Sciences

BERNARD KOLMAN AND ROBERT E. BECK. Elementary Linear Programming with Applications

CLIVE L. DYM AND ELIZABETH S. IVEY. Principles of Mathematical Modeling

ERNEST L. HALL. Computer Image Processing and Recognition

ALLEN B. TUCKER, JR., Text Processing: Algorithms, Languages, and Applications

MARTIN CHARLES GOLUMBIC. Algorithmic Graph Theory and Perfect Graphs

In preparation

GABOR T. HERMAN. Image Reconstruction from Projections: The Fundamentals of Computerized Tomography

WEBB MILLER AND CELIA WRATHALL. Software from Roundoff Analysis of Matrix Algorithms

Contents

1 Basic Tools of Probabilistic Modelling

1.1 General background

On a certain level of abstraction, computer systems belong to the same family as, for example, job-shops, supermarkets, hairdressing salons and airport terminals; all these are sometimes described as "mass service systems" and more often as "queueing systems". Customers (or tasks, or jobs, or machine parts) arrive according to some random pattern; they require a variety of services (execution of arithmetic and logical operations, transfer of information, seat reservations) of random durations. Services are provided by one or more servers, perhaps at different speeds. The order of service is determined by a set of rules which constitutes the "scheduling strategy", or "service discipline".

The mathematical analysis of such systems is the subject of queueing theory. Since A. K. Erlang's studies of telephone switching systems, in 1917–1918, that theory has progressed considerably; today it boasts an impressive collection of results, methods and techniques. Interest in queueing theory has always been stimulated by problems with practical applications. In particular, most of the theoretical advances of the last decade are directly attributable to developments in the area of computer systems performance evaluation.

Because customer interarrival times and the demands placed on the various servers are random, the state $S(t)$ of a queueing system at time t of its operation is a random variable. The set of these random variables $\{S(t), t \geq 0\}$ is a stochastic process. A particular realisation of the random variables—that is, a particular realisation of all arrival events, service

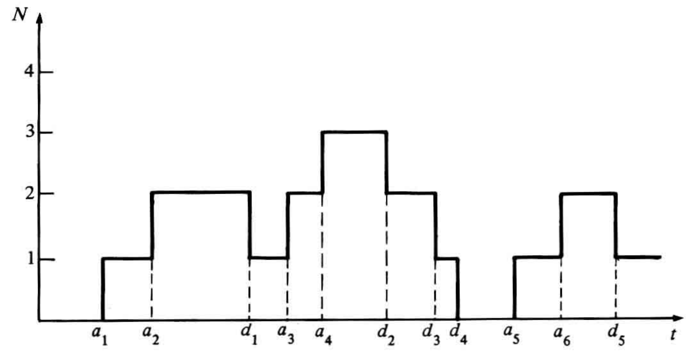


FIG. 1.1

demands, etc.—is a "sample path" of the stochastic process. For example, in a single-server queueing system where all customers are of the same type, one might be interested in the stochastic process $\{N(t), t \geq 0\}$, where $N(t)$ is the number of customers waiting and/or being served at time t. A portion of a possible sample path for this process is shown in Fig. 1.1: customers arrive at moments $a_1, a_2, \ldots$ and depart at moments $d_1, d_2, \ldots$

An examination of the sample paths of a queueing process can disclose some general relations between different quantities associated with a given path. For instance, in the single-server system, if $N(t_1) = N(t_2)$ for some $t_1 < t_2$, and there are k arrivals in the interval (t_1, t_2), then there are k departures in that interval. Since a sample path represents a system in operation, relations of the above type are sometimes called "operational laws" or "operational identities" (Buzen [1]). We shall derive some operational identities in section 1.7. Because they apply to individual sample paths, these identities are independent of any probabilistic assumptions governing the underlying stochastic process. Thus, the operational approach to performance evaluation is free from the necessity to make such assumptions. It is, however, tied to specific sample paths and hence to specific runs of an existing system where measurements can be taken.

The probabilistic approach involves studying the stochastic process which represents the system. The results of such a study necessarily depend on the probabilistic assumptions governing the process. These results are themselves probabilistic in nature and concern the population of all possible sample paths. They are not associated with a particular run of an existing system, or with any existing system at all. It is often desirable to evaluate not only the expected performance of a system, but also the likely deviations from that expected performance. Dealing with probability distributions makes this possible, at least in principle.

We shall be concerned mainly with steady-state system behaviour—that is, with the characteristics of a process which has been running for a long time and has settled down into a "statistical equilibrium regime". Long-run performance measures are important because they are stable; being independent of the early history of the process, and independent of time, they are also much easier to deal with. We shall, of course, be interested in the conditions which ensure the existence of steady-state.

This chapter introduces the reader to the rudiments of stochastic processes and queueing theory. Results used later in the book will be derived here, with the emphasis on explaining important methods and ideas rather than on rigorous proofs. In discussing queueing systems, we shall use the classic descriptive notation devised by D. G. Kendall:

·/·/·

(arrival pattern) (service pattern) (number of servers)

e.g. $D/M/2$ describes a queueing system with Deterministic (constant) interarrival times, Markov (exponential) service times and 2 servers.

1.2 Markov processes. The exponential distribution

Let $S(t)$ be a random variable depending on a continuous parameter t ($t \geq 0$) and taking values in the set of non-negative integers $\{0, 1, 2, \ldots\}$. We think of t as time and of $S(t)$ as the system state at time t. The requirement that the states should be represented as positive integers is not important; it is essential that they should be denumberable. Later, we shall have occasions to use vectors of integers as state descriptors.

The collection of random variables $\{S(t), t \geq 0\}$ is a stochastic process. That collection is said to be a "Markov process" if the probability distribution of the state at time $t+y$ depends only on the state at time t and not on the process history prior to t:

$$P(S(t+y)=j \mid S(u); u \leq t) = P(S(t+y)=j \mid S(t)), \qquad t, y \geq 0, j = 0, 1, \ldots . \tag{1.1}$$

The right-hand side of (1.1) may depend on t, y, j and the value of $S(t)$. If, in addition, it is independent of t, i.e. if

$$P(S(t+y)=j \mid S(t)=i) = p_{i,j}(y) \text{ for all } t, \tag{1.2}$$