Lecture Notes in
Control and
Information Sciences

Edited by A.V. Balakrishnan and M. Thoma

202

26

D. L. Iglehart · G. S. Shedler

Regenerative Simulation of Response Times in Networks of Queues

Springer-Verlag
Berlin · Heidelberg · New York

Lecture Notes in Control and Information Sciences

Edited by A.V. Balakrishnan and M. Thoma

26

D. L. Iglehart · G. S. Shedler

Regenerative Simulation of Response Times in Networks of Queues

Springer-Verlag
Berlin Heidelberg New York 1980

Authors
Donald L. Iglehart
Department of Operations Research
Stanford University
Stanford, California 94305
and
Gerald S. Shedler
IBM Research Laboratory
San Jose, California 95193

ISBN 3-540-09942-5 Springer-Verlag Berlin Heidelberg NewYork
ISBN 0-387-09942-5 Springer-Verlag NewYork Heidelberg Berlin

Printing and binding: Beltz Offsetdruck, Hemsbach/Bergstr.
2060/3020-543210

PREFACE

Discrete event digital simulation of stochastic models has been one of the most important practical tools of systems analysis for well over twenty years. The complexity of models for most real systems is such that we are unable to study them analytically. Computer simulation is our only alternative and we must seek theoretically sound and computationally efficient methods for carrying out the simulation. While a great deal of effort has been devoted to the development of simulation programming languages and programs, relatively little has been done to develop theoretical foundations to justify the estimation methods implemented in the simulations. Typically, simulation program packages compute by rote and report a large variety of point estimates for various characteristics of the model being simulated. Seldom do these reports indicate the variability or statistical precision of the point estimates.

This monograph deals with probabilistic and statistical methods for discrete event simulation of networks of queues. The emphasis is on the use of underlying stochastic structure for the design of simulation experiments and the analysis of simulation output. We focus on recently developed methods for estimation of general characteristics of "passage times" in closed networks of queues. Informally, a passage time is the time for a job to traverse a portion of a network. Such quantities are important in computer and communication system models where they represent job response times. In this context, expected values as well as other characteristics of response times are of interest.

The presentation is self contained. We have attempted to make this material accessible to simulation practitioners as well as to students and researchers interested in the methodology of discrete event simulation. For this reason, we have provided a number of examples and have separated the exposition of the estimation procedures from the derivations. Some knowledge of elementary probability theory, statistics, and stochastic models is sufficient to understand the estimation procedures and the examples. The derivations use results often contained in first year graduate courses in stochastic processes. These sections can be omitted by the reader interested primarily in application of the procedures.

We have benefitted from the comments of Peter Glynn, Lily Jow, Austin Lemoine, and Gerard Scallan who read an earlier draft of the monograph. It is a pleasure to acknowledge our gratitude to Julie Countryman and Blanca Gallegos of the Manuscript Processing Center at the IBM San Jose Research Laboratory for their expert typing of a difficult manuscript. We are also indebted to Jean Chen and Jack DeLany of Central Scientific Services at IBM Research for their careful preparation of the figures.

We are both grateful to the National Science Foundation for support under Grant MCS-7909139. In addition, one of us (D.L.I.) gratefully acknowledges partial support under National Science Foundation Grant MCS-7523607 and Office of Naval Research Contract N00014-76-C-0578 (NR 042-343).

Donald L. Iglehart
Gerald S. Shedler

Stanford University
Stanford, California

IBM Research Laboratory
San Jose, California

November, 1979

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1.0. INTRODUCTION

Networks of queues occur frequently in diverse applications. In particular, they are widely used in studies of computer and communication system performance as models for the interactions among system resources. This monograph deals with mathematical and statistical methods for discrete event simulation of networks of queues. The emphasis is on methods for the estimation of general characteristics of "passage times" in closed networks. Informally, a passage time is the time for a job to traverse a portion of a network. Such quantities, calculated as random sums of queueing times, are important in computer and communication system models where they represent job response times.

By simulation we mean observation of the behavior of a stochastic system of interest by artificial sampling on a digital computer. With discrete event simulation, stochastic changes of the system state occur only at a set of increasing time points. Simulation is a tool which can be used to study complex stochastic systems when analytic and/or numerical techniques do not suffice; in connection with the study of complex networks of queues encountered in applications, this is often the case.

When simulating, we experiment with a stochastic system and observe its behavior. During the simulation we measure certain quantities in the system and, using statistical techniques, draw inferences about characteristics of well defined random variables. The most obvious methodological advantage of simulation is that in principle it is applicable to stochastic systems of arbitrary complexity. In practice, however, it is often a decidedly nontrivial matter to obtain from a simulation information which is both useful and accurate, and to obtain it in an efficient manner. These difficulties arise primarily from the inherent variability in a stochastic system, and it is necessary to seek theoretically sound and computationally efficient methods for carrying out the simulation. Apart from implementation considerations, important concerns for simulation relate to efficient methods for generation of realizations (sample paths) of the stochastic system under study, the design of simulation experiments, and the analysis of simulation output. It is fundamental for simulation, since results are based on observation of a stochastic system, that some assessment of the precision of results be provided.

Assessing the statistical precision of a point estimate requires careful design of the simulation experiments and analysis of the simulation output. In general, the desired statistical precision takes the form of a confidence interval for the quantity of interest. Among the issues the simulator must face are the initial conditions for the system being simulated, the length of the simulation run, the number of replications of the experiments, and the length of the confidence interval. Over the last five years, there has been increased attention paid to these issues, and a theory of simulation analysis (the regenerative method) has been developed which, when applicable, provides some measure of statistical precision. The regenerative method, which is based on limit theorems developed for regenerative stochastic processes, plays a key role in our discussion of simulation methods for passage times in networks of queues.

Under the usual queueing-theoretic (independent and identically distributed service and interarrival time) assumptions, analyses based on a "numbers-in-queue" and "stages-of-service" state vector can be carried out. Typically it is necessary to assume that all service and interarrival time distributions are exponential or have a Cox-phase (exponential stage) representation. Under these assumptions, expressions suitable for numerical evaluation are obtainable for queue length distributions. Other measures of system performance (calculated as random sums of queueing times) involve the times, here called passage times, for a job to traverse a portion of the network. Often when such quantities arise in computer and communication system models, they represent job response times. In this context, expected values as well as other characteristics of passage times (e.g., percentiles) are of interest. The analyses based on the numbers-in-queue, stages-of-service state vector yield expected values for passage times, but do not yield other passage time characteristics of interest. Moreover, alternative analyses to provide these measures of the variability of system response are in general not available, and it is necessary to resort to simulation. Although the usual process of numbers-in-queue and stages-of-service is a regenerative process (in fact a Markov chain) under the probabilistic assumptions that we make here, the regenerative method cannot be applied directly to this process to estimate general passage time characteristics. This is essentially because passage times are not totally contained within cycles of the numbers-in-queue, stages-of-service process.

The organization of the presentation is as follows. This initial section provides some motivation for study of simulation methods for passage times in networks of queues, a brief overview of some of the methodological considerations for simulation, and a summary of the discussion which appears in subsequent sections.

The estimation methods developed here for passage times in networks of queues use the regenerative method for analysis of simulation output. Based on a single simulation run, these methods provide (strongly consistent) point estimates and (asymptotically) valid confidence intervals for general characteristics of limiting passage times. Section 2 provides a review of the regenerative method. The section contains a brief discussion of the underlying theory of regenerative stochastic processes along with some examples of regenerative processes in networks of queues.

Section 3 provides a specification of the basic class of closed networks of queues with which we deal, and the probabilistic assumptions therein. Initially, we restrict attention to networks with stochastically identical jobs and give a state vector definition based on a linear job stack. The section also contains the formal definition of passage time in a network of queues.

The notion of a distinguished "marked" job is fundamental to the method for estimation of passage time characteristics described and developed in Section 4. The approach is to consider a Markov renewal process arising from a continuous time Markov chain defined by the usual numbers-in-queue, stages-of-service state vector augmented by information sufficient to track the marked job. We arbitrarily select a job to serve as the marked job and measure its passage times during the simulation. They key steps in the derivation of this marked job method are identification of an appropriate regenerative process in discrete time and development of a ratio formula from which point estimates and confidence intervals can be obtained for quantities associated with the limiting passage time.

In Section 5 we consider application of the marked job method to two particular closed networks of queues, and display some numerical results. The first example is a relatively simple network. Despite the apparent structural simplicity of this network, it exhibits the essence of the passage time simulation problem. The second and more complex network arises as a model for a computer data base management system. This model illustrates the representation of complex congestion phenomena in the framework of Section 3.

The extension of the marked job method to certain finite capacity open networks of queues is the subject of Section 6. Particular stochastic point processes associated with a Markov renewal process generate arrivals to the networks, and there are two formulations of the finite capacity constraint. The network structure we permit is essentially the same as that described in Section 3 except that here the networks are open. To estimate passage times in these networks, we track an appropriate sequence of typical jobs, based on the idea of a marked job. These are to be typical jobs in the sense that the sequence of passage times for the marked