BOOK REVIEWS

BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 8, Number 1, January 1983

0273-0979/82/0000-0830/$01.75

Matrix-geometric solutions in stochastic models, an algorithmic approach, by Marcel F. Neuts, The Johns Hopkins University Press, Baltimore, 1981, xiii + 332 pp., $32.50.

This book is concerned with steady-state solutions to queueing systems. In general, such systems involve customers (not necessarily people) arriving at some service facility, waiting for service if it is not immediately available, and leaving the system after having been served. They are characterized by the arrival pattern of customers (given in terms of the distribution of the number of arrivals or of interarrival times), the service pattern of servers (given by service distributions), the queue discipline (e.g., first come, first served; last in, first out; selection for service in random order independent of the time of arrival to the queue; and so on), the system capacity (size of the waiting room), the number of service channels, and the number of service stages (one or multiple stages, with or without recycling of the departing customers).

Queueing systems can be deterministic, probabilistic or both. An example of one which is both is a queue in which the customers arrive by appointment or at fixed intervals (a deterministic input) but their service times may differ (probabilistic service). Most of the work done in this field is usually of a probabilistic nature, i.e., the interarrival and service time variables are assumed to be random variables. Their efficiency can be measured in terms of the length of time a customer might be forced to wait (the waiting time), the number of customers which may accumulate (length of the queue), and the busy (or idle) period of the servers.

The classical approach to the analysis of queues is to assume that the variables representing the interarrival and service times follow some specific distributions, and define the state of the system by the number of customers in the system (in queue and in service) at a given time. The object is to find the probability distribution of the number of customers in the system from which the desired measures of efficiency can be derived. One way of doing this is to formulate a system of differential-difference equations to represent the behavior of the queue in time. The unknowns are the probabilities that the system is in a certain state at a given time. The solution to this system of equations is known as "the transient solution". Of particular interest is the solution to the system of equations when time tends to infinity—the steady-state solution. This is the main subject of the book by Marcel F. Neuts.

There are alternative methods (generating functions, operators, etc.) for solving the system of differential-difference equations that represent the behavior of a queueing system in the steady-state. When a queueing system is such that its future state depends only on its present and not on its past state the model is said to be Markovian, and the Chapman-Kolmogorov backward and forward equations are used to describe the system. When the system is non-Markovian, these equations do not apply. However, for many non-Markovian queues, one can identify a Markov chain (referred to as the embedded Markov chain) to which some of the theory of Markov chains is applicable. Here, there is a matrix of transition (from state to state) probabilities which is raised to infinite powers to obtain its limiting form and used to construct the steady-state solution.

This book, the first of its kind, deals entirely with steady-state solutions to queueing systems of the non-Markovian type within the framework of the embedded Markov chain. The first chapter contains existence and uniqueness theorems for solutions derived from the embedded Markov chain for a single server queue of two types: (a) one which has a negative exponential interarrival-times distribution, and a general identical (same for all customers) distribution of service times; and (b) one which has a general independent, identically-distributed interarrival times, and a negative exponential service times distribution.

The second chapter includes a comprehensive up-to-date study of Phase Type (PH) distributions, usually introduced to deal with input and service distributions of the hyperexponential type (i.e., obtained from a mixture of negative exponential distributions with the property that the standard deviation of the mixture exceeds the mean) or the Erlangian type (i.e., obtained from the sum of independent and identically distributed negative exponential random variables). According to the author, PH-distributions "lend themselves to the representation of certain qualitative features of data, such as bimodality or increased tail probabilities."

In Chapter 3 the author deals with quasi-birth-and-death processes useful in the analysis of busy periods and in the development of waiting-time expressions for queues whose interarrival and/or service times distributions are PH-distributions. Marcel F. Neuts writes in p. 132, "The relationship of quasi-birth-and-death processes to systems of linear differential equations with constant coefficients is used to construct efficient, if occasionally time-consuming, algorithms." A quasi-birth-and-death process is a continuous parameter Markov process with an infinitesimal operator which has a block tridiagonal structure after an appropriate ordering of states. In these processes the steady-state equations can be replaced by a matrix-vector-valued difference equation of the form $P_nA + P_{n+1}B + P_{n+2}C = 0$ with initial conditions of the form $P_0E + P_1F = 0$, where P_n is the probability that the system is in state n in the steady-state. As the author points out, they are of interest because "All references to complex analysis and roots of transcendental equations are avoided." Using these processes, the author studies, in Chapter 4, the queue which has a general independent, identically-distributed interarrival times, and whose service times distribution is a PH-distribution, and related models.

Buffer models or networks with blocking are studied in Chapter 5. The simplest example of such a network consists of two queues in series so that customers departing the first queue arrive to wait at the second queue which has a limited waiting room capacity. When the capacity of the queue of the second system is reached, the first queue momentarily stops serving. The limited capacity of the second queue creates a buffer for the two-unit system. Quasi-birth-and-death processes are used here to model the behavior of the system and compute the steady-state solution for a more general network consisting of two units separated by a buffer, but each unit is a set of parallel exponential servers. There is a single queue before each of the two units. Extensions and variations of this model are also briefly discussed. The author also includes a very abbreviated account of important works on networks with and without buffers. Of special interest is §5.4 which gives an application of buffer models to computer operations. It is called the multiprogramming-multiprocessor system. It has an infinite input queue to which customers arrive according to a homogeneous Poisson process (i.e., negative exponential inter-arrival times distribution). The network consists of two queues (units), each of which is a set of parallel exponential servers. A customer arriving at the queue, when allowed to enter the network, is processed by one of the exponential servers of the first unit. Upon completion of service, the customer either leaves the system or joins the second unit. In the latter case, the customer must return to the first unit to be processed once more. The capacity constraint imposed by the buffer requires that the total number of customers in the network (in queue and in service) must not exceed a positive integer that in many cases is greater than or equal to the total number of servers of the two units together. The first unit is referred to as the input-output device, while the second one is referred to as the central processing unit.

Finally, in Chapter 6, quasi-birth-and-death processes are used to study simple queueing systems characterized by distributions whose parameters are random variables. The author deals with queues whose interarrival and service time distribution parameters are Markovian.

The book is elegantly written. Its title could mislead the reader to think that it includes an analysis of general stochastic models rather than queues. The author has succeeded well in following Leibnitz's exhortation which he has placed on a separate page in front:

> "It is unworthy of excellent men to lose hours like slaves in the labor of calculation which could safely be relegated to anyone if machines were used."

However, without detailed illustrations, this kind of aloofness in an applied field like queueing theory could create problems for the uninitiated to see how the author's results can be tested and used in practice. The book is a useful contribution to queueing theory.

Luis G. Vargas