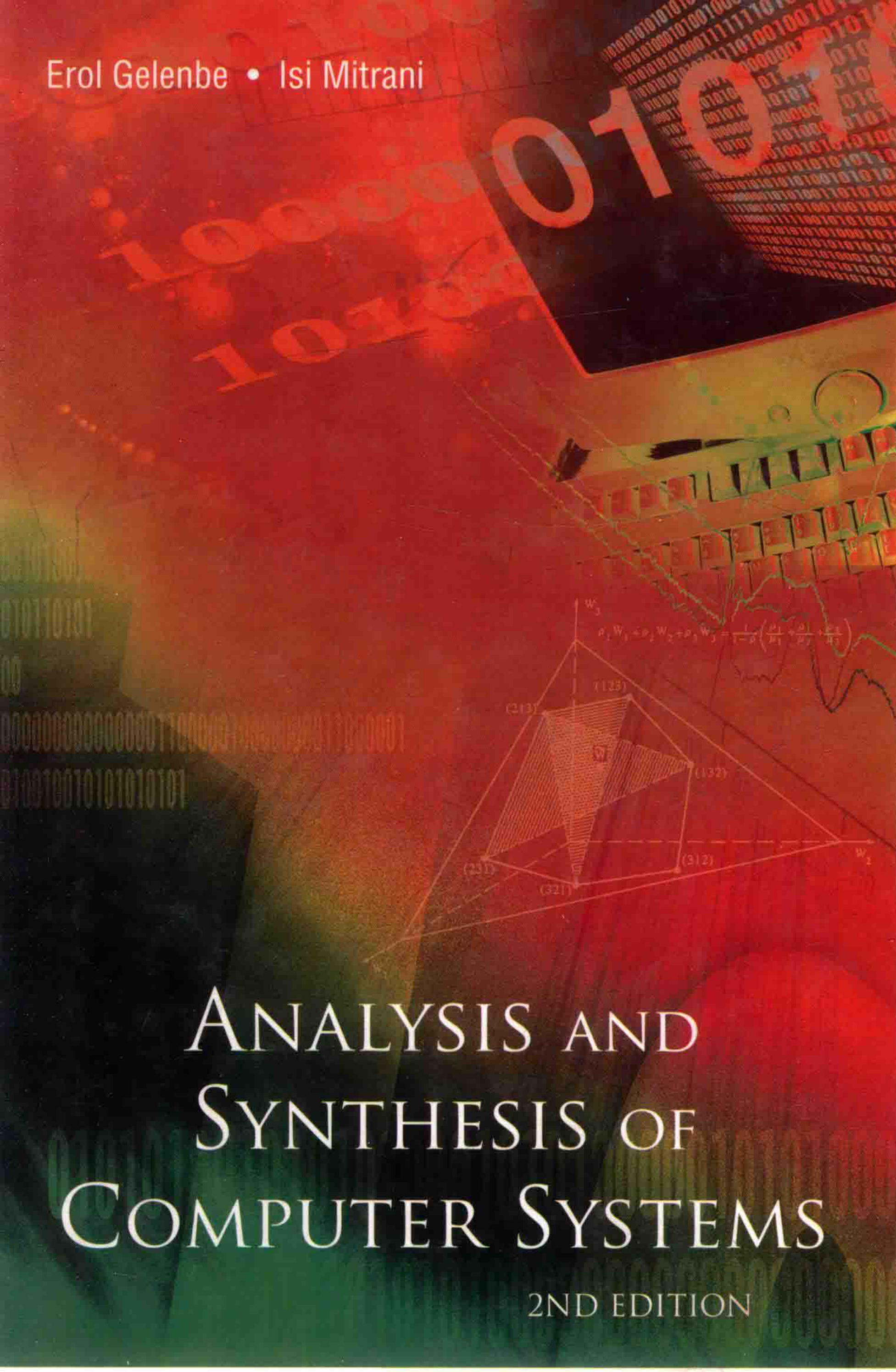
Erol Gelenbe • Isi Mitrani



ANALYSIS AND SYNTHESIS OF COMPUTER SYSTEMS

2ND EDITION

Imperial College Press

Vol. 4

Advances in Computer Science and Engineering: Texts

Advances in Computer Science and Engineering: Texts Vol. 4

Analysis and Synthesis of Computer Systems

2nd Edition

Erol Gelenbe
Imperial College, UK

Isi Mitrani
University of Newcastle upon Tyne, UK

ICP Imperial College Press

Published by

Imperial College Press
57 Shelton Street
Covent Garden
London WC2H 9HE

Distributed by

World Scientific Publishing Co. Pte. Ltd.
5 Toh Tuck Link, Singapore 596224
USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601
UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

ANALYSIS AND SYNTHESIS OF COMPUTER SYSTEMS (2nd Edition)
Advances in Computer Science and Engineering: Texts — Vol. 4

Desk Editor: Tjan Kwang Wei

ISBN-13 978-1-84816-395-9

Typeset by Stallion Press
Email: enquiries@stallionpress.com

Printed in Singapore by World Scientific Printers.

Analysis and Synthesis of Computer Systems

2nd Edition

Advances in Computer Science and Engineering: Texts

Editor-in-Chief: Erol Gelenbe *(Imperial College)*
Advisory Editors: Manfred Broy *(Technische Universitaet Muenchen)*
Gérard Huet *(INRIA)*

Published

Vol. 1 Computer System Performance Modeling in Perspective: A Tribute to the Work of Professor Kenneth C. Sevcik
edited by E. Gelenbe (Imperial College London, UK)

Vol. 2 Residue Number Systems: Theory and Implementation
by A. Omondi (Yonsei University, South Korea) and
B. Premkumar (Nanyang Technological University, Singapore)

Vol. 3: Fundamental Concepts in Computer Science
edited by E. Gelenbe (Imperial College Londo, UK) and
J.-P. Kahane (Université de Paris Sud - Orsay, France)

Vol. 4: Analysis and Synthesis of Computer Systems (2nd Edition)
by Erol Gelenbe (Imperial College, UK) and
Isi Mitrani (University of Newcastle upon Tyne, UK)

Preface to the Second Edition

The book has been revised and extended, in order to reflect important developments in the field of probabilistic modelling and performance evaluation since the first edition. Notable among these is the introduction of queueing network models with positive and negative customers. A large class of such models, together with their solutions and applications, is described in Chapter 4. Another recent development concerns the solution of models where the evolution of a queue is controlled by a Markovian environment. These Markov-modulated queues occur in many different contexts; their exact and approximate solution is the subject of Chapter 5. Finally, the queue with a server of walking type described in Chapter 2 is given a more general treatment in Chapter 10.

Erol Gelenbe
Isi Mitrani
February 2010

Contents

Chapter 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 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

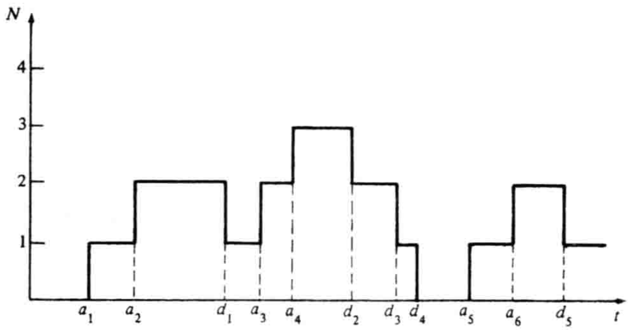


Fig. 1.1.

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:

$\cdot/\cdot/\cdot$

(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:

$$\begin{aligned} P(S(t+y) &= j|S(u); u \leq t) \\ &= P(S(t+y) = j|S(t)), \quad t, y \geq 0,\ j = 0, 1, \ldots. \end{aligned} \tag{1.1}$$

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

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

then the Markov process is said to be "time-homogeneous" (for an excellent treatment of stochastic processes see Cinlar [3]). From now on, whenever we talk of a Markov process, we shall assume that it is time-homogeneous.

Thus, for a Markov process, the probability $p_{i,j}(y)$ of moving from state i to state j in time y is independent of the time at which the process was in state i and of anything that happened before that time. This very important property will be referred to as the "memoryless property".

The probability $p_{i,j}(y)$, regarded as a function of y, is called the "transition probability function". The memoryless property immediately implies the following set of functional equations:

$$p_{i,j}(x+y) = \sum_{k=0}^{\infty} p_{i,j}(x)p_{k,j}(y), \quad x, y \geq 0, \; i, j = 0, 1, \ldots . \tag{1.3}$$

These equations express simply the fact that, in order to move from state i to state j in time $x + y$, the process has to be in some state k after time x and then move to state j in time y (and the second transition does not depend on i and x). They are the Chapman–Kolmogorov equations of the Markov process. Introducing the infinite matrix $\mathbf{P}(\mathrm{y})$ of transition functions $p_{i,j}(y)$, we can rewrite (1.3) as

$$\mathbf{P}(x+y) = \mathbf{P}(x)\mathbf{P}(y), \quad x, y \geq 0. \tag{1.4}$$

We shall assume that the functions $p_{i,j}(y)$ are continuous at $y = 0$:

$$\lim_{y \to 0} p_{i,j}(y) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases} \tag{1.5}$$

That assumption, together with (1.3), ensures that $p_{i,j}(y)$ is continuous, and has a continuous derivative, for all $y \geq 0$; $i, j = 0, 1, \ldots$ (we state this without proof).

A special role is played by the derivatives $a_{i,j}$ of the transition functions at $t = 0$. By definition,

$$\begin{aligned} a_{i,i} &= \lim_{y \to 0} \frac{p_{i,i}(y) - 1}{y}, \quad i = 0, 1, \ldots \\ a_{i,j} &= \lim_{y \to 0} \frac{p_{i,j}(y)}{y}, \quad i \neq j = 0, 1, \ldots . \end{aligned} \tag{1.6}$$

Hence, if h is small,

$$p_{i,j}(h) = a_{i,j}h + o(h), \quad i \neq j = 0, 1, \ldots, \tag{1.7}$$

where $o(x)$ is a function such that $\lim_{x \to 0}[o(x)/x] = 0$.

In other words, if the Markov process is in state i at some moment t, then the probability that at time $t + h$ it is in state j is nearly proportional to h, with coefficient of proportionality $a_{i,j}$. That is why $a_{i,j}$ is called the "instantaneous transition rate from state i to state j", $i \neq j$. The probability that the process leaves state i by $t+h$ is approximately equal to

$$1 - p_{i,i}(h) = -a_{i,i}h + o(h), \quad i = 0, 1, \ldots, \tag{1.8}$$

so $-a_{i,i}$ is the instantaneous rate of transition out of state i. Of course, we must have

$$-a_{i,i} = \sum_{\substack{j=0 \\ j \neq i}}^{\infty} a_{i,j}. \tag{1.9}$$

In fact, since $\mathbf{P}(y)$ is a stochastic matrix (its rows sum up to 1), the rows of $\mathbf{P}'(y)$ must sum up to 0 for all $y \geq 0$.

Let $\mathbf{A} = [a_{i,j}]$, $i, j = 0, 1, \ldots$ be the matrix of instantaneous transition rates. Differentiating (1.4) with respect to x and then letting $x \to 0$ yields a system of equations known as the Chapman–Kolmogorov backward differential equations:

$$\mathbf{P}'(y) = \mathbf{A}\mathbf{P}(y). \tag{1.10}$$

Similarly, differentiating (1.4) with respect to y and letting $y \to 0$ yields the Chapman–Kolmogorov forward differential equations

$$\mathbf{P}'(x) = \mathbf{P}(x)\mathbf{A}. \tag{1.11}$$

Either (1.10) or (1.11) can be solved for the transition probability functions, subject to the initial conditions $\mathbf{P}(0) = \mathbf{I}$ (the identity matrix) and $\mathbf{P}'(0) = \mathbf{A}$. In a purely formal way, treating $\mathbf{P}(y)$ as a numerically valued function and $\mathbf{A}$ as a constant, (1.10) and (1.11) are satisfied by

$$\mathbf{P}(y) = \mathrm{e}^{\mathbf{A}y}. \tag{1.12}$$

This turns out, indeed, to be the solution, provided that (1.12) is interpreted as

$$\mathbf{P}(y) = \sum_{n=0}^{\infty} \frac{y^n}{n!}\mathbf{A}^n, \quad y \geq 0. \tag{1.13}$$

Thus, the transition probability functions are completely determined by their derivatives at $y = 0$. It should be clear, however, that to find them in practice is by no means a trivial operation. The matrix $\mathbf{P}(y)$, for finite values of y, is referred to as the "transient solution" of the Markov process. As far as closed-form expressions are concerned, transient solutions are unobtainable for all but a few very simple Markov processes.

Let $\{S(t), t \geq 0\}$ be a Markov process with instantaneous transition rate matrix $\mathbf{A}$. Suppose that at time t the process is in state i. What is the distribution of the interval η_i until the first exit from state i (that interval is called the "holding time")? And what is the probability $q_{i,j}$ that the next state to be entered will be state j? According to the memoryless property, the answers to both these questions are independent of t and of the process history prior to t. In particular, they are independent of how long the process has already spent in state i. Consider first the holding time; denote by $\hat{H}_i(x)$ the complementary distribution function of η_i : $\hat{H}_i(x) = P(n_i > x)$. From the memoryless property, if the process stays in state i for time x, the probability that it will remain there for at least another interval y is independent of x. Therefore,

$$\hat{H}_i(x+y) = \hat{H}_i(x)\hat{H}_i(y), \quad x, y \geq 0. \tag{1.14}$$

Any distribution function which satisfies (1.14) must fall into one of the following three categories:

(i) $\hat{H}_i(x) = 1$ for all $x \geq 0$. If this is the case, once the process enters state i it remains there forever (properly speaking, the holding time does not have a distribution function then). States of this type are called "absorbing".
(ii) $\hat{H}_i(x) = 0$ for all $x \geq 0$. In this case the process bounces out of state i as soon as it enters it. Such states are called "instantaneous".
(iii) $\hat{H}_i(x)$ is monotone decreasing from 1 to 0 on the interval $[0, \infty)$ and is differentiable. States in this category are called "stable".

From now on, we shall assume that all states are stable. Differentiating Eq. (1.4) with respect to y and letting $y \to 0$ we obtain $\hat{H}'_i(x) = -\lambda_i \hat{H}_i(x)$, where $\lambda_i = -\hat{H}'_i(0)$. Hence

$$\hat{H}_i(x) = \mathrm{e}^{-\lambda_i x}, \quad x \geq 0,$$

and the distribution function $H_i(x) = P(\eta_i \leq x)$ is given by

$$H_i(x) = 1 - \mathrm{e}^{-\lambda_i x}, \quad x \geq 0. \tag{1.15}$$

To determine the parameter λ_i in terms of the matrix $\mathbf{A}$, note that according to (1.15) the probability of leaving state i in a small interval h is equal to $H_i(h) = \lambda_i h + o(h)$. Comparing this with (1.8) shows that λ_i is exactly the instantaneous transition rate out of state i:

$$\lambda_i = -a_{i,i}, \quad i = 0, 1, \ldots. \tag{1.16}$$

From (1.15), (1.7) and the memoryless property it follows that the probability that the process remains in state i for time x and then moves to state j in the infinitesimal interval $(x, x + dx)$ is equal to

$$\mathrm{e}^{-\lambda_i x} a_{i,j} \mathrm{d}x, \quad x \geq 0, \ j \neq i.$$

Integrating this expression over all $x \geq 0$ gives us the probability that the next state to be entered will be state j:

$$q_{i,j} = \int_0^\infty \mathrm{e}^{-\lambda_i x} a_{i,j} \mathrm{d}x = \frac{a_{i,j}}{\lambda_i} = -\frac{a_{i,j}}{a_{i,i}}, \quad i \neq j = 0, 1, \ldots. \tag{1.17}$$

We derived (1.15) and (1.17) under the assumption that the Markov process was observed at some arbitrary, but fixed, moment t. These results continue to hold if, for example, the process is observed just after it enters state i. Moreover, a stronger assertion can be made (we state it without proof): given that the process has just entered state i, the time it spends there and the state it enters next are mutually independent.

The behaviour of a Markov process can thus be described as follows: at time $t = 0$ the process starts in some state, say i; it remains there for an interval of time distributed exponentially with parameter λ_i (average length $1/\lambda_i$); the process then enters state j with probability $q_{i,j}$, remains there for an exponentially distributed interval with mean $1/\lambda_j$, enters state k with probability $q_{j,k}$, etc. The successive states visited by the process form a "Markov chain" — that is, the next state depends on the one immediately before it, but not on all the previous ones and not on the number of moves made so far. This Markov chain is said to be "embedded" in the Markov process.

We shall conclude this section by examining a little more closely the exponential distribution defined in (1.15). That distribution plays a central role in most probabilistic models that are analytically tractable. It owes its preeminent position to the memoryless property. If the duration η of a certain activity is distributed exponentially with parameter λ, and if that activity is observed at time x after its beginning, then the remaining duration of the activity is independent of x and is also distributed exponentially with parameter λ:

$$P(\eta > x + y \mid \eta > x) = \frac{P(\eta > x + y)}{P(\eta > x)} = \frac{\mathrm{e}^{-\lambda(x+y)}}{\mathrm{e}^{-\lambda x}} = \mathrm{e}^{-\lambda y} = P(\eta > y). \tag{1.18}$$