Queueing Networks and Product Forms

A Systems Approach

NICO M. VAN DIJK

WILEY–INTERSCIENCE SERIES IN SYSTEMS AND OPTIMIZATION

Queueing Networks and Product Forms

A Systems Approach

Nico M. van Dijk

University of Amsterdam, The Netherlands

JOHN WILEY & SONS

Chichester · New York · Brisbane · Toronto · Singapore

Other Wiley Editorial Offices

John Wiley & Sons, Inc., 605 Third Avenue,
New York, NY 10158-0012, USA

Jacaranda Wiley Ltd, G.P.O. Box 859, Brisbane,
Queensland 4001, Australia

John Wiley & Sons (Canada) Ltd, 22 Worcester Road,
Rexdale, Ontario M9W 1L1, Canada

John Wiley & Sons (SEA) Pte Ltd, 37 Jalan Pemimpin #05-04,
Block B, Union Industrial Building, Singapore 2057

Library of Congress Cataloging-in-Publication Data

Dijk, N.M. van.
Queueing networks and product forms : a systems approach / Nico M. van Dijk.
p. cm.—(Wiley–Interscience series in systems and optimization)
Includes bibliographical references and index.
ISBN 0 471 92848 8
1. Queueing theory. I. Title. II. Series.
T57.9.D54 1993 92-41673
519.8'2—dc20 CIP

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0 471 92848 8

Typeset in 10/12 Times by Thomson Press (India) Ltd, New Delhi
Printed and bound in Great Britain by Biddes Ltd, Guildford, Surrey

This book is dedicated to the memory
of my father-in-law

Preface

Why write another book on queueing networks when a number of excellent texts on queueing or stochastic service networks, and applications such as computer, telecommunication, data or manufacturing networks, have appeared over the last decade (e.g. Kleinrock 1976, König and Stoyan 1976, Kelly 1979, Bruell and Balbo 1980, Gelenbe and Mitrani 1980, Sauer and Chandy 1981, Lavenberg 1982, Trivedi 1982, Hayes 1984, Whittle 1986, Bertsekas and Gallager 1987, Mitrani 1987, Schwartz 1987, Perros and Altiok 1988, Walrand 1988, King 1990)? And why a book on just the single aspect of *product forms*, which is generally perceived as narrow from a practical point of view?

For one thing, this book intends to indicate that this sub-area is not that narrow from neither a theoretical or a practical point of view. In fact, were it not for Jackson's well-known product form and later extensions, queueing or stochastic analysis would most certainly not have been as popular in the above application areas as it is today. Furthermore, none of the existing books is devoted only to *product-form results for queueing networks*. The most general and most highly recommended books in this area are by Kelly (1979) and Whittle (1986). These two books, however, reveal product-form results in a genuine and abstract setting such as migration and spatial processes, on the one hand, but, on the other, pay relatively little attention to practical queueing network features such as blocking or interference phenomena, and give few practical insights into detecting and using product forms. Today, a wide range of material on such topics is available and deserves attention. This book aims in particular to shed a somewhat different light on this topic, or at least to advocate common-sense insights from which both researchers and practitioners may benefit for recognizing and exploiting product-form results.

BACKGROUND AND FURTHER MOTIVATION

Queueing networks have gained a wide popularity over the last few decades for modelling and performance-evaluation purposes in telecommunications, computer technology and manufacturing. Much of the success of queueing networks can probably be attributed not only to their flexible and generic modelling capacities, on the one hand, but also to the simple Jacksonian product-form type expressions that are available under sufficiently simplifying assumptions, on the other.

As for closed-form expressions, an extensive literature has appeared since Jackson's pioneering papers (Jackson 1975, 1963), providing extensions of product-form results for job-shop or migration networks. From a practical point of view, some results worth noting here are the inclusion of different queueing disciplines such as in the well-known BCMP paper for computer application, (Baskett *et al.* 1976) and Kelly's related results in a wider context (Kelly 1975, 1976, 1979, the extension to non-exponential service or handling tims as developed most notably by König *et al.* (1976) and Schassberger (1977, 1978) and the extension to systems with blocking as in Cohen (1957), Kelly (1979), Pittel (1980) and Hordijk and Van Dijk (1981a,b). Unfortunately, the underlying conditions for such product-form expressions to hold seemed to be rather restricitive for practical applications. For example, priority queueing does not seem to be included, first-come first-served service scheduling appears to require exponential and indistinguishable services and results for systems with blocking seem limited to reversible structures. More generally, typical practical phenomena such as non-reversible blocking, dynamic routing, job priorities, concurrent servicing and server breakdowns seem to exclude simple closed-form expressions for steady-state measures.

Thus we appeared to be faced with a discrepancy between elegant though quite limited simple closed product-form expressions, on the one hand, but phenomena which may destroy these expressions and which are to be dealt with in practice on the other. How can this discrepancy be overcome and how can we benefit from the results obtained for advising in engineering situations? Only insight on a *theoretical* and a *practical* level can be the key.

INSIGHTS

A general practical insight as to why and to what extent product-form results will hold appears to be lacking. Such an insight, even though with negative direct results, can be quite useful for practical purposes such as:

(1) To determine quickly whether one can expect a simple expression and which type of expressions to look for (when positive); or (when negative)
(2) To realize which particular system protocol or characteristic will destroy the closed-form result, so as to possibly obtain a reasonable approximation, heuristic, or a secure bound.

In this light, it is interesting to note that telecommunications technology has been rapidly evolving over the last few decades and a large expansion is to be expected in the future. Though positive from a development point of view, this trend also introduces a new problem. While today's telephone networks can build upon more than 60 years of research, experiments and detailed knowledge, the communications system of tomorrow is too complex for detailed analysis yet it is to be designed in the short term and also on a relatively short operational basis.

Simple robust performance estimates or secure bounds that can be obtained at low computational expense to achieve a rapid first indication of a system performance are thus of interest. Simple insights at system protocol level as to when exact product-form results can be obtained or not and a knowledge of their flexibility and limitations will certainly be helpful (if not essential) to develop such tools. Further, the recognition of which simplifying assumptions will lead to product-form results does provide a reference framework which can often be seen as a first-order modelling. For example, complicating factors such as propagation delays in communications systems or breakdowns in a service network are typically second order. Despite the extensive literature, however, this aspect of product-form results (that is, a simple unifying insight at system protocol level and its potential uses for approximating or bounding purposes) has seldom been advocated as a practical tool.

MAIN RESULTS

This book aims to provide a systematic and common-sense approach to obtaining product-form results based on insights into principles of partial balance which can usually be expressed in terms of the underlying system protocols. Using this approach, one may be able to obtain either of the following two results:

(1) The system will have a product form and the nature of that form,
(2) The system will not have a product form and the reasons for its not having one.

Various concepts of partial balances in relation to product-form results have been reported in the literature (for example, partial, detailed, local, centre and station balance). These partially overlap but are also often unrelated and practitioners as well as scientists may still be rather confused about the interrelationships or which to apply or check in a practical situation. Furthermore, such concepts have been used mainly to formalize product-form results of a particular form. However, they have seldom been advocated or utilized as a non-mathematical tool by which one can directly obtain a closed-form expression.

Balance hierarchy

This book provides a number of different partial balances which can all be interpreted physically and used as a practical tool. Most notably, the following hierarchy of physical partial principles will hereby be established:

- Per station (station balance)
- Per job class at a station (job-class balance)
- Per job or position at a station (job-local balance)
- Per specified group of jobs at stations (group-local balance)

As station balance is the most generally used because it involves practical phenomena such as blocking due to finite capacity constraints, service interruptions caused by breakdowns or dynamic routing as a result of overflows, this book will be primarily devoted to the concept of station balance.

Practical relevance

All known partial balances are unified here but are also extended to allow for:

- Finite capacity constraints at individual stations or clusters of stations;
- Multiclass first-come first-served parallel queues with common capacity constraints;
- Service mechanisms which take job priorities into account; or
- Systems with batch or discrete-time servicing.

This hierarchy is bidirectional with more restrictive conditions on system protocols, on the one hand, but more detailed product-form results and more flexible service distributional assumptions, on the other. These hierarchical insights will be useful in engineering situations to determine:

- What level to work in;
- What conditions to be checked, and
- What assumptions to be made.

Both practitioners and researchers may benefit from these 'quick engineering insights'. First, they will be able to quickly evaluate the performance of a system when a product form can be obtained (for instance, to compare the system under different workloads or parameter values). Secondly, they can determine by which system protocols a product-form expression is violated so that further numerical or approximate computations can be developed. However, in this case, analysis of the violation may suggest appropriate approximations or, as will also be illustrated in this book, simple, inaccurate but secure bounds.

Finally 'novel' product-form results are still obtained for practical applications in computer communications, broadcasting and manufacturing. These results are usually far from trivial and, although they are not covered in the literature, they are worth reporting in their own right. However, experience has shown that such results can usually be recognized and unified by one of these simple partial balance key concepts.

Special applications

Separate chapters will therefore pay special attention to specific applications of present-day interest, and there will be a variety of generic structures for

practical applications throughout all chapters. This book contains a separate chapter which is solely devoted to: communication structures such as CSMA and circuit switching.

It is shown that product-form expressions also apply to such networks while using only straightforward and balance equations. Moreover, it also serves as a first illustration of how exponentiality conditions can be avoided by a concept of an appropriate detailed partial balance.

OUTLINE OF THIS BOOK

The structure is as follows. First, in a preliminary Chapter 0 some practical motivation is provided by a brief discussion of typical features that arise in computer, manufacturing and telecommunications applications. These special features will be dealt with in more detail in the course of the book.

Chapter 1 contains an informal introduction to and discussion of the use of balance equations and particularly their reductions to concepts of partial balance equations. A cash-balance example is presented to illustrate the actual use and results of these concepts in terms of concrete system protocols. A detailed hierarchy of partial balances will be discussed and established.

Chapter 2 presents basic properties of exponentiality that can be used to model the behaviour of queueing network applications by rates. Next, the use of balance is formally justified by a brief but self-contained presentation of different continuous-time Markov chains. Finally, a convenient discrete-time renewal result is given and a communications example is studied in detail to give a first illustration of how the combination of an appropriate concept of partial balance equations and the discrete renewal result may lead to a simple product-form expression also without exponentiality assumptions (insensitivity).

Chapter 3 introduces and illustrates the properties of the concept of station balance by means of studying:

- Classical Erlang and Engset models
- A two-station assembly-line example
- Closed and open cyclic queueing network models
- Closed and open Jackson networks with random routing.

Next, the phenomenon of blocking is introduced. First, simple rules are extracted from a two-station example in order to obtain a product form. Then, by a three-station counter-example it is shown why these rules are usually violated when blocking is involved. This example also illustrates, however, that by simple physical insights one can obtain an appropriate product-form modification for non-product-form systems.

Chapter 4 concentrates on exploring this latter idea in more detail. A bounding methodology is illustrated for such non-product-form features as:

- Breakdowns
- Dynamic routing
- Finite capacity constraints

Numerical support as well as an optimal design application are included. This chapter is presented on a primarily non-mathematical level and could be read before Chapters 2 and 3.

Chapter 5 studies the concept of station balance in a more formal way for arbitrary networks with state (e.g. load-dependent) routing and servicing. A characterization of station balance is derived in terms of local solutions of state-dependent traffic equations. These local solutions, in turn, can often be expressed in concrete system parameters. The general result is illustrated by a large number of examples.

Chapter 6 shows that the station principle can also be operated on a more global level to obtain product-form results for queueing networks with finite capacity constraints on clusters of stations rather than on individual stations.

Chapter 7 is the applications chapter devoted to communications and broadcasting networks. This chapter can be read directly after Chapter 2.

STRUCTURING OF CHAPTERS

As the aim of this book is to be practical and instructive but, at the same time, general and illustrative, the following structure has been adopted. First, an example is worked out in detail. Next, a general framework is set up and general conditions are derived. These conditions are then illustrated by a number of representative examples. Finally, special results, consequences or applications will be presented.

TECHNICAL LEVEL

Chapters 1–4, 6 and 7 can be read as a complete and self-contained body at a rather low technical level, and no more than basic calculus is required. These chapters will give a substantial flavour of when to expect and how to use product-form results. Chapter 5 is not more difficult but operates on a rather more abstract level.

PRECEDENCE

The following precedence scheme to read the book or selected sections may suggest several sub-section that can be read or taught separately. The series of Chapters 0-1-2-3-4, for example, gives a complete first insight into the consequences and practical potential of product-form results for single-class

queueing networks at a low technical and practical level. The series of Chapters 0-1-2-3-7 is particularly tailored to practical applications.

The book is aimed at students, researchers and practitioners in:

- Mathematics
- Computer science
- Telecommunications
- Systems engineering

However, I wish to apologize to mathematicians, on the one hand, for heuristic steps without mentioning formal technicalities such as issues of existence and uniqueness and to practitioners, on the other, for not going into detailed technicalities that are certainly as important for actual modelling and performance evaluation in practice. At the same time, I also wish to express my hope that both theoreticians and practitioners can appreciate the primary interest of this book to provide a common-sense but, at the same time, unifying insight into the phenomenon of product forms and their potential. The book seems convenient to teach from since it can be used, following the above precedence scheme, on different levels and in self-contained sub-sections.

Acknowledgements

First, I am indebted to numerous colleagues in the field of queueing and performance evaluation for stimulating discussions at conferences, workvisits and other occasions. These discussions have always had a substantial motivating impact on me. Next, I wish to acknowledge Robert Cooper for his lucidly written book that introduced me to the area of queueing; Frank Kelly, for his first papers and his book on queueing networks that awoke my research interests in this area; Arie Hordijk, for our many stimulating discussions which made me appreciate both intuitive and formal approaches; and Henk Tijms for his practical viewpoints during our collaboration. Last but not least, my special thanks go to Mary Tan for her marvellous typing job and for permanently 'being available' without which this book would most certainly not have been written, and to the Faculty of Economics and Econometrics at the Free University Amsterdam for the support and facilities provided.

Contents

CHAPTER 0

Practical motivation

0.1 INTRODUCTION

Queueing network modelling has experienced a very rapid growth over the past few decades with applications in many disciplines. Among these are traditional areas such as biology (migration and population models), electrical engineering (electrical flow models) and chemistry (polymerization and clustering models). However, especially over the last fifteen years, its most prevalent applications are found in the areas of:

- Manufacturing
- Computer networking
- Telecommunications and broadcasting

Much of this success can be attributed to the early applications of computer network design in the 1970s, and as a result of present-day technological developments associated with the integration of production, information and communication systems, applications arising from each of these areas will be described and further motivation for more complex queueing network analysis provided.

From a performance evaluation point of view, excellent references for each of these application areas are available and at the end of this chapter we give a far from exhaustive selection. For detailed descriptions of specific applications in these areas the reader is referred to the references or related literature mentioned therein. This list illustrates that the different application areas lead to special features, not only from a technological point of view but also for performance-evaluation purposes. In particular, specific differences are found in:

(1) *Configuration*
- Series of stations (manufacturing)
- Central server models (computer networking)
- Direct end-to-end communications (telecommunications)

(2) *Blocking or interference phenomena*
- Finite storage buffers (manufacturing)
- Common resource contentions (computer networking)
- Source interactions or message collisions (telecommunications)

(3) *Service requirements and service disciplines*

- Deterministic and FCFS or LCFS (manufacturing)
- Packetized and processor sharing (computer networking)
- Arbitrary transmissions and multi-channel (telecommunications)

From a modelling point of view these different aspects can be unified into a single framework. To obtain closed-form expressions, however, as is the primary intent of this book, these distinctions will play a crucial role and will require specialized results. This preliminary chapter aims to shed some light on these differences and to provide some practical underlying motivation for the various queueing network aspects that will be investigated in detail later such as *blocking, state-dependent routing, capacity limitations, breakdowns, service disciplines, insensitivity phenomena* and *discrete-time analysis.* To this end, each of the three major application areas mentioned will be briefly discussed to highlight special features and generic structures.

0.2 MANUFACTURING SYSTEMS

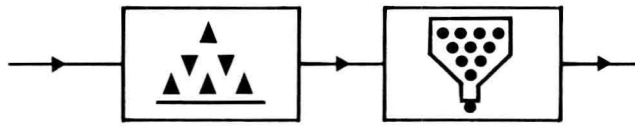

Figure 0.1

In manufacturing systems, pallets with parts, production tools or raw materials move along various machines or workstations to be assembled and worked upon or processed for certain random (e.g. constant) amounts of time. Typically, multiple-part types, machines that are only applicable for specific types, selective allocation of parts, finite storage limitations and possible machine breakdowns are involved.

0.2.1 Assembly lines

The most simple but also most common generic structure of a manufacturing system is that of an assembly line with a number of workstations in series (Figure 0.1).

0.2.2 Finite buffers

Here, a finite capacity constraint on the total number of parts at some of the individual stations is most common. More precisely, storage buffers between successive workstations can be modelled as finite workstations. For example, a storage buffer of size N between two workstations with a single machine for